Morphological Prototyping for Unsupervised Slide Representation Learning in Computational Pathology

Andrew H. Song[1,2,*], Richard J. Chen[1,2,*], Tong Ding[1,2], Drew F.K. Williamson[1,2,†],
Guillaume Jaume[1,2], Faisal Mahmood[1,2]
[1]Mass General Brigham and [2]Harvard University
asong@bwh.harvard.edu, richardchen@g.harvard.edu, faisalmahmood@bwh.harvard.edu

Abstract

Representation learning of pathology whole-slide images (WSIs) has been has primarily relied on weak supervision with Multiple Instance Learning (MIL). However, the slide representations resulting from this approach are highly tailored to specific clinical tasks, which limits their expressivity and generalization, particularly in scenarios with limited data. Instead, we hypothesize that morphological redundancy in tissue can be leveraged to build a task-agnostic slide representation in an unsupervised fashion. To this end, we introduce PANTHER*, a prototype-based approach rooted in the Gaussian mixture model that summarizes the set of WSI patches into a much smaller set of morphological prototypes. Specifically, each patch is assumed to have been generated from a mixture distribution, where each mixture component represents a morphological exemplar. Utilizing the estimated mixture parameters, we then construct a compact slide representation that can be readily used for a wide range of downstream tasks. By performing an extensive evaluation of* PANTHER *on subtyping and survival tasks using 13 datasets, we show that 1)* PANTHER *outperforms or is on par with supervised MIL baselines and 2) the analysis of morphological prototypes brings new qualitative and quantitative insights into model interpretability. The code is available at* https://github.com/mahmoodlab/Panther.

1. Introduction

Representation learning of whole-slide images (WSIs) is a fundamental task in Computational Pathology (CPath) [77]. Given a WSI, the goal is to learn a slide-level representation that can be used for various downstream tasks, such as diagnosis, prognostication, and therapeutic response prediction. The standard approach is weakly supervised learning based on Multiple Instance Learning (MIL) [13, 29, 40], in which the gigapixel WSI is tokenized into a large set of patch embeddings ($N > 10,000$) with a pretrained vision encoder, followed by aggregation of the embeddings [40]. Current advances in CPath have examined: (1) learning better patch embeddings with domain-specific vision encoders based on self-supervised learning (SSL) [1, 18, 21, 31, 39, 44, 46, 58, 83] and (2) developing new aggregation strategies for pooling patch embeddings into a slide representation [55, 56, 60, 73]. As many histology datasets have limited samples for supervised MIL, an emerging goal in CPath is to (3) shift slide representation learning from *weakly-supervised* to *unsupervised* [17, 41, 52, 69, 81], which may help mitigating data and label scarcity and improving generalization.

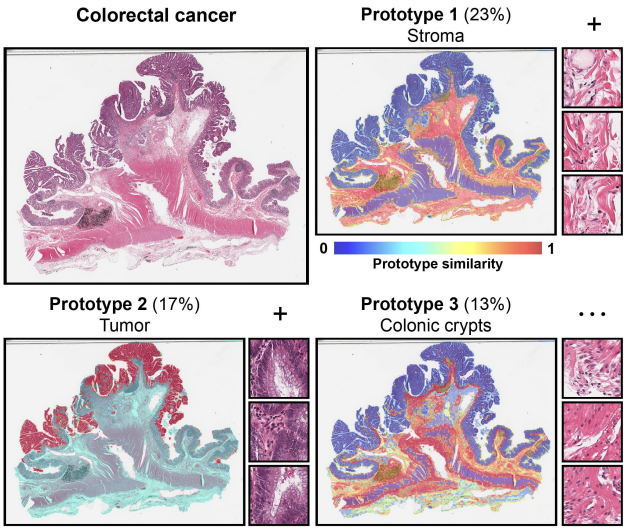

Figure 1. **Slide decomposition into morphological prototypes** Due to morphological redundancy across and within the tissue, a slide can be decomposed into prototypes. We introduce PANTHER, a method that can identify and extract morphological prototypes to form a compact and unsupervised slide representation.

*Equal contribution
†Presently at Emory University School of Medicine

We postulate that such models are particularly suited for fine-grained classification tasks, such as survival outcome prediction that require holistic modeling of morphologies found in the tissue microenvironment [88, 96]. In contrast to "needle-in-a-haystack" tasks (*e.g.,* micro-metastasis detection) that require localizing tumor patches, "panoramic" tasks require integrating spatial heterogeneity (diversity of distinct tumor populations) [35, 61], interactions and context (immune infiltrate near invasive tumor margin) [3, 71], and size (number and size of masses) [6]. Attention-based architectures [40, 43, 60] demonstrate clinical-grade performance in the former task (selectively focusing on diagnostic patches of a single visual concept) [4]; however, they have limited expressivity in the panoramic tasks that benefit from understanding proportions and mixtures of visual concepts [7, 15, 20].

Based on these insights, we propose an *unsupervised* slide representation framework that can accurately capture the proportions and mixtures of morphological visual concepts. Specifically, we build on the observation that WSI patches show *morphological redundancy* and thus a handful of morphological patterns are repeated (**Fig.** 1). Formally, we hypothesize that a concise set of key descriptors (prototypes), coupled with distribution characterizing the extent and variation of each descriptor, would comprehensively summarize the WSI. As this summary only relies on the statistical characteristics of each WSI, this yields a generic and unbiased slide embedding applicable across multiple downstream tasks. To faithfully summarize the WSI, the goal becomes to construct (1) a mapping between each patch and the prototypes and (2) a slide embedding with the learned mapping that includes representation of each prototype and its extent, *i.e.,* its cardinality.

To this end, we introduce PANTHER, a **P**rototype **A**ggregatio**N**-based framework for compac**T** **HE**terogenous slide set **R**epresentation (PANTHER). Inspired by previous work in prototype-based set representation learning [26, 49, 64, 89], PANTHER builds an *unsupervised* slide embedding by assuming that each patch embedding is generated from the Gaussian mixture model (GMM), with each morphological prototype representing a mixture component. By employing GMM, the two aforementioned goals are seamlessly satisfied through the parameter estimation procedure with Expectation-Maximization (EM) [27, 49]: (1) the estimated posterior probability of mixture assignment for each patch defines the mapping between a patch embedding and a prototype, and (2) mixture probability represents cardinality, and mixture mean & covariance represent the representation of corresponding morphological pattern. The slide representation is formed as a concatenation of the estimated GMM parameters across all prototypes. Since the representation can be decomposed along each prototype, this allows for per-prototype nonlinear modeling and interpretation of the slide centered around each histology visual concept.

To summarize, our contributions are (1) the first prototypical framework for learning compact and unsupervised *slide representations* in WSI based on GMM; (2) a comprehensive evaluation of four diagnostic and nine prognostic tasks, demonstrating the outperformance against nearly all unsupervised and supervised baselines; (3) post-hoc interpretability with the quantification and visualization of morphological prototype spread within the tissue.

2. Related work

2.1. Multiple instance learning in CPath

While initial histology-based outcome prediction was centered on pathologist-annotated region-of-interests [11, 47, 65], later works have utilized WSIs for clinical prediction tasks with MIL [13, 17, 22, 40, 75]. There is a sustained effort for new MIL schemes, with a focus on new patch aggregation strategies to learn more representative and task-specific slide embedding, towards better predictive accuracy [55, 59, 73, 78, 84, 94] or interpretability [43, 79]. MIL methods based on multiscale representation slides have recently shown promise for "panoramic" tasks [5, 33, 37]. PANTHER is similar to MIL in that the slide-level embedding is constructed from patches for outcome prediction. However, PANTHER constructs an *unsupervised* set embedding and is agnostic to downstream tasks, in contrast to supervised MIL frameworks.

2.2. Quantification of distance between sets

Measuring the distance between two distinct sets (Wasserstein distance), *e.g.,* between supports of empirical probability distributions, has received increasing attention from a diverse range of disciplines such as signal processing [50], vision-language tasks [68], and computational biology [2, 10, 72]. Given a similarity (or cost) metric between set elements such as $\mathcal{L}_2$ distance, the Wasserstein distance is defined as transporting (or matching) elements of one set to the other, incurring a minimum cost. Computing such distance is commonly called the optimal transport (OT) [25].

It is natural to extend this idea to CPath, where the biological entities of interest (e.g., WSI and genomic data) are typically modeled as a set of biological concepts. Different from MIL setting where a set of WSI patches is matched to a patient-level clinical label, this concerns matching between two sets: 1) sets of WSI patches for slide retrieval [24] or domain adaptation [30], 2) different cancer datasets to quantify morphological distance between different cancer types [90], and 3) a set of WSI patches and a set of genomic tokens to learn optimal fusion for improved prognosis [85]. Similarly, PANTHER can be seen as the matching problem between a set of WSI patches and prototypes.

2.3. Prototype-based set representation

Prototypes are representative examples of data points that share the same class, usually formulated as centroids from clustering that describe unique human-interpretable concepts and other semantic information [16, 19, 76]. Recently, it has been applied to compactly represent large set data in bioinformatics and NLP [26, 49, 54, 64]. The desiderata for prototype-based set representation is to model: (1) cardinality, i.e., how many elements in the set are associated with a prototype, and (2) description, i.e., prototype identity.

Posed also in many related forms such as signatures [53, 95] and bag of visual words (BoVW) [12, 23, 74], learning prototypical representations is a natural problem in CPath as repeating histology patterns often reflect the same morphology [36, 45, 67, 82, 86, 87, 91]. Recent prototypical MIL approaches for pathology include AttnMISL [88], which aggregates patch embeddings within the same cluster, followed by aggregating the pooled cluster embeddings. Following recent advances in visual self-supervised learning, prototypes have been used for constructing unsupervised slide features via pooling similar patch embeddings into a concatenated representation (H2T [81]) or measuring the proportion of prototypes assignments in WSIs (HPL [69]). We note that H2T and HPL have limitations in not encoding cardinality or not including deep visual features, which are relevant for interpretability and solving panoramic tasks. Moreover, as many pretrained vision encoders in CPath are pretrained on TCGA, prototypical patterns lack extensive evaluation of out-of-domain performance.

3. Methods

We present PANTHER, an unsupervised slide representation learning framework based on a compact set of prototypes with GMM (**Fig.** 2). We first explain the GMM setup and its connection to a slide embedding (Section 3.1). We then present how it is used for downstream tasks (Section 3.2) and prototype-based interpretation (Section 3.3).

3.1. Prototype-based aggregation

Given a WSI for subject j, we tessellate it into small non-overlapping patches $\mathbf{X}^j = \{\mathbf{x}_1^j, \cdots, \mathbf{x}_{N_j}^j\}$ with $\mathbf{x}_n^j \in \mathbb{R}^{W\times H\times 3}$. We then employ a feature extractor $f_{\text{enc}}(\cdot)$ pretrained on large archives of histopathology images [18], to extract a representative and compressed embedding from each patch. The set of extracted embeddings $\mathbf{Z}^j = \{\mathbf{z}_1^j, \cdots, \mathbf{z}_{N_j}^j\}$ with $\mathbf{z}_n^j = f_{\text{enc}}(\mathbf{x}_n^j) \in \mathbb{R}^d$ is then *aggregated* to construct a slide embedding $\mathbf{z}_{\text{WSI}}^j$.

We aim to represent the set $\mathbf{Z}^j$ with a small set of prototypes $\mathbf{H} = \{\mathbf{h}_1, \cdots, \mathbf{h}_C\}$ with $\mathbf{h}_c \in \mathbb{R}^d$, $C \ll N_j$, without compromising essential morphological information. Using the prototypes as *references*, each patch is aggregated (or mapped) to the references and form $\mathbf{z}_{\text{WSI}}^j \in \mathbb{R}^{C\cdot M}$ [49, 64]

$$\mathbf{z}_{\text{WSI}}^j = [\sum_{n=1}^{N_j} \phi^j(\mathbf{z}_n^j, \mathbf{h}_1), \cdots, \sum_{n=1}^{N_j} \phi^j(\mathbf{z}_n^j, \mathbf{h}_C)], \quad (1)$$

where $\phi^j(\cdot,\cdot) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^M$ is a function that maps a pair of patch embedding and reference, based on the similarity between the two, to a post-aggregation prototype embedding. In our work, the typical dimensions are $C = 8 \sim 32$ and $N_j = 10,000 \sim 20,000$. The $\mathbf{z}_{\text{WSI}}^j$ dimension is fixed such that a variable-length set of N_j features can always be represented in fixed-length.

To define and estimate the mapping function ϕ^j, we introduce a probabilistic framework for patch embedding distribution and assume each $\mathbf{z}_n^j$ is generated from a GMM,

$$\begin{aligned} p(\mathbf{z}_n^j;\theta^j) &= \sum_{c=1}^{C} p(c_n^j = c;\theta^j)\cdot p(\mathbf{z}_n^j | c_n^j = c;\theta^j) \\ &= \sum_{c=1}^{C} \pi_c^j \cdot \mathcal{N}(\mathbf{z}_n^j; \boldsymbol{\mu}_c^j, \Sigma_c^j),\ s.t. \sum_{c=1}^{C} \pi_c^j = 1, \end{aligned} \quad (2)$$

where π_c^j refers to the mixture probability of component c in WSI set j and θ^j refers to the set of parameters to be estimated, i.e., $\theta^j = \{\pi_c^j, \boldsymbol{\mu}_c^j, \Sigma_c^j\}_{c=1}^C$. For ease of computation, we use the diagonal covariance Σ_c^j. Intuitively, Eq. 2 states that a morphological prototype and its variations correspond to a mixture component, with π_c^j indicating the extent to which the pattern manifests in the j^{th} WSI.

We formulate the slide embedding construction as that of estimating $\widehat{\theta^j}$ from $\mathbf{Z}^j$. To this end, we maximize the log-likelihood (or log-posterior if prior is introduced for θ^j)

$$\max_{\theta^j} \log p(\mathbf{Z}^j;\theta^j) = \max_{\theta^j} \sum_{n=1}^{N_j} \log p(\mathbf{z}_n^j;\theta^j). \quad (3)$$

We use the expectation-maximization (EM) algorithm to obtain the maximum likelihood estimate θ^j [27, 49], with derivation provided in **Supplementary Information**. The algorithm produces a posterior distribution $q(c_n^j = c|\mathbf{z}_n^j)$, which represents the probability that $\mathbf{z}_n^j$ is associated with prototype c. At EM iteration $t+1$, it is given as

$$q^{(t+1)}(c_n^j = c|\mathbf{z}_n^j) = \frac{\pi_c^{j,(t)} \mathcal{N}(\mathbf{z}_n^j; \boldsymbol{\mu}_c^{j,(t)}, \Sigma_c^{j,(t)})}{\sum_{c=1}^{C} \pi_c^{j,(t)} \mathcal{N}(\mathbf{z}_n^j; \boldsymbol{\mu}_c^{j,(t)}, \Sigma_c^{j,(t)})}. \quad (4)$$

Eq. (4) indicates that a patch is assigned to a certain prototype c that is most similar, with the similarity measured in terms of weighted $\mathcal{L}_2$ distance. Moreover, the soft prototype assignment, i.e., $q^{(t+1)}(c_n^j = c|\mathbf{z}_n^j) > 0$, $\forall c$, allows each patch to contribute towards all prototypes, in contrast to hard prototype assignment approaches [69, 81, 88, 91].

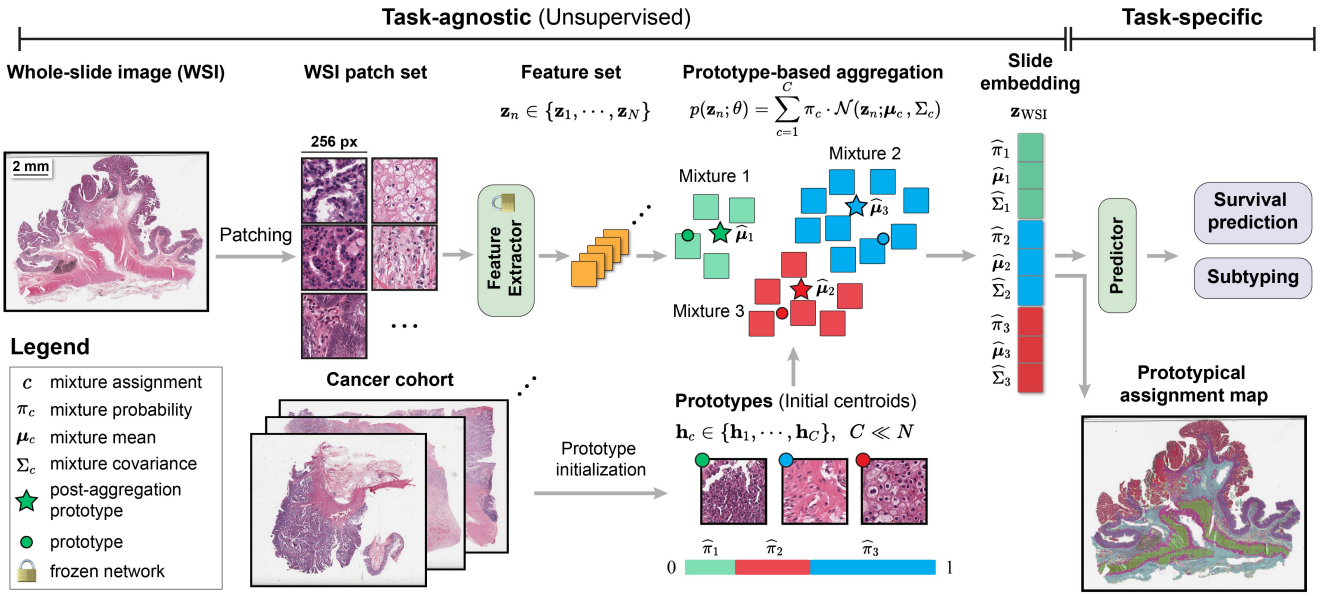

Figure 2. **Overview of PANTHER workflow**. Whole-slide image (WSI) is segmented and patched into a set of WSI patches. A compressed feature for each patch is encoded through a feature extractor pretrained on a large histopathology dataset. PANTHER uses the Gaussian mixture model for patch embedding distribution, with each mixture corresponding to a morphologically distinct prototype. The estimated model parameters are concatenated to form the slide representation, which can be used as input to a predictor module for clinical downstream tasks and visualized as a prototypical assignment map.

Using $q_{n,c}^{j,(t+1)} = q^{(t+1)}(c_n^j = c|\mathbf{z}_n^j)$ for notational simplicity, we can estimate $\theta^{j,(t+1)}$ as

$$\pi_c^{j,(t+1)} = \frac{\sum_{n=1}^{N_j} q_{n,c}^{j,(t+1)}}{N_j}, \quad \boldsymbol{\mu}_c^{j,(t+1)} = \frac{\sum_{n=1}^{N_j} q_{n,c}^{j,(t+1)} \cdot \mathbf{z}_n^j}{\sum_{n=1}^{N_j} q_{n,c}^{j,(t+1)}}$$
$$\Sigma_c^{j,(t+1)} = \frac{\sum_{n=1}^{N_j} q_{n,c}^{j,(t+1)} \cdot (\mathbf{z}_n^j - \boldsymbol{\mu}_c^{j,(t+1)})^2}{\sum_{n=1}^{N_j} q_{n,c}^{j,(t+1)}} \tag{5}$$

We set $\pi_c^{j,(0)} = 1/C$, $\boldsymbol{\mu}_c^{j,(0)} = \mathbf{h}_c$, and $\Sigma_c^{j,(0)} = \mathbf{I}$, which serves as a morphology-aware initialization. Due to its iterative nature, EM can be placed as a neural network module. The initialization for $\{\mathbf{h}_c\}_{c=1}^C$ is performed with K-means clustering on the entire patch training set, constructed by aggregating patch embeddings from all training slides in a cohort.

Based on the final estimate $\widehat{\theta^j}$ after the EM convergence, the slide embedding $\mathbf{z}_{\text{WSI}}^j \in \mathbb{R}^{C \cdot M}$ with $M = 1 + 2d$ can be represented as a concatenation of the elements in $\widehat{\theta^j}$, following set representation learning literature [49, 64],

$$\mathbf{z}_{\text{WSI}}^j = [\mathbf{z}_{\text{WSI},1}^j, \cdots, \mathbf{z}_{\text{WSI},C}^j]$$
$$= [\underbrace{\widehat{\pi}_1^j, \widehat{\boldsymbol{\mu}}_1^j, \widehat{\Sigma}_1^j}_{\sum_{n=1}^{N_j} \phi^j(\mathbf{z}_n^j, \mathbf{h}_1)}, \cdots, \underbrace{\widehat{\pi}_C^j, \widehat{\boldsymbol{\mu}}_C^j, \widehat{\Sigma}_C^j}_{\sum_{n=1}^{N_j} \phi^j(\mathbf{z}_n^j, \mathbf{h}_C)}]. \tag{6}$$

We emphasize that while the prototypes $\{\mathbf{h}_c\}_{c=1}^C$ are shared across different WSIs, the parameter estimation and the slide embedding construction are performed per WSI. Overall, the embedding $\mathbf{z}_{\text{WSI}}^j$ satisfies two essential principles for a faithful WSI representation and good downstream performance. First, it accounts for the cardinality of each prototype explicitly through $\widehat{\pi}_c^j$ and implicitly through $\widehat{\boldsymbol{\mu}}_c^j$ and $\widehat{\Sigma}_c^j$. In addition, by concatenating the features (rather than averaging), feature vector for each morphological prototype is directly accessible for downstream tasks.

3.1.1 Connection to optimal transport

The aggregation in PANTHER can be seen as matching the empirical distribution of patch embeddings and the prototypes, defined as $\hat{p}_j$ and $\hat{q}_j$ respectively. Specifically, we have $\hat{p}_j = \sum_{n=1}^{N_j} a_n^j \cdot \delta_{\mathbf{z}_n^j}$ and $\hat{q}_j = \sum_{c=1}^{C} \pi_c^j \cdot \delta_{\mathbf{h}_c}$, where $\sum_{n=1}^{N_j} a_n^j = \sum_{c=1}^{C} \pi_c^j = 1$. OT aggregates $\{\mathbf{z}_n^j\}_{n=1}^{N_j}$ to $\{\mathbf{h}_c\}_{c=1}^C$ by minimizing the Wasserstein distance between $\hat{p}_j$ and $\hat{q}_j$, typically assuming uniform distribution for $\{a_n^j\}_{n=1}^{N_j}$ and $\{\pi_c^j\}_{c=1}^C$, i.e., $a_n^j = 1/N_j$ and $\pi_c^j = 1/C$ [25, 64]. While PANTHER also assumes $a_n^j = 1/N_j$, the mixture probability π_c^j is estimated (Eq. 5). Furthermore, the OT solution can be seen as a special case of GMM solution with uniform prototype distribution [49].

3.2. Downstream evaluation

The embedding $\mathbf{z}_{\text{WSI}}^j$ can be used as the input to a *predictor* module $g(\cdot)$ for various downstream tasks. The predictor

$g(\cdot)$ can be a linear layer (linear probing), or implemented as a multilayer perceptron (MLP). Instead of using the entire $\mathbf{z}_{\text{WSI}}^j$ as an input to a MLP, we propose a structured MLP

$$\begin{aligned} \mathbf{z}'_{\text{WSI}} &= [g_1^{\text{indiv.}}(\mathbf{z}_{\text{WSI},1}), \ldots, g_C^{\text{indiv.}}(\mathbf{z}_{\text{WSI},C})] \\ g(\mathbf{z}_{\text{WSI}}) &= g^{\text{pred.}}(\mathbf{z}'_{\text{WSI}}), \end{aligned} \tag{7}$$

where j is dropped for notational simplicity, and $g^{\text{pred.}}$ and $\{g_c^{\text{indiv.}}\}_{c=1}^C$ assume one of $\{\text{Identity}, \text{Linear}, \text{MLP}\}$. Eq. 7 leverages that $\mathbf{z}_{\text{WSI}}^j$, which is a concatenation of mixture estimates, can be decomposed along each prototype and learns per-prototype nonlinear mapping $g_c^{\text{indiv.}}$. This is not possible with a typical MIL: First, the large and variable size of N prohibits the learning of $\{g_n^{\text{indiv.}}\}_{n=1}^N$. Moreover, the permutation invariance makes finding an "appropriate" function $g_n^{\text{indiv.}}$ for a given patch embedding $\mathbf{z}_n$ non-trivial.

3.3. Interpretability

Based on the estimated prototype assignment probability q, we propose two approaches for interpretability. First, for each patch embedding $\mathbf{z}_n^j$, we assign the prototype with the highest posterior probability,

$$c_n^j = \text{argmax}_c \, q(c_n^j = c \mid \mathbf{z}_n^j), \tag{8}$$

and overlay the prototype assignments onto WSI, visualizing how pathology visual concepts are distributed within each WSI (prototypical assignment map, **Fig.** 3), with $\widehat{\pi}_c^j$ quantifying the extent of the distribution. For a specific prototype c', we can also visualize how morphologically similar each patch embedding is to the prototype using $q(c_n^j = c' \mid \mathbf{z}_n^j)$ (**Fig.** 1).

4. Experiments

4.1. Datasets

Subtyping We evaluate PANTHER on four different subtyping tasks: EBRAINS fine subtyping (30 classes) and coarse subtyping (12 classes) for rare brain cancer types [32, 70], Non-Small Cell Lung Carcinoma (NSCLC) subtyping on TCGA and CPTAC (2 classes), and ISUP grading based on Prostate cancer grade assessment (PANDA) challenge (6 classes) [8, 9]. We use balanced accuracy and weighted F1 metrics for evaluation on EBRAINS and NSCLC subtyping task and Cohen's κ for the ISUP grading task.

Survival We evaluate PANTHER on TCGA across several cancer types: Breast Invasive Carcinoma (BRCA), Colon and Rectum Adenocarcinoma (CRC), Bladder Urothelial Carcinoma (BLCA), Uterine corpus endometrial carcinoma (UCEC), Kidney renal clear cell carcinoma (KIRC), and Lung adenocarcinoma (LUAD). For TCGA, we use the 5-fold site-stratified cross-validation. For cancer types with external validation datasets (**KIRC**: CPTAC, **LUAD**: CPTAC, NLST), we use the models trained on TCGA and evaluate on the external dataset. We use the concordance index (c-index) for evaluation. To address the shortcomings of overall survival accounting for non-cancerous deaths [14, 42, 57], we use disease-specific survival (DSS). Additional details can be found in **Supplementary Information**.

4.2. Baselines

We employ 1) *unsupervised* baselines, which use unsupervised slide representation followed by the task-specific linear network, and 2) *supervised* baselines, which construct supervised slide representation for each task. For the *unsupervised* baselines, we use the following: 1) **DeepSets** [93] The slide embedding $\mathbf{z}_{\text{WSI}}^j \in \mathbb{R}^d$ is formed by averaging all the features in the set. 2) **ProtoCounts** [69, 91] K-means clustering is performed on the cohort-aggregated set of features. The slide embedding $\mathbf{z}_{\text{WSI}}^j \in \mathbb{R}^C$ is a count vector of the number of patches assigned to each cluster. 3) **H2T** [81] The patch embeddings are clustered and averaged within each cluster. The averaged cluster centroids are concatenated, with $\mathbf{z}_{\text{WSI}}^j \in \mathbb{R}^{C \cdot d}$. 4) **Optimal Transport (OT)** [64] The patch features of a WSI is aggregated to a set of prototypes with OT [25], with $\mathbf{z}_{\text{WSI}}^j \in \mathbb{R}^{C \cdot d}$. OT assumes uniform mixture probability, i.e., $\pi_c^j = 1/C, \forall c$.

We also implement the following supervised baselines: Attention-based MIL (ABMIL) [40], Transformer-based MIL (TransMIL) [73], Prototype-clustering based MIL (AttnMISL) [88], and low-rank MIL (ILRA) [84].

For PANTHER, we experiment with variations of $\mathbf{z}_{\text{WSI}}^j$ to better understand our model: 1) **All** (original): All mixture parameters are concatenated, $\mathbf{z}_{\text{WSI}}^j \in \mathbb{R}^{C(1+2d)}$. 2) **Weighted avg.** (WA): $\boldsymbol{\mu}_c$ and Σ_c weighted-averaged by π_c and concatenated, $\mathbf{z}_{\text{WSI}}^j \in \mathbb{R}^{2d}$. 3) **Top (Bottom)**: Parameters for mixture component with the largest (smallest) $\widehat{\pi}_c^j$ is selected, $\mathbf{z}_{\text{WSI}}^j \in \mathbb{R}^{(1+2d)}$. We use either linear (+lin.) or nonlinear head (+MLP) on top of $\mathbf{z}_{\text{WSI}}^j$.

For the feature extractor $f_{\text{enc}}(\cdot)$, which is the same for all baselines used in this work, we used UNI [18], a ViT-L/16 DINOv2 [28, 66] that was pre-trained on a large internal histology dataset of 1×10^8 patches from 1×10^6 WSIs*. We also experiment with other feature encoders, CTransPath [83] and ResNet50 [34], the results of which can be found in **Supplementary Information**.

4.3. Implementation

WSIs at $20\times$ magnification ($0.5\,\mu m$/pixel) are patched with non-overlapping patches of 256×256 pixels. For each WSI, we use all patches without sampling. We found that a single EM step is sufficient for convergence. The prototypes $\mathbf{H}$ are constructed from K-means clustering on the set of patches aggregated from the training cohort (all slides) for each task. The same $\mathbf{H}$ is used for AttnMISL, ProtoCounts, H2T, OT, and PANTHER. Additional details on training and loss functions can be found in **Supplementary Information**.

*Accessible at: https://github.com/mahmoodlab/UNI

Table 1. **Subtyping prediction** Results of PANTHER and baselines for four different subtyping tasks. All methods use UNI features [18]. Best performance in **bold**, second best underlined. AttnMISL, ProtoCounts, H2T, OT, and PANTHER use $C = 16$ prototypes.

	Train on	EBRAINS (fine, 32 classes)		EBRAINS (coarse, 12 classes)		TCGA-NSCLC (2 classes)		PANDA (6 classes)	
	Test on	EBRAINS		EBRAINS		TCGA	CPTAC	Karolinska	Radboud
		(Bal. acc.)	(F1)	(Bal. acc.)	(F1)	(Bal. acc.)	(Bal. acc.)	(Cohen's κ)	(Cohen's κ)
Supervised.	ABMIL [40]	0.674	0.744	0.834	0.906	0.949	0.904	0.935	0.918
	TransMIL [73]	**0.701**	0.758	0.848	**0.921**	0.959	0.867	**0.942**	0.922
	DSMIL [55]	0.648	0.698	0.824	0.882	**0.980**	0.791	0.909	0.911
	AttnMISL [88]	0.534	0.636	0.647	0.823	0.888	0.823	0.882	0.894
	ILRA [84]	0.618	0.695	0.820	0.896	0.939	0.887	0.931	0.925
Unsup.	DeepSets [93]	0.033	0.073	0.082	0.2	0.571	0.707	< 0	< 0
	ProtoCounts [69, 91]	0.038	0.018	0.097	0.079	0.429	0.569	< 0	0.13
	H2T [81]	0.117	0.223	0.181	0.421	0.929	0.821	0.457	0.755
	OT [64]	0.700	0.756	0.837	0.915	0.950	0.867	0.817	0.883
Ours	PANTHER$_{\text{Top}}$ + lin.	0.471	0.571	0.554	0.758	0.857	0.833	0.631	0.689
	PANTHER$_{\text{Bot.}}$ + lin.	0.038	0.080	0.138	0.329	0.602	0.705	0.071	0.000
	PANTHER$_{\text{WA}}$ + lin.	0.497	0.598	0.569	0.784	0.888	0.860	0.663	0.787
	PANTHER$_{\text{All}}$ + lin.	0.691	0.756	0.829	0.904	0.939	0.882	0.866	0.909
	PANTHER$_{\text{All}}$ + MLP	0.693	**0.760**	**0.854**	0.908	**0.980**	**0.906**	0.923	**0.931**

5. Results

5.1. Subtyping and survival prediction

Subtyping and survival prediction results are shown in Table 1 and Table 2. Overall, PANTHER consistently outperforms or is on par with all supervised and unsupervised baselines. We highlight key insights and provide hypotheses for the high performance of PANTHER.

PANTHER *vs.* supervised MIL PANTHER$_{\text{All}}$+MLP outperforms or is on par with the best-performing supervised baseline on subtyping (TransMIL) and survival prediction tasks (mix). With linear probing, PANTHER$_{\text{All}}$+lin. remains competitive against MIL on subtyping and performs better on most cancer types in survival prediction, demonstrating the strong representation quality of $\mathbf{z}^j_{\text{WSI}}$. This is encouraging as PANTHER builds a slide representation in an unsupervised fashion, unlike MIL which learns a patch aggregation end-to-end with the downstream tasks. Interestingly, despite relying on a similar prototype construction as PANTHER, AttnMISL [88] performs consistently lower than PANTHER. We attribute this difference to AttnMISL *averaging* the prototypes with attention weights, whereas PANTHER builds a slide embedding by *concatenating* them. Our baseline PANTHER$_{\text{WA}}$+lin. further confirms this by showing that averaging leads to a consistently lower performance.

PANTHER vs. unsupervised baselines We observe that PANTHER$_{\text{All}}$+MLP outperforms most unsupervised baselines on subtyping and survival prediction tasks. We attribute this gain to two design principles behind PANTHER: (1) prototypes are represented as low-dimensional feature vectors, and (2) the resulting slide embedding encodes the cardinality of each prototype, *i.e.,* their extent in the WSI. In comparison, ProtoCounts only encodes the count information, leading to poor performance. Interestingly, DeepSets, which builds slide embeddings as the sum of all patch embeddings, and similarly our baseline PANTHER$_{\text{WA}}$+lin., which takes weighted averaging of the prototype features, lead to poor subtyping performance despite implicitly encoding both deep patch representations and cardinality. We hypothesize that subtyping requires a mechanism to "isolate" discriminative information, which can be implemented using attention (as in ABMIL and TransMIL), or using prototype concatenation as in PANTHER.

Unsurprisingly, H2T and OT come closest to PANTHER on subtyping tasks as they also aggregate the patches to the prototypes (albeit with different mechanisms from PANTHER) and use concatenation. However, on ISUP grading in prostate cancer which is clinically assessed using the primary and secondary Gleason patterns, PANTHER sees significant performance increases. Lastly, H2T and OT lack explicit mechanisms to incorporate cardinality into slide representation, an important feature of PANTHER for interpretability (Section 5.2).

In this context, PANTHER appears as a comprehensive unsupervised slide representation method that concatenates deep prototype representations along with their cardinality.

PANTHER ablations We further ablate PANTHER by retaining only a single component with the highest (lowest) $\widehat{\pi}_c$, *i.e.,* PANTHER$_{\text{Top}}$ (PANTHER$_{\text{Bot.}}$). We observe that both PANTHER$_{\text{Top}}$ and PANTHER$_{\text{Bot.}}$ performs consistently poorly. This reaffirms that capturing morphologi-

Table 2. **Survival prediction** Results of PANTHER and baselines for measuring patient disease-specific survival based on c-index. All methods use UNI features [18]. Best performance in **bold**, second best underlined. All models and prototypes are trained on TCGA. AttnMISL, ProtoCounts, H2T, OT, and PANTHER use $C = 16$ prototypes. Standard deviation (in parentheses) are reported over five runs.

	Dataset	BRCA	CRC	BLCA	UCEC	KIRC		LUAD		
	Test on	TCGA	TCGA	TCGA	TCGA	TCGA	CPTAC	TCGA	CPTAC	NLST
Supervised	ABMIL [40]	0.644 (± 0.05)	0.608 (± 0.09)	0.550 (± 0.06)	0.669 (± 0.07)	0.684 (± 0.06)	0.613 (± 0.06)	0.654 (± 0.06)	0.572 (± 0.03)	0.519 (± 0.04)
	TransMIL [73]	0.612 (± 0.07)	**0.684** (± 0.06)	0.595 (± 0.06)	0.695 (± 0.08)	0.671 (± 0.10)	0.639 (± 0.04)	0.665 (± 0.10)	0.555 (± 0.03)	0.484 (± 0.05)
	DSMIL [55]	0.496 (± 0.00)	0.5 (± 0.00)	0.501 (± 0.00)	0.497 (± 0.00)	0.5 (± 0.00)	0.5 (± 0.00)	0.501 (± 0.00)	0.502 (± 0.00)	0.5 (± 0.00)
	AttnMISL [88]	0.627 (± 0.08)	0.639 (± 0.10)	0.485 (± 0.06)	0.581 (± 0.12)	0.649 (± 0.09)	0.608 (± 0.06)	0.673 (± 0.10)	0.632 (± 0.03)	0.577 (± 0.04)
	ILRA [84]	0.649 (± 0.10)	0.555 (± 0.10)	0.550 (± 0.04)	0.632 (± 0.02)	0.637 (± 0.14)	0.611 (± 0.03)	0.586 (± 0.06)	0.651 (± 0.05)	0.482 (± 0.01)
Unsupervised	DeepSets [93]	0.673 (± 0.11)	0.563 (± 0.10)	0.581 (± 0.05)	0.730 (± 0.05)	0.715 (± 0.08)	0.634 (± 0.01)	0.652 (± 0.05)	0.550 (± 0.01)	0.509 (± 0.04)
	ProtoCounts [69, 91]	0.490 (± 0.11)	0.552 (± 0.06)	0.533 (± 0.09)	0.441 (± 0.06)	0.461 (± 0.06)	0.503 (± 0.09)	0.460 (± 0.11)	0.577 (± 0.11)	0.500 (± 0.01)
	H2T [81]	0.672 (± 0.07)	0.639 (± 0.11)	0.566 (± 0.05)	0.715 (± 0.09)	0.703 (± 0.11)	0.631 (± 0.04)	0.662 (± 0.09)	0.583 (± 0.03)	0.603 (± 0.04)
	OT [64]	0.755 (± 0.06)	0.622 (± 0.09)	0.603 (± 0.04)	0.747 (± 0.08)	0.695 (± 0.09)	0.650 (± 0.02)	**0.687** (± 0.08)	0.641 (± 0.02)	0.495 (± 0.04)
Ours	PANTHER$_{\text{Top}}$ + lin.	0.718 (± 0.11)	0.534 (± 0.08)	0.543 (± 0.05)	0.707 (±0.08)	**0.741** (± 0.11)	0.608 (± 0.01)	0.575 (± 0.05)	0.607 (± 0.02)	0.417 (± 0.06)
	PANTHER$_{\text{Bot.}}$ + lin.	0.578 (± 0.13)	0.452 (± 0.08)	0.494 (± 0.08)	0.570 (± 0.11)	0.524 (± 0.14)	0.532 (± 0.08)	0.592 (± 0.10)	0.539 (± 0.12)	0.519 (± 0.02)
	PANTHER$_{\text{WA}}$ + lin.	0.670 (± 0.09)	0.647 (± 0.12)	0.586 (± 0.04)	0.753 (± 0.09)	0.730 (± 0.07)	0.623 (± 0.01)	0.654 (± 0.07)	0.461 (± 0.01)	0.482 (± 0.06)
	PANTHER$_{\text{All}}$ + lin.	0.722 (± 0.07)	0.645 (± 0.07)	0.602 (± 0.05)	0.751 (± 0.11)	0.703 (± 0.13)	0.649 (± 0.04)	0.672 (± 0.06)	0.568 (± 0.05)	0.623 (± 0.07)
	PANTHER$_{\text{All}}$ + MLP	**0.758** (± 0.06)	0.665 (± 0.10)	**0.612** (± 0.07)	**0.757** (± 0.10)	0.716 (± 0.10)	**0.691** (± 0.03)	0.685 (± 0.06)	**0.653** (± 0.04)	**0.634** (± 0.04)

cal heterogeneity is crucial for accurate prediction of a patient's clinical outcome [63, 80]. That PANTHER$_{\text{Bot.}}$ is the lowest-performing agrees with our intuition, as subtypes and grades are most often determined by pathologists based on visual cues that must be integrated across the entirety of the tumor, rather than utilizing only a particular region or morphology within the whole. Interestingly, PANTHER$_{\text{Top}}$ performs relatively well for the NSCLC subtyping task, which we attribute to it being a relatively simple binary classification task and the most populous component (highest $\widehat{\pi}_c$) for the NSCLC WSIs on average being tumor.

PANTHER with linear vs. non-linear head In all tasks, PANTHER$_{\text{All}}$+MLP consistently boosts the performance over PANTHER$_{\text{All}}$+lin., demonstrating additional predictive capability enabled by per-prototype non-linearity modeling.

5.2. Interpretability

To understand which prototypes are used in learning slide representations, we visualize (1) the patch-level prototype assignments per WSI via probability $q(c_n^j = c \mid \mathbf{z}_n^j)$, and (2) the distribution of mixture components $\widehat{\pi}_c^j$ within and across all WSIs in the cohort.

Overall, we find that GMMs are a simple yet powerful framework for mapping the spatial organization of histologic visual concepts in the tissue microenvironment. Visual assessment by a board-certified pathologist (D.F.K.W.) revealed that prototypical patterns reflect distinct morphological phenotypes of tumor-, tumor-associated stromal, and immune-cell populations as well as normal tissue components (**Fig.** 3C). In NSCLC, we discover prototypes that correspond to adenocarcinoma (turquoise, C2 and C15) and squamous cell carcinoma (orange, C12) patterns (**Fig.** 3A,B). Analysis of $\widehat{\pi}_c$ shows that these patterns almost exclusively appeared in LUAD and LUSC slides (**Fig.** 3C). In CRC, the distribution of prototypical patterns had strong concordance with existing tissue annotations for CRC tissue types (CRC-100K) [47], with further visualizations presented in the **Supplementary Information**.

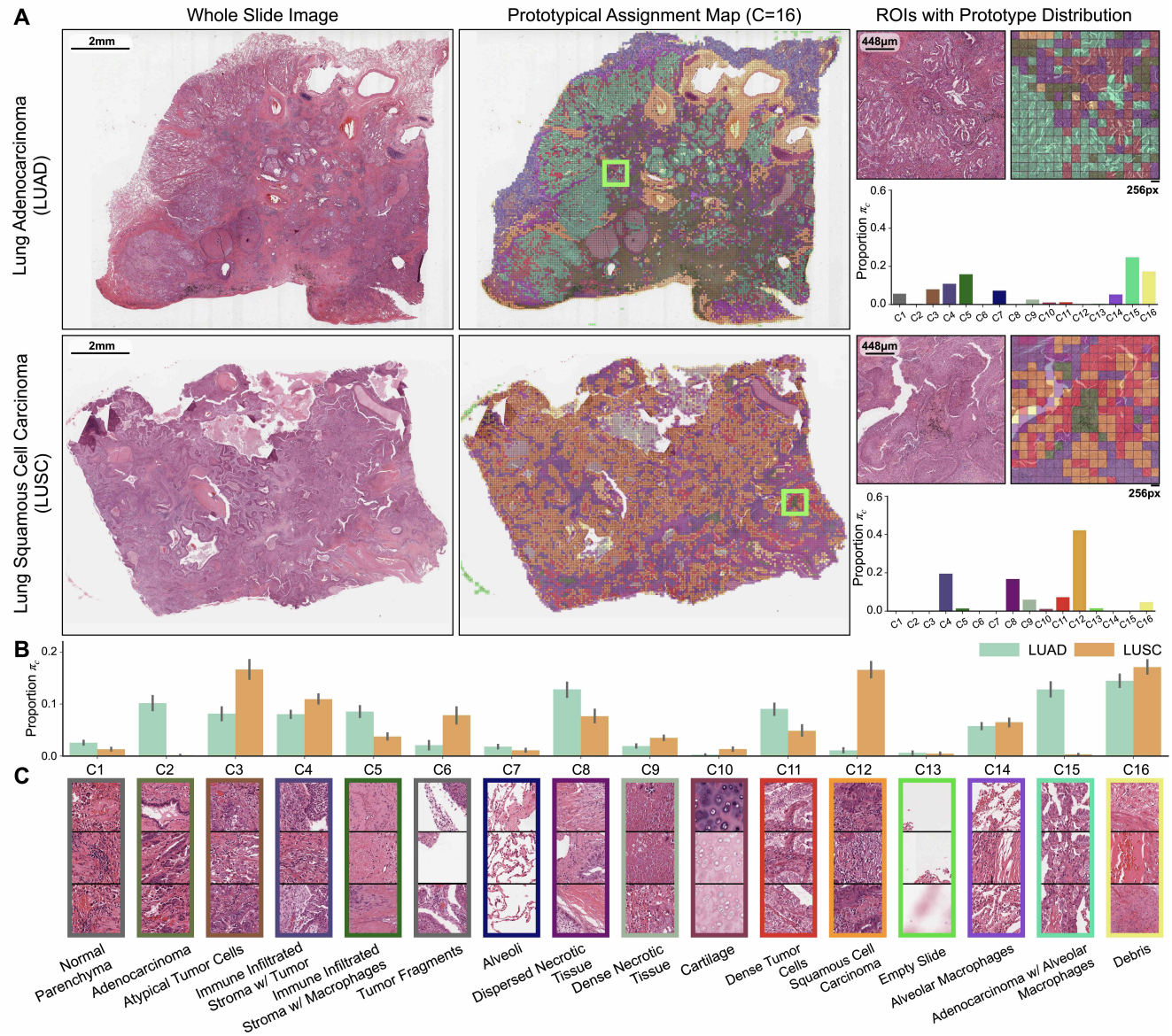

Figure 3. **Prototype-oriented heatmap interpretation**. (A) Examples of WSIs and prototypical assignment maps from LUAD and LUSC, with estimated prototype distribution $\hat{\pi}_c$ for each WSI. (B) Prototype distribution and morphological annotations by a board-certified pathologist in the NSCLC cohort. The adenocarcinoma prototypes (C2, C15) and squamous cell carcinoma (C12) appear exclusively in LUAD and LUSC respectively, showing that PANTHER can correctly capture essential morphological cues in the tissue.

5.3. Further ablations

We run further ablation studies and present additional insights in **Supplementary Information**. Overall, PANTHER is robust across different choices of the number of prototypes C, survival loss functions, and feature encoders.

6. Conclusion and limitations

We present PANTHER, a prototype-based aggregation framework for learning unsupervised slide representations with Gaussian mixtures as the patch distribution. We believe this is an important addition to the emerging group of slide representation studies, with the unsupervised nature of PANTHER making it readily applicable to diverse tasks. Limitations include using $C = 16$ for all tasks, which may lead to over- or under-clustering for certain cancers. Future work includes introducing more expressive mixture models for patch distributions, determining the number of prototypes in a data-driven manner, and evaluating on rare cancer cohorts with small sample sizes.

References

[1] Shekoofeh Azizi, Laura Culp, Jan Freyberg, Basil Mustafa, Sebastien Baur, Simon Kornblith, Ting Chen, Nenad Tomasev, Jovana Mitrović, Patricia Strachan, et al. Robust and data-efficient generalization of self-supervised machine learning for diagnostic imaging. *Nature Biomedical Engineering*, pages 1–24, 2023. 1

[2] Saurav Basu, Soheil Kolouri, and Gustavo K Rohde. Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry. *Proceedings of the National Academy of Sciences*, 111(9):3448–3453, 2014. 2

[3] Andrew H Beck, Ankur R Sangoi, Samuel Leung, Robert J Marinelli, Torsten O Nielsen, Marc J Van De Vijver, Robert B West, Matt Van De Rijn, and Daphne Koller. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. *Science Translational Medicine*, 3(108):108ra113–108ra113, 2011. 2

[4] Babak Ehteshami Bejnordi, Mitko Veta, Paul Johannes Van Diest, Bram Van Ginneken, Nico Karssemeijer, Geert Litjens, Jeroen AWM Van Der Laak, Meyke Hermsen, Quirine F Manson, Maschenka Balkenhol, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, 318(22):2199–2210, 2017. 2

[5] Gianpaolo Bontempo, Angelo Porrello, Federico Bolelli, Simone Calderara, and Elisa Ficarra. Das-mil: Distilling across scales for mil classification of histological wsis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 248–258. Springer, 2023. 2

[6] James D Brierley, Mary K Gospodarowicz, and Christian Wittekind. *TNM classification of malignant tumours*. John Wiley & Sons, 2017. 2

[7] Christian Bueno and Alan Hylton. On the representation power of set pooling networks. *Advances in Neural Information Processing Systems*, 34:17170–17182, 2021. 2

[8] Wouter Bulten, Kimmo Kartasalo, Po-Hsuan Cameron Chen, Peter Ström, Hans Pinckaers, Kunal Nagpal, Yuannan Cai, David F Steiner, Hester van Boven, Robert Vink, et al. Artificial intelligence for diagnosis and gleason grading of prostate cancer: the panda challenge. *Nature Medicine*, 28(1):154–163, 2022. 5, 2

[9] Wouter Bulten, Hans Pinckaers, Hester van Boven, Robert Vink, Thomas de Bel, Bram van Ginneken, Jeroen van der Laak, Christina Hulsbergen-van de Kaa, and Geert Litjens. Automated deep-learning system for gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology*, 21(2):233–241, 2020. 5, 2

[10] Charlotte Bunne, Stefan G Stark, Gabriele Gut, Jacobo Sarabia Del Castillo, Mitch Levesque, Kjong-Van Lehmann, Lucas Pelkmans, Andreas Krause, and Gunnar Rätsch. Learning single-cell perturbation responses using neural optimal transport. *Nature methods*, pages 1–10, 2023. 2

[11] Dmitrii Bychkov, Nina Linder, Riku Turkki, Stig Nordling, Panu E Kovanen, Clare Verrill, Margarita Walliander, Mikael Lundin, Caj Haglund, and Johan Lundin. Deep learning based tissue analysis predicts outcome in colorectal cancer. *Scientific Reports*, 8(1), 2018. 2

[12] Juan C Caicedo, Angel Cruz, and Fabio A Gonzalez. Histopathology image classification using bag of features and kernel functions. In *Artificial Intelligence in Medicine: 12th Conference on Artificial Intelligence in Medicine, AIME 2009, Verona, Italy, July 18-22, 2009. Proceedings 12*, pages 126–135. Springer, 2009. 3

[13] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8):1301–1309, 2019. 1, 2

[14] Iain Carmichael, Andrew H Song, Richard J Chen, Drew FK Williamson, Tiffany Y Chen, and Faisal Mahmood. Incorporating intratumoral heterogeneity into weakly-supervised deep learning models via variance pooling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 387–397. Springer, 2022. 5

[15] Lyndon Chan, Mahdi S Hosseini, Corwyn Rowsell, Konstantinos N Plataniotis, and Savvas Damaskinos. Histosegnet: Semantic segmentation of histological tissue type in whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10662–10671, 2019. 2

[16] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019. 3

[17] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022. 1, 2

[18] Richard J. Chen, Tong Ding, Ming Y. Lu, Drew F. K. Williamson, Guillaume Jaume, Andrew H. Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, Mane Williams, Lukas Oldenburg, Luca L. Weishaupt, Judy J. Wang, Anurag Vaidya, Long Phi Le, Georg Gerber, Sharifa Sahai, Walt Williams, and Faisal Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 2024. 1, 3, 5, 6, 7, 2

[19] Yixin Chen and James Z Wang. Image categorization by learning and reasoning with regions. *The Journal of Machine Learning Research*, 5:913–939, 2004. 3

[20] Zixiang Chen, Yihe Deng, Yue Wu, Quanquan Gu, and Yuanzhi Li. Towards understanding the mixture-of-experts layer in deep learning. *Advances in neural information processing systems*, 35:23049–23062, 2022. 2

[21] Ozan Ciga, Tony Xu, and Anne Louise Martel. Self supervised contrastive learning for digital histopathology. *Machine Learning with Applications*, 7:100198, 2022. 1

[22] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567, 2018. 2

[23] Angel Cruz-Roa, Juan C Caicedo, and Fabio A González. Visual pattern mining in histology image collections using bag of features. *Artificial intelligence in medicine*, 52(2):91–106, 2011. 3

[24] Yufei Cui, Ziquan Liu, Yixin Chen, Yuchen Lu, Xinyue Yu, Xue Liu, Tei-Wei Kuo, Miguel R. D. Rodrigues, Chun Jason Xue, and Antoni B. Chan. Retrieval-augmented multiple instance learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2

[25] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013. 2, 4, 5

[26] Dan dan Guo, Long Tian, Minghe Zhang, Mingyuan Zhou, and Hongyuan Zha. Learning prototype-oriented set representations for meta-learning. In *International Conference on Learning Representations*, 2022. 2, 3

[27] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977. 2, 3, 1

[28] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 5

[29] M Murat Dundar, Sunil Badve, Vikas C Raykar, Rohit K Jain, Olcay Sertel, and Metin N Gurcan. A multiple instance learning approach toward optimal classification of pathology slides. In *2010 20th International Conference on Pattern Recognition*, pages 2732–2735. IEEE, 2010. 1

[30] Kianoush Falahkheirkhah, Alex Xijie Lu, David Alvarez-Melis, and Grace Huynh. Domain adaptation using optimal transport for invariant learning using histopathology datasets. In *Medical Imaging with Deep Learning*, 2023. 2

[31] Alexandre Filiot, Ridouane Ghermi, Antoine Olivier, Paul Jacob, Lucas Fidon, Alice Mac Kain, Charlie Saillard, and Jean-Baptiste Schiratti. Scaling self-supervised learning for histopathology with masked image modeling. *medRxiv*, pages 2023–07, 2023. 1

[32] Gemma Gatta, Riccardo Capocaccia, Laura Botta, Sandra Mallone, Roberta De Angelis, Eva Ardanaz, Harry Comber, Nadya Dimitrova, Maarit K Leinonen, Sabine Siesling, et al. Burden and centralised treatment in europe of rare tumours: results of rarecarenet—a population-based study. *The Lancet Oncology*, 18(8):1022–1039, 2017. 5

[33] Noriaki Hashimoto, Daisuke Fukushima, Ryoichi Koga, Yusuke Takagi, Kaho Ko, Kei Kohno, Masato Nakaguro, Shigeo Nakamura, Hidekata Hontani, and Ichiro Takeuchi. Multi-scale domain-adversarial multiple-instance cnn for cancer subtype classification with unannotated histopathological images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3852–3861, 2020. 2

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[35] Andreas Heindl, Sidra Nawaz, and Yinyin Yuan. Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology. *Laboratory investigation*, 95(4):377–384, 2015. 2

[36] Le Hou, Dimitris Samaras, Tahsin M Kurc, Yi Gao, James E Davis, and Joel H Saltz. Patch-based convolutional neural network for whole slide tissue image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2424–2433, 2016. 3

[37] Wentai Hou, Lequan Yu, Chengxuan Lin, Helong Huang, Rongshan Yu, Jing Qin, and Liansheng Wang. Hˆ 2-mil: exploring hierarchical representation with heterogeneous multiple instance learning for whole slide image analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 933–941, 2022. 2

[38] Frederick M Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo I Olopade, Jakob N Kather, et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature communications*, 12(1):4423, 2021. 2

[39] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas J Montine, and James Zou. A visual–language foundation model for pathology image analysis using medical twitter. *Nature medicine*, 29(9):2307–2316, 2023. 1

[40] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018. 1, 2, 5, 6, 7

[41] Guillaume Jaume, Lukas Oldenburg, Anurag J Vaidya, Richard J Chen, Drew FK Williamson, Thomas Peeters, Andrew H Song, and Faisal Mahmood. Transcriptomics-guided slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 1

[42] Guillaume Jaume, Anurag Vaidya, Richard Chen, Drew Williamson, Paul Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5

[43] Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive mil: intrinsically interpretable multiple instance learning for pathology. *Advances in Neural Information Processing Systems*, 35:20689–20702, 2022. 2

[44] Cheng Jiang, Xinhai Hou, Akhil Kondepudi, Asadur Chowdury, Christian W Freudiger, Daniel A Orringer, Honglak Lee, and Todd C Hollon. Hierarchical discriminative learning improves visual representations of biomedical microscopy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19798–19808, 2023. 1

[45] Shivam Kalra, Hamid R Tizhoosh, Charles Choi, Sultaan Shah, Phedias Diamandis, Clinton JV Campbell, and Liron Pantanowitz. Yottixel–an image search engine for large archives of histopathology whole slide images. *Medical Image Analysis*, 65:101757, 2020. 3

[46] Mingu Kang, Heon Song, Seonwook Park, Donggeun Yoo,

and Sérgio Pereira. Benchmarking self-supervised learning on diverse pathology datasets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3344–3354, 2023. 1

[47] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS Medicine*, 16(1), 2019. 2, 7, 3

[48] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):1–12, 2018. 2

[49] Minyoung Kim. Differentiable expectation-maximization for set representation learning. In *International Conference on Learning Representations*, 2022. 2, 3, 4, 1

[50] Soheil Kolouri, Se Rim Park, Matthew Thorpe, Dejan Slepcev, and Gustavo K. Rohde. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017. 2

[51] Daisuke Komura, Akihiro Kawabe, Keisuke Fukuta, Kyohei Sano, Toshikazu Umezaki, Hirotomo Koda, Ryohei Suzuki, Ken Tominaga, Mieko Ochi, Hiroki Konishi, et al. Universal encoding of pan-cancer histology by deep texture representations. *Cell Reports*, 38(9), 2022. 3

[52] Tristan Lazard, Marvin Lerousseau, Etienne Decencière, and Thomas Walter. Giga-ssl: Self-supervised learning for gigapixel images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4304–4313, 2023. 1

[53] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A sparse texture representation using local affine regions. *IEEE transactions on pattern analysis and machine intelligence*, 27(8):1265–1278, 2005. 3

[54] Juho Lee, Yoonho Lee, Jungtaek Kim, Adam Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pages 3744–3753. PMLR, 2019. 3

[55] Bin Li, Yin Li, and Kevin W Eliceiri. Dual-stream multiple instance learning network for whole slide image classification with self-supervised contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2021. 1, 2, 6, 7, 5

[56] Tiancheng Lin, Zhimiao Yu, Hongyu Hu, Yi Xu, and Chang-Wen Chen. Interventional bag multi-instance learning on whole-slide pathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19830–19839, 2023. 1

[57] Jianfang Liu, Tara Lichtenberg, Katherine A Hoadley, Laila M Poisson, Alexander J Lazar, Andrew D Cherniack, Albert J Kovatich, Christopher C Benz, Douglas A Levine, Adrian V Lee, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018. 5

[58] Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, et al. A visual-language foundation model for computational pathology. *Nature Medicine*, pages 1–12, 2024. 1

[59] Ming Y Lu, Bowen Chen, Andrew Zhang, Drew FK Williamson, Richard J Chen, Tong Ding, Long Phi Le, Yung-Sung Chuang, and Faisal Mahmood. Visual language pretrained multiple instance zero-shot transfer for histopathology images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19764–19775, 2023. 2

[60] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021. 1, 2

[61] Wenqi Lu, Simon Graham, Mohsin Bilal, Nasir Rajpoot, and Fayyaz Minhas. Capturing cellular topology in multi-gigapixel pathology images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 260–261, 2020. 2

[62] Margaux Luck, Tristan Sylvain, Joseph Paul Cohen, Heloise Cardinal, Andrea Lodi, and Yoshua Bengio. Learning to rank for censored survival data. *arXiv preprint arXiv:1806.01984*, 2018. 2

[63] Andriy Marusyk, Michalina Janiszewska, and Kornelia Polyak. Intratumor heterogeneity: the rosetta stone of therapy resistance. *Cancer cell*, 37(4):471–484, 2020. 7

[64] Grégoire Mialon, Dexiong Chen, Alexandre d'Aspremont, and Julien Mairal. A trainable optimal transport embedding for feature aggregation and its relationship to attention. In *International Conference on Learning Representations*, 2021. 2, 3, 4, 5, 6, 7

[65] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018. 2

[66] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 5

[67] Wentao Pan, Jiangpeng Yan, Hanbo Chen, Jiawei Yang, Zhe Xu, Xiu Li, and Jianhua Yao. Human-machine interactive tissue prototype learning for label-efficient histopathology image segmentation. In *International Conference on Information Processing in Medical Imaging*, pages 679–691. Springer, 2023. 3

[68] Shraman Pramanick, Li Jing, Sayan Nag, Jiachen Zhu, Hardik J Shah, Yann LeCun, and Rama Chellappa. VoLTA: Vision-language transformer with weakly-supervised local-feature alignment. *Transactions on Machine Learning Research*, 2023. 2

[69] Adalberto Claudio Quiros, Nicolas Coudray, Anna Yeaton, Xinyu Yang, Bojing Liu, Hortense Le, Luis Chiriboga, Afreen Karimkhan, Navneet Narula, David A. Moore, Christopher Y. Park, Harvey Pass, Andre L. Moreira, John Le Quesne, Aristotelis Tsirigos, and Ke Yuan. Mapping the landscape of histomorphological cancer phenotypes using self-supervised learning on unlabeled, unannotated pathology slides, 2023. 1, 3, 5, 6, 7

[70] Thomas Roetzer-Pejrimovsky, Anna-Christina Moser, Baran Atli, Clemens Christian Vogel, Petra A Mercea, Romana Prihoda, Ellen Gelpi, Christine Haberler, Romana Höftberger, Johannes A Hainfellner, et al. The digital brain tumour atlas, an open histopathology resource. *Scientific Data*, 9(1):55, 2022. 5, 2

[71] Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, et al. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell reports*, 23(1):181–193, 2018. 2

[72] Geoffrey Schiebinger, Jian Shu, Marcin Tabaka, Brian Cleary, Vidya Subramanian, Aryeh Solomon, Joshua Gould, Siyan Liu, Stacie Lin, Peter Berube, et al. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943, 2019. 2

[73] Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021. 1, 2, 5, 6, 7

[74] Sivic and Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proceedings ninth IEEE international conference on computer vision*, pages 1470–1477. IEEE, 2003. 3

[75] Ole-Johan Skrede, Sepp De Raedt, Andreas Kleppe, Tarjei S Hveem, Knut Liestøl, John Maddison, Hanne A Askautrud, Manohar Pradhan, John Arne Nesheim, Fritz Albregtsen, et al. Deep learning for prediction of colorectal cancer outcome: a discovery and validation study. *The Lancet*, 395(10221):350–360, 2020. 2

[76] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 3

[77] Andrew H Song, Guillaume Jaume, Drew FK Williamson, Ming Y Lu, Anurag Vaidya, Tiffany R Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12):930–949, 2023. 1

[78] Wenhao Tang, Sheng Huang, Xiaoxian Zhang, Fengtao Zhou, Yi Zhang, and Bo Liu. Multiple instance learning framework with masked hard instance mining for whole slide image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4078–4087, 2023. 2

[79] Kevin Thandiackal, Boqi Chen, Pushpak Pati, Guillaume Jaume, Drew FK Williamson, Maria Gabrani, and Orcun Goksel. Differentiable zooming for multiple instance learning on whole-slide images. In *European Conference on Computer Vision*, pages 699–715. Springer, 2022. 2

[80] Ilio Vitale, Efrat Shema, Sherene Loi, and Lorenzo Galluzzi. Intratumoral heterogeneity in cancer progression and response to immunotherapy. *Nature Medicine*, 27(2):212–224, 2021. 7

[81] Quoc Dang Vu, Kashif Rajpoot, Shan E. Ahmed Raza, and Nasir Rajpoot. Handcrafted histological transformer (h2t): Unsupervised representation of whole slide images. *Medical Image Analysis*, 85:102743, 2023. 1, 3, 5, 6, 7

[82] Tiep Huu Vu, Hojjat Seyed Mousavi, Vishal Monga, Ganesh Rao, and UK Arvind Rao. Histopathological image classification using discriminative feature-oriented dictionary learning. *IEEE transactions on medical imaging*, 35(3):738–751, 2015. 3

[83] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022. 1, 5, 2

[84] Jinxi Xiang and Jun Zhang. Exploring low-rank property in multiple instance learning for whole slide image classification. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 5, 6, 7

[85] Yingxue Xu and Hao Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21241–21251, 2023. 2

[86] Yan Xu, Jun-Yan Zhu, Eric Chang, and Zhuowen Tu. Multiple clustered instance learning for histopathology cancer image classification, segmentation and clustering. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 964–971. IEEE, 2012. 3

[87] Litao Yang, Deval Mehta, Sidong Liu, Dwarikanath Mahapatra, Antonio Di Ieva, and Zongyuan Ge. TPMIL: Trainable prototype enhanced multiple instance learning for whole slide image classification. In *Medical Imaging with Deep Learning*, 2023. 3

[88] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020. 2, 3, 5, 6, 7

[89] Hangting Ye, Wei Fan, Xiaozhuang Song, Shun Zheng, He Zhao, Dan dan Guo, and Yi Chang. PTaRL: Prototype-based tabular representation learning via space calibration. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[90] Anna Yeaton, Rahul G Krishnan, Rebecca Mieloszyk, David Alvarez-Melis, and Grace Huynh. Hierarchical optimal transport for comparing histopathology datasets. In *Medical Imaging with Deep Learning*, 2022. 2

[91] Jin-Gang Yu, Zihao Wu, Yu Ming, Shule Deng, Yuanqing Li, Caifeng Ou, Chunjiang He, Baiye Wang, Pusheng Zhang, and Yu Wang. Prototypical multiple instance learning for predicting lymph node metastasis of breast cancer from whole-slide pathological images. *Medical Image Analysis*, 85:102748, 2023. 3, 5, 6, 7

[92] Shekoufeh Gorgi Zadeh and Matthias Schmid. Bias in cross-

entropy-based training of deep survival networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3126–3137, 2020. 2

[93] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017. 5, 6, 7

[94] Hongrun Zhang, Yanda Meng, Yitian Zhao, Yihong Qiao, Xiaoyun Yang, Sarah E Coupland, and Yalin Zheng. Dtfd-mil: Double-tier feature distillation multiple instance learning for histopathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18802–18812, 2022. 2

[95] Jianguo Zhang, Marcin Marszałek, Svetlana Lazebnik, and Cordelia Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *International journal of computer vision*, 73:213–238, 2007. 3

[96] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7234–7242, 2017. 2