Multimodal Prototyping for cancer survival prediction

Andrew H. Song [1 2] **Richard J. Chen** [1 2] **Guillaume Jaume** [1 2] **Anurag Vaidya** [1 2 3] **Alexander S. Baras** [4] **Faisal Mahmood** [1 2]

Abstract

Multimodal survival methods combining gigapixel histology whole-slide images (WSIs) and transcriptomic profiles are particularly promising for patient prognostication and stratification. Current approaches involve tokenizing the WSIs into smaller patches ($> 10^4$ patches) and transcriptomics into gene groups, which are then integrated using a Transformer for predicting outcomes. However, this process generates many tokens, which leads to high memory requirements for computing attention and complicates post-hoc interpretability analyses. Instead, we hypothesize that we can: (1) effectively summarize the morphological content of a WSI by condensing its constituting tokens using *morphological* prototypes, achieving more than $300\times$ compression; and (2) accurately characterize cellular functions by encoding the transcriptomic profile with *biological pathway* prototypes, all in an unsupervised fashion. The resulting multimodal tokens are then processed by a fusion network, either with a Transformer or an optimal transport cross-alignment, which now operates with a small and fixed number of tokens without approximations. Extensive evaluation on six cancer types shows that our framework outperforms state-of-the-art methods with much less computation while unlocking new interpretability analyses. The code is available at `https://github.com/mahmoodlab/MMP`.

[1]Mass General Brigham, Boston, MA, USA. [2]Harvard Medical School, Boston, MA, USA. [3]Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Johns Hopkins University School of Medicine, Baltimore, MD, USA.. Correspondence to: Andrew H. Song <asong@bwh.harvard.edu>, Faisal Mahmood <faisalmahmood@bwh.harvard.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

1. Introduction

Patient prognostication – the task of predicting the progression of a disease – is a cornerstone of clinical research and can help identify novel biomarkers indicative of disease progression (Song et al., 2023; 2024b). Prognostication is often cast as predicting survival based on a series of assays describing the patient's medical state. Due to the complexity and diverse aspects of prognostication, multimodal approaches that combine histology and omics data, such as transcriptomics (Acosta et al., 2022), are particularly promising. Histology is represented through whole-slide images (WSIs), which offer a detailed spatial depiction of the tissue, such as a tumor, with resolutions that can exceed $10^5 \times 10^5$ pixels. Differently, transcriptomics is often delineated through bulk RNA sequencing, which provides insights into gene expression. The complementary information in both modalities was shown to be predictive of survival and can be used to inform disease progression (Chen et al., 2022; Lipkova et al., 2022; Steyaert et al., 2023). However, the distinct characteristics of each modality present challenges in effectively integrating them together.

WSI modeling is typically done with multiple instance learning (MIL) (Ilse et al., 2018; Campanella et al., 2019; Lu et al., 2021). This method involves (1) extracting the set of patches that constitute the WSI ($> 10^4$ per WSI), (2) feeding them through a pre-trained patch encoder to generate patch embeddings, and (3) aggregating the patch embeddings with a pooling network. In contrast, transcriptomics modeling can be done using a feed-forward neural network, treating each gene expression as a tabular data entry, or by grouping them into coarse gene families (Chen et al., 2021; Zhou & Chen, 2023; Xu & Chen, 2023) or *biological pathways* (Elmarakeby et al., 2021; Jaume et al., 2024). The set of patch embeddings and gene groups can then be seen as tokens, which can be fed to a Transformer to derive a multimodal representation used for outcome prediction.

However, fusing large numbers of tokens with a Transformer is computationally expensive, and most approaches resort to attention approximation (Shao et al., 2021; Jaume et al., 2024), cross-attention (Chen et al., 2021; Zhou & Chen, 2023; Xu & Chen, 2023), or token subsampling (Wulczyn et al., 2020; Xu & Chen, 2023). Even when employing

alternatives to Transformers, such as optimal transport (OT) cross-modal alignment (Duan et al., 2022; Pramanick et al., 2022), addressing a set of tokens remains challenging. The small size of multimodal cohorts, often just a few hundred samples, intensifies this issue, resulting in a Large-p (large input dimensionality), Small-n (small sample size) problem. Moreover, interpreting how thousands of tokens interact and contribute to patient-level prediction, which is crucial for clinical insights, presents a significant challenge.

Instead, we hypothesize that we can summarize the patch embeddings using *morphological* prototypes. Indeed, due to inherent morphological redundancy in human tissue, the histology patches that constitute the WSI can be assumed as variations of key *morphologies*, *e.g.,* clear cell tumor, necrosis, benign stroma, etc, which we can extract and encode. In molecular pathology, decades of research have identified *biological pathways* that encode specific cellular functions (Liberzon et al., 2015; Elmarakeby et al., 2021), which we can leverage to define *pathway* prototypes. This drastically reduces the number of tokens before multimodal fusion, thereby opening up possibilities for seamless integration of diverse fusion strategies, with interpretability greatly simplified. The challenge then revolves around the extraction and encoding of meaningful multimodal prototypes.

Here, we introduce a **M**ulti**M**odal **P**rototyping framework for patient prognostication (MMP). Inspired by prototype-based aggregation (Mialon et al., 2021; Kim, 2022; Song et al., 2024a), we construct an *unsupervised* and *compact* WSI representation with a Gaussian mixture model, where the mixture parameters define the slide summary, each mapping to a morphological prototype (16 to 32 prototypes). Following existing work (Jaume et al., 2024), we transform transcriptomics into a set of 50 Cancer Hallmark pathway prototypes (Liberzon et al., 2015). With significantly fewer tokens, we show that multimodal Transformers can be readily applied to the joint set of histology and pathway tokens without relying on approximations. In addition, we establish a connection between Optimal Transport (OT) cross-alignment, a popular alternative for cross-modal alignment, and the Transformer cross-attention, thereby unifying both under a single framework. On six cancer cohorts from The Cancer Genome Atlas (TCGA), MMP outperforms nearly all uni- and multimodal baselines with a much smaller number of operations, demonstrating the predictive performance and efficiency of prototype-based approaches. Finally, the tractable number of tokens allows visualization of bi-directional interactions between the morphological and pathway prototypes, different from previous multimodal frameworks relying on uni-directional interpretation.

To summarize, our contributions are (1) a method for summarizing slides using morphological prototypes and summarizing transcriptomic profiles using established biological pathway prototypes; (2) a unified and memory-efficient multimodal fusion framework; (3) extensive evaluation and ablation experiments on six cancer cohorts highlighting the predictive power of the MMP; (4) a novel multimodal patient representation that enables novel interpretability analyses.

2. Related Work

2.1. Representing sets with prototypes

With NLP and bioinformatics producing more datasets represented as sets, recent approaches have explored representation learning of sets with *prototypes*, *i.e.,* interpretable exemplars that can encode distinct concepts (Snell et al., 2017; Lee et al., 2019; Mialon et al., 2021; Kim, 2022; dan Guo et al., 2022; Lee et al., 2024). In computational pathology, AttnMISL (Yao et al., 2020) and H2T (Vu et al., 2023) perform K-means clustering within each WSI and use the cluster centroid embeddings as prototype (hard clustering). The proportion of patches in the cluster has also been used to represent the cluster (Quiros et al., 2023). MMP extends PANTHER (Song et al., 2024a) to a multimodal setting and formalizes the prototype-based set representation with a mathematical treatise and generalizes this concept to include Gaussian mixture models, OT, and clustering.

2.2. Prognostication with multimodal fusion

Late fusion. Early works exploring multimodal survival employed late fusion techniques based on merging unimodal representations, for instance using concatenation or Kronecker product (Chen et al., 2020b). While initial frameworks incorporated histology with small region-of-interests (Mobadersany et al., 2018; Wang et al., 2021), the development of MIL has enabled slide-level prognostication studies combined with omics (Chen et al., 2022; Ding et al., 2023; Volinsky-Fremond et al., 2024). However, late fusion methods are limited in modeling local cross-modal interactions potentially predictive of prognosis.

Transformer fusion. Transformers have facilitated progress in multimodal fusion by modeling interactions between cross-modal tokens (or "early fusion"). To address the computational complexity of dealing with a large number of tokens, token subsampling can be performed (Wulczyn et al., 2020; Xu & Chen, 2023), or fusion can be simplified to cross-attention (Chen et al., 2021; Jaume et al., 2024; Zhou & Chen, 2023). Present strategies allow the use of a single Transformer for merging tokens and facilitating early-fusion across various modalities (Jaegle et al., 2022; Girdhar et al., 2022; Wang et al., 2023; Liang et al., 2023; Wang et al., 2024). However, these works remain limited by the large number of histology tokens ($> 10^4$). In contrast, MMP does not require approximation due to the prototype-based formulation significantly reducing the number of tokens. A recent work (Zhang et al., 2024) employs prototypes to re-

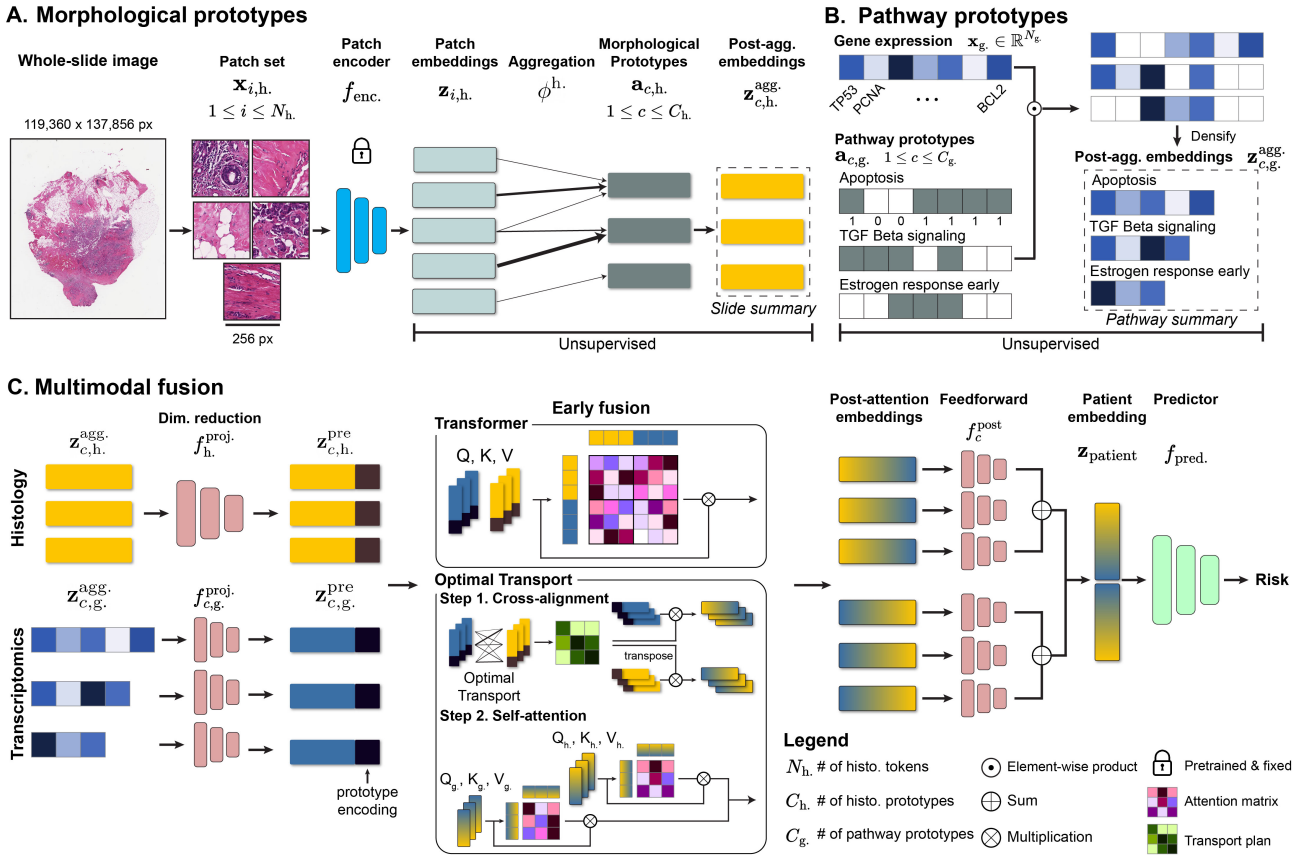

Figure 1. **Overview of MMP**. (A) The tessellated WSI patches (tokens) are projected to low-dimensional embeddings with a pretrained patch encoder. The patch embeddings ($N_{\text{h.}} > 10^4$) are aggregated to *slide summary* using a small set of prototypes ($C_{\text{h.}}$ <32). (B) The transcriptomics data is projected onto a set of binary vectors indicating the presence of specific genes in each pathway, forming *pathway summary*. (C) The post-aggregation embeddings from both modalities are first matched to the same dimension. Cross-modal interactions between histology and transcriptomics are learned with a Transformer or an Optimal Transport, with intra-modal interactions learned with Transformer-based self-attention. The attended embeddings are aggregated to form a patient-level embedding used for risk prediction.

move intra- and inter-modality redundancy, but is specific to a time-discretized survival formulation. In contrast, MMP allows flexible survival problem formulation.

Optimal Transport fusion. OT-based cross-alignment between sets of multimodal tokens (Chen et al., 2020a; Cao et al., 2022; Duan et al., 2022; Pramanick et al., 2023) has gained interest as an alternative to Transformers. In multimodal prognosis, MOTCat (Xu et al., 2023) and MMP replace Transformer-based cross-attention with an OT.

3. Methods

We introduce MMP, a **M**ulti**M**odal **P**rototyping framework for survival prediction. We describe the construction of morphological and pathway prototypes (Section 3.1) and the multimodal fusion mechanism (Section 3.2). We finish by describing survival prediction (Section 3.3) and prototype-specific designs (Section 3.4).

As for notations, z represents a scalar, $\mathbf{z}$ a vector, and $\mathbf{Z}$ a matrix. For a set of vectors $\{\mathbf{z}_c\}_{c=1}^{C} \in \mathbb{R}^d$, $\mathbf{Z} \in \mathbb{R}^{C \times d}$ represents the corresponding matrix with $\mathbf{z}_c$ as the cth row entry. The notation $[\mathbf{X}, \mathbf{Z}]$ is used to indicate concatenation.

3.1. Prototype-based encoding

3.1.1. Morphological prototypes (Histology)

Preprocessing. Given a WSI, we divide it into $N_{\text{h.}}$ non-overlapping patches at 20× magnification (0.5μm/pixel), forming a set of histology patches $\{\mathbf{x}_{i,\text{h.}}\}_{i=1}^{N_{\text{h.}}}$ of varying cardinality, with typically $N_{\text{h.}} > 10^4$. Each $\mathbf{x}_{i,\text{h.}}$ is then mapped to a low-dimensional embedding using a pretrained patch encoder $f_{\text{enc}}(\cdot)$, such that $\mathbf{z}_{i,\text{h.}} = f_{\text{enc}}(\mathbf{x}_{i,\text{h.}}) \in \mathbb{R}^D$.

Aggregation. The standard practice for WSI-based outcome prediction uses Multiple Instance Learning (MIL) to aggregate patch embeddings into a slide embedding, where patch embeddings $\mathbb{S}_{\text{h.}} = \{\mathbf{z}_{i,\text{h.}}\}_{i=1}^{N_{\text{h.}}}$ are pooled using a learnable

function $\phi^{\text{h.}}(\cdot) : \mathbb{S}_{\text{h.}} \to \mathbb{R}^D$, and forms a post-aggregation slide embedding $\mathbf{z}_{\text{h.}}^{\text{agg.}} = \phi^{\text{h.}}(\mathbb{S}_{\text{h.}})$. In the prototype-based approach, we take an alternate route by defining $C_{\text{h.}}$ aggregation functions $\phi_c^{\text{h.}}(\cdot) : \mathbb{S}_{\text{h.}} \to \mathbb{R}^{d_{\text{h.}}}$, such that $\mathbf{z}_{c,\text{h.}}^{\text{agg.}} = \phi_c^{\text{h.}}(\mathbb{S}_{\text{h.}})$, where $C_{\text{h.}}$ is the number of prototypes. This produces a set $\mathbb{S}_{\text{slide}} = \{\mathbf{z}_{c,\text{h.}}^{\text{agg.}}\}_{c=1}^{C_{\text{h.}}}$, referred to as *slide summary*.

Prototypes. We define *prototypes*, denoted as $\{\mathbf{a}_{c,\text{h.}}\}_{c=1}^{C_{\text{h.}}}$, $\mathbf{a}_{c,\text{h.}} \in \mathbb{R}^{d_{\text{h.}}}$, such that each prototype exemplifies a unique morphology from the training set. Specifically, we apply K-means clustering on *all* the patch embeddings from the training dataset to extract the $C_{\text{h.}}$ cluster centroids. The prototypes $\{\mathbf{a}_{c,\text{h.}}\}_{c=1}^{C_{\text{h.}}}$ are then used as parameters for the prototype-specific aggregation function $\phi_c^{\text{h.}}$ (**Fig. 1A**).

Slide summary. Given the prototype $\mathbf{a}_{c,\text{h.}}$ and patch embeddings $\mathbb{S}_{\text{h.}}$, we can express $\phi_c^{\text{h.}}$ as

$$\mathbf{z}_{c,\text{h.}}^{\text{agg.}} = \phi_c^{\text{h.}}(\mathbb{S}_{\text{h.}}, \mathbf{a}_{c,\text{h.}}) = \sum_{i=1}^{N_{\text{h.}}} g(\mathbf{z}_{i,\text{h.}}, \mathbf{a}_{c,\text{h.}}),\ \forall c. \quad (1)$$

This allows the large variable-length set of patch embeddings to be represented using a small fixed-length set comprised of prototypical tokens. We use $C_{\text{h.}} \leq 32$ with typically $N_{\text{h.}} > 10^4$, achieving more than 300× reduction. Eq. 1 implies that aggregation is performed by summing the contribution from all embeddings in $\mathbb{S}_{\text{h.}}$, the exact manner of which is determined by the mapping g.

We explore three strategies for defining g: hard clustering (HC), OT, and our preferred choice, Gaussian Mixture Models (GMM). We briefly explain GMM, similar to the developments in PANTHER (Song et al., 2024a), and defer the detailed explanations of other strategies in **Appendix A**.

GMM-based slide summarization. With GMM as the generative model, the probability distribution for $\mathbf{z}_{i,\text{h.}}$ is

$$\begin{aligned} p(\mathbf{z}_{i,\text{h.}}; \theta) &= \sum_{c=1}^{C_{\text{h.}}} p(c_i = c) \cdot p(\mathbf{z}_{i,\text{h.}} | c_i = c) \\ &= \sum_{c=1}^{C_{\text{h.}}} \pi_c \cdot \mathcal{N}(\mathbf{z}_{i,\text{h.}}; \boldsymbol{\mu}_c, \Sigma_c), \end{aligned} \quad (2)$$

where c_i denotes the mixture identity of $\mathbf{z}_{i,\text{h.}}$ and $\theta = \{\pi_c, \boldsymbol{\mu}_c, \Sigma_c\}_{c=1}^{C_{\text{h.}}}$ denotes mixture probability, mean, and diagonal covariance. Intuitively, $\boldsymbol{\mu}_c$ and π_c represent a morphological exemplar and the proportion of similar patterns in WSI, respectively. The posterior distribution $p(c_i = c|\mathbf{z}_{i,\text{h.}})$ indirectly represents the distance between $\mathbf{z}_{i,\text{h.}}$ and $\mathbf{a}_{c,\text{h.}}$ and consequently its contribution towards each element of the *slide summary*. We obtain the maximum-likelihood estimate, $\widehat{\theta} = \arg\max_\theta \sum_{i=1}^{N_{\text{h.}}} \log p(\mathbf{z}_{i,\text{h.}}; \theta)$, via expectation-maximization (EM), which can be performed as a feedforward network operation (Dempster et al., 1977; Kim, 2022).

The estimated GMM parameters are concatenated to form a post-aggregation embedding, $\mathbf{z}_{c,\text{h.}}^{\text{agg.}} = [\widehat{\pi}_c, \widehat{\boldsymbol{\mu}}_c, \widehat{\Sigma}_c] \in \mathbb{R}^{d_{\text{h.}}}$ with $d_{\text{h.}} = 2D + 1$. During EM-based inference, the prototypes $\mathbf{a}_{c,\text{h.}}$ are used as the initial parameters for the mixture means, $\boldsymbol{\mu}_c^{(0)} = \mathbf{a}_{c,\text{h.}}$ with Σ_c as the identity matrix. Owing to GMM's soft clustering nature, the distribution $p(c_i = c|\mathbf{z}_{i,\text{h.}})$ is non-zero, implying that all elements of $\mathbb{S}_{\text{h.}}$ contribute to $\mathbf{z}_{c,\text{h.}}^{\text{agg.}}$ (Eq. 1). We note that deriving the *slide summary* ($\mathbb{S}_{\text{slide}}$ or $\mathbf{Z}_{\text{h.}}^{\text{agg.}} \in \mathbb{R}^{C_{\text{h.}} \times d_{\text{h.}}}$) from the patch embeddings is done in an *unsupervised* manner, and drastically reduces the input size to the multimodal fusion model.

3.1.2. Pathway prototypes (Genomics)

We aim to define a similar compact prototypical representation of the transcriptomic profile. The transcriptomic profile spans $N_{\text{g.}}$ gene expressions for each tissue, and is described as $\{x_{i,\text{g.}}\}_{i=1}^{N_{\text{g.}}}$, with $x_{i,\text{g.}} \in \mathbb{R}$.

Prototypes. We tokenize gene expression into biological pathway prototypes, *i.e.,* into groups of genes that interact in certain ways to implement previously described cellular processes (Liberzon et al., 2015; Reimand et al., 2019). Unlike histology, the number $C_{\text{g.}}$ (*e.g.,* $C_{\text{g.}}$=50 for Hallmark pathways) and composition of prototypes is fixed and can be defined using existing biological pathway databases.

We define the prototypes as $\{\mathbf{a}_{c,\text{g.}}\}_{c=1}^{C_{\text{g.}}}$, where the binary vector $\mathbf{a}_{c,\text{g.}} \in \{0,1\}^{N_{\text{g.}}}$ with 1 and 0 indicates the presence and absence of a specific gene in the pathway c.

Pathway summary. Denoting $N_{c,\text{g.}}$ as the number of genes in pathway c, we can construct $\mathbf{z}_{c,\text{g.}}^{\text{agg.}} \in \mathbb{R}^{N_{c,\text{g.}}}$, and *pathway summary* $\mathbb{S}_{\text{path.}} = \{\mathbf{z}_{c,\text{g.}}^{\text{agg.}}\}_{c=1}^{C_{\text{g.}}}$ as,

$$\mathbf{z}_{c,\text{g.}}^{\text{agg.}} = \phi^{\text{g.}}(\mathbf{x}_{\text{g.}}, \mathbf{a}_{c,\text{g.}}) = R(\mathbf{x}_{\text{g.}} \odot \mathbf{a}_{c,\text{g.}}) \in \mathbb{R}^{N_{c,\text{g.}}}, \quad (3)$$

where $\mathbf{x}_g \in \mathbb{R}^{N_{\text{g.}}}$ is the vector representation of $\{x_{i,g}\}_{i=1}^{N_{\text{g.}}}$, $\odot$ denotes element-wise multiplication, and R densifies the pathway representation by removing zero elements (**Fig. 1B**). In our work, $N_g \simeq 3 \times 10^4$ and $N_{c,\text{g.}} < 200$, achieving more than 20× reduction.

To summarize, morphological and pathway prototypes are used to extract a *slide summary* $\mathbb{S}_{\text{slide}}$ and a *pathway summary* $\mathbb{S}_{\text{path.}}$. The morphological prototypes are defined in the patch embedding space with fixed-length, $\mathbf{z}_{c,\text{h.}}^{\text{agg.}} \in \mathbb{R}^{d_{\text{h.}}}$, and encode distinct morphological attributes. Differently, the pathway prototypes are defined in the raw data space with variable-length $\mathbf{z}_{c,\text{g.}}^{\text{agg.}} \in \mathbb{R}^{N_{c,\text{g.}}}$, and encode specific biological pathways. We note that both approaches are *unsupervised* and thus not require patient outcomes.

3.2. Multimodal fusion

3.2.1. Token dimension matching

Prior to multimodal fusion, we first match the dimensions of tokens from each modality. We use a linear projection $\mathbf{z}_{c,\text{h.}}^{\text{pre}} = f_{\text{h.}}^{\text{pre}}(\mathbf{z}_{c,\text{h.}}^{\text{agg.}}) \in \mathbb{R}^d$ for histology. For pathways, we use

an MLP or self-normalizing neural networks (SNN) (Klambauer et al., 2017) $f^{\text{pre}}_{c,\text{g.}}$ per prototype to map variable-length representations to a common length, $\mathbf{z}^{\text{pre}}_{c,\text{g.}} = f^{\text{pre}}_{c,\text{g.}}(\mathbf{z}^{\text{agg.}}_{c,\text{g.}}) \in \mathbb{R}^d$. The parameters of $f^{\text{pre}}_{\text{h.}}$ and $\{f^{\text{pre}}_{c,\text{g.}}\}_{c=1}^{C_{\text{g.}}}$ are learned for each downstream task.

3.2.2. Multimodal fusion

Inspired by multimodal early fusion methods, we learn dense intra- and cross-modal interactions between the histology and pathway tokens. We explore two strategies: Transformer attention and OT cross-alignment (**Fig. 1C**).

Transformer attention. We introduce three learnable query, key, value matrices $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{d\times d}$. Denoting $\mathbf{Q} = (\mathbf{Q}^{\text{T}}_{\text{g.}}\ \mathbf{Q}^{\text{T}}_{\text{h.}})^{\text{T}} = \left(\mathbf{Z}^{\text{pre,T}}_{\text{g.}}\ \mathbf{Z}^{\text{pre,T}}_{\text{h.}}\right)^{\text{T}}\mathbf{W}_Q \in \mathbb{R}^{(C_{\text{g.}}+C_{\text{h.}})\times d}$, and likewise for $\mathbf{K}$ and $\mathbf{V}$, we can define the standard Transformer attention (Vaswani et al., 2017; Xu et al., 2023)

$$\begin{aligned}\mathbf{Z}^{\text{post}}_{\text{g.+h.}} &= \begin{pmatrix}\mathbf{Z}^{\text{post}}_{\text{g.}}\\ \mathbf{Z}^{\text{post}}_{\text{h.}}\end{pmatrix} = \sigma\left(\frac{\mathbf{Q}\mathbf{K}^{\text{T}}}{\sqrt{d}}\right)\mathbf{V} \in \mathbb{R}^{(C_{\text{g.}}+C_{\text{h.}})\times d}\\ &= \sigma\left(\frac{1}{\sqrt{d}}\begin{pmatrix}\mathbf{Q}_{\text{g.}}\mathbf{K}^{\text{T}}_{\text{g.}} & \mathbf{Q}_{\text{g.}}\mathbf{K}^{\text{T}}_{\text{h.}}\\ \mathbf{Q}_{\text{h.}}\mathbf{K}^{\text{T}}_{\text{g.}} & \mathbf{Q}_{\text{h.}}\mathbf{K}^{\text{T}}_{\text{h.}}\end{pmatrix}\right)\begin{pmatrix}\mathbf{V}_{\text{g.}}\\ \mathbf{V}_{\text{h.}}\end{pmatrix},\end{aligned} \tag{4}$$

where $\sigma(\cdot)$ denotes row-wise softmax. Eq. 4 illustrates how multimodal attention can be decomposed into the intra-modal self-attention (g. $\rightarrow$ g., h. $\rightarrow$ h.) and cross-modal cross-attention (g. $\rightarrow$ h., h. $\rightarrow$ g.). In MMP, the complexity of computing attention is simplified to $\mathcal{O}((C_{\text{g.}} + C_{\text{h.}})^2)$, a considerable reduction from $\mathcal{O}((N_{\text{g.}} + N_{\text{h.}})^2)$ in most multimodal fusion methods that do not use prototyping.

Optimal Transport cross-alignment Modeling cross-modal interactions can also be approached from the point of view of OT, where we aim to learn the transport plan $\mathbf{T} \in \mathbb{R}_+^{C_{\text{g.}}\times C_{\text{h.}}}$ with the minimal total cost between the empirical distributions $\hat{p}(\mathbf{z}^{\text{pre}}_{\text{g.}}) = \frac{1}{C_{\text{g.}}}\sum_{c=1}^{C_{\text{g.}}}\delta(\mathbf{z}^{\text{pre}}_{c,\text{g.}})$ and $\hat{p}(\mathbf{z}^{\text{pre}}_{\text{h.}}) = \frac{1}{C_{\text{h.}}}\sum_{c'=1}^{C_{\text{h.}}}\delta(\mathbf{z}^{\text{pre}}_{c',\text{h.}})$, and where $\delta(\cdot)$ is a delta function. The pairwise cost $\mathbf{D}_{c,c'}$ between the two tokens is typically computed using a L_2 distance or negative dot product. The estimate $\widehat{\mathbf{T}}$ is given as the solution to the entropic-regularized OT problem (Kolouri et al., 2017),

$$\begin{aligned}&\min_{\mathbf{T}}\sum_{c,c'}\mathbf{D}_{c,c'}\cdot\mathbf{T}_{c,c'} + \varepsilon\mathbf{T}_{c,c'}\log\mathbf{T}_{c,c'}\\ &\text{s.t.}\sum_{c=1}^{C_{\text{g.}}}\mathbf{T}_{c,c'} = 1/C_{\text{h.}} \quad\text{and}\quad \sum_{c'=1}^{C_{\text{h.}}}\mathbf{T}_{c,c'} = 1/C_{\text{g.}},\end{aligned} \tag{5}$$

where ε is the regularization parameter. The optimal plan $\widehat{\mathbf{T}}$ can be obtained with the Sinkhorn algorithm (Cuturi, 2013), which can be differentiated (Genevay et al., 2018). This enables joint learning of $f^{\text{pre}}_{\text{h.}}$ and $\{f^{\text{pre}}_{c,\text{g.}}\}_{c=1}^{C_{\text{g.}}}$ along with the plan $\widehat{\mathbf{T}}$. Cross-alignment with $\widehat{\mathbf{T}}$, *i.e.*, $\widehat{\mathbf{T}}\mathbf{Z}^{\text{pre}}_{\text{h.}}$, performs h. $\rightarrow$ g. attention, while $\widehat{\mathbf{T}}^{\text{T}}$ performs g. $\rightarrow$ h. attention, $\widehat{\mathbf{T}}^{\text{T}}\mathbf{Z}^{\text{pre}}_{\text{g.}}$. After the alignment, we learn intra-modal interactions using the Transformer self-attention. Denoting $\mathbf{Q}_{\text{g.}} = (\widehat{\mathbf{T}}\mathbf{Z}^{\text{pre}}_{\text{h.}})\mathbf{W}_Q$ and $\mathbf{Q}_{\text{h.}} = (\widehat{\mathbf{T}}^{\text{T}}\mathbf{Z}^{\text{pre}}_{\text{g.}})\mathbf{W}_Q$, and likewise for $\mathbf{K}_{\text{g.}}, \mathbf{V}_{\text{g.}}, \mathbf{K}_{\text{h.}}, \mathbf{V}_{\text{h.}}$, we obtain

$$\mathbf{Z}^{\text{post}}_{\text{h.}} = \sigma\left(\frac{\mathbf{Q}_{\text{h.}}\mathbf{K}^{\text{T}}_{\text{h.}}}{\sqrt{d}}\right)\mathbf{V}_{\text{h.}},\ \mathbf{Z}^{\text{post}}_{\text{g.}} = \sigma\left(\frac{\mathbf{Q}_{\text{g.}}\mathbf{K}^{\text{T}}_{\text{g.}}}{\sqrt{d}}\right)\mathbf{V}_{\text{g.}}. \tag{6}$$

3.2.3. Connection between Transformer and Optimal transport cross-alignment

The Transformer cross-attention and the OT cross-alignment exhibit similarities in the way attention and the transport plan are being modeled. We can formalize these similarities to demonstrate the connection between the two. Specifically, we show that the Transformer cross-attention is similar to OT cross-alignment, under certain conditions.

Lemma 3.1. *Let $\mathbf{Z}_{g.} \in \mathbb{R}^{C_{g.}\times d}$ and $\mathbf{Z}_{h.} \in \mathbb{R}^{C_{h.}\times d}$ be the matrix representation of the token sets $\{\mathbf{z}_{i,g.}\}_{i=1}^{C_{g.}}$ and $\{\mathbf{z}_{k,h.}\}_{k=1}^{C_{h.}}$. Let $\mathbf{Z}_{g.}\mathbf{W}_Q^T \in \mathbb{R}^{C_{g.}\times d}$ and $\mathbf{Z}_{h.}\mathbf{W}^T \in \mathbb{R}^{C_{h.}\times d}$ be the linear projections of both sets. Let $\widehat{\mathbf{T}} \in \mathbb{R}_+^{C_{g.}\times C_{h.}}$ be the optimal transport plan, i.e., the solution to the entropic-regularized, unbalanced optimal transport problem between the two projected sets. Then, $\widehat{\mathbf{T}}$ is equivalent to the Transformer cross-attention matrix, $\sigma(\mathbf{Z}_{g.}\mathbf{W}_Q^T\mathbf{W}\mathbf{Z}_{h.}^T/\sqrt{d})$, up to a multiplicative factor where $\sigma(\cdot)$ denotes row-wise softmax, $\{\mathbf{W}_Q\mathbf{z}_{i,g.}\}_{i=1}^{C_{g.}}$ are queries, and $\{\mathbf{W}\mathbf{z}_{k,h.}\}_{k=1}^{C_{h.}}$ are keys.*

Proof. The detailed derivation can be found in **Appendix B**. □

This lays the groundwork for MMP to integrate both approaches within a single framework, rather than regarding them as fundamentally distinct approaches. This offers a platform for future innovations in multimodal strategies.

3.3. Survival prediction

The post-attention embeddings are subject to a sequence of post-attention feedforward network f^{post} with layer normalization (LN), averaging within each modality, and concatenation to form a patient-level embedding $\mathbf{z}_{\text{patient}} = \left[\sum_{c=1}^{C_{\text{g.}}}\text{LN}(f^{\text{post}}(\mathbf{z}^{\text{post}}_{c,\text{g.}})), \sum_{c=1}^{C_{\text{h.}}}\text{LN}(f^{\text{post}}(\mathbf{z}^{\text{post}}_{c,\text{h.}}))\right]$. The resulting embedding is fed through a linear predictor $f_{\text{pred.}}$ for patient-level risk prediction.

We use the Cox proportional hazards loss (Cox, 1972; Katzman et al., 2018; Carmichael et al., 2022), which requires training in batches to preserve the risk order within a patient group. Due to the large $N_{\text{h.}}$, it is computationally challenging to form a batch for non-prototype approaches. **Appendix C** provides a detailed explanation of losses.

3.4. Enhancing prototypes

Given that the identities of the prototypes remain consistent across patients – *e.g.,* prototype c consistently represents the same morphological concept or pathway – we can additionally inject this property into model design considerations. Specifically, we incorporate (1) a prototype-specific encoding and (2) a post-attention feed-forward network.

Prototype encoding, denoted as $\mathbf{e}_c$, can be connected to modality-specific encodings (Jaegle et al., 2022; Liang et al., 2023). Specifically, we append the encodings to the embeddings before feeding them to the fusion network. We experiment with two approaches: 1) fixed one-hot encoding $\mathbf{e}_c \in \{0,1\}^{d_e}$ with $d_e = C_{\text{g.}} + C_{\text{h.}}$ and 2) random-initialized and learnable embedding $\mathbf{e}_c \in \mathbb{R}^{d_e}$ with $d_e = 32$. The modified embeddings are then given as $\mathbf{z}^{\text{pre}}_{c,\text{h.}} = [\mathbf{z}^{\text{pre}}_{c,\text{h.}}, \mathbf{e}_c] \in \mathbb{R}^{d+d_e}$, and same for $\mathbf{z}^{\text{pre}}_{c,\text{g.}}$.

We also employ prototype-specific feedforward network (FFN) f^{post}_c to the post-attention embeddings to learn additional nonlinearity per prototype. This differs from previous works that share f^{post}, which might limit expressivity. These components cannot be used for non-prototype frameworks, since 1) the patch identity is not preserved across patients and 2) large $N_{\text{h.}}$ makes the use of $\mathbf{e}_c$ and f^{post}_c infeasible.

4. Experiments

4.1. Datasets

We use publicly available The Cancer Genome Atlas (TCGA) to evaluate MMP across six cancer types: Bladder urothelial carcinoma (BLCA) ($n = 359$), Breast invasive carcinoma (BRCA) ($n = 868$), Lung adenocarcinoma (LUAD) ($n = 412$), Stomach adenocarcinoma (STAD) ($n = 318$), Colon and Rectum adenocarcinoma (CRC) ($n = 296$), and Kidney renal clear cell carcinoma (KIRC) ($n = 340$). We train the models to predict risks for disease-specific survival (DSS) (Liu et al., 2018). Following standard practice, we use 5-fold site-stratified cross-validation to mitigate batch effect (Howard et al., 2021). We evaluate MMP with the concordance index (C-Index), which measures the concordance between the ordering based on patients' survival days and the predicted risks.

Log-2 transformed transcripts per million bulk RNA sequencing expression for all TCGA cohorts is accessed through UCSC Xena database (Goldman et al., 2020). The $C_{\text{g.}} = 50$ Hallmark gene sets from the Molecular Signatures Database (MSigDB) (Subramanian et al., 2005; Liberzon et al., 2015) are used to select and organize genes into biological pathways. Hallmark gene sets (pathways) represent well-defined biological states in cancer. After organizing genes into Hallmark pathways, we obtained 4,241 unique genes across the 50 pathways, with a minimum and maximum pathway size of 31 and 199, respectively. More dataset details can be found in **Appendix D**.

4.2. Baselines

Histology. We employ Attention-based MIL (ABMIL) (Ilse et al., 2018), ABMIL with information bottleneck (ABMIL-IB) (Li et al., 2023), Transformer-based MIL (Trans-MIL) (Shao et al., 2021), low-rank MIL (ILRA) (Xiang & Zhang, 2023), and prototype-based MIL (AttnMISL) (Yao et al., 2020). We also use the unimodal version of MMP.
Transcriptomics. We employ a feed-forward neural network (2-layer MLP) (non-prototype) and a baseline with pathway-specific SNNs (Jaume et al., 2024; Zhang et al., 2024), followed by concatenation.
Multimodal. We use MCAT (Chen et al., 2021), SurvPath (Jaume et al., 2024), MOTCat (Xu & Chen, 2023), and CMTA (Zhou & Chen, 2023), which all use multimodal tokenization to derive histology and omics tokens (pathways in our evaluation), followed by co-attention Transformer.
MMP variants. We test MMP with a Transformer cross-attention ($\text{MMP}_{\text{Trans.}}$) and OT cross-alignment (MMP_{OT}). The rest of the model comprises GMM histology aggregation with $C_{\text{h.}} = 16$, learnable random prototype encoding, and prototype-specific feedforward networks.

For the patch encoder, we use UNI (Chen et al., 2024), a DINOv2-based ViT-Large (Dosovitskiy et al., 2021; Oquab et al., 2023) pretrained on 1×10^8 patches sampled across 1×10^5 WSIs from Mass General Brigham. We also ablate with CTransPath (Wang et al., 2022), a Swin Transformer pretrained on 3.2×10^4 WSIs from the TCGA, and ResNet50 pretrained on Imagenet (Deng et al., 2009). Further information on all baselines can be found in **Appendix E**.

4.3. Implementation

All models are trained with a 1×10^{-4} learning rate with cosine decay scheduler, AdamW optimizer, and 1×10^{-5} weight decay for 20 epochs. MMP uses the Cox loss with a batch size of 64. Non-prototype baselines are trained with the NLL survival loss (Zadeh & Schmid, 2020) with a batch size of 1. During training, in MCAT, SurvPath, MOTCat, and CMTA, we randomly sample 4,096 patches per WSI to increase diversity and reduce memory. During inference, the whole WSI is used. All prototype baselines use $C_{\text{h.}} = 16$.

5. Results

5.1. Survival prediction

The results are shown in **Table 1**. Overall, MMP outperforms all baselines (+5.4% and +7.8% avg. over the next-best multimodal and unimodal models) and ranks within top-2 for 5 out of 6 diseases. We highlight the main findings.

Table 1. **Survival prediction** Results for MMP and other baselines for measuring disease-specific survival with C-Index. The clinical baseline includes age, sex, and cancer grade as reported in the TCGA cohort. We use the same histology feature encoder, UNI, a ViT-L/16 model pretrained on an internal histology dataset (Chen et al., 2024). All histology prototype-based methods share the same set of morphological prototypes with $C_{\text{h.}} = 16$. Standard deviation is reported over five runs. m.p. and p.p. denote morphological prototype and pathway prototype, respectively. The best and second-best performances are denoted by **bold** and underlined, respectively.

	Dataset	m.p.	p.p.	BRCA	BLCA	LUAD	STAD	CRC	KIRC	Avg. (↑)
	Clinical			0.563±0.055	0.570±0.033	0.528±0.028	0.592±0.044	0.655±0.119	0.602±0.066	0.585
gene	Gene exp.			0.638±0.090	0.627±0.055	0.577±0.057	0.562±0.083	0.588±0.105	0.681±0.072	0.612
gene	Pathways		✓	0.615±0.054	0.606±0.084	0.626±0.077	0.566±0.080	0.590±0.104	0.681±0.090	0.614
histology	ABMIL			0.570±0.086	0.550±0.039	0.571±0.036	0.559±0.059	0.660±0.096	0.684±0.115	0.599
histology	TransMIL			0.601±0.110	0.584±0.057	0.547±0.054	0.487±0.057	0.555±0.059	0.678±0.191	0.575
histology	AttnMISL	✓		0.599±0.117	0.493±0.064	0.627±0.076	0.533±0.040	**0.728**±0.110	0.648±0.102	0.605
histology	IB-MIL			0.511±0.068	0.524±0.051	0.578±0.067	0.525±0.061	0.576±0.129	0.702±0.081	0.569
histology	ILRA			0.597±0.124	0.581±0.055	0.511±0.077	0.550±0.094	0.643±0.124	0.651±0.164	0.589
histology	MMP	✓		0.669±0.119	0.593±0.062	0.600±0.039	0.488 ±0.093	0.646±0.111	0.701±0.177	0.611
Multimodal	MCAT		✓	0.648±0.100	0.619±0.048	0.615±0.072	0.528±0.114	0.578±0.136	0.670±0.235	0.610
Multimodal	SurvPath		✓	0.709±0.062	0.619±0.052	0.612±0.060	0.556±0.136	0.539±0.150	0.738±0.131	0.629
Multimodal	MOTCat		✓	0.717±0.029	0.622±0.064	0.589±0.059	0.561±0.075	0.590±0.130	0.708±0.104	0.631
Multimodal	CMTA		✓	0.687±0.077	0.605±0.076	0.622±0.059	0.547±0.088	0.559±0.195	0.720±0.124	0.623
Multimodal	**MMP$_{\text{OT}}$**	✓	✓	**0.753**±0.069	0.628±0.064	**0.643**±0.013	0.580±0.071	0.636±0.120	**0.748**±0.099	**0.665**
Multimodal	**MMP$_{\text{Trans.}}$**	✓	✓	0.738±0.069	**0.635**±0.051	0.642±0.037	**0.598**±0.051	0.630±0.125	0.747±0.106	**0.665**

Comparison with clinical baseline. All multimodal baselines perform superior to the clinical baseline comprised of important prognostic variables – age, sex, and cancer grade (Bonnier et al., 1995; Rakha et al., 2010; Tas et al., 2013). This demonstrates the clinical potential of multimodal frameworks for enhanced patient prognostication. Additional univariate clinical baselines are in **Appendix F**.

Unimodal vs. Multimodal. All multimodal baselines (excluding MCAT) outperform the unimodal baselines (histology and transcriptomics). This aligns with previous multimodal literature showing that histology and transcriptomics contain complementary information to be leveraged for better prognostication. In addition for CRC, we observe unimodal histology baselines outperforming multimodal baselines, indicating that challenges remain in multimodal training dynamics of histology-omic models (Gat et al., 2020; Wang et al., 2020).

Prototypes vs. non-prototypes. MMP significantly outperforms all multimodal approaches that are based on prototyping (+5.4% avg. over the next-best model, MOTCat). While every multimodal baseline utilizes early fusion, MCAT and MOTCat learn only the uni-directional cross-modal interaction from transcriptomics to histology. Conversely, SurvPath omits histology-to-histology interactions in self-attention computation to reduce computational requirements. Overall, we attribute the superior performance of MMP to our ability to 1) retain and encode morphological information predicted of prognosis in the morphological prototypes, 2) model both the intra- and cross-modal interactions without approximations, and 3) employ the Cox survival loss. The quality of the morphological prototypes is reaffirmed in the unimodal setting, where the prototype-based AttnMISL and the unimodal MMP are the two best-performing models, outperforming all other approaches.

Transformer vs. OT-based cross-attention. We observe that the performance of MMP$_{\text{Trans.}}$ and MMP$_{\text{OT}}$ are on the same level, empirically confirming the connection between both approaches highlighted in Section 3.2.3.

5.2. Risk stratification

We perform log-rank tests (Bland & Altman, 2004) between the high-risk and low-risk cohorts, stratified at 50% percentile of the risks predicted by MMP and MOTCat, the next-best performing model. Specifically, we aggregate the predicted risks across all test folds to construct the cohort-level risk set. **Table 2** shows the p-values for the log-rank tests on all 6 cancer types. With a statistical significance threshold of 0.05, we observe that MMP can significantly stratify the high- and low-risk groups for all 6 cancer types, whereas MOTCat was significant for 3 cancer types. This demonstrates the strength of MMP for risk stratification over other baselines and reaffirms its clinical potential.

5.3. Ablation study

We perform extensive ablations of MMP (**Table 3**). We summarize our findings below. (1) **Number of morphological prototypes**: Larger number of morphological prototypes ($C_{\text{h.}}$) generally yields better performance up to 16, then performance stagnates. To facilitate easier interpretation with fewer exemplars, we set $C_{\text{h.}} = 16$. (2) **Feature encoder**: UNI, a DINOv2-pretrained ViT-L encoder, substantially im-

Table 2. **Risk stratification.** We report log-rank p-values for high- and low-risk patient cohorts for MMP and MOTCat. p-values below 0.05 are considered statistically significant.

	BRCA	BLCA	LUAD
MOTCAT	6.16×10^{-5}	8.60×10^{-5}	9.65×10^{-1}
MMP$_{\text{Trans.}}$	3.08×10^{-5}	4.50×10^{-2}	8.37×10^{-5}
	STAD	**CRC**	**KIRC**
MOTCAT	5.70×10^{-2}	7.04×10^{-1}	2.40×10^{-4}
MMP$_{\text{Trans.}}$	1.40×10^{-2}	3.60×10^{-2}	2.59×10^{-8}

proves compared to CTransPath and ResNet50 pretrained on ImageNet. This underscores the importance of a powerful vision encoder trained on large histology datasets. (3) **Histology aggregation**: Aggregation based on a GMM yields the best performance over optimal transport (OT) and hard clustering (HC). We hypothesize that this is due to GMM explicitly capturing sufficient statistics of patch embedding distribution, *e.g.,* mixture probability and covariance, which other approaches cannot readily integrate. (4) **Prototype encoding**: Adding prototype encoding, $\mathbf{e}_c$ and f_c^{post}, leads to better performance. This suggests the benefits of a prototype-specific measure that leverages the consistent prototype identity. (5) **Fusion stage**: Early-fusion of tokens via cross-attention (MMP) outperforms late-fusion, which concatenates the self-attended embeddings averaged within each modality, without capturing cross-modal interactions.

We also perform unimodal MMP ablations in **Appendix G**, to isolate the impact of histology-related design choices.

Table 3. **Ablation study.** C-Index and its change against MMP as a single model component is modified, averaged across six cohorts.

Ablation	Model			Avg.
Full model	MMP			**0.665**
Number of histo. proto.	$C_{\text{h.}} = 16$	$\Rightarrow$	$C_{\text{h.}} = 8$	0.655 (−1.5%)
			$C_{\text{h.}} = 32$	0.662 (−0.5%)
Histo. enc. f_{enc}	UNI	$\Rightarrow$	ResNet50	0.620 (−6.8%)
			CTransPath	0.643 (−3.3%)
Histo. agg. $\phi_c^{\text{h.}}$	GMM	$\Rightarrow$	OT	0.658 (−1.1%)
			HC	0.629 (−5.4%)
Proto. embed. $\mathbf{e}_c$	random	$\Rightarrow$	None	0.652 (−2.0%)
			One-hot	0.660 (−0.8%)
FFN f_c^{post}	Indiv.	$\Rightarrow$	Shared	0.658 (−1.1%)
Co-attention	Trans.	$\Rightarrow$	OT	**0.665** (−0.0%)
Fusion	Early	$\Rightarrow$	Late	0.646 (−2.9%)

5.4. Loss function

We evaluate the performance of MMP trained using either the Cox or the NLL loss (**Table 9** in **Appendix H**). The NLL loss has widely been used for multimodal frameworks as it accommodates training with a batch of a single patient, a necessity for managing many tokens. Conversely, the Cox loss requires a batch size greater than a single patient, which involves ordering patients within the batch. Applying both losses is feasible due to the reduced computational requirements in MMP. We observe that the Cox loss surpasses NLL loss overall (average C-Index of 0.665 *vs.* 0.644). Furthermore, increasing the batch size (bs) with NLL leads to enhanced performance (0.621 with bs=1 *vs.* 0.644 with bs=16), emphasizing the benefit of the reduced tokens.

5.5. Computational complexity

To assess the computational benefits of MMP, we measure the number of floating-point operations (FLOPs) for cross-attention baselines (**Table 4**). MMP achieves at least 5× fewer giga-FLOPS, demonstrating the superior efficiency of prototyping. We observe that the aggregation (MMP$_{\text{agg.}}$), which maps $N_{\text{h.}}$ tokens to $C_{\text{h.}}$ prototypes, constitutes most of the MMP operations, with the fusion (MMP$_{\text{fusion}}$) requiring significantly less due to condensed token set.

Table 4. **Computational complexity.** Average number of tokens per WSI and average number of giga-FLOPs per patient.

	LUAD		KIRC	
	tokens	GFLOPs (↓)	tokens	GFLOPs (↓)
MCAT	4,714	2.10	12,802	5.49
SurvPath	4,714	2.00	12,802	5.41
CMTA	4,714	17.2	12,802	40.1
MMP$_{\text{agg.}}$	4,714	0.309	12,802	0.839
MMP$_{\text{fusion}}$	16	0.025	16	0.025
MMP$_{\text{total}}$	·	**0.334**	·	**0.864**

6. Interpretability

Unimodal: As WSIs are represented with a compact set of 16 prototypes in MMP, we can directly visualize a prototype heatmap for prototype c that corresponds to the most similar patches, by relying on the posterior $p(c_i = c|\mathbf{z}_{i,\text{h.}})$ (**Fig. 2A, B**, with additional examples in **Appendix I**). The prototype assignment map can clearly show how all the prototypes are distributed in a given WSI (**Fig. 2C**). For a prototype c, we can also visualize the most representative patches, by querying the patch embeddings closest to $\widehat{\boldsymbol{\mu}}_c$, and its proportion in the WSI, with $\widehat{\pi}_c$ (**Fig. 2D**).

Multimodal: With a tractable number of histology tokens, we can visualize the cross-modal attention interactions based on cross-attention scores, from histology to pathways (h. → g.) and pathways to histology (g. → h.). In contrast to MCAT and MOTCat, which only model and visualize g. → h. interactions, *i.e.,* which patches correspond to the queried pathway (histology importance), MMP can also

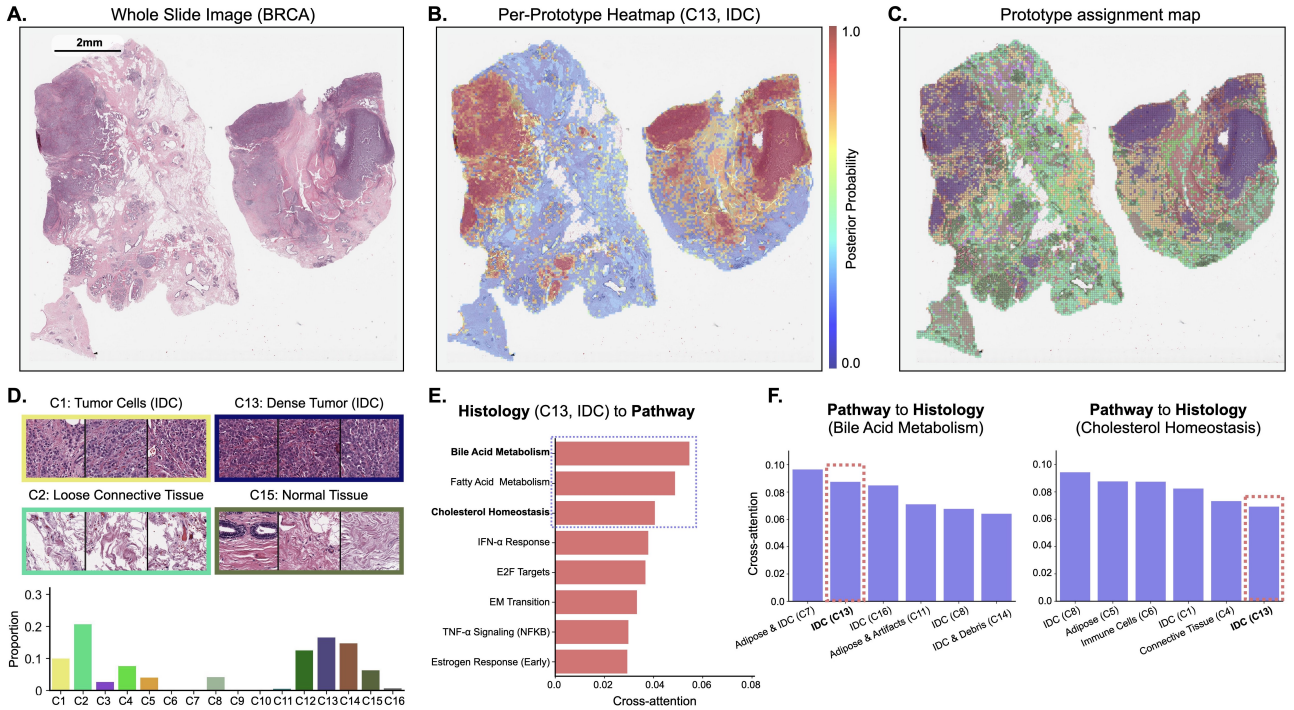

Figure 2. **Cross-modal interaction visualization**. (A) A WSI for a BRCA patient. (B) The morphological prototype heatmap for $c = 13$ (C13), representing *invasive ductal carcinoma* (IDC), based on the posterior distribution for C13. (C) Prototype assignment map showing the closest morphological prototype for each patch in the WSI. (D) Top-3 patches for each morphological prototype and proportion of each prototype in the WSI. (E) The top-10 pathways attending to C13. (F) Top-6 morphological prototypes attending to the pathways in (E).

visualize h. $\rightarrow$ g. interactions, *i.e.,* which pathways correspond to the queried prototype (pathway importance). While SurvPath also models h. $\rightarrow$ g. via cross-attention, since the histology patch tokens are redundant and not prototypical, visualizing h. $\rightarrow$ g. is intractable.

As an example, for the prototype c =13 (C13), which represents the dominant *invasive ductal carcinoma* morphology in the BRCA WSI, we can visualize its highly-attended pathways (h. $\rightarrow$ g.) – bile acid metabolism, fatty acid metabolism, and cholesterol homeostasis, being important oncogenic pathways in BRCA (**Fig. 2E**). This agrees with the literature that highlights the association between these pathways and breast cancer prognosis (Nelson et al., 2014; Koundouros & Poulogiannis, 2020; Režen et al., 2022). We can also visualize the highly-attended morphological prototypes for these pathways (g. $\rightarrow$ h.), with C13 highly attended by bile acid metabolism (**Fig. 2F**). Other IDC variations, such as C1 and C8, are also highly attended by these pathways. By virtue of bi-directional visualization capability, MMP can elucidate tightly-linked relationships, characterized by strong bi-directional cross-attention values (C13 and bile acid metabolism), which is a unique capability over other methods that have only visualized (g. $\rightarrow$ h.) (**Fig. 2F**) but not (h. $\rightarrow$ g.) (**Fig. 2E**). Further discussion and visualizations are available in the **Appendix I**.

7. Conclusion & Future works

We introduced MMP, a prototype-based multimodal fusion framework for survival prediction in computational pathology. This framework introduces a prototype-based tokenization method that effectively reduces the number of tokens and the associated computational complexity common in multimodal fusion frameworks. Such reduction leads to improved overall prognostic performance and allows a bi-directional concept-based interpretation of how morphology and transcriptomes interact.

We consider this an essential step forward for future multimodal prognosis research, which we believe can be extended and validated in different ways (also further detailed in **Appendix J**). First, the number of prototypes can be determined in a data-driven manner, *e.g.,* using frameworks in Dirichlet processes (Lee et al., 2020; Li et al., 2022). Next, instead of relying on shallow MLP for transcriptomics modeling, we can leverage the latest advances in single-cell foundation models (Rosen et al., 2023; Theodoris et al., 2023; Cui et al., 2024). Finally, a validation with different outcomes, such as progression-free interval and recurrence risk (Liu et al., 2018), as well as application to rare diseases for which not overfitting to a small cohort is paramount, will bring MMP closer to clinical translation.

Acknowledgements

The authors would like to thank Minyoung Kim at Samsung AI Center for the helpful advice on prototype-based aggregation; Ming Y. Lu and Tong Ding for setting up supervised MIL benchmarks. The authors acknowledge funding support from the Brigham and Women's Hospital (BWH) President's Fund, Mass General Hospital (MGH) Pathology and by the National Institute of Health (NIH) National Institute of General Medical Sciences (NIGMS) through R35GM138216.

Impact Statement

This manuscript details efforts to enhance cancer prognosis through the integration of whole-slide imaging and gene expression profiling. Our study utilizes data from The Cancer Genome Atlas public database, which has been ethically and institutionally approved by all contributing sites. Although our research may yield societal impacts, it is important to emphasize that this study is designed solely for research applications and not yet intended for clinical use. Further larger external cohort validation will be required for realizing the clinical potential of our work.

References

Acosta, J. N., Falcone, G. J., Rajpurkar, P., and Topol, E. J. Multimodal biomedical AI. *Nature Medicine*, 28(9):1773–1784, 2022.

Benamou, J.-D. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: Mathematical Modelling and Numerical Analysis*, 37(5):851–868, 2003.

Bland, J. M. and Altman, D. G. The logrank test. *Bmj*, 328 (7447):1073, 2004.

Bonnier, P., Romain, S., Charpin, C., Lejeune, C., Tubiana, N., Martin, P.-M., and Piana, L. Age as a prognostic factor in breast cancer: relationship to pathologic and biologic features. *International journal of cancer*, 62(2): 138–144, 1995.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8): 1301–1309, 2019.

Cao, Z., Xu, Q., Yang, Z., He, Y., Cao, X., and Huang, Q. Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in Neural Information Processing Systems*, 35:39090–39102, 2022.

Carmichael, I., Song, A. H., Chen, R. J., Williamson, D. F., Chen, T. Y., and Mahmood, F. Incorporating intratumoral heterogeneity into weakly-supervised deep learning models via variance pooling. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 387–397. Springer, 2022.

Chen, L., Gan, Z., Cheng, Y., Li, L., Carin, L., and Liu, J. Graph optimal transport for cross-domain alignment. In *International Conference on Machine Learning*, pp. 1542–1553. PMLR, 2020a.

Chen, R. J., Lu, M. Y., Wang, J., Williamson, D. F., Rodig, S. J., Lindeman, N. I., and Mahmood, F. Pathomic fusion: an integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2020b.

Chen, R. J., Lu, M. Y., Weng, W.-H., Chen, T. Y., Williamson, D. F., Manz, T., Shady, M., and Mahmood, F. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4025, 2021.

Chen, R. J., Lu, M. Y., Williamson, D. F., Chen, T. Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40 (8):865–878, 2022.

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

Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F.-X. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, pp. 1–11, 2024.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

dan Guo, D., Tian, L., Zhang, M., Zhou, M., and Zha, H. Learning prototype-oriented set representations for meta-learning. In *International Conference on Learning Representations*, 2022.

Dempster, A. P., Laird, N. M., and Rubin, D. B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22, 1977.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Ding, K., Zhou, M., Metaxas, D. N., and Zhang, S. Pathology-and-genomics multimodal transformer for survival outcome prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 622–631. Springer, 2023.

Dongre, A. and Weinberg, R. A. New insights into the mechanisms of epithelial–mesenchymal transition and implications for cancer. *Nature reviews Molecular cell biology*, 20(2):69–84, 2019.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Duan, J., Chen, L., Tran, S., Yang, J., Xu, Y., Zeng, B., and Chilimbi, T. Multi-modal alignment using representation codebook. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15651–15660, 2022.

Elmarakeby, H. A., Hwang, J., Arafeh, R., Crowdis, J., Gang, S., Liu, D., AlDubayan, S. H., Salari, K., Kregel, S., Richter, C., et al. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880): 348–352, 2021.

Gat, I., Schwartz, I., Schwing, A., and Hazan, T. Removing bias in multi-modal classifiers: Regularization by maximizing functional entropies. *Advances in Neural Information Processing Systems*, 33:3197–3208, 2020.

Genevay, A., Peyre, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1608–1617. PMLR, 09–11 Apr 2018.

Girdhar, R., Singh, M., Ravi, N., van der Maaten, L., Joulin, A., and Misra, I. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16102–16112, 2022.

Giudetti, A. M., De Domenico, S., Ragusa, A., Lunetti, P., Gaballo, A., Franck, J., Simeone, P., Nicolardi, G., De Nuccio, F., Santino, A., et al. A specific lipid metabolic profile is associated with the epithelial mesenchymal transition program. *Biochimica et Biophysica Acta (BBA)-Molecular and Cell Biology of Lipids*, 1864 (3):344–357, 2019.

Goldman, M. J., Craft, B., Hastie, M., Repečka, K., McDade, F., Kamath, A., Banerjee, A., Luo, Y., Rogers, D., Brooks, A. N., et al. Visualizing and interpreting cancer genomics data via the xena platform. *Nature biotechnology*, 38(6):675–678, 2020.

Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., and Rosati, R. A. Evaluating the yield of medical tests. *Jama*, 247(18):2543–2546, 1982.

Howard, F. M., Dolezal, J., Kochanny, S., Schulte, J., Chen, H., Heij, L., Huo, D., Nanda, R., Olopade, O. I., Kather, J. N., et al. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature communications*, 12(1):4423, 2021.

Ilse, M., Tomczak, J., and Welling, M. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. PMLR, 2018.

Ishay-Ronen, D., Diepenbruck, M., Kalathur, R. K. R., Sugiyama, N., Tiede, S., Ivanek, R., Bantug, G., Morini, M. F., Wang, J., Hess, C., et al. Gain fat—lose metastasis: converting invasive breast cancer cells into adipocytes inhibits cancer metastasis. *Cancer cell*, 35(1):17–32, 2019.

Jaegle, A., Borgeaud, S., Alayrac, J.-B., Doersch, C., Ionescu, C., Ding, D., Koppula, S., Zoran, D., Brock, A., Shelhamer, E., Henaff, O. J., Botvinick, M., Zisserman, A., Vinyals, O., and Carreira, J. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022.

Jaume, G., Vaidya, A., Chen, R., Williamson, D., Liang, P., and Mahmood, F. 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.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep

neural network. *BMC medical research methodology*, 18 (1):1–12, 2018.

Kim, M. Differentiable expectation-maximization for set representation learning. In *International Conference on Learning Representations*, 2022.

Klambauer, G., Unterthiner, T., Mayr, A., and Hochreiter, S. Self-normalizing neural networks. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34(4):43–59, 2017. doi: 10.1109/MSP.2017.2695801.

Koundouros, N. and Poulogiannis, G. Reprogramming of fatty acid metabolism in cancer. *British journal of cancer*, 122(1):4–22, 2020.

Kvamme, H., Borgan, Ø., and Scheel, I. Time-to-event prediction with neural networks and cox regression. *arXiv preprint arXiv:1907.00825*, 2019.

Lee, D. B., Lee, S., Ko, J., Kawaguchi, K., Lee, J., and Hwang, S. J. Self-supervised dataset distillation for transfer learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. W. Set transformer: A framework for attention-based permutation-invariant neural networks. In *International conference on machine learning*, pp. 3744–3753. PMLR, 2019.

Lee, S., Ha, J., Zhang, D., and Kim, G. A neural dirichlet process mixture model for task-free continual learning. In *International Conference on Learning Representations*, 2020.

Li, B. and Dewey, C. N. RSEM: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics*, 12:1–16, 2011.

Li, H., Zhu, C., Zhang, Y., Sun, Y., Shui, Z., Kuang, W., Zheng, S., and Yang, L. Task-specific fine-tuning via variational information bottleneck for weakly-supervised pathology whole slide image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7454–7463, 2023.

Li, N., Li, W., Jiang, Y., and Xia, S.-T. Deep dirichlet process mixture models. In *Uncertainty in Artificial Intelligence*, pp. 1138–1147. PMLR, 2022.

Liang, P. P., Lyu, Y., Fan, X., Tsaw, J., Liu, Y., Mo, S., Yogatama, D., Morency, L.-P., and Salakhutdinov, R. High-modality multimodal transformer: Quantifying modality & interaction heterogeneity for high-modality representation learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., and Tamayo, P. The molecular signatures database hallmark gene set collection. *Cell systems*, 1(6): 417–425, 2015.

Lipkova, J., Chen, R. J., Chen, B., Lu, M. Y., Barbieri, M., Shao, D., Vaidya, A. J., Chen, C., Zhuang, L., Williamson, D. F., et al. Artificial intelligence for multimodal data integration in oncology. *Cancer cell*, 40(10):1095–1110, 2022.

Liu, J., Lichtenberg, T., Hoadley, K. A., Poisson, L. M., Lazar, A. J., Cherniack, A. D., Kovatich, A. J., Benz, C. C., Levine, D. A., Lee, A. V., et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. *Cell*, 173(2):400–416, 2018.

Loo, S. Y., Toh, L. P., Xie, W. H., Pathak, E., Tan, W., Ma, S., Lee, M. Y., Shatishwaran, S., Yeo, J. Z. Z., Yuan, J., et al. Fatty acid oxidation is a druggable gateway regulating cellular plasticity for driving metastasis in breast cancer. *Science Advances*, 7(41):eabh2443, 2021.

Lu, M. Y., Williamson, D. F., Chen, T. Y., Chen, R. J., Barbieri, M., and Mahmood, F. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature biomedical engineering*, 5(6):555–570, 2021.

Mialon, G., Chen, D., d'Aspremont, A., and Mairal, J. A trainable optimal transport embedding for feature aggregation and its relationship to attention. In *International Conference on Learning Representations*, 2021.

Mobadersany, P., Yousefi, S., Amgad, M., Gutman, D. A., Barnholtz-Sloan, J. S., Velázquez Vega, J. E., Brat, D. J., and Cooper, L. A. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115 (13):E2970–E2979, 2018.

Nelson, E. R., Chang, C.-y., and McDonnell, D. P. Cholesterol and breast cancer pathophysiology. *Trends in Endocrinology & Metabolism*, 25(12):649–655, 2014.

Olea-Flores, M., Juárez-Cruz, J. C., Mendoza-Catalán, M. A., Padilla-Benavides, T., and Navarro-Tito, N. Signaling pathways induced by leptin during epithelial–mesenchymal transition in breast cancer. *International journal of molecular sciences*, 19(11):3493, 2018.

Olea-Flores, M., Juárez-Cruz, J. C., Zuñiga-Eulogio, M. D., Acosta, E., García-Rodríguez, E., Zacapala-Gomez, A. E., Mendoza-Catalán, M. A., Ortiz-Ortiz, J., Ortuño-Pineda, C., and Navarro-Tito, N. New actors driving the epithelial–mesenchymal transition in cancer: The role of leptin. *Biomolecules*, 10(12):1676, 2020.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Pölsterl, S. scikit-survival: A library for time-to-event analysis built on top of scikit-learn. *Journal of Machine Learning Research*, 21(212):1–6, 2020.

Pramanick, S., Roy, A., and Patel, V. M. Multimodal learning using optimal transport for sarcasm and humor detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3930–3940, 2022.

Pramanick, S., Jing, L., Nag, S., Zhu, J., Shah, H. J., LeCun, Y., and Chellappa, R. VoLTA: Vision-language transformer with weakly-supervised local-feature alignment. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

Quiros, A. C., Coudray, N., Yeaton, A., Yang, X., Liu, B., Le, H., Chiriboga, L., Karimkhan, A., Narula, N., Moore, D. A., Park, C. Y., Pass, H., Moreira, A. L., Quesne, J. L., Tsirigos, A., and Yuan, K. Mapping the landscape of histomorphological cancer phenotypes using self-supervised learning on unlabeled, unannotated pathology slides, 2023.

Rakha, E. A., Reis-Filho, J. S., Baehner, F., Dabbs, D. J., Decker, T., Eusebi, V., Fox, S. B., Ichihara, S., Jacquemier, J., Lakhani, S. R., et al. Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast cancer research*, 12:1–12, 2010.

Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., et al. Pathway enrichment analysis and visualization of omics data using g: Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature protocols*, 14(2):482–517, 2019.

Režen, T., Rozman, D., Kovács, T., Kovács, P., Sipos, A., Bai, P., and Mikó, E. The role of bile acids in carcinogenesis. *Cellular and molecular life sciences*, 79(5):243, 2022.

Rosen, Y., Roohani, Y., Agrawal, A., Samotorcan, L., Consortium, T. S., Quake, S. R., and Leskovec, J. Universal cell embeddings: A foundation model for cell biology. *bioRxiv*, pp. 2023–11, 2023.

Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in Neural Information Processing Systems*, 34:2136–2147, 2021.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Song, A. H., Jaume, G., Williamson, D. F., Lu, M. Y., Vaidya, A., Miller, T. R., and Mahmood, F. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 1(12):930–949, 2023.

Song, A. H., Chen, R. J., Ding, T., Williamson, D. F., Jaume, G., and Mahmood, F. Morphological prototyping for unsupervised slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Song, A. H., Williams, M., Williamson, D. F., Chow, S. S., Jaume, G., Gao, G., Zhang, A., Chen, B., Baras, A. S., Serafin, R., et al. Analysis of 3D pathology samples using weakly supervised AI. *Cell*, 187(10):2502–2520, 2024b.

Steyaert, S., Pizurica, M., Nagaraj, D., Khandelwal, P., Hernandez-Boussard, T., Gentles, A. J., and Gevaert, O. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nature Machine Intelligence*, 5(4): 351–362, 2023.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

Tas, F., Ciftci, R., Kilic, L., and Karabulut, S. Age is a prognostic factor affecting survival in lung cancer patients. *Oncology letters*, 6(5):1507–1513, 2013.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., et al. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023.

Uno, H., Cai, T., Pencina, M. J., D'Agostino, R. B., and Wei, L.-J. On the c-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data. *Statistics in medicine*, 30(10):1105–1117, 2011.

Vaidya, A., Chen, R. J., Williamson, D. F., Song, A. H., Jaume, G., Yang, Y., Hartvigsen, T., Dyer, E. C., Lu, M. Y., Lipkova, J., et al. Demographic bias in misdiagnosis by computational pathology models. *Nature Medicine*, 30(4):1174–1190, 2024.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Volinsky-Fremond, S., Horeweg, N., Andani, S., Barkey Wolf, J., Lafarge, M. W., de Kroon, C. D., Ørtoft, G., Høgdall, E., Dijkstra, J., Jobsen, J. J., et al. Prediction of recurrence risk in endometrial cancer with multimodal deep learning. *Nature Medicine*, pp. 1–12, 2024.

Vu, Q. D., Rajpoot, K., Raza, S. E. A., and Rajpoot, N. Handcrafted Histological Transformer (H2T): Unsupervised representation of whole slide images. *Medical Image Analysis*, 85:102743, 2023. ISSN 1361-8415.

Wang, W., Tran, D., and Feiszli, M. What makes training multi-modal classification networks hard? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12695–12705, 2020.

Wang, W., Bao, H., Dong, L., Bjorck, J., Peng, Z., Liu, Q., Aggarwal, K., Mohammed, O. K., Singhal, S., Som, S., et al. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19175–19186, 2023.

Wang, X., Yang, S., Zhang, J., Wang, M., Zhang, J., Yang, W., Huang, J., and Han, X. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical image analysis*, 81:102559, 2022.

Wang, Y.-Y., Lehuédé, C., Laurent, V., Dirat, B., Dauvillier, S., Bochet, L., Le Gonidec, S., Escourrou, G., Valet, P., and Muller, C. Adipose tissue and breast epithelial cells: a dangerous dynamic duo in breast cancer. *Cancer letters*, 324(2):142–151, 2012.

Wang, Z., Li, R., Wang, M., and Li, A. GPDBN: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction. *Bioinformatics*, 37(18):2963–2970, 2021.

Wang, Z., Zhang, Y., Xu, Y., Imoto, S., Chen, H., and Song, J. Histo-genomic knowledge distillation for cancer prognosis from histopathology whole slide images, 2024.

Wong, W. H. Theory of partial likelihood. *The Annals of statistics*, pp. 88–123, 1986.

Wu, Y.-d. and Zhou, B. TNF-α/NF-κB/Snail pathway in cancer cell migration and invasion. *British journal of cancer*, 102(4):639–644, 2010.

Wulczyn, E., Steiner, D. F., Xu, Z., Sadhwani, A., Wang, H., Flament-Auvigne, I., Mermel, C. H., Chen, P.-H. C., Liu, Y., and Stumpe, M. C. Deep learning-based survival prediction for multiple cancer types using histopathology images. *PloS ONE*, 15(6), 2020.

Xiang, J. and Zhang, J. Exploring low-rank property in multiple instance learning for whole slide image classification. In *The Eleventh International Conference on Learning Representations*, 2022.

Xiang, J. and Zhang, J. Exploring low-rank property in multiple instance learning for whole slide image classification. In *The Eleventh International Conference on Learning Representations*, 2023.

Xiong, Y., Zeng, Z., Chakraborty, R., Tan, M., Fung, G., Li, Y., and Singh, V. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14138–14148, 2021.

Xu, P., Zhu, X., and Clifton, D. A. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

Xu, Y. and Chen, H. 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 (ICCV)*, pp. 21241–21251, October 2023.

Yao, J., Zhu, X., and Huang, J. Deep multi-instance learning for survival prediction from whole slide images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 496–504. Springer, 2019.

Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., and Huang, J. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

Yu, X. Q., Yap, M. L., Cheng, E. S., Ngo, P. J., Vaneckova, P., Karikios, D., Canfell, K., and Weber, M. F. Evaluating prognostic factors for sex differences in lung cancer survival: findings from a large australian cohort. *Journal of Thoracic Oncology*, 17(5):688–699, 2022.

Zadeh, S. G. and Schmid, M. Bias in cross-entropy-based training of deep survival networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(9):3126–3137, 2020.

Zadeh, S. G. and Schmid, M. Bias in cross-entropy-based training of deep survival networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9):3126–3137, 2021. doi: 10.1109/TPAMI.2020.2979450.

Zhang, Y., Xu, Y., Chen, J., Xie, F., and Chen, H. Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. In *The Twelfth International Conference on Learning Representations*, 2024.

Zhou, F. and Chen, H. Cross-modal translation and alignment for survival analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21485–21494, 2023.

A. Prototype-based histopathology baselines

In this section, we present the three histology prototype aggregation approaches that can be used by MMP, with particular emphasis on the Gaussian mixture model (GMM). The following prototype-based aggregation schemes can be embedded as a feed-forward module in our models.

A.1. Hard clustering (HC)

For each $\mathbf{z}_{i,\text{h.}}$, we identify the closest prototype $\mathbf{a}_{c,\text{h.}}$ evaluated with the $\mathcal{L}_2$ distance, *i.e.,* $c_i = \arg\max_c \|\mathbf{z}_{i,\text{h.}} - \mathbf{a}_{c,\text{h.}}\|_2$ to determine the cluster assignment. The post-aggregation embedding $\mathbf{z}^{\text{agg.}}_{c,\text{h.}}$ is an average of all embeddings assigned to c,

$$\mathbf{z}^{\text{agg.}}_{c,\text{h.}} = \sum_{i=1}^{N_{\text{h.}}} \mathbf{1}_{c_i=c} \cdot \mathbf{z}_{i,\text{h.}} / \sum_{i=1}^{N_{\text{h.}}} \mathbf{1}_{c_i=c}. \tag{7}$$

where $\mathbf{1}$ is the indicator function.

A.2. Optimal transport (OT)

We can formulate aggregation as that of transporting from the empirical distribution of $\hat{p}(\mathbf{z}_{\text{h.}}) = 1/N_{\text{h.}} \cdot \sum_{i=1}^{N_{\text{h.}}} \delta(\mathbf{z}_{i,h})$ to $\hat{p}(\mathbf{a}_{\text{h.}}) = 1/C_{\text{h.}} \cdot \sum_{i=1}^{C_{\text{h.}}} \delta(\mathbf{a}_{c,\text{h.}})$. The transport plan $\mathbf{T} \in \mathbb{R}_+^{N_{\text{h.}} \times C_{\text{h.}}}$ is given as the solution to the following entropic-regularized optimal transport problem (Cuturi, 2013; Kolouri et al., 2017),

$$\min_{\mathbf{T}} \sum_{i,c} \|\mathbf{z}_{i,\text{h.}} - \mathbf{a}_{c,\text{h.}}\|_2 \cdot \mathbf{T}_{i,c} + \epsilon \cdot \mathbf{T}_{i,c} \log \mathbf{T}_{i,c}, \quad \text{such that } \sum_{i=1}^{N_{\text{h.}}} \mathbf{T}_{i,c} = 1/C_{\text{h.}} \text{ and } \sum_{i=1}^{C_{\text{h.}}} \mathbf{T}_{i,c} = 1/N_{\text{h.}}, \tag{8}$$

where ε is the regularization parameter. Based on the optimal transport plan $\widehat{\mathbf{T}}$ obtained by the widely-used Sinkhorn algorithm (Cuturi, 2013), the post-aggregation embedding is given as $\mathbf{z}^{\text{agg.}}_{c,\text{h.}} = \sum_{i=1}^{N_{\text{h.}}} \widehat{\mathbf{T}}_{i,c} \cdot \mathbf{z}_{i,\text{h.}}$.

A.3. Gaussian Mixture Models

With the Gaussian mixture model (GMM) as the generative model for each token embedding, we provide a detailed derivation for estimation of 1) the posterior probability for the prototype assignment $q(c|\mathbf{z}_{i,\text{h.}};\theta)$ and 2) the GMM parameters $\theta = \{\pi_c, \boldsymbol{\mu}_c, \Sigma_c\}$. Given the GMM specification,

$$\begin{aligned} p(\mathbf{z}_{i,\text{h.}};\theta) &= \sum_{c=1}^{C_{\text{h.}}} p(c_i = c;\theta) \cdot p(\mathbf{z}_{i,\text{h.}}|c_i = c;\theta) \\ &= \sum_{c=1}^{C_{\text{h.}}} \pi_c \cdot \mathcal{N}(\mathbf{z}_{i,\text{h.}};\boldsymbol{\mu}_c, \Sigma_c), \;\; s.t. \sum_{c=1}^{C_{\text{h.}}} \pi_c = 1, \end{aligned} \tag{9}$$

the goal is to estimate θ that maximizes the log-likelihood $\max_\theta \sum_{i=1}^{N_{\text{h.}}} \log p(\mathbf{z}_{i,\text{h.}};\theta) = \max_\theta \sum_{n=1}^{N_{\text{h.}}} \log p(\mathbf{z}_{i,\text{h.}};\theta)$. We now present a detailed walkthrough of the expectation-maximization (EM) algorithm (Dempster et al., 1977; Kim, 2022; Song et al., 2024a) and how these ultimately lead to $\mathbf{z}^{\text{agg.}}_{c,\text{h.}}$.

Using Jensen's inequality, we can lower-bound the log-likelihood as follows,

$$\begin{aligned} \sum_{i=1}^{N_{\text{h.}}} \log p(\mathbf{z}_{i,\text{h.}};\theta) &= \sum_{i=1}^{N_{\text{h.}}} \log \sum_{c=1}^{C_{\text{h.}}} p(\mathbf{z}_{i,\text{h.}}, c_i = c;\theta) \\ &= \sum_{i=1}^{N_{\text{h.}}} \log \sum_{c=1}^{C_{\text{h.}}} q(c_i = c|\mathbf{z}_{i,\text{h.}};\theta_{\text{old}}) \cdot \frac{p(\mathbf{z}_{i,\text{h.}}, c_i = c;\theta)}{q(c_i = c|\mathbf{z}_{i,\text{h.}};\theta_{\text{old}})} \\ &\geq \sum_{i=1}^{N_{\text{h.}}} \sum_{c=1}^{C_{\text{h.}}} q(c_i = c|\mathbf{z}_{i,\text{h.}};\theta_{\text{old}}) \log \frac{p(\mathbf{z}_{i,\text{h.}}, c_i = c;\theta)}{q(c_i = c|\mathbf{z}_{i,\text{h.}};\theta_{\text{old}})} \\ &= \sum_{i=1}^{N_{\text{h.}}} \underbrace{E_{q(c_i=c|\mathbf{z}_{i,\text{h.}};\theta_{\text{old}})} \left[\log p(\mathbf{z}_{i,\text{h.}}, c_i = c;\theta)\right]}_{Q(\theta;\theta_{\text{old}})} - \sum_{i=1}^{N_{\text{h.}}} \underbrace{E_{q(c_i=c|\mathbf{z}_{i,\text{h.}};\theta_{\text{old}})} \left[q(c_i = c|\mathbf{z}_{i,\text{h.}};\theta_{\text{old}})\right]}_{-H(C;\theta_{\text{old}})}. \end{aligned} \tag{10}$$

Instead of maximizing the log-likelihood directly, we can now maximize a surrogate function, which is the lower bound given by Jensen's inequality. It can be shown that increasing this lower bound with respect to θ leads to monotonically increasing the actual log-likelihood (Dempster et al., 1977). The optimization procedure involves iterative alternating steps of the E-step and the M-step and is thus referred to as the Expectation-Maximization (EM) algorithm.

The surrogate function consists of two terms, $Q(\theta;\theta_{\text{old}})$ and $H(C;\theta_{\text{old}})$, which are expectations with respect to the posterior probability of prototype assignment, $q(c_i = c|\mathbf{z}_{i,\text{h.}};\theta_{\text{old}})$. In the E-step, we can use Bayes' rule to compute the posterior probability and, consequently the expectations,

$$\begin{aligned} q(c_i = c|\mathbf{z}_{i,\text{h.}};\theta_{\text{old}}) &= \frac{q(\mathbf{z}_{i,\text{h.}}|c_i = c;\theta_{\text{old}}) \cdot q(c_i = c;\theta_{\text{old}})}{q(\mathbf{z}_{i,\text{h.}};\theta_{\text{old}})} \\ &= \frac{q(\mathbf{z}_{i,\text{h.}}|c_i = c;\theta_{\text{old}}) \cdot q(c_i = c;\theta_{\text{old}})}{\sum_{c=1}^{C_{\text{h.}}} q(\mathbf{z}_{i,\text{h.}}|c_i = c;\theta_{\text{old}}) \cdot q(c_i = c;\theta_{\text{old}})} \\ &= \frac{\pi_c \cdot \mathcal{N}(\mathbf{z}_{i,\text{h.}};\boldsymbol{\mu}_c,\Sigma_c)}{\sum_{c=1}^{C_{\text{h.}}} \pi_c \cdot \mathcal{N}(\mathbf{z}_{i,\text{h.}};\boldsymbol{\mu}_c,\Sigma_c)}. \end{aligned} \tag{11}$$

In the M-step, we find θ_{new} that maximizes the surrogate function based on the posterior probability computed from the E-step. Since the term $H(C;\theta_{\text{old}})$ is not a function of θ and therefore a constant (it is a function of θ_{old}), we only need to optimize the term $Q(\theta;\theta_{\text{old}})$ by taking the derivative with respect to θ,

$$\begin{aligned} \sum_{i=1}^{N_{\text{h.}}} \frac{\partial Q(\theta;\theta_{\text{old}})}{\partial \pi_c} = 0 \Rightarrow \pi_c^{\text{new}} &= \frac{\sum_{i=1}^{N_{\text{h.}}} q(c_i = c|\mathbf{z}_{i,\text{h.}};\theta_{\text{old}})}{N_{\text{h.}}} \\ \sum_{i=1}^{N_{\text{h.}}} \frac{\partial Q(\theta;\theta_{\text{old}})}{\partial \boldsymbol{\mu}_c} = 0 \Rightarrow \boldsymbol{\mu}_c^{\text{new}} &= \frac{\sum_{i=1}^{N_{\text{h.}}} q(c_i = c|\mathbf{z}_{i,\text{h.}};\theta_{\text{old}}) \cdot \mathbf{z}_{i,\text{h.}}}{\sum_{i=1}^{N_{\text{h.}}} q(c_i = c|\mathbf{z}_{i,\text{h.}};\theta_{\text{old}})} \\ \sum_{i=1}^{N_{\text{h.}}} \frac{\partial Q(\theta;\theta_{\text{old}})}{\partial \Sigma_c} = 0 \Rightarrow \Sigma_c^{\text{new}} &= \frac{\sum_{i=1}^{N_{\text{h.}}} q(c_i = c|\mathbf{z}_{i,\text{h.}};\theta_{\text{old}}) \cdot (\mathbf{z}_{i,\text{h.}} - \boldsymbol{\mu}_c^{\text{new}})^2}{\sum_{i=1}^{N_{\text{h.}}} q(c_i = c|\mathbf{z}_{i,\text{h.}};\theta_{\text{old}})}. \end{aligned} \tag{12}$$

The E-step and M-step alternate until convergence is reached. In our setting, we usually found one round of EM iteration sufficient. As for the initial parameters, we set $\pi_c^{(0)} = 1/C_{\text{h.}}$, $\boldsymbol{\mu}_c^{(0)} = \mathbf{a}_{c,\text{h.}}$, and $\Sigma_c^{(0)} = \mathbf{I}$, which serves as a morphology-aware initialization for the algorithm. The initialization for $\{\mathbf{a}_{c,\text{h.}}\}_{c=1}^{C_{\text{h.}}}$ is performed with K-means clustering on the training set of patches. This is constructed by aggregating token embeddings from all training slides in a disease cohort.

Once $\widehat{\theta}$ is estimated, the post-aggregation embedding $\mathbf{z}_{c,\text{h.}}^{\text{agg.}} \in \mathbb{R}^{d_{\text{h.}}}$ with $d_{\text{h.}} = 1 + 2D$, can be represented as a concatenation $\mathbf{z}_{c,\text{h.}}^{\text{agg.}} = [\widehat{\pi}_c, \widehat{\boldsymbol{\mu}}_c, \widehat{\Sigma}_c]$. Denoting $q_i = q(c_i = c|\mathbf{z}_{i,\text{h.}};\theta_{\text{old}})$, we can express $\mathbf{z}_{c,\text{h.}}^{\text{agg.}}$ as in Eq. 1,

$$\mathbf{z}_{c,\text{h.}}^{\text{agg.}} = \sum_{i=1}^{N_{\text{h.}}} \left[q_i/N_{\text{h.}},\ \ q_i\mathbf{z}_{i,\text{h.}}/(\sum_{i=1}^{N_{\text{h.}}} q_i),\ \ q_i \left(z_{i,\text{h.}} - \sum_{i=1}^{N_{\text{h.}}} q_i\mathbf{z}_{i,\text{h.}}/(\sum_{i=1}^{N_{\text{h.}}} q_i) \right)^2 /(\sum_{i=1}^{N_{\text{h.}}} q_i) \right], \tag{13}$$

which can indeed expressed as a sum of the mapping function g (albeit non-trivial to write out the full expression due to the iterative nature of EM) over $N_{\text{h.}}$ elements.

B. Proof for similarity between OT-based cross-alignment and Transformer-based cross-attention

Lemma B.1. *Let $\mathbf{Z}_{g.} \in \mathbb{R}^{C_{g.} \times d}$ and $\mathbf{Z}_{h.} \in \mathbb{R}^{C_{h.} \times d}$ be the matrix representation of the token sets $\{\mathbf{z}_{i,g.}\}_{i=1}^{C_{g.}}$ and $\{\mathbf{z}_{k,h.}\}_{k=1}^{C_{h.}}$. Let $\mathbf{Z}_{g.}\mathbf{W}_Q^T \in \mathbb{R}^{C_{g.} \times d}$ and $\mathbf{Z}_{h.}\mathbf{W}^T \in \mathbb{R}^{C_{h.} \times d}$ be the linear projections of both sets. Let $\widehat{\mathbf{T}} \in \mathbb{R}_+^{C_{g.} \times C_{h.}}$ be the optimal transport plan, i.e., the solution to the entropic-regularized, unbalanced optimal transport problem between the two projected sets. Then, $\widehat{\mathbf{T}}$ is equivalent to the Transformer cross-attention matrix, $\sigma(\mathbf{Z}_{g.}\mathbf{W}_Q^T\mathbf{W}\mathbf{Z}_{h.}^T/\sqrt{d})$, up to a multiplicative factor where $\sigma(\cdot)$ denotes row-wise softmax, $\{\mathbf{W}_Q\mathbf{z}_{i,g.}\}_{i=1}^{C_{g.}}$ are queries, and $\{\mathbf{W}\mathbf{z}_{k,h.}\}_{k=1}^{C_{h.}}$ are keys.*

Proof. This proof is an extension and adaption of a lemma from (Kim, 2022) to our application. We use the negative dot-product similarity as the cost between two sets of linearly-projected tokens $\{\mathbf{W}_Q\mathbf{z}_{i,\text{g.}}\}_{i=1}^{C_{\text{g.}}}$ and $\{\mathbf{W}\mathbf{z}_{k,\text{h.}}\}_{k=1}^{C_{\text{h.}}}$ as

$\mathbf{D}_{i,k} = -\mathbf{z}_{i,\text{g.}}^{\text{T}}\mathbf{W}_Q^{\text{T}}\mathbf{W}\mathbf{z}_{k,\text{h.}}$. We can formulate the entropic-regularized optimal transport problem for optimizing the transport plan $\mathbf{T} \in \mathbb{R}_+^{C_{\text{g.}} \times C_{\text{h.}}}$,

$$\min_{\mathbf{T}} \sum_{i,k} \mathbf{D}_{i,k} \cdot \mathbf{T}_{i,k} + \varepsilon \mathbf{T}_{i,k} \log \mathbf{T}_{i,k}, \quad \text{s.t.} \sum_{k=1}^{C_{\text{h.}}} \mathbf{T}_{i,k} = \frac{1}{C_{\text{g.}}}, \forall i, \tag{14}$$

without the constraint $\sum_{i=1}^{C_{\text{g.}}} \mathbf{T}_{i,k} = 1/C_{\text{h.}}$. Note that this can be considered as an unbalanced OT problem (Benamou, 2003; Chizat et al., 2018), as Eq. 14 can be written as

$$\min_{\mathbf{T}} \sum_{i,k} (\mathbf{D}_{i,k} \cdot \mathbf{T}_{i,k} + \varepsilon \mathbf{T}_{i,k} \log \mathbf{T}_{i,k}) + \lambda_1 \cdot \text{Div.}(\mathbf{T}^{\text{T}} \cdot \mathbf{1}_{C_{\text{g.}}}, 1/C_{\text{h.}} \cdot \mathbf{1}_{C_{\text{h.}}}) + \lambda_2 \cdot \text{Div.}(\mathbf{T} \cdot \mathbf{1}_{C_{\text{h.}}}, 1/C_{\text{g.}} \cdot \mathbf{1}_{C_{\text{g.}}}), \tag{15}$$

with $\lambda_1 \to 0$ and $\lambda_2 \to \infty$, where Div. is some divergence measure and $\mathbf{1}_{C_{\text{g.}}}$ is a $C_{\text{g.}}$-length vector of ones. We now take Eq. 14 and solve it by using Lagrange multiplier,

$$\mathcal{L} = \sum_{i,k} (\mathbf{D}_{i,k} \cdot \mathbf{T}_{i,k} + \varepsilon \mathbf{T}_{i,k} \log \mathbf{T}_{i,k}) + \sum_{i=1}^{C_{\text{g.}}} \beta_i \left(\sum_{k=1}^{C_{\text{h.}}} \mathbf{T}_{i,k} - \frac{1}{C_{\text{g.}}} \right). \tag{16}$$

We proceed by taking the derivative of $\mathcal{L}$ with respect to $\mathbf{T}_{i,k}$ and setting it to 0,

$$\frac{\partial \mathcal{L}}{\partial \mathbf{T}_{i,k}} = \mathbf{D}_{i,k} + \varepsilon(\log \mathbf{T}_{i,k} + 1) + \beta_i = 0 \Rightarrow \mathbf{T}_{i,k} = \exp\left(-\mathbf{D}_{i,k}/\varepsilon + \gamma_i\right), \tag{17}$$

where $\gamma_i = -(\beta_i/\varepsilon + 1)$ is some constant. To solve for γ_i, we can use the constraint $\sum_{k=1}^{C_{\text{h.}}} \mathbf{T}_{i,k} = \frac{1}{C_{\text{g.}}}$,

$$\exp(\gamma_i) \sum_{k=1}^{C_{\text{h.}}} \exp(-\mathbf{D}_{i,k}/\varepsilon) = \frac{1}{C_{\text{g.}}} \Rightarrow \exp(\gamma_i) = \frac{1}{C_{\text{g.}} \cdot \sum_{k=1}^{C_{\text{h.}}} \exp(-\mathbf{D}_{i,k}/\varepsilon)}, \tag{18}$$

and obtain $\widehat{\mathbf{T}}_{i,k}$ (by also setting $\varepsilon = \sqrt{d}$),

$$\widehat{\mathbf{T}}_{i,k} = \frac{\exp(-\mathbf{D}_{i,k}/\sqrt{d})}{C_{\text{g.}} \cdot \sum_{k=1}^{C_{\text{h.}}} \exp(-\mathbf{D}_{i,k}/\sqrt{d})} = \frac{\exp(\mathbf{z}_{i,\text{g.}}^{\text{T}}\mathbf{W}_Q^{\text{T}}\mathbf{W}\mathbf{z}_{k,\text{h.}}/\sqrt{d})}{C_{\text{g.}} \cdot \sum_{k=1}^{C_{\text{h.}}} \exp(\mathbf{z}_{i,\text{g.}}^{\text{T}}\mathbf{W}_Q^{\text{T}}\mathbf{W}\mathbf{z}_{k,\text{h.}}/\sqrt{d})}, \tag{19}$$

with the softmax term appearing as an entry for $\widehat{\mathbf{T}}$. This is the same as the Transformer-based cross-attention operation up to a multiplicative factor of $1/C_{\text{g.}}$. □

C. Survival loss functions

Survival analysis models the time to an event, where the event outcome is not always observed (*i.e.,* censored). In cancer survival outcome prediction, a censored event refers to patient survival or last known follow-up time, whereas an uncensored event is a patient death. Let T be a continuous random variable representing patient survival time, and the survival function $S(t) = P(T \geq t_0)$ be the probability of a patient surviving longer than time t_0. The goal of survival analysis is to estimate the hazard function, $\lambda(t)$, which denotes the probability of an event occurring instantaneously at time $t > t_0$ (Cox, 1972). We now detail the Cox proportional Hazards and Negative log-likelihood survival losses.

C.1. Cox proportional hazards loss

Cox proportional hazards model parameterizes the hazard function as an exponential linear function $\lambda(t|x) = \lambda_0(t) \exp^{\theta x}$. λ_0 is the baseline hazard function describing how the risk of an event changes over time. θ are the model parameters describing how the hazards vary with the features of a patient, $\bar{\mathbf{x}}_{\text{patient}} \in \mathbb{R}^{2d}$. To express the likelihood of an event to be observed at time t with model parameters θ, the Cox partial log-likelihood can be used (Wong, 1986):

$$l(\theta, \bar{\mathbf{x}}_{\text{patient}}) = -\sum_{i \in U} \left(\bar{\mathbf{x}}_{\text{patient},i}\theta - \log(\sum_{j \in R_i} \exp(\bar{\mathbf{x}}_{\text{patient},j}\theta)) \right) \tag{20}$$

$$\frac{\partial l(\theta, \bar{\mathbf{x}}_{\text{patient}})}{\partial \bar{\mathbf{x}}_{\text{patient},i}} = \delta(i)\theta - \sum_{i,j \in C_j, U} \frac{\theta \exp(\bar{\mathbf{x}}_{\text{patient},i}\theta)}{\sum_{k \in C_j} \exp(\bar{\mathbf{x}}_{\text{patient},k}\theta)} \tag{21}$$

where U is the set of uncensored patients, C is the set of censored patients, R_i is the set of patients whose last time of follow-up or time of death is after i, and $\delta(i)$ signifies if event outcome is observed or if censored.

C.2. Negative log-likelihood loss

The Negative log-likelihood (NLL) survival loss (Zadeh & Schmid, 2021) generalizes the NLL to censored data. The aim is to predict the survival of a patient from the learned patient level embedding $\bar{\mathbf{x}}_{\text{patient}} \in \mathbb{R}^{2d}$. In accordance with previous work (Zadeh & Schmid, 2021), the patient's survival state is defined by: (1) censorship status c, where $c = 0$ represents an observed patient death due to disease and $c = 1$ corresponds to the patient's last known follow-up, and (2) a time-to-event t_i, which corresponds to the time between the patient's diagnosis and observed death if $c = 0$, or the last follow-up if $c = 1$. Instead of predicting the observed time-to-event t_i, we *discretize* it by defining non-overlapping time intervals $(t_{j-1}, t_j),\ j \in [1, ..., n]$ based on the quartiles of survival time values, and denote as y_j. The setup simplifies to a classification problem with censorship information, where each patient is now defined by $(\bar{\mathbf{x}}_{\text{slide}}, y_j, c)$. Next, we build a classifier such that each output logit $\hat{y}_j$ corresponds to a time interval. Then, we define the discrete hazard function $f_{\text{hazard}}(y_j|\bar{\mathbf{x}}_{\text{patient}}) = S(\hat{y}_j)$ where S is the sigmoid activation. Intuitively, $f_{\text{hazard}}(y_j|\bar{\mathbf{x}}_{\text{patient}})$ represents the probability that the patient dies during time interval (t_{j-1}, t_j). Additionally, we define the discrete survival function $f_{\text{surv}}(y_j|\bar{\mathbf{x}}_{\text{patient}}) = \prod_{k=1}^{j} \left(1 - f_{\text{hazard}}(y_k|\bar{\mathbf{x}}_{\text{patient}})\right)$ that represents the probability that the patient survives up to time interval (t_{j-1}, t_j). Now, the NLL survival loss can be formally defined as:

$$\mathcal{L}\Big(\{\bar{\mathbf{x}}^{(i)}_{\text{patient}}, y^{(i)}_j, c^{(i)}\}_{i=1}^{N_D}\Big) = \tag{22}$$

$$\sum_{i=1}^{N_D} -c^{(i)} \log(f_{\text{surv}}(y^{(i)}_j|\bar{\mathbf{x}}^{(i)}_{\text{patient}})) \tag{23}$$

$$+ (1 - c^{(i)}) \log(f_{\text{surv}}(y^{(i)}_j - 1|\bar{\mathbf{x}}^{(i)}_{\text{patient}})) \tag{24}$$

$$+ (1 - c^{(i)}) \log(f_{\text{hazard}}(y^{(i)}_j|\bar{\mathbf{x}}^{(i)}_{\text{patient}})) \tag{25}$$

where N_D is the number of samples in the dataset. Eq. 23 enforces high survival probability for patients alive after the final follow-up, Eq. 24 enforces high survival up to the time stamp where death was observed for patients that died, and Eq. 25 ensures correct timestamp is predicted for patients with observed death. As NLL does not require a set of patients for training, unlike Cox loss, it has been the de-facto loss function for cancer survival prediction with histology data, with the large number of tokens rendering the formation of patient batch infeasible.

C.3. Concordance Index

The Concordance Index (C-Index) (Harrell et al., 1982) is a popular metric to measure the performance of survival prediction model (Chen et al., 2022; Jaume et al., 2024) and measures the rank correlation between the predicted risk scores and observed time points t. In prognosis prediction, the C-Index can be conceptually understood as a metric that assesses the accuracy of a model in predicting a higher risk of adverse outcomes for patients with shorter survival times. Formally, C-Index is defined as the ratio of concordant pairs to total comparable pairs. Two patients i and j are comparable if the patient with the lower observed time experienced an event (*i.e.,* if $t_i > t_j$ then $\delta_j = 1$, where δ is a binary indicator of whether event is observed or if last follow up time is known). A comparable pair (i, j) is considered concordant if the risk predicted by a survival model $\hat{f}_{\text{risk}}$ is larger for the patient with the smaller event time, *i.e.,* $\hat{f}_{\text{risk},j} > \hat{f}_{\text{risk},i}$ given $t_j < t_i$. Otherwise, the pair is considered discordant (Pölsterl, 2020). While C-Index allows for easy comparisons between models, known limitations exist, such as it is overly optimistic for increasing censorship in datasets (Uno et al., 2011).

D. Datasets

D.1. TCGA cohort

We evaluate all baselines on 6 cancer cohorts from TCGA: Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Lung adenocarcinoma (LUAD), Stomach adenocarcinoma (STAD), Colon and Rectum adenocarcinoma (CRC), Kidney renal clear cell carcinoma (KIRC), and low-grade gliomas (LGG). **Table 5** contains representative statistics of the dataset. A WSI is tessellated into nonoverlapping patches (tokens) of 256×256 pixels at $20\times$ magnification ($0.5\mu m$/pixel).

Table 5. **TCGA cohort statistics** The number of patients, total WSIs, and the average number of patches (tokens) in a WSI. A single patient can have multiple WSIs.

	Num. of patients	Num. of slides	Avg. set size
BLCA	359	423	16,312
BRCA	868	928	11,565
LUAD	412	463	4,714
STAD	318	318	10,955
CRC	296	300	9,127
KIRC	340	346	12,802

D.2. RNA-seq expression data

Bulk RNA-seq expression for all TCGA cohorts—accessed from UCSC Xena database (Goldman et al., 2020)—is measured by Illumina HiSeq 2000 RNA Sequencing platform and then $\log_2(x+1)$ transformed RSEM normalized (Li & Dewey, 2011). The $C_{\text{g.}} = 50$ Hallmark gene sets from Molecular Signatures Database (MSigDB) (Subramanian et al., 2005; Liberzon et al., 2015) are used to select and organize genes into biological pathways. Hallmark gene sets represent well-defined biological states in cancer. After organizing genes into Hallmark gene sets, we had 4,241 unique genes across the 50 gene sets. The average length of the gene sets is 142, with the minimum and maximum of 31 and 199.

E. Baselines

E.1. Unimodal baselines

In this section, we explain the unimodal MIL baselines that we compare our proposed framework with.

1. **ABMIL** (Ilse et al., 2018): Attention-based multiple instance learning (ABMIL) first assigns patch-level importance scores through a local attention mechanism, where the score for one patch only depends on the contents of that patch. The attention-weighted sum of patches is used as the slide-level representation. The independence assumption of ABMIL neglects correlations between different patches.

2. **TransMIL** (Shao et al., 2021): Since ABMIL is unable to learn patch-level correlations, Transformer-based multiple instance learning (TransMIL) has been proposed. TransMIL first squares the sequence of low dimensional representations, then applies a Pyramidal Positional Encoding module to encode spatial knowledge, and finally uses Nystrom attention (Xiong et al., 2021) to approximate self-attention scores between patches. The CLS token is taken as the slide-level representation.

3. **Low-rank MIL** (Xiang & Zhang, 2022): While TransMIL tries to learn slide-level representations by encoding patch correlations, it does not leverage the redundancy in WSI, which (Xiang & Zhang, 2022) used to propose iterative low-rank attention (ILRA). Each ILRA block consists of two layers: one aims to project the sequence of patch representations to a low-rank space by cross-attending it with a latent matrix, and the second reconstructs the input. Max-pooling over the output of k such layers yields a low-rank slide-level representation.

4. **AttnMISL** (Yao et al., 2020): In contrast with ABMIL, TransMIL, and ILRA, which learn slide-level representations using patch representations, AttnMISL first clusters patches into morphological prototypes using K-means clustering. Next, each prototype is encoded using prototype-specific fully convolutional Siamese networks (Yao et al., 2019). The slide-level representation is then created using local attention pooling over the prototypes.

5. **Information Bottleneck MIL** (Li et al., 2023): Information bottlenecks (IB) are used to compress a WSI by removing irrelevant instances. IB aims to find patch instances that minimize the mutual information between the distribution of patches and patch representations. By only keeping such instances, (Li et al., 2023) argue that most informative patches are retained, which can then be aggregated into a compact representation of WSI.

6. **Unimodal MMP**: This is similar to PANTHER (Song et al., 2024a) in that GMM is used to map each histology patch embeddings into a pre-defined set of morphological prototypes. However, whereas PANTHER concatenates the post-aggregation embeddings to form the slide representation, the unimodal MMP employs $f_{\text{h.}}^{\text{pre}}$, $f_{c,\text{g.}}^{\text{pre}}$, and f_c^{post} that are learned along with the downstream tasks.

E.2. Multimodal baselines

We compare our proposed method MMP with several early-fusion multimodal survival baselines.

1. **MCAT** (Chen et al., 2021): Multimodal Co-Attention Transformer (MCAT) is an early fusion technique that learns a dense co-attention mapping between histology and omic tokens. This mapping is then used to calculate omic-guided histology features, which are concatenated with omics to predict patient survival. MCAT uses omic prototypes because it groups genes into 6 functional families.

2. **MOTCat** (Xu & Chen, 2023): Multimodal Optimal Transport-based Co-attention Transformer (MOTCat) uses Optimal Transport to learn an optimal plan between histology tokens and genes grouped into 6 functional groups, similar to MCAT. The estimated optimal transport plan is then used for selecting the most informative histology tokens.

3. **SurvPath** (Jaume et al., 2024): Unlike MCAT and MOTCat, which are limited to six gene families, SurvPath introduces a transcriptomics tokenizer to encode genes into biological pathways that represent known cellular functions. The pathway tokens are then fused with histology patches via a memory-efficient transformer, which learns interactions between pathways and those between pathways and histology, but does not learn histology-to-histology interactions.

4. **CMTA** (Zhou & Chen, 2023): Cross-Modal Translation and Alignment (CMTA) framework uses two parallel Transformer encoder-decoder modules. Encoders are used to extract intra-modal representations for each modality. Decoders generate cross-modal representations. A cross-modal attention module between the two encoders facilitates learning the cross-modal relations.

F. Clinical baselines

We assess how MMP and other survival prediction frameworks perform against basic clinical information included in patient metadata. Based on age, sex, and grade, empirically shown as crucial prognostic factors (Bonnier et al., 1995; Rakha et al., 2010; Tas et al., 2013; Yu et al., 2022), we perform univariate/multivariate linear Cox regression to obtain the baseline. We observe that MMP outperforms the baseline overall, hinting at its clinical potential for patient prognosis.

Table 6. **Survival prediction with clinical variables** The clinical variables for the TCGA cohort were downloaded from cBioPortal. All denotes the combination of age, sex, and grade.

Dataset	BRCA	BLCA	LUAD	STAD	CRC	KIRC	Avg. (↑)
Age	0.496$\pm$0.086	0.578$\pm$0.056	0.533$\pm$0.063	0.449$\pm$0.055	0.357$\pm$0.161	0.554$\pm$0.147	0.495
Sex	0.490$\pm$0.011	0.489$\pm$0.028	0.480$\pm$0.049	0.529$\pm$0.069	0.542$\pm$0.070	0.437$\pm$0.057	0.495
Grade	0.597$\pm$0.078	0.515$\pm$0.018	N/A	0.552$\pm$0.055	N/A	0.594$\pm$0.083	N/A
All	0.563$\pm$0.055	0.570$\pm$0.033	0.528$\pm$0.028	0.592$\pm$0.044	0.655$\pm$0.119	0.602$\pm$0.066	0.585
MMP$_{\text{Trans.}}$	0.738$\pm$0.069	0.635$\pm$0.051	0.642$\pm$0.037	0.598$\pm$0.051	0.630$\pm$0.125	0.747$\pm$0.106	0.665

G. Histology ablations

We perform additional experiments in four cancer types, varying: (1) the number of histology prototypes ($C_h = 8, 16, 32$) and (2) the pretrained encoder (ResNet50, CTransPath, and UNI). The results are shown in **Tables** 7, 8.

We observe that performance with UNI features is relatively consistent across C_h, with $C_h = 32$ being the weakest. The choice of $C_h = 16$ was influenced by two factors: (1) This gave the best overall performance in multimodal evaluation.

(2) $C_h = 8$ sometimes fails to distinguish between two similar but subtly different morphological exemplars (by grouping them into a single cluster), whereas $C_h = 32$ induces harder morphological interpretation due to an excessive number of exemplars. $C_h = 16$ offered the best trade-off.

Table 7. **Ablation on the number of histology prototypes** Unimodal MMP was trained on varying number of histology prototypes $C_{\text{h.}}$ for select cancer types.

Dataset	BRCA	BLCA	LUAD	CRC	Avg. (↑)
$C_{\text{h.}} = 8$	0.720± 0.06	0.601± 0.04	0.592± 0.04	0.641± 0.11	0.639
$C_{\text{h.}} = 16$	0.669± 0.12	0.593 ± 0.06	0.600± 0.04	0.646± 0.11	0.627
$C_{\text{h.}} = 32$	0.680 ± 0.09	0.590± 0.05	0.587± 0.04	0.617± 0.13	0.619

Table 8. **Ablation on the histology encoder** Unimodal MMP was trained on different histology encoders for select cancer types.

Dataset	BRCA	BLCA	LUAD	CRC	Avg. (↑)
ResNet50	0.574± 0.11	0.511± 0.05	0.600± 0.06	0.534± 0.18	0.555
CTransPath	0.653± 0.10	0.566± 0.05	0.578 ± 0.02	0.574 ± 0.14	0.593
UNI	0.669 ± 0.12	0.593 ± 0.06	0.600 ± 0.04	0.646 ± 0.11	0.627

H. Survival loss ablation experiment

We assess how the train batch size affects the performance, using the Cox proportional hazards loss (Cox, 1972) and NLL survival loss (Zadeh & Schmid, 2020) (**Table 9**). To this end, we use the MMP full model. We observe that the C-Index increases with a larger batch size until it reaches the peak and starts to decline, regardless of the loss function (peak for Cox loss: 0.665 with batch size 64 and NLL loss: 0.644 with batch size 16). The increase can be attributed to stable training from having more patients in each batch to compare the predicted risks against (Kvamme et al., 2019). The decrease is likely due to a smaller number of parameter updates within the same number of epochs. This suggests the benefits of batch-based training for survival prediction, which does not apply to non-prototype-based approaches as they rely on the NLL survival loss with a single patient batch due to large $N_{\text{h.}}$. We also observe that employing the Cox loss gives a better overall performance, which illustrates another benefit of forming a batch of patients in MMP with fewer tokens. We attribute the lower performance of NLL survival loss to the discretization of time into non-overlapping coarse bins, which might result in discarding valuable survival information.

Table 9. **Batch size ablation.** Average C-Index across 5 cross-validation folds with varying batch sizes of patients with Cox and NLL loss. A batch of a single patient cannot be used for Cox loss.

		BRCA	BLCA	LUAD	STAD	CRC	KIRC	Avg.(↑)
Cox	$B = 1$	N/A	N/A	N/A	N/A	N/A	N/A	N/A
	$B = 16$	0.711	**0.642**	0.648	0.558	**0.635**	0.730	0.654
	$B = 32$	0.729	0.636	**0.648**	0.584	0.627	0.735	0.660
	$B = 64$	**0.738**	0.635	0.645	**0.598**	0.630	**0.744**	**0.665**
	$B = 128$	0.729	0.622	0.644	0.586	0.617	0.731	0.655
NLL	$B = 1$	0.664	0.602	0.616	0.508	0.627	**0.712**	0.621
	$B = 16$	**0.662**	**0.635**	**0.656**	0.561	**0.660**	0.691	**0.644**
	$B = 32$	0.618	0.622	0.646	**0.570**	0.554	0.690	0.617
	$B = 64$	0.590	0.616	0.635	0.556	0.574	0.678	0.608
	$B = 128$	0.587	0.610	0.623	0.523	0.523	0.640	0.584

I. Additional interpretability results

Unimodal: Compared to **Fig. 2**, **Fig. 3A** visualizes multiple prototype heatmaps, illustrating how different prototypes reflect distinct morphological tissue patterns in the tumor microenvironment. Using their nearest histology patches, a board-certified pathologist qualitatively assessed and captioned each prototype with a general morphological description. We found that using $C_{\text{h.}} = 16$ may still have redundancy in unique prototypes, as multiple prototypes are found to describe

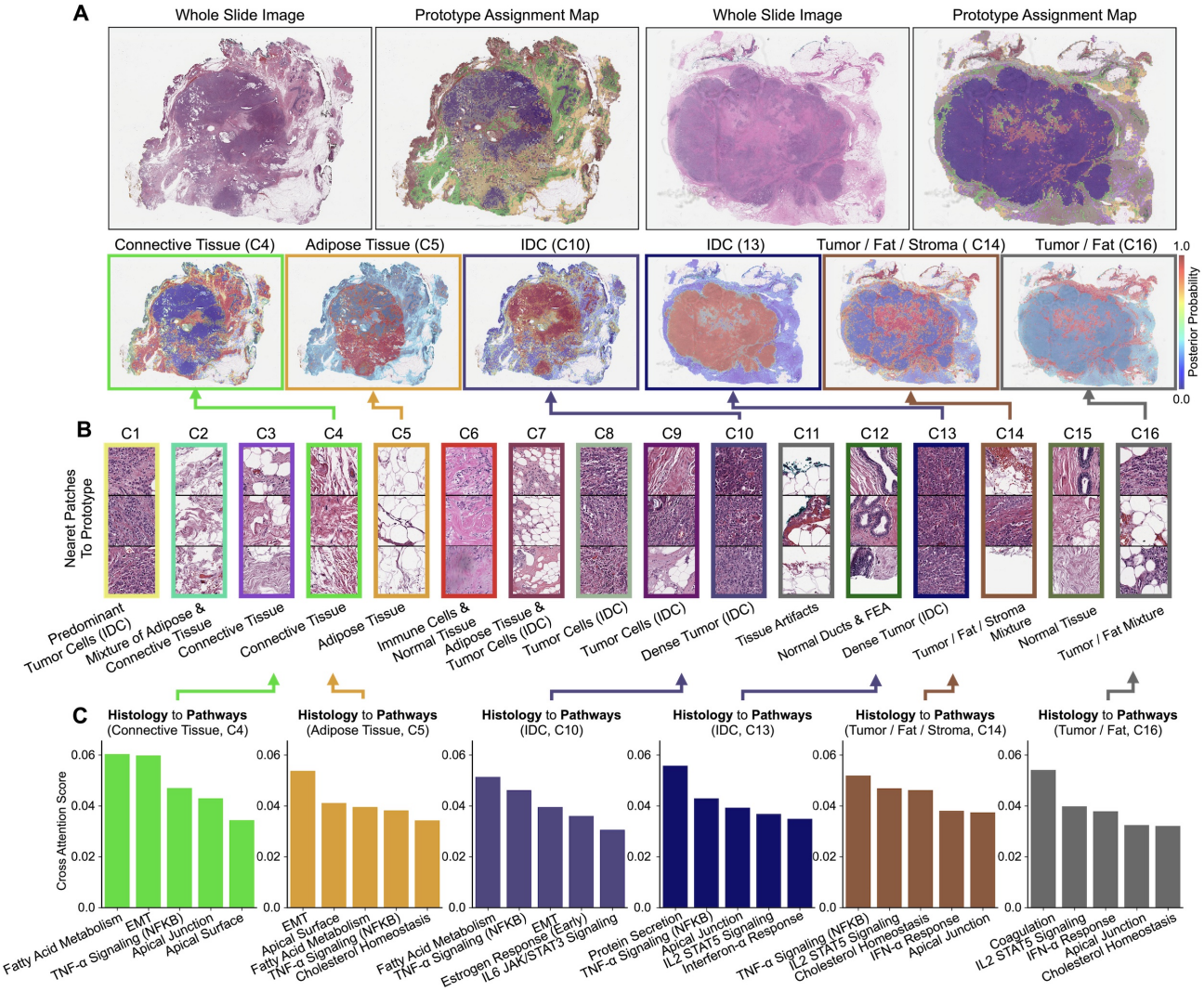

Figure 3. **Cross-modal interaction visualization**. (A) BRCA WSIs with their prototype assignment map (categorical assignment of each histology patch to their nearest prototype), and prototype heatmaps of the the top-3 prominent tissue patterns in the WSI. (B) Morphological annotations provided by a board-certified pathologist of the nearest histology patches for each prototype. (C) For each prototype visualized in (A), we can visualize its most highly-attended pathways (h. $\rightarrow$ g.), *i.e.*, which pathways correspond to the queried prototype (pathway importance).

IDC presence (C1, C8, C9, C10, C13). In general, however, each prototype was still found to be semantic in delineating general tumor tissue, normal connective tissue and stroma, adipose tissue, and tissue with immune cell presence, which is reflected in the high performance of unimodal MMP over other histology baselines found in **Table 1** (with MMP using $C_{\text{h.}} = 8$ and $C_{\text{h.}} = 32$ having worse performance in **Table 3**).

Multimodal: In **Fig. 3C**, we further visualize cross-modal histology to pathway interactions (h. $\rightarrow$ g.) in BRCA, a unique capability in MMP compared to other works which have only visualized pathway to histology interactions (g. $\rightarrow$ h.) (Chen et al., 2021; Xu et al., 2023; Jaume et al., 2024). Across all (h. $\rightarrow$ g.) visualizations for the prototypes shown in **Fig. 3C**, fatty acid metabolism and cholesterol homeostasis were conserved in having high cross-attention scores, which corroborates with biomedical literature on how cancer cells hijack these pathways for exogenous energy uptake from the tissue microenvironment (enabling tumorigenesis and cancer progression) (Nelson et al., 2014; Koundouros & Poulogiannis, 2020). Other conserved and highly-attended pathways include tumor necrosis factor (TNF)-α signaling and epithelial-mesenchymal transition (EMT), which are canonical markers related to tumor proliferation and invasion (Wu & Zhou, 2010; Dongre & Weinberg, 2019). We note that SurvPath also found EMT to have high importance, however,

we note a subtle difference in that EMT importance was derived from attribution-based interpretability with respect to predicted survival risk, and not via cross-attention that pinpoints a relationship with an exact morphological pattern. In MMP, we find that EMT not only attends to invasive tumor (C10), but is also the most highly-attended pathway to adipose tissue (C5), which corroborates with recent and accumulating evidence of adipose tissue being more than a causal observer in contributing to inflammation and tumor progression (Wang et al., 2012; Olea-Flores et al., 2018; Giudetti et al., 2019; Ishay-Ronen et al., 2019; Olea-Flores et al., 2020; Loo et al., 2021).

J. Limitations & recommendations for future directions

Multimodal interpretability: Due to potential redundancy of prototypes (corresponding to unique morphological patterns), queries for (h. $\rightarrow$ g.) are not unique, with many prototypes associated with tumor cell presence of IDC morphology and thus querying similar pathways. In **Fig. 2**, we also note potential asymmetrical relationships in histology-pathway correspondences, in which cholesterol homeostasis highly attends to C13 (top 3 pathways out of 50) but C13 does not attend as highly to cholesterol homeostasis (top 6 prototypes out of 16). Again, this may be due to the redundancy of prototypes, with other IDC-related prototypes (C8) highly attending to cholesterol homeostasis instead. We note that though pathologist annotation found many clusters to correspond to similar morphological patterns for tumors, there may exist subtle differences in fine-grained features such as tumor grade, tumor invasiveness, tumor colocalization with stroma, adipose tissue, and immune cells which may have fine-grained interactions to pathways. Future directions include developing approaches that would narrow down the number of unique prototypes, which may improve both survival modeling and cross-modal interpretability.

Study designs involving TCGA: The TCGA is the largest publicly-available pan-cancer atlas with paired histology-omic samples, and has been an immeasurable resource for the CPath community in building computational tools for unimodal and multimodal cancer prognosis. Still, the TCGA has several limitations which we provide caution. First, in addition to issues such as site-specific H&E intensity bias (Howard et al., 2021), and demographic bias (Vaidya et al., 2024), pretrained encoders developed on the TCGA should also be avoided when evaluating multimodal cancer prognosis tasks due to potential issues in data contamination. Though UNI was not pretrained on TCGA (Chen et al., 2024), using UNI (or any pretrained ROI encoder) as a part of non-parametric methods such as K-means clustering or GMMs may still lead to instances where all patches can be assigned to a single prototype, as demonstrated in PANTHER (Song et al., 2024a). Second, important consideration must be taken in utilizing the different survival endpoints available for each TCGA cohort. For instance, the median time-to-event and time-to-censor for disease-specific survival TCGA-BRCA is 26 and 25 months respectively, meaning that the follow-up time is too short to see breast cancer-specific deaths. Other works which have assessed the suitability of DSS as a survival endpoint in TCGA-BRCA were able to still show statistically significant differences between ER+ and ER- tumors, while also acknowledging its shortcomings (Liu et al., 2018).

Unimodal versus multimodal survival analysis: As emphasized in the **Introduction** and **Related Work** sections, multimodal survival analysis is a challenging clinical task that has seen significant interest in the biomedical, computer vision, and machine learning communities. Though multimodal integration generally outperforms unimodal baselines, we note that the development of better unimodal baselines may (or may not) close the performance gap for certain cancer types, which is an area of further exploration. In PORPOISE (Chen et al., 2022) and MCAT (Chen et al., 2021), multimodal integration (using ResNet50 features transferred from ImageNet for histology and gene families from MutSigDB for genomics) was found to improve in 9 out of 14 cancer types in the TCGA, with genomics generally outperforming histology in unimodal baselines. In SurvPath (Jaume et al., 2024), MOTCat (Xu et al., 2023) and PIBD (Zhang et al., 2024), which improved unimodal baselines in MCAT using CTransPath features and hallmark gene family features, also found very similar trends with multimodal improvement. Interestingly, MCAT was shown to lag behind unimodal genomics in the analysis of SurvPath, which may be attributed to not only stronger gene features used, but also higher computational complexity with the increased number of omics tokens used for Transformer attention (thus necessitating computational efficiency). In MMP, which improves the unimodal histology baseline further using UNI features, we observe that the unimodal ablation of MMP (based on GMM, 0.611 overall C-Index) is able to catch up with unimodal genomics baselines (0.612 to 0.614 overall C-Index) and also with multimodal baselines like MCAT (0.610 overall C-Index) (**Table 1**). We hypothesize that this is due to the simplicity of GMMs in representing WSIs as a fixed set of prototypes, which thus allows supervision using the Cox loss instead of the negative log-likelihood loss. As better unimodal baselines are developed, we envision new types of multimodal fusion techniques will also be needed that would emphasise simplicity and interpretability in developing easy-to-train survival methods in high-dimensional, low-sample size regimes for cancer prognostication.