Modeling Dense Multimodal Interactions Between Biological Pathways and Histology for Survival Prediction

Guillaume Jaume[1,2,*], Anurag Vaidya[1,2,*], Richard J. Chen[1,2], Drew F.K. Williamson[1,2,†]
Paul Pu Liang[3], Faisal Mahmood[1,2]
[1]Mass General Brigham, [2]Harvard University and [3]CMU
gjaume@bwh.harvard.edu, ajvaidya@bwh.harvard.edu, faisalmahmood@bwh.harvard.edu

Abstract

Integrating whole-slide images (WSIs) and bulk transcriptomics for predicting patient survival can improve our understanding of patient prognosis. However, this multimodal task is particularly challenging due to the different nature of these data: WSIs represent a very high-dimensional spatial description of a tumor, while bulk transcriptomics represent a global description of gene expression levels within that tumor. In this context, our work aims to address two key challenges: (1) how can we tokenize transcriptomics in a semantically meaningful and interpretable way?, and (2) how can we capture dense multimodal interactions between these two modalities? Here, we propose to learn biological pathway tokens from transcriptomics that can encode specific cellular functions. Together with histology patch tokens that encode the slide morphology, we argue that they form appropriate reasoning units for interpretability. We fuse both modalities using a memory-efficient multimodal Transformer that can model interactions between pathway and histology patch tokens. Our model, SurvPath*, achieves state-of-the-art performance when evaluated against unimodal and multimodal baselines on five datasets from The Cancer Genome Atlas. Our interpretability framework identifies key multimodal prognostic factors, and, as such, can provide valuable insights into the interaction between genotype and phenotype. Code available at https://github.com/mahmoodlab/SurvPath.*

1. Introduction

Predicting patient prognosis is a fundamental task in computational pathology (CPATH) that aims to utilize histology whole-slide images (WSIs) for automated risk assessment, patient stratification, and response-to-treatment prediction [3, 5, 6, 13, 61, 62]. Patient prognostication is often framed as a survival task, in which the goal is to learn risk estimates that correctly rank the survival time from the primary diagnostic WSI(s) [34, 36, 71, 81, 85]. As WSIs can be as large as 100,000 × 100,000 pixels, weakly supervised methods such as multiple instance learning (MIL) are often employed for survival prediction. In MIL, WSIs are tokenized into small patches, from which features are extracted and fed into pooling networks, such as attention networks, for downstream classification [30, 59].

While histology provides phenotypic information about cell types and their organization into tissues, alternate modalities can provide complementary signals that may in-

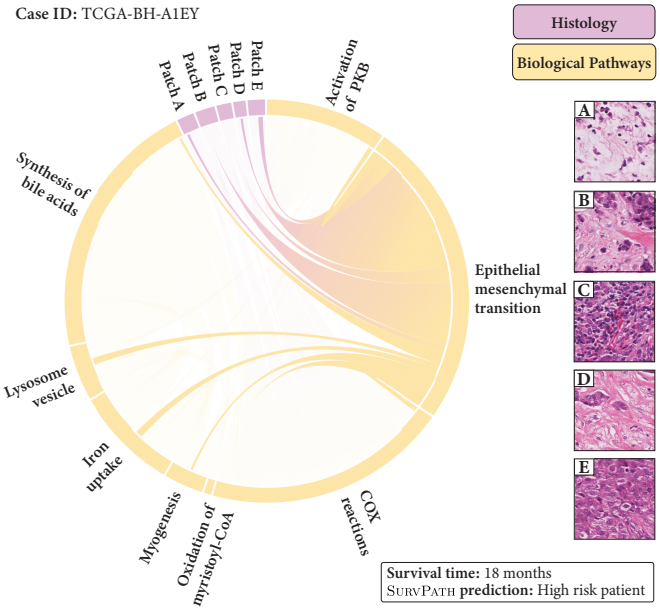

Figure 1. **Multimodal interpretability with SurvPath.** SurvPath enables visualization of multimodal interactions via a Transformer cross-attention between *biological pathways* and *morphological patterns*, here exemplified in a high-risk breast cancer. The chord thickness denotes attention weight.

*Equal contribution
†Presently at Emory University School of Medicine

dependently be linked to prognosis. For instance, bulk transcriptomics, which represents the average gene expression in a tissue, can reveal a richer global landscape of cell types and cell states [37, 77] and has been shown to be a strong predictor of patient survival [24, 51, 56]. By combining both modalities, we can integrate the global information provided by bulk transcriptomics with the spatial information from the WSI. While most existing methods adopt *late fusion* mechanisms [11, 37] (*i.e.,* fusing modality-level representations), we design an *early fusion* method that can explicitly model fine-grained cross-modal relationships between local morphological patterns and transcriptomics. In comparison with widely employed vision-language models [2, 55, 70], multimodal fusion of transcriptomics and histology presents two key technical challenges:

1. *Tokenizing transcriptomics modality:* Modalities based on image and text can be unequivocally tokenized into object regions and word tokens [38, 70], however, tokenizing transcriptomics in a semantically meaningful and interpretable way is challenging. As transcriptomics data is already naturally represented as a feature vector, many prior studies ignore tokenization and directly concatenate the entire feature with other modalities, which limits multimodal learning to *late fusion* operations [37, 77]. Alternatively, genes can be partitioned into coarse functional sets that represent different gene families (*e.g.,* tumor-suppressor genes and oncogenes) that can be used as tokens [10]. Nevertheless, such sets provide a rudimentary and incomplete depiction of intracellular interactions as one gene family can be involved in different cellular functions. Consequently, they may lack semantic correspondence with fine-grained morphologies. Instead, we propose tokenizing genes according to established *biological pathways* [22, 40, 64]. Pathways are gene sets with known interactions that relate to *specific* cellular functions, such as the TGF-β signaling cascade, which contributes to the epithelial-mesenchymal transition in breast cancer [75]. Compared to coarse sets (*e.g.,* $N_{\mathcal{P}} = 6$ [10]), pathway-based gene grouping can yield hundreds to thousands of tokens that represent unique molecular processes ($N_{\mathcal{P}} = 331$ in our work), which we hypothesize are more suitable representations for multimodal fusion with histology. In addition, as pathways represent unique cellular functions, they constitute appropriate basic reasoning units for interpretability (see Fig. 1).

2. *Capturing dense multimodal interactions:* Early fusion of histology and pathway tokens can be done with a Transformer that uses self-attention to capture pairwise similarities between all tokens [69]. However, modeling pairwise interactions between large sets of histology patch tokens (*e.g.,* $N_{\mathcal{H}} = 15,000$) and pathway tokens ($N_{\mathcal{P}} = 331$) poses scalability challenges for fusion. Due to the quadratic complexity of the Transformer attention, modeling all possible interactions imposes substantial computational and memory requirements. To tackle this issue, we introduce a new unified, memory-efficient attention mechanism that can model patch-to-pathway, pathway-to-patch, and pathway-to-pathway interactions. Modeling these three forms of interaction is achieved by the following: (1) designing the queries, keys, and values to share parameters across token types [31, 32], and (2) simplifying the attention layer to ignore patch-to-patch interactions, which we find through experimentation to be not as effective for survival analysis.

To summarize, our contributions are (1) a transcriptomics tokenizer that leverages existing knowledge of cellular biology to generate *biological pathway* tokens; (2) SURVPATH, a memory-efficient and resolution agnostic, multimodal Transformer formulation that integrates transcriptomics and patch tokens for predicting patient survival; (3) a multi-level interpretability framework that enables deriving unimodal and cross-modal insights about the prediction; (4) a series of experiments and ablations showing the predictive power of SURVPATH, using five datasets from The Cancer Genome Atlas Program (TCGA) and benchmarked against both unimodal and multimodal fusion methods.

2. Related Work

2.1. Survival Analysis on WSIs

Recently, several histology-based survival models have been proposed [60, 63, 79, 85, 89]. Most contributions have been dedicated to modeling tumor heterogeneity and the tumor microenvironment using MIL. To this end, several MIL pooling strategies have been proposed, such as using graph neural networks to model local patch interactions [16, 36, 46], accounting for the variance between patch embeddings [57], or adopting multi-magnification patch representations [42].

2.2. Multimodal Transformers and Interpretability

In parallel, the use of Transformers for multimodal fusion has gained significant attention in classification and generative tasks [58, 67, 83]. Multimodal tokens can be concatenated and fed to a regular Transformer [18, 69], a hierarchical Transformer [41], or a cross-attention Transformer [43, 49, 52]. As the number and dimensionality of modalities increase, the typical sequence length can become too large to be fed to vanilla Transformers, hence the need for low-complexity methods. Several models have proposed re-formulations of self-attention to reduce memory and computational requirements [4, 12, 14, 15, 29, 31, 80, 82], for instance, by approximating self-attention with a low-rank decomposition [44, 82], using latent bottleneck distillation [31, 32, 50], by optimizing GPU reads/writes [14, 15] or using sparse attention patterns [4, 53]. Recently, interpretable multimodal models or post-hoc interpretation

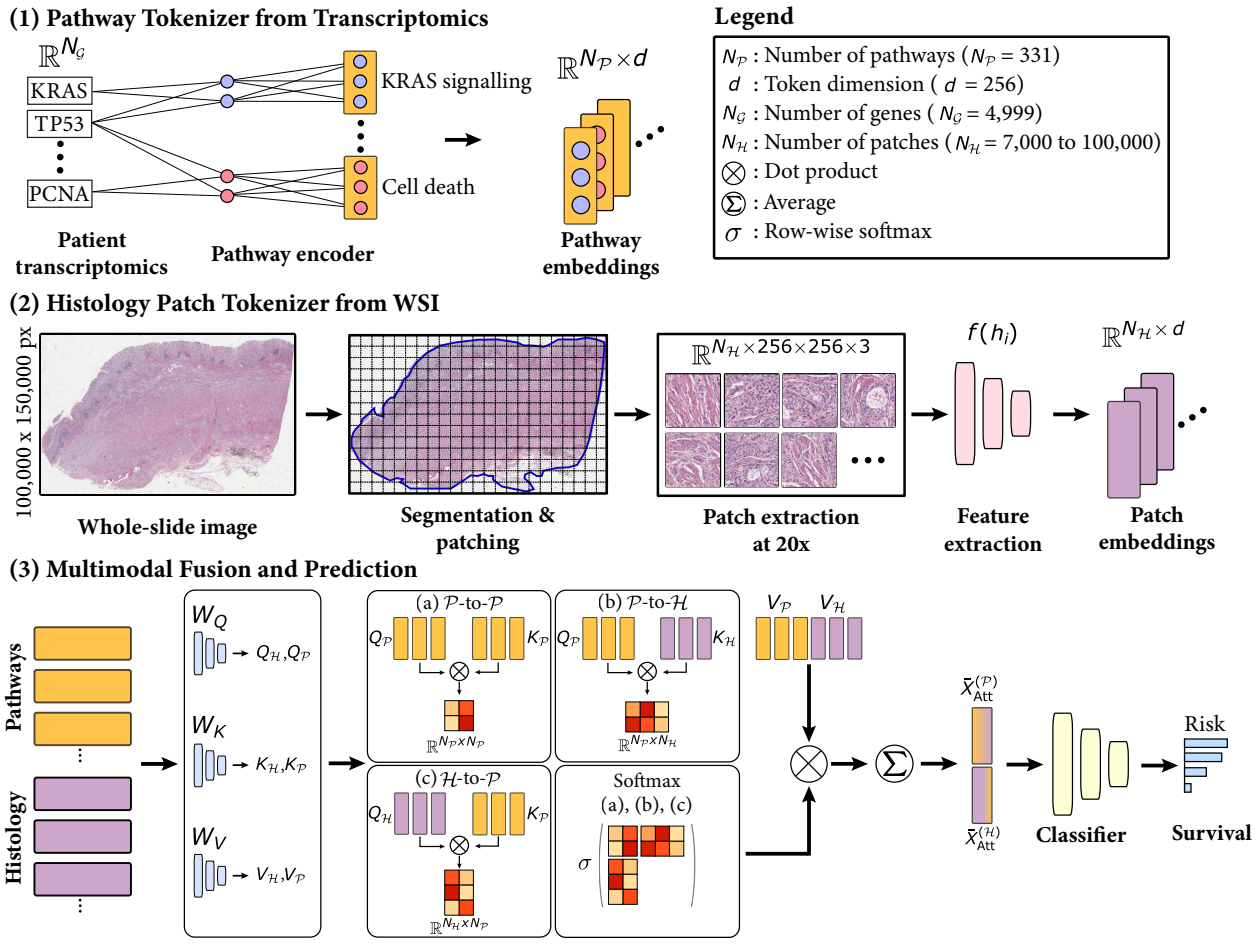

Figure 2. **Block diagram of SurvPath.** (1) We tokenize transcriptomics into *biological pathway* tokens that are semantically meaningful, interpretable, and end-to-end learnable. (2) We further tokenize the corresponding histology whole-slide image into patch tokens using an SSL pre-trained feature extractor. (3) We combine pathway and patch tokens using a memory-efficient multimodal Transformer for survival outcome prediction.

methods [1, 39, 72] have also emerged as a critical area of research, especially in sensitive human-AI collaborative decision-making scenarios such as healthcare and human-computer interactions.

2.3. Multimodal Survival Analysis

Multimodal integration is an important objective in cancer prognosis [61], as combining histology and omics data such as genomics or transcriptomics is the current clinical practice for many cancer types. The majority of these works employ *late fusion* mechanisms [9, 68], and mostly differ in the way modality fusion is operated. Fusion can be based on vector concatenation [48], modality-level alignment [7], bilinear pooling (*i.e.,* Kronecker product) [9, 77], or factorized bilinear pooling [37, 54].

Differently, *early fusion* mechanisms can be employed, in which cross-modal interactions between individual constituents of the input are modeled [10, 17, 84, 88]. Our work builds off MCAT [10], which uses a cross-attention module to model the attention of histology patches (keys, values) toward gene sets (queries). However, MCAT has several limitations: (1) cross-attention being one-sided and models only patch-to-genes interactions, (2) transcriptomics tokenization using coarse sets that do not reflect actual molecular processes, and (3) significant gene overlap between sets, which leads to redundant cross-attention heatmaps.

3. Method

Here, we present SurvPath, our proposed method for multimodal survival prediction based on histology and transcriptomics. Sec. 3.1 presents the transcriptomics encoder to build biological pathway tokens, Sec. 3.2 presents the histology encoder to build patch tokens, Sec. 3.3 presents our Transformer-based multimodal aggregation, and Sec. 3.4

presents its application to survival prediction (see Fig. 2). Finally, Sec. 3.5 introduces our multi-level interpretability framework.

3.1. Pathway Tokenizer from Transcriptomics

Composing pathways: Selecting the appropriate reasoning unit for transcriptomics analysis is challenging, owing to the intricate and hierarchical nature of cellular processes. Pathways, consisting of a group of genes or subpathways involved in a particular biological process, represent a natural reasoning unit for this analysis. A comparison may be drawn to action recognition, where an action (*i.e.,* a biological pathway) can be described by a series of movements captured by sensors (*i.e.,* transcriptomics measurements of a group of genes).

Encoding pathways: Given a set of transcriptomics measurements of $N_{\mathcal{G}}$ genes, denoted as $\mathbf{g} \in \mathbb{R}^{N_{\mathcal{G}}}$, and the composition of each pathway, we aim to build pathway-level tokens $\mathbf{X}^{(\mathcal{P})} \in \mathbb{R}^{N_{\mathcal{P}} \times d}$, where d denotes the token dimension. Transcriptomics can be seen as tabular data, which can be efficiently encoded with multilayer perceptrons (MLPs). Specifically, we are learning pathway-specific weights ϕ_i, *i.e.,* $\mathbf{x}_i^{(\mathcal{P})} = \phi_i(\mathbf{g}_{\mathcal{P}_i})$, where $\mathbf{g}_{\mathcal{P}_i}$ is the gene set present in pathway $\mathcal{P}_i$. This can be viewed as learning a *sparse* multi-layer perceptron (S-MLP) [20, 25, 45] that maps transcriptomics $\mathbf{g} \in \mathbb{R}^{N_{\mathcal{G}}}$ to tokens $\mathbf{x}^{(\mathcal{P})} \in \mathbb{R}^{N_{\mathcal{P}} d}$. The network sparsity is controlled by the gene-to-pathway connectivity embedded in the S-MLP weights. By simply reshaping $\mathbf{x}^{(\mathcal{P})} \in \mathbb{R}^{N_{\mathcal{P}} d}$ into $\mathbf{X}^{(\mathcal{P})} \in \mathbb{R}^{N_{\mathcal{P}} \times d}$, we define pathway tokens that can be used by the Transformer. Each pathway token corresponds to a deep representation of the gene-level transcriptomics that comprises it, which is both (1) interpretable as it encodes a specific biological function and (2) learnable in an end-to-end fashion with respect to the prediction task.

3.2. Histology Patch Tokenizer from WSIs

Given an input WSI, we aim to derive low-dimensional patch-level embeddings defining patch tokens. We start by identifying tissue regions to ensure that the background, which carries no biological meaning, is disregarded. Then, we decompose the identified tissue regions into a set of $N_{\mathcal{H}}$ non-overlapping patches at 20× magnification (or $\sim 0.5\,\mu m$/pixel resolution), that we denote as $\mathbf{H} = \{\mathbf{h}_1, ..., \mathbf{h}_{N_{\mathcal{H}}}\}$. Due to the large number of patches per WSI (*e.g.,* can be > 50,000 patches or 78 GB as floats), patch embeddings need to be extracted prior to model training to reduce the overall memory requirements. Formally, we employ a pre-trained feature extractor $f(\cdot)$ to map each patch $\mathbf{h}_i$ into a patch embedding as $\mathbf{x}_i^{(\mathcal{H})} = f(\mathbf{h}_i)$. In this work, we use a Swin Transformer encoder that was pre-trained via contrastive learning on more than 15 million pan-cancer histopathology patches [73, 74]. The resulting patch embeddings represent compressed representations of the patches (compression ratio of 256), that we further pass through a learnable linear transform to match the token dimension d, yielding patch tokens $\mathbf{X}^{(\mathcal{H})} \in \mathbb{R}^{N_{\mathcal{H}} \times d}$.

3.3. Multimodal Fusion

We aim to design an early fusion mechanism to model dense multimodal interactions between pathway and patch tokens. We employ Transformer attention [69] that measures and aggregates pair-wise interactions between multimodal tokens. Specifically, we define a multimodal sequence by concatenating the pathway and patch tokens resulting in $(N_{\mathcal{H}} + N_{\mathcal{P}})$ tokens of dimensions d, and denoted as $\mathbf{X} \in \mathbb{R}^{(N_{\mathcal{P}}+N_{\mathcal{H}}) \times d}$. Following the self-attention terminology [69], we define three linear projections of the tokens using learnable matrices, denoted as $\mathbf{W}_Q \in \mathbb{R}^{d \times d_q}$, $\mathbf{W}_K \in \mathbb{R}^{d \times d_k}$, and $\mathbf{W}_V \in \mathbb{R}^{d \times d_v}$ to extract the queries ($\mathbf{Q}$), keys ($\mathbf{K}$), values ($\mathbf{V}$), and self-attention $\mathbf{A}$, setting $d = d_k = d_q = d_v$. Transformer attention is then defined as:

$$\mathbf{X}_{\text{Att}} = \sigma\Big(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\Big)\mathbf{V} = \begin{pmatrix} \mathbf{A}_{\mathcal{P}\to\mathcal{P}} & \mathbf{A}_{\mathcal{P}\to\mathcal{H}} \\ \mathbf{A}_{\mathcal{H}\to\mathcal{P}} & \mathbf{A}_{\mathcal{H}\to\mathcal{H}} \end{pmatrix} \begin{pmatrix} \mathbf{V}_{\mathcal{P}} \\ \mathbf{V}_{\mathcal{H}} \end{pmatrix} \tag{1}$$

where σ is the row-wise softmax. The term $\mathbf{Q}\mathbf{K}^T$ has memory requirements $\mathcal{O}\big((N_{\mathcal{H}} + N_{\mathcal{P}})^2\big)$, which for long sequences becomes expensive to compute. This constitutes a major bottleneck as a WSI can have $N_{\mathcal{H}} > 50,000$ patches, making this computation challenging on most hardware. Instead, we propose to decompose the multimodal Transformer attention into four intra- and cross-modality terms: (1) the intra-modal pathway self-attention encoding pathway-to-pathway interactions $\mathbf{A}_{\mathcal{P}\to\mathcal{P}} \in \mathbb{R}^{N_{\mathcal{P}} \times N_{\mathcal{P}}}$, (2) the cross-modal pathway-guided cross-attention encoding pathway-to-patch interactions $\mathbf{A}_{\mathcal{P}\to\mathcal{H}} \in \mathbb{R}^{N_{\mathcal{P}} \times N_{\mathcal{H}}}$, (3) the cross-modal histology-guided cross attention encoding patch-to-pathway interactions $\mathbf{A}_{\mathcal{H}\to\mathcal{P}} \in \mathbb{R}^{N_{\mathcal{H}} \times N_{\mathcal{P}}}$, and (4) the intra-modal full histology self-attention encoding patch-to-patch interactions $\mathbf{A}_{\mathcal{H}\to\mathcal{H}} \in \mathbb{R}^{N_{\mathcal{H}} \times N_{\mathcal{H}}}$.

As the number of patch tokens is much larger than the number of pathways, *i.e.,* $N_{\mathcal{H}} >> N_{\mathcal{P}}$, most memory requirements come from computing and storing $\mathbf{A}_{\mathcal{H}\to\mathcal{H}}$. To address this bottleneck, we approximate Transformer attention as:

$$\hat{\mathbf{X}}_{\text{Att}} = \begin{pmatrix} \mathbf{X}_{\text{Att}}^{(\mathcal{P})} \\ \hat{\mathbf{X}}_{\text{Att}}^{(\mathcal{H})} \end{pmatrix} = \sigma\left[\frac{1}{\sqrt{d}} \begin{pmatrix} \mathbf{Q}_{\mathcal{P}}\mathbf{K}_{\mathcal{P}}^T & \mathbf{Q}_{\mathcal{P}}\mathbf{K}_{\mathcal{H}}^T \\ \mathbf{Q}_{\mathcal{H}}\mathbf{K}_{\mathcal{P}}^T & -\infty \end{pmatrix}\right]\mathbf{V} \tag{2}$$

where $\mathbf{Q}_{\mathcal{P}}$ (respectively $\mathbf{K}_{\mathcal{P}}$) and $\mathbf{Q}_{\mathcal{H}}$ (respectively $\mathbf{K}_{\mathcal{H}}$) denotes the subset of pathway and histology queries and keys. Setting pre-softmax patch-to-patch interactions to

$-\infty$ is equivalent to ignoring these interactions. Expanding Eq. 2, we obtain that $\mathbf{X}_{\text{Att}}^{(\mathcal{P})} = \sigma\left(\frac{\mathbf{Q}_{\mathcal{P}}\mathbf{K}^T}{\sqrt{d}}\right)\mathbf{V}_{\mathcal{P}}$, and $\hat{\mathbf{X}}_{\text{Att}}^{(\mathcal{H})} = \sigma\left(\frac{\mathbf{Q}_{\mathcal{H}}\mathbf{K}_{\mathcal{P}}^T}{\sqrt{d}}\right)\mathbf{V}_{\mathcal{H}}$. The number of interactions becomes drastically smaller, enabling computing $\hat{\mathbf{A}}$ with limited memory. This formulation can be seen as a sparse attention pattern [4] on a multimodal sequence, where sparsity is imposed between patch tokens. This formulation is parameter-efficient as a unique set of keys, queries, and values is learned for encoding both modalities. Additionally, this formulation resembles a graph neural network on a graph where pathways interconnect, and each pathway links to all patches. After passing $\hat{\mathbf{X}}_{\text{Att}}$ through a feed-forward layer with layer normalization, we take the mean representation of the post-attention pathway and patch tokens denoted as $\bar{\mathbf{x}}_{\text{Att}}^{\mathcal{P}}$ and $\bar{\mathbf{x}}_{\text{Att}}^{\mathcal{H}}$, respectively. The final representation $\bar{\mathbf{x}}_{\text{Att}}$, is then defined by the concatenation of $\bar{\mathbf{x}}_{\text{Att}}^{\mathcal{P}}$ and $\bar{\mathbf{x}}_{\text{Att}}^{\mathcal{H}}$.

3.4. Survival Prediction

Using the multimodal embedding $\bar{\mathbf{x}}_{\text{Att}} \in \mathbb{R}^{2d}$, our supervised objective is to predict patient survival. Following previous work [87], we define the patient's survival state by: (1) censorship status c, where $c = 0$ represents an observed patient death 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 diagnostic and observed death if $c = 0$, or the last follow-up if $c = 1$. Instead of directly predicting the observed time of event t, we approximate 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 denoted as y_j. The problem simplifies to classification, where each patient is now defined by $(\bar{\mathbf{x}}_{\text{Att}}, y_j, c)$. We define our classifier such that each output logit (after sigmoid activation) $\sigma(\hat{y}_j)$ represents the probability that the patient dies during time interval (t_{j-1}, t_j). We further take the cumulative product of the logits as $\prod_{k=1}^{j}\left(1 - \sigma(\hat{y}_k)\right)$ to represent the probability that the patient survives up to time interval (t_{j-1}, t_j). Finally, by taking the negative of the sum of all logits, we can define a patient-level risk used for training the network. More information are provided in the **Supplemental**.

3.5. Multi-Level Interpretability

We propose an interpretability framework that operates across multiple levels to derive transcriptomics, histology, and cross-modal interpretability (see **Supplemental**).

Transcriptomics: We employ Integrated Gradient (IG) [65] to identify the influence of *pathways* and *genes*, resulting in a score describing the degree to which each pathway, respectively gene, contributes to predicting the risk. A negative IG score corresponds to a pathway/gene associated with a lower risk, while a positive IG score indicates an association with a higher risk. A very small score denotes negligible influence. Such interpretability analysis serves two purposes: (1) validation of known genes and pathways associated with prognosis and (2) identification of novel gene and pathway candidates that could predict prognosis.

Histology: We process analogously with IG to derive *patch-level* influence that enables studying the morphology of low and high-risk-associated patches.

Cross-modal interactions: Finally, we can study *pathway-to-patch* and *patch-to-pathway* interactions using the learned Transformer attention matrix $\hat{\mathbf{A}}$. Specifically, we define the importance of patch j (respectively pathway) with respect to pathway i (respectively patch) as $\hat{\mathbf{A}}_{ij}$ (respectively $\hat{\mathbf{A}}_{ji}$). This enables building heatmaps correlating a pathway and corresponding morphological features. This interpretability property is unique to our framework and enables studying how specific cellular functions described by a pathway interact with histology.

4. Experiments

4.1. Dataset and Implementation

We evaluate SURVPATH on five datasets from TCGA: Bladder Urothelial Carcinoma (BLCA) (n=359), Breast Invasive Carcinoma (BRCA) (n=869), Stomach Adenocarcinoma (STAD) (n=317), Colon and Rectum Adenocarcinoma (COADREAD) (n=296), and Head and Neck Squamous Cell Carcinoma (HNSC) (n=392). Prior studies have focused on predicting overall survival (OS) [9], however, this approach risks overestimating the proportion of cancer-related deaths as patients may have succumbed to other causes. Instead, we predict disease-specific survival (DSS) as a more accurate representation of the patient's disease status.

Pathway collection: We used the Xena database [23] to access raw transcriptomics from TCGA ($N_{\mathcal{G}} = 60,499$ in total) along with DSS labels. We extracted pathways from two resources: Reactome [22] and the Human Molecular Signatures Database (MSigDB) – Hallmarks [40, 64]. Reactome and MSigDB–Hallmarks comprise 1,281 and 50 human biological pathways, respectively. We further selected pathways for which at least 90% of the transcriptomics are available, resulting in 331 pathways derived from 4,999 different genes (281 Reactome pathways from 1,577 genes and 50 Hallmarks pathways from 4,241 genes).

Histology collection: We collected all diagnostic WSIs used for primary diagnosis, resulting in 2,233 WSIs with an average of 14,509 patches per WSI at 20× (assuming 256×256 patches). In total, we collected over 2.86 TB of raw image data, comprising around 32.4 million patches.

Implementation: We used 5-fold cross-validation to train all models. Each split was stratified according to the sample site to mitigate potential batch artifacts [28]. To

Table 1. Results of SURVPATH and baselines in predicting disease-specific patient survival measured with c-Index (at 20×). Best performance in **bold**, second best underlined. Cat refers to concatenation, KP refers to Kronecker product. All omics and multimodal baselines were trained with the Reactome and Hallmark pathway sets.

	Model/Study	BRCA (↑)	BLCA (↑)	COADREAD (↑)	HNSC (↑)	STAD (↑)	Overall (↑)
WSI	ABMIL [30]	0.493±0.126	0.518±0.078	0.630±0.102	0.580±0.019	0.550±0.077	0.554
	AMISL [85]	0.500±0.000	0.500±0.000	0.500±0.000	0.518±0.015	0.506±0.014	0.508
	TransMIL [59]	0.530±0.057	0.551±0.091	0.632±0.143	0.523±0.043	0.544±0.080	0.556
Omics	MLP	0.611±0.080	0.627±0.062	0.625±0.060	0.548±0.045	0.586±0.098	0.599
	SNN [35]	0.528±0.094	0.584±0.113	0.521±0.109	0.550±0.065	0.565±0.080	0.550
	S-MLP [20]	0.512±0.028	0.595±0.114	0.581±0.066	0.542±0.052	0.515±0.081	0.549
Multimodal	ABMIL (Cat) [48]	0.541±0.158	0.562±0.067	0.592±0.102	0.580±0.089	0.523±0.098	0.560
	ABMIL (KP) [11]	0.615±0.083	0.566±0.038	0.584±0.109	0.566±0.066	0.525±0.140	0.571
	AMISL (Cat) [85]	0.462±0.179	0.518±0.055	0.510±0.137	0.478±0.051	0.538±0.025	0.501
	AMISL (KP) [85]	0.533±0.106	0.554±0.055	0.567±0.182	0.516±0.068	0.552±0.035	0.544
	TransMIL (Cat) [59]	0.598±0.087	**0.630**±0.047	0.539±0.189	0.542±0.091	0.536±0.090	0.569
	TransMIL (KP) [59]	0.629±0.144	0.625±0.079	0.566±0.081	0.515±0.116	0.552±0.035	0.577
	MOTCat [84]	0.600±0.095	0.596±0.079	0.641±0.182	0.560±0.062	0.550±0.103	0.589
	MCAT [10]	0.652±0.117	0.598±0.094	0.634±0.204	0.531±0.049	0.557±0.101	0.594
	SURVPATH (Ours)	**0.655**±0.089	0.625±0.056	**0.673**±0.170	**0.600**±0.061	**0.592**±0.047	**0.629**

increase variability during training, we randomly sampled 4,096 patches from the WSI. At test time, all patches were used to yield the final prediction (see **Supplemental**).

4.2. Baselines

We group as: (1) unimodal histology methods, (2) unimodal transcriptomics methods, and (3) multimodal methods (further sub-categorized into early *vs.* late fusion methods).

Histology baselines: All baselines use the same pretrained feature extractor as SURVPATH based on [73]. We compare with *ABMIL* [30], which uses a gated-attention pooling, *AMISL* [85], which first clusters patch embeddings using K-means before attention, and *TransMIL* [59], that approximates patch self-attention with Nyström method [82].

Transcriptomics baselines: All baselines use the same input defined by aggregating Reactome and Hallmarks transcriptomics. (a) *MLP* [27] uses a 4-layer MLP, (b) *SNN* [9, 27] supplements *MLP* with additional alpha dropout layers, and (c) *S-MLP* [20, 25] uses a 2-layer sparse pathway-aware MLP followed by a dense 2-layer MLP. This baseline shares similarities with our transcriptomics encoder.

Multimodal baselines: (a) **Late fusion:** We combine ABMIL [30], AMISL [85], and TransMIL [59] with an S-MLP using concatenation [48], denoted as *ABMIL (Cat)*, *AMISL (Cat)*, and *TransMIL (Cat)*, and Kronecker product [9, 21, 78, 86], denoted as *ABMIL (KP)*, *AMISL (KP)*, and *TransMIL (KP)*. (b) **Early fusion:** *MCAT* [10] which uses genomic-guided cross-attention followed by modality-specific self-attention blocks, and *MOTCat* [84] which uses Optimal Transport (OT) for matching the patch token and genomic token distributions.

4.3. Survival Prediction Results

Table 1 present results of SURVPATH and baselines evaluated at 20× magnification (see **Supplemental** for 10× analysis). SURVPATH reaches best overall performance, outperforming unimodal and multimodal baselines at both 20× and 10×. At 20×, SURVPATH reaches +7.3% compared to TransMIL, +3.0% compared to MLP, and +3.5 compared to MCAT. We attribute the high performance of SURVPATH to (1) the use of both modalities, (2) a unified, simple, and parameter-efficient fusion model, and (3) a semantically meaningful transcriptomics tokenizer.

Transcriptomics *vs.* Histology *vs.* Multimodal: Multimodal baselines significantly outperform histology baselines. Interestingly, a simple MLP trained on our set of transcriptomics constitutes a strong baseline that outperforms several multimodal methods. This highlights the challenge of performing robust feature selection and integrating heterogeneous and high-dimensional data modalities. In addition, the relatively small dataset size further complicates the learning of complex models and risk over-fitting. Comparisons against clinical variables are provided in **Supplemental**.

Context *vs.* No context: ABMIL and TransMIL perform similarly despite TransMIL modeling path-to-patch interactions using Nyström attention. This observation supports our design choice of disregarding patch-to-patch interactions. In addition, SURVPATH performance is similar across magnifications (0.629 overall c-index in both cases). This observation also holds for most histology and multimodal baselines.

Sparse *vs.* dense transcriptomics encoders: A dense MLP yields better performance than a sparse pathway-

Table 2. Studying design choices for tokenization (top) and fusion (bottom) in SURVPATH at 20× magnification. **Top:** *Single* refers to no tokenization, using tabular transcriptomics features as a single token. *Families* refers to the set of six gene families in MutSigDB, as used in [10]. *React.+Hallmarks* refers to the main SURVPATH model reported in Table 1. **Bottom:** $A_{\mathcal{P}\to\mathcal{P}}$ and $A_{\mathcal{P}\leftrightarrow\mathcal{H}}$ refers to pathway-to-pathway, pathway-to-patch, and patch-to-pathway interactions, which is the main SURVPATH model reported in Table 1. $\tilde{\mathbf{A}}$ refers to using Nyström attention to approximate $\mathbf{A}$.

	Model/Study	BRCA (↑)	BLCA (↑)	COADREAD (↑)	HNSC (↑)	STAD (↑)	Overall (↑)
Tokenizer	Single	0.625±0.149	0.560±0.086	0.604±0.176	0.580±0.075	0.563±0.140	0.586
	Families	0.620±0.094	0.613±0.061	0.671±0.111	0.600±0.076	0.540±0.071	0.609
	Hallmarks	0.645±0.039	**0.635**±0.093	0.633±0.151	0.589±0.076	0.581±0.039	0.615
	Reactome	0.579±0.044	0.604±0.080	0.639±0.200	0.574±0.061	**0.619**±0.047	0.602
	React.+Hallmarks	**0.655**±0.089	0.625±0.056	**0.673**±0.170	**0.600**±0.061	0.592±0.047	**0.629**
Fusion	$\mathbf{A}_{\mathcal{P}\to\mathcal{P}}$, $\mathbf{A}_{\mathcal{P}\to\mathcal{H}}$	0.446±0.116	0.603±0.038	0.565±0.166	0.526±0.030	0.582±0.053	0.544
	$\mathbf{A}_{\mathcal{P}\to\mathcal{P}}$, $\mathbf{A}_{\mathcal{H}\to\mathcal{P}}$	0.546±0.118	0.589±0.037	0.633±0.130	0.498±0.037	0.480±0.083	0.549
	$\mathbf{A}_{\mathcal{P}\to\mathcal{P}}$, $\mathbf{A}_{\mathcal{H}\to\mathcal{P}}$, $\mathbf{A}_{\mathcal{P}\to\mathcal{H}}$	**0.655**±0.089	0.625±0.056	**0.673**±0.170	**0.600**±0.061	0.592±0.047	**0.629**
	$\tilde{\mathbf{A}}$ [82]	0.555±0.066	0.565±0.101	0.612±0.194	0.508±0.032	0.493±0.086	0.547

aware MLP. However, sparse networks have shown to be particularly parameter-efficient when the number of genes considered drastically increases and are more interpretable than regular MLPs [20]. As the number of genes increases, this trend might evolve.

Early *vs.* Late fusion: *Early fusion* methods (MCAT [10], MOTCat [84] and SURVPATH) outperform all late fusion methods. We attribute this observation to the creation of a joint feature space that can model fined-grained interactions between transcriptomics and histology tokens. Overall, these findings justify the need for (1) modeling dense interactions between pathway and patch tokens and (2) unifying fusion in a single Transformer attention.

4.4. Ablation Study

To evaluate our design choices, we performed a series of ablations studying different *Tokenizers* and *Fusion* schemes.

Tokenizer: SURVPATH employs the Reactome and Hallmarks databases as sources of biological pathways. We assess the model performance when using each database in isolation, as well as using all genes assigned to one token (*Single*) and the gene families used in [10]. With increased granularity of transcriptomics tokens, the overall performance increases, showing that building semantic tokens brings interpretability properties and improves performance. We attribute this observation to the fact that each token encodes more and more specific biological functions, enabling better cross-modal modeling.

Fusion: We ablate SURVPATH by further simplifying Transformer attention to its left part considering $A_{\mathcal{P}\to\mathcal{P}}$ and $A_{\mathcal{H}\to\mathcal{P}}$, and to its top part $A_{\mathcal{P}\to\mathcal{P}}$ and $A_{\mathcal{P}\to\mathcal{H}}$ (this design resembles MCAT [10] where a single, shared multimodal attention layer is learned). Both branches bring complementary information (observed decrease of -5.6% and -7.5% in c-index), justifying the need to model both pathway-to-patch and patch-to-pathways interactions. We further adapt SURVPATH with Nyström attention that enables training on very long sequences by simplifying self-attention with a low-rank approximation. This yields significantly worse performance -6.9%. We hypothesize that the "true full attention" has low-entropy, making it more challenging to be approximated by low-rank methods [8], and that sparse attention patterns offer better approximations.

4.5. Interpretability

Examination of the multi-level interpretability can lead to novel biological insight regarding the interplay between pathways and histology in determining a patient's risk. Here, we compare a low (top) and high (bottom) risk case of breast invasive carcinoma (BRCA) (Fig. 3) and bladder urothelial carcinoma (BLCA) (see **Supplemental**).

In analyzing Fig. 3, we observe that several pathways have high absolute importance scores in the low and high-risk cases, most notably the Hallmark Epithelial-Mesenchymal Transition (EMT) [76] and COX Reactions pathways [47], both of which are known to be involved in breast cancer. EMT is thought to underlie tumor cells' ability to invade and metastasize [33], and the inverse importance of this pathway for the low- and high-risk cases is compatible with this analysis. This finding is enforced by studying the cross-modal interpretability that highlights the association of EMT with nests of tumor cells invading stroma. Members of the COX family of cyclooxygenases, especially COX-2, have also been implicated in breast carcinogenesis and are being investigated as a component of therapeutic regimens [26]. Cross-modal interpretability demonstrates stromal and immune cells in both cases. Though there is some overlap between important pathways in the two cases in Fig. 3, the majority differ between the two. For instance, in the high-risk case, a pathway relating to iron metabolism (a known contributor to breast carcinogenesis and prognosis [66]) was iden-

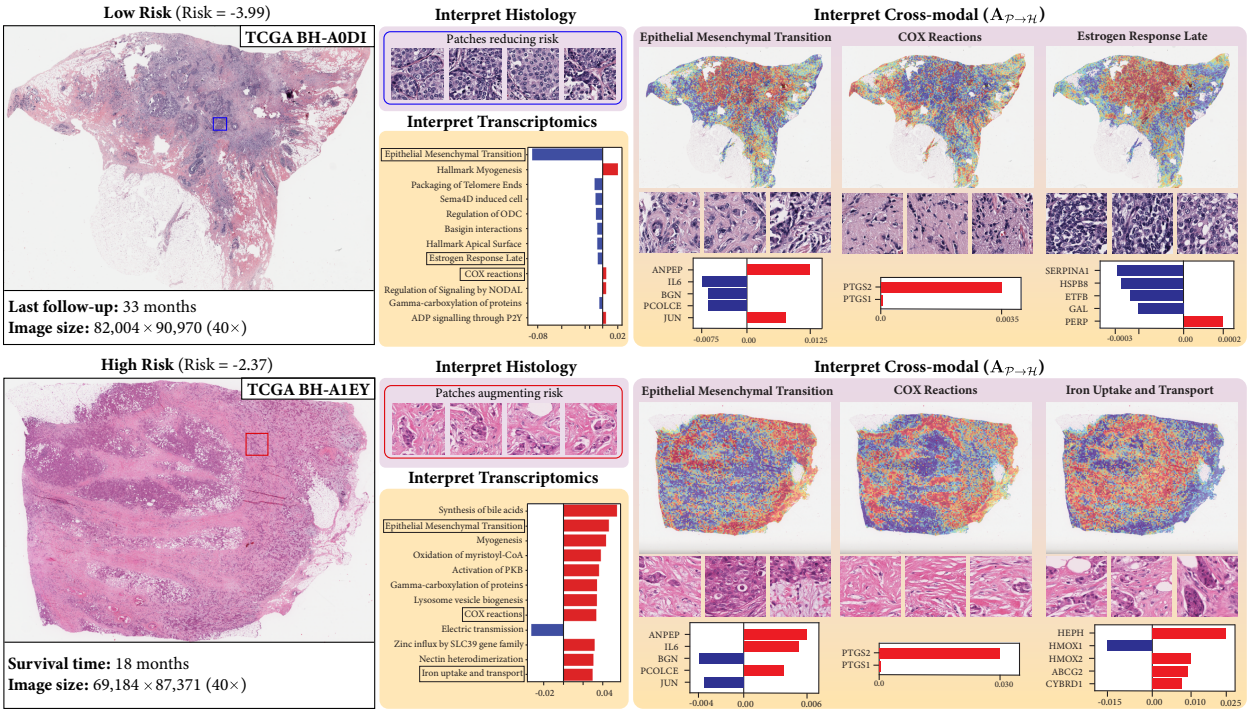

Figure 3. **Multi-level interpretability visualization in a breast cancer patient. Top:** Low-risk patient. **Bottom:** High-risk patient. Genes and pathways in red increase risk, and those in blue decrease risk. Heatmap colors indicate importance, with red indicating high importance and blue indicating low importance. The pathways and morphologies identified as important in these cases generally correspond well with patterns that have been previously described in invasive breast cancer (*e.g.* Estrogen Response Late).

tified, with patches showing small nests of tumor cells invading through a dense stroma. In the low-risk case, a pathway relating to the cellular response to estrogen was found to be important, with corresponding patches demonstrating lower-grade invasive carcinoma or carcinoma in situ morphologies, consistent with others' observation that hormone-positive breast cancers tend to be lower grade and have longer survival times [19]. Interestingly, the Hallmark Myogenesis pathway is assigned relatively high positive importance for both cases in Fig. 3. Myogenesis has not been extensively studied in breast cancer, but it is plausible that tumor cells either themselves express genes involved in this pathway as part of their epithelial-mesenchymal transition or they induce stromal cells to do so. This highlights the ability of our method to drive novel biological insight for subsequent investigation.

The flexibility of our approach in providing unimodal and cross-modal interpretability allows us to uncover novel multimodal biomarkers of prognosis that could conceivably be used to design better cancer therapies. As our understanding of the molecular underpinnings of disease grows, the interpretability of SURVPATH may spur research into the possibility of targeting specific combinations of morphologies and pathways.

5. Conclusion

This paper addresses two challenges posed by the multimodal fusion of transcriptomics and histology: (1) we address the challenge of transcriptomics tokenization by defining *biological pathway* tokens that encode semantically meaningful and interpretable functions, and (2) we overcome the computational challenge of integrating long multimodal sequences by designing a multimodal Transformer with sparse modality-specific attention patterns. Our model achieves state-of-the-art survival performance when tested on five datasets from TCGA. In addition, our interpretability framework reveals known and candidate prognostic features. While our interpretability framework enables identifying prognostic features, these findings remain qualitative. Future work could focus on interpretability metrics that generalize findings at dataset-level, *e.g.,* with quantitative morphological characterizations of specific pathways. In addition, our findings suggest that including patch-to-patch interactions does not lead to improved performance. Nonetheless, the absence of a performance boost should not be an evidence that patch-to-patch interactions are unnecessary, but rather that modeling such interactions is a challenging problem that remains to be solved.

References

[1] Estelle Aflalo, Meng Du, Shao-Yen Tseng, Yongfei Liu, Chenfei Wu, Nan Duan, and Vasudev Lal. Vl-interpret: An interactive visualization tool for interpreting vision-language transformers. In *CVPR*, pages 21406–21415, 2022. 3

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2

[3] Peter Bandi, Oscar Geessink, Quirine Manson, Marcory Van Dijk, Maschenka Balkenhol, Meyke Hermsen, Babak Ehteshami Bejnordi, Byungjae Lee, Kyunghyun Paeng, Aoxiao Zhong, et al. From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge. *IEEE transactions on medical imaging*, 38(2):550–560, 2018. 1

[4] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020. 2, 5

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

[6] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature Medicine*, 2019. 1

[7] Anika Cheerla and Olivier Gevaert. Deep learning with multimodal representation for pancancer prognosis prediction. *Bioinformatics*, 35(14):i446–i454, 2019. 3

[8] Beidi Chen, Tri Dao, Eric Winsor, Zhao Song, Atri Rudra, and Christopher Ré. Scatterbrain: Unifying sparse and low-rank attention. In *Advances in Neural Information Processing Systems*, 2021. 7

[9] Richard J Chen, Ming Y Lu, Jingwen Wang, Drew FK Williamson, Scott J Rodig, Neal I Lindeman, and Faisal Mahmood. 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, 2020. 3, 5, 6

[10] Richard J Chen, Ming Y Lu, Wei-Hung Weng, Tiffany Y Chen, Drew FK Williamson, Trevor Manz, Maha Shady, and Faisal Mahmood. Multimodal co-attention transformer for survival prediction in gigapixel whole slide images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4025, 2021. 2, 3, 6, 7

[11] Richard J. Chen, Ming Y. Lu, Drew F.K. Williamson, Tiffany Y. Chen, Jana Lipkova, Zahra Noor, Muhammad Shaban, Maha Shady, Mane Williams, Bumjin Joo, and Faisal Mahmood. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878, 2022. 2, 6

[12] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamás Sarlós, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy J. Colwell, and Adrian Weller. Rethinking attention with performers. In *ICLR*, 2021. 2

[13] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature Medicine*, 24(10):1559–1567, 2018. 1

[14] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. 2023. 2

[15] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022. 2

[16] Donglin Di, Changqing Zou, Yifan Feng, Haiyan Zhou, Rongrong Ji, Qionghai Dai, and Yue Gao. Generating hypergraph-based high-order representations of whole-slide histopathological images for survival prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–16, 2022. 2

[17] Kexin Ding, Mu Zhou, Dimitris Metaxas, and Shaoting Zhang. Pathology-and-genomics multimodal transformer for survival outcome prediction. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, pages 622–631, 2023. 3

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[19] Lisa K Dunnwald, Mary Anne Rossing, and Christopher I Li. Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. *Breast cancer research*, 9:1–10, 2007. 8

[20] Haitham A Elmarakeby, Justin Hwang, Rand Arafeh, Jett Crowdis, Sydney Gang, David Liu, Saud H AlDubayan, Keyan Salari, Steven Kregel, Camden Richter, et al. Biologically informed deep neural network for prostate cancer discovery. *Nature*, 598(7880):348–352, 2021. 4, 6, 7

[21] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. *arXiv preprint arXiv:1606.01847*, 2016. 6

[22] Marc Gillespie, Bijay Jassal, Ralf Stephan, Marija Milacic, Karen Rothfels, Andrea Senff-Ribeiro, Johannes Griss, Cristoffer Sevilla, Lisa Matthews, Chuqiao Gong, Chuan Deng, Thawfeek Varusai, Eliot Ragueneau, Yusra Haider, Bruce May, Veronica Shamovsky, Joel Weiser, Timothy Brunson, Nasim Sanati, Liam Beckman, Xiang Shao, Antonio Fabregat, Konstantinos Sidiropoulos, Julieth Murillo, Guilherme Viteri, Justin Cook, Solomon Shorser, Gary Bader, Emek Demir, Chris Sander, Robin Haw, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter

D'Eustachio. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D687–D692, 2021. 2, 5

[23] Mary J Goldman, Brian Craft, Mim Hastie, Kristupas Repečka, Fran McDade, Akhil Kamath, Ayan Banerjee, Yunhai Luo, Dave Rogers, Angela N Brooks, Jingchun Zhu, and David Haussler. Visualizing and interpreting cancer genomics data via the xena platform. *Nature biotechnology*, 38 (6):675–678, 2020. 5

[24] Balazs Gyorffy. Survival analysis across the entire transcriptome identifies biomarkers with the highest prognostic power in breast cancer. *Computational and Structural Biotechnology Journal*, 19:4101–4109, 2021. 2

[25] Jie Hao, Youngsoon Kim, Tae-Kyung Kim, and Mingon Kang. Pasnet: Pathway-associated sparse deepneural network for prognosis prediction from high-throughput data. *BMC Bioinformatics*, 19, 2018. 4, 6

[26] Randall E Harris, Bruce C Casto, and Zachary M Harris. Cyclooxygenase-2 and the inflammogenesis of breast cancer. *World journal of clinical oncology*, 5(4):677, 2014. 7

[27] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice Hall PTR, 1994. 6

[28] Frederick Howard, James Dolezal, Sara Kochanny, Jefree Schulte, Heather Chen, Lara Heij, Dezheng Huo, Rita Nanda, Olufunmilayo Olopade, Jakob Kather, Nicole Cipriani, Robert Grossman, and Alexander Pearson. The impact of site-specific digital histology signatures on deep learning model accuracy and bias. *Nature Communications*, 12, 2021. 5

[29] Weizhe Hua, Zihang Dai, Hanxiao Liu, and Quoc Le. Transformer quality in linear time. In *Proceedings of the 39th International Conference on Machine Learning*, pages 9099–9117. PMLR, 2022. 2

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

[31] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the 38th International Conference on Machine Learning*, pages 4651–4664, 2021. 2

[32] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier J Henaff, Matthew Botvinick, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver IO: A general architecture for structured inputs & outputs. In *International Conference on Learning Representations*, 2022. 2

[33] Raghu Kalluri, Robert A Weinberg, et al. The basics of epithelial-mesenchymal transition. *The Journal of clinical investigation*, 119(6):1420–1428, 2009. 7

[34] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A. Valous, Dyke Ferber, Lina Jansen, Constantino Carlos Reyes-Aldasoro, Inka Zörnig, Dirk Jäger, Hermann Brenner, Jenny Chang-Claude, Michael Hoffmeister, and Niels Halama. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16(1):1–22, 2019. 1

[35] Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, page 972–981, Red Hook, NY, USA, 2017. Curran Associates Inc. 6

[36] Yongju Lee, Jeong Hwan Park, Sohee Oh, Kyoungseob Shin, Jiyu Sun, Minsun Jung, Cheol Lee, Hyojin Kim, Jin-Haeng Chung, Kyung Chul Moon, et al. Derivation of prognostic contextual histopathological features from whole-slide images of tumours via graph deep learning. *Nature Biomedical Engineering*, 2022. 1, 2

[37] Ruiqing Li, Xingqi Wu, Ao Li, and Minghui Wang. HFB-Surv: hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. *Bioinformatics*, 38(9): 2587–2594, 2022. 2, 3

[38] Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430*, 2022. 2

[39] Paul Pu Liang, Yiwei Lyu, Gunjan Chhablani, Nihal Jain, Zihao Deng, et al. Multiviz: Towards visualizing and understanding multimodal models. In *ICLR*, 2023. 3

[40] Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, Mahmoud Ghandi, Jill P. Mesirov, and Pablo Tamayo. The molecular signatures database (msigdb) hallmark gene set collection. *Cell systems*, 1 6:417–425, 2015. 2, 5

[41] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. Interbert: Vision-and-language interaction for multi-modal pretraining. *ArXiv*, abs/2003.13198, 2020. 2

[42] Huidong Liu and Tahsin Kurc. Deep learning for survival analysis in breast cancer with whole slide image data. *Bioinformatics*, 38(14):3629–3637, 2022. 2

[43] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2019. Curran Associates Inc. 2

[44] Jiachen Lu, Jinghan Yao, Junge Zhang, Xiatian Zhu, Hang Xu, Weiguo Gao, Chunjing Xu, Tao Xiang, and Li Zhang. Soft: Softmax-free transformer with linear complexity. In *NeurIPS*, 2021. 2

[45] Jianzhu Ma, Michael Ku Yu, Samson H. Fong, Keiichiro Ono, Eric Sage, Barry Demchak, Roded Sharan, and Trey Ideker. Using deep learning to model the hierarchical structure and function of a cell. *Nature methods*, 15:290 – 298, 2018. 4

[46] Callum Christopher Mackenzie, Muhammad Dawood, Simon Graham, Mark Eastwood, and Fayyaz ul Amir Afsar Minhas. Neural graph modelling of whole slide images for survival ranking. In *The First Learning on Graphs Conference*, 2022. 2

[47] D Mazhar, Richard Ang, and J Waxman. Cox inhibitors and breast cancer. *British journal of cancer*, 94:346–50, 2006. 7

[48] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A Gutman, Jill S Barnholtz-Sloan, José E Velázquez Vega, Daniel J Brat, and Lee AD Cooper. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018. 3, 6

[49] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *Computer Vision – ECCV 2020*, pages 336–352, Cham, 2020. Springer International Publishing. 2

[50] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems*, pages 14200–14213. Curran Associates, Inc., 2021. 2

[51] Ádám Nagy, Gyöngyi Munkácsy, and Balázs Győrffy. Pancancer survival analysis of cancer hallmark genes. *Scientific Reports*, 11, 2020. 2

[52] Kien Nguyen and Takayuki Okatani. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6087–6096, 2018. 2

[53] Jiezhong Qiu, Hao Ma, Omer Levy, Wen-tau Yih, Sinong Wang, and Jie Tang. Blockwise self-attention for long document understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2555–2565, Online, 2020. Association for Computational Linguistics. 2

[54] Lin Qiu, Aminollah Khormali, and Kai Liu. Deep biological pathway informed pathology-genomic multimodal survival prediction, 2023. 3

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[56] Srivatsan Raghavan, Peter S. Winter, Andrew W. Navia, Hannah L. Williams, Alan DenAdel, Kristen E. Lowder, Jennyfer Galvez-Reyes, Radha L. Kalekar, Nolawit Mulugeta, Kevin S. Kapner, Manisha S. Raghavan, Ashir A. Borah, Nuo Liu, Sara A. Väyrynen, Andressa Dias Costa, Raymond W.S. Ng, Junning Wang, Emma K. Hill, Dorisanne Y. Ragon, Lauren K. Brais, Alex M. Jaeger, Liam F. Spurr, Yvonne Y. Li, Andrew D. Cherniack, Matthew A. Booker, Elizabeth F. Cohen, Michael Y. Tolstorukov, Isaac Wakiro, Asaf Rotem, Bruce E. Johnson, James M. McFarland, Ewa T. Sicinska, Tyler E. Jacks, Ryan J. Sullivan, Geoffrey I. Shapiro, Thomas E. Clancy, Kimberly Perez, Douglas A. Rubinson, Kimmie Ng, James M. Cleary, Lorin Crawford, Scott R. Manalis, Jonathan A. Nowak, Brian M. Wolpin, William C. Hahn, Andrew J. Aguirre, and Alex K. Shalek. Microenvironment drives cell state, plasticity, and drug response in pancreatic cancer. *Cell*, 184(25):6119–6137.e26, 2021. 2

[57] Yoni Schirris, Efstratios Gavves, Iris Nederlof, Hugo Mark Horlings, and Jonas Teuwen. Deepsmile: Contrastive self-supervised pre-training benefits msi and hrd classification directly from h&e whole-slide images in colorectal and breast cancer. *Medical Image Analysis*, 79:102464, 2022. 2

[58] Fahad Shamshad, Salman Hameed Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and H. Fu. Transformers in medical imaging: A survey. *ArXiv*, abs/2201.09873, 2022. 2

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

[60] Zhuchen Shao, Yang Chen, Hao Bian, Jian Zhang, Guojun Liu, and Yongbing Zhang. Hvtsurv: Hierarchical vision transformer for patient-level survival prediction from whole slide image. In *AAAI*, 2023. 2

[61] Artem Shmatko, Narmin Ghaffari Laleh, Moritz Gerstung, and Jakob Nikolas Kather. Artificial intelligence in histopathology: enhancing cancer research and clinical oncology. *Nature Cancer*, 3(9):1026–1038, 2022. 1, 3

[62] Andrew H. Song, Guillaume Jaume, Drew F. K. Williamson, Ming Y. Lu, Anurag Vaidya, Tiffany R. Miller, and Faisal Mahmood. Artificial intelligence for digital and computational pathology. *Nature Reviews Bioengineering*, 2023. 1

[63] Andrew H Song, Richard J Chen, Tong Ding, Drew FK Williamson, Guillaume Jaume, and Faisal Mahmood. Morphological prototyping for unsupervised slide representation learning in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[64] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and Jill P. Mesirov. 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. 2, 5

[65] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, page 3319–3328. JMLR.org, 2017. 5

[66] Suzy V Torti and Frank M Torti. Cellular iron metabolism in prognosis and therapy of breast cancer. *Critical Reviews™ in Oncogenesis*, 18(5), 2013. 7

[67] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, 2019. Association for Computational Linguistics. 2

[68] Luís A. Vale-Silva and Karl Rohr. Multisurv: Long-term cancer survival prediction using multimodal deep learning. *medRxiv*, 2020. 3

[69] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *Neural Information Processing Systems (NeurIPS)*, 2017. 2, 4

[70] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022. 2

[71] Xiaodong Wang, Ying Chen, Yunshu Gao, Huiqing Zhang, Zehui Guan, Zhou Dong, Yuxuan Zheng, Jiarui Jiang, Haoqing Yang, Liming Wang, Xianming Huang, Lirong Ai, Wenlong Yu, Hongwei Li, Changsheng Dong, Zhou Zhou, Xiyang Liu, and Guanzhen Yu. Predicting gastric cancer outcome from resected lymph node histopathology images using deep learning. *Nature Communications*, 12, 2021. 1

[72] Xingbo Wang, Jianben He, Zhihua Jin, et al. M2lens: Visualizing and explaining multimodal models for sentiment analysis. *IEEE Transactions on Visualization and Computer Graphics*, 2021. 3

[73] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Junzhou Huang, Wei Yang, and Xiao Han. Transpath: Transformer-based self-supervised learning for histopathological image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 186–195. Springer, 2021. 4, 6

[74] Xiyue Wang, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 2022. 4

[75] Yifan Wang and Binhua P Zhou. Epithelial-mesenchymal transition—a hallmark of breast cancer metastasis. *Cancer hallmarks*, 1(1):38–49, 2013. 2

[76] Yifan Wang and Binhua Peter Zhou. Epithelial-mesenchymal transition—a hallmark of breast cancer metastasis. *Cancer hallmarks*, 1 1:38–49, 2013. 7

[77] Zhiqin Wang, Ruiqing Li, Minghui Wang, and Ao Li. GPDBN: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction. *Bioinformatics*, 37(18):2963–2970, 2021. 2, 3

[78] Wei-Hung Weng, Yuannan Cai, Angela Lin, Fraser Tan, and Po-Hsuan Cameron Chen. Multimodal multitask representation learning for pathology biobank metadata prediction. *arXiv preprint arXiv:1909.07846*, 2019. 6

[79] Suzanne Wetstein, Vincent Jong, Nikolas Stathonikos, Mark Opdam, Gwen Dackus, Josien Pluim, Paul Diest, and Mitko Veta. Deep learning-based breast cancer grading and survival analysis on whole-slide histopathology images. *Scientific Reports*, 12:15102, 2022. 2

[80] Haixu Wu, Jialong Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Flowformer: Linearizing transformers with conservation flows. In *International Conference on Machine Learning*, 2022. 2

[81] Ellery Wulczyn, David Steiner, Zhaoyang Xu, Apaar Sadhwani, Hongwu Wang, Isabelle Flament, Craig Mermel, Po-Hsuan Chen, Yun Liu, and Martin Stumpe. Deep learning-based survival prediction for multiple cancer types using histopathology images. *Plos ONE*, 2019. 1

[82] Y. Xiong, Z. Zeng, R. Chakraborty, M. Tan, G. Fung, Y. Li, and V. Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021. 2, 6, 7

[83] Peng Xu, Xiatian Zhu, and David Clifton. Multimodal learning with transformers: A survey, 2022. 2

[84] Yingxue Xu and Hao Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3, 6, 7

[85] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65: 101789, 2020. 1, 2, 6

[86] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250*, 2017. 6

[87] Shekoufeh Gorgi Zadeh and Matthias Schmid. Bias in cross-entropy-based training of deep survival networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9):3126–3137, 2021. 5

[88] Fengtao Zhou and Hao Chen. Cross-modal translation and alignment for survival analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21485–21494, 2023. 3

[89] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2