POMP: Pathology-omics Multimodal Pre-training Framework for Cancer Survival Prediction

Suixue Wang[1] , **Shilin Zhang**[2] , **Huiyuan Lai**[3] , **Weiliang Huo**[1] and **Qingchen Zhang**[4,*]
[1] School of Information and Communication Engineering, Hainan University
[2] College of Intelligence and Computing, Tianjin University
[3] University of Groningen
[4] School of Computer Science and Technology, Hainan University
{wangsuixue, wlhuo, zhangqingchen}@hainanu.edu.cn, zhang_shilin_sd@163.com, h.lai@rug.nl

Abstract

Cancer survival prediction is an important direction in precision medicine, aiming to help clinicians tailor treatment regimens for patients. With the rapid development of high-throughput sequencing and computational pathology technologies, survival prediction has shifted from clinical features to joint modeling of multi-omics data and pathology images. However, existing multimodal learning methods struggle to effectively learn pathology-omics interactions due to the lack of proper alignment of multimodal data before fusion. In this paper, we propose POMP, a pathology-omics multimodal pre-training framework jointly learned with three training tasks for integrating pathological images and omics data for cancer survival prediction. To better perform cross-modal learning, we introduce a pathology-omics contrastive learning method to align the pathology and omics information. POMP leverages the principle of pre-trained models and explores the benefit of aligning multimodal information from the same patient, achieving state-of-the-art results on six cancer datasets from the Cancer Genome Atlas (TCGA). We also show that our contrastive learning method allows us to exploit the cosine similarity of pathological images and omics data as the survival risk score, which can further boost prediction performance compared with other commonly used methods. The code is available at https://github.com/SuixueWang/POMP.

1 Introduction

Cancer is widely recognized as a major global public health concern. For instance, there were 19.3 million newly diagnosed cases of cancer in 2020, and approximately 10.0 million deaths attributed to cancer [Sung *et al.*, 2021]. Cancer survival prediction is an important direction in precision medicine, which aims to predict the time from cancer diagnosis to death events in cancer patients, thereby helping clinicians to tailor optimal treatment regimens [Vale-Silva and Rohr, 2021]. Traditional survival prediction methods mainly relied on clinical features like symptoms, signs, tumor biomarkers, and biochemical tests, as well as follow-up notes recording recurrence, metastasis, and response to treatment [Gensheimer *et al.*, 2019; Chicco and Jurman, 2020]. Unfortunately, these methods are laborious and unfeasible when applied in clinical practice [Wang *et al.*, 2019]. With the swift advancements in whole slide imaging and high-throughput sequencing technologies, survival prediction has transitioned from clinical notes to multi-omics data and pathology images, which are also known as whole slide images (WSIs) [Herrmann *et al.*, 2021; Srinidhi *et al.*, 2021]. Computational pathology methods can extract gold-standard information on survival prediction from pathology images regarding tumor cells and their microenvironment [Herrmann *et al.*, 2021], and integrate multi-omics data such as mRNA, miRNA, and DNA methylation to enhance prediction from a molecular perspective [Srinidhi *et al.*, 2021].

In recent years, multimodal learning has made promising progress in survival prediction by capturing complementary information from various modal perspectives [Cui *et al.*, 2023]. However, most works simply concatenate representations extracted from multimodal data, and employ cross-modal attention mechanism [Chen *et al.*, 2021; Wang *et al.*, 2023; Li *et al.*, 2022] or tensor fusion method [Chen *et al.*, 2020; Wang *et al.*, 2021a] to learn potential interaction information. Due to the lack of multimodal alignment, these methods prevent the models from fully utilizing the information between different modalities. Specifically, the survival prediction model is trained to learn pathology representation and omics representation from their own spaces, which has two main limitations: (i) it poses a challenge for the multimodal fusion module to fuse different modalities from different semantic spaces; (ii) this would lose the pathology-histology alignment information that could potentially be used as a signal for survival prediction.

In this work, we break with traditional approaches to study how to align multimodal information from the same patient before fusion while serving as survival predictions. To do so, we propose a pathology-omics multimodal pre-training (POMP) framework for cancer survival prediction, leveraging the principle of pre-trained models and exploring the benefit

*Corresponding author: Qingchen Zhang.

of aligning multimodal information from the same patient. In practice, we first encode the pathology image and multi-omics data independently with a pathology encoder and an omics encoder, and then use a multimodal encoder to fuse representations of image and omics data. In particular, we introduce a pathology-omics contrastive learning method based on cosine similarity to align multimodal information, making the two modal data from the same patient closer in latent space. To fully learn the potential interaction information between different modalities, POMP is jointly pre-trained with three tasks based on the self-supervised learning paradigm. During fine-tuning, we leverage the cosine similarity, calculated in the same way as the pathology-omics contrastive learning, as the survival risk score for cancer survival prediction.

We conduct extensive experiments on six cancer datasets from the Cancer Genome Atlas (TCGA) [Tomczak *et al.*, 2015]. The experimental results show that POMP consistently outperforms existing state-of-the-art methods across six datasets.

Our primary contributions are summarized as follows:

1. We propose a novel multimodal pre-training framework jointly learned with three training tasks to integrate pathological images and multi-omics data for cancer survival prediction.
2. We introduce a pathology-omics contrastive learning method to align the multi-modalities before fusion. This enables POMP to pull pathology images and omics data from the same patients closer together in the latent space, thereby better performing cross-modal learning.
3. POMP achieves the best results on six cancer datasets; Extensive experiments show that leveraging the cosine similarity, calculated in the same way as the pathology-omics contrast learning, as the survival risk score achieves better performance compared to other commonly used survival risk computational methods.

2 Related Work

2.1 Multimodal Pre-training

Multimodal pre-training aims to improve the performance of downstream tasks by jointly pre-training the model with multiple pre-training objectives, garnering considerable attention in AI research [Xu *et al.*, 2023a; Pei *et al.*, 2024]. A substantial amount of work has been presented across various domains, for example, vision-language pre-training (VLP) [Radford *et al.*, 2021; Kim *et al.*, 2021; Li *et al.*, 2021; Yin *et al.*, 2024], vision-audio pre-training (VAP) [Deshmukh *et al.*, 2023; Xu *et al.*, 2023b], audio-language pre-training (ALP) [Elizalde *et al.*, 2023; Wu *et al.*, 2023b], audio-visual-text Pre-training (AVTP) [Guzhov *et al.*, 2022; Wu *et al.*, 2022; Xu *et al.*, 2024]. Among them, VLP is a major research problem in this field. The pre-training tasks commonly used in VLP include masked language modeling (MLM) [Li *et al.*, 2021], masked image modeling (MIM) [Kim *et al.*, 2021], image-text matching (ITM) [Li *et al.*, 2021], image-text contrastive learning (ITC) [Bao *et al.*, 2022; Cheng *et al.*, 2021; Jin *et al.*, 2023], and word patch alignment (WPA) [Kim *et al.*, 2021]. While these multimodal pre-training methods have achieved great success in natural images (with low pixels and uniform size) and texts, it is not feasible to directly copy them to the scenario of cancer survival prediction since the pathological image has gigapixels, non-uniform sizes, and large surfaces of worthless background regions.

2.2 Survival Prediction using Multi-modality

Recently, survival prediction using multi-modality has achieved notable advancement, attributed to the success of deep learning-based multimodal fusion technologies [Cui *et al.*, 2023; Wang *et al.*, 2024]. For instance, DPDBN [Wang *et al.*, 2021b] and CAMR [Wu *et al.*, 2023a] concatenate representations learned from multi-modality, where the input features of pathological images, such as sizes, shapes, intensity distributions, textures, and brightness levels, are extracted through the hand-crafted CellProfiler tool [McQuin *et al.*, 2018]. DeepCorrSurv [Yao *et al.*, 2017] maximizes the correlation to learn the shared representation from two modalities, while PathOmics [Ding *et al.*, 2023] minimizes the mean square error to strengthen the interaction between different modalities. Pathomic Fusion [Chen *et al.*, 2020] introduces a tensor fusion method, Kronecker product, to fuse multi-modality. MCAT, CMTA, and HC-MAE [Chen *et al.*, 2021; Zhou and Chen, 2023; Wang *et al.*, 2023] employ a cross-modal attention mechanism to integrate multi-modality. However, most methods lack proper alignment of multimodal data before fusion, which makes it difficult to effectively learn pathology-omics interactions. We aim to overcome this by leveraging self-supervised learning and contrastive learning, where our framework POMP is jointly pre-trained with three tasks to align multi-modal information.

3 Method

We propose POMP, a pathology-omics multimodal pre-training framework for cancer survival prediction, as illustrated in Figure 1. Given a patient sample $\boldsymbol{x}_i$ from the data set $\mathbf{x} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$, containing a gigapixel pathological image $\boldsymbol{P}_i$, multi-omics data $\boldsymbol{O}_i$, survival status $\boldsymbol{e}_i$, and survival time $\boldsymbol{t}_i$, we aim to train POMP to predict survival risk scores. This section first introduces data preprocessing, and then details POMP's key components, followed by its pre-training and fine-tuning.

3.1 Data Preprocessing

Pathological Images. Building upon the previous works [Zhou and Chen, 2023; Wang *et al.*, 2023], we start by employing CLAM [Lu *et al.*, 2021] to automatically identify the high-value region and eliminate the background area within the pathological image. In computational pathology, gigapixel WSI can reach dimensions of 150,000 $\times$ 150,000 pixels at $20\times$ magnification. A straightforward way is to crop WSIs into 256×256 pixel patches, but this will generate a substantial number of segments, greatly increasing the computational load for self-attention in Pathology-Transformer. We divide the high-value region into M non-overlapping sub-regions with 4096×4096 pixels, which are then downsampled by a factor of 16 to the image patches $\{\boldsymbol{i}_k\}_{k=1}^{M}$ with 256×256

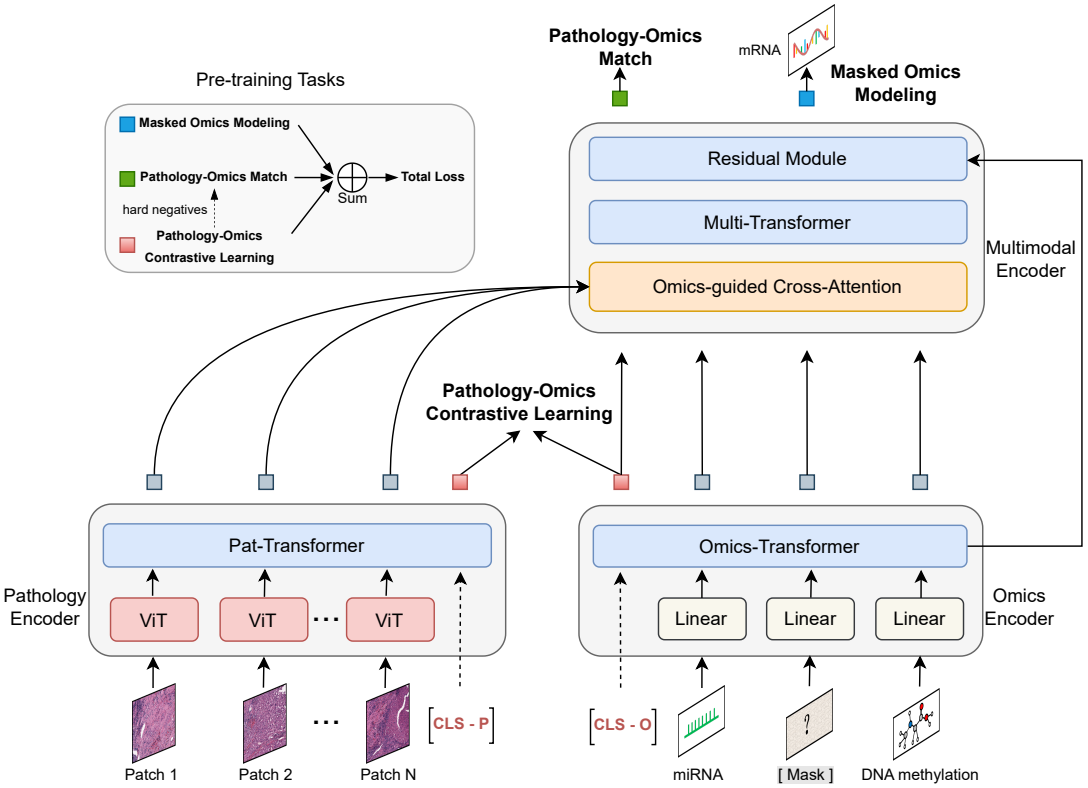


Figure 1: Overview of our framework POMP. It consists of a pathology encoder, an omics encoder, and a multimodal encoder. We pre-train POMP with three tasks: pathology-omics contrastive learning, pathology-omics match, and masked omics modeling; during fine-tuning, the cosine similarity of pathology representation and omics representation is taken as a survival risk score.

pixels. Although this might lead to the loss of fine-grained information, Pathology-Transformer can efficiently capture inter-patch relationships.

Multi-omics data. For multi-omics data including RNA-Seq, miRNA, and DNA methylation, we first employ the K-nearest neighbor interpolation method [Troyanskaya *et al.*, 2001] to fill in missing values. Then, we calculate the variance of each gene across all patient samples and remove the genes with zero variance. To perform differential gene expression analysis, we use the pydeseq2 package [Muzellec *et al.*, 2023] to select the genes that exhibit significant changes. Finally, for each omics type in each cancer dataset, we use the random survival forest (RSF) [Pölsterl, 2020] to calculate the feature importance of each gene across all patient samples and retain the top 300 genes for survival analysis.

3.2 POMP Architecture

Given a patient sample containing pathological images and multi-omics data, we encode them separately into vector representations using two encoders, which are then fed into a multimodal encoder to fuse them.

Pathology Encoder. Pathology encoder consists of M Vision Transformer (ViT) [Dosovitskiy *et al.*, 2020] and a Pat-Transformer. We start by embedding the image patches $\{\boldsymbol{i}_k\}_{k=1}^{M}$ with 256×256 pixels by M ViTs with shared weights, which can be written as:

$$\boldsymbol{I}_{pat} = \text{AvgPool}\left(\text{ViT}\left(\{\boldsymbol{i}_k\}_{k=1}^{M}\right)\right) \tag{1}$$

where AvgPool is average pooling used to generate the image patch embeddings $\boldsymbol{I}_{pat} \in \mathbb{R}^{M \times d}$. The lower level patch size in ViT is 16×16.

Afterward, we randomly initialize a [CLS-P] token $\boldsymbol{I}_{cls} \in \mathbb{R}^{1 \times d}$, followed by concatenating $\boldsymbol{I}_{pat}$ and $\boldsymbol{I}_{cls}$, and then feed them to the Pathology-Transformer for fully learning the correlation between image patches within WSI, thereby obtaining the final representation of WSI:

$$\boldsymbol{I}'_{final} = \text{Pat-Transformer}\left([\boldsymbol{I}_{pat} \oplus \boldsymbol{I}_{cls}] + \boldsymbol{E}_{pos}\right) \tag{2}$$

where $\boldsymbol{E}_{pos}$ denotes the relative position embedding of image patches. $\boldsymbol{I}'_{final}$ consists of a sequence of vectors: $\{\boldsymbol{h}_1^v, \ldots, \boldsymbol{h}_M^v, \boldsymbol{v}\}$, where $\boldsymbol{v}$ is the final representation of the [CLS-P] token. We define $\boldsymbol{I}_{final} = \{\boldsymbol{h}_1^v, \ldots, \boldsymbol{h}_M^v\}$, which corresponds exclusively to image patch tokens.

Omics Encoder. Let $\boldsymbol{O}_{pre} \in \mathbb{R}^{3 \times 300}$ denotes preprocessed multi-omics data, including mRNA, miRNA, and DNA methylation. We first feed the omics data to a shared linear network to learn the intrinsic information of the omics. Subsequently, the outputs of the shared linear network, together with a randomly initialized vector of [CLS-O] token, are sent into a 2-layer Transformer to learn the interactive information between multi-omics data, thereby generating the

final representations of multi-omics data:

$$\boldsymbol{O}_{final} = \text{Omics-Transformer}\left([\boldsymbol{O}_{cls} \oplus \text{Linear}\left(\boldsymbol{O}_{pre}\right)]\right) \quad (3)$$

Where $\boldsymbol{O}_{final}$ consists of four vectors: $\{\boldsymbol{w}, \boldsymbol{h}_1^w, \boldsymbol{h}_2^w, \boldsymbol{h}_3^w\}$, which are the final representations of [CLS-O] token, mRNA, miRNA, and DNA methylation.

Multimodal Encoder. We propose a multimodal encoder, which is implemented by an omics-guided cross-attention module, a 2-layer Transformer, and a residual module, to fuse the representations of two modalities, thereby learning the final multimodal representations:

$$\begin{aligned}\boldsymbol{H}_{cross} &= \text{CrossAttn}\left(\mathbf{W}_q\boldsymbol{O}_{final}, \mathbf{W}_k\boldsymbol{I}_{final}, \mathbf{W}_v\boldsymbol{I}_{final}\right) \\ &= \text{Softmax}\left(\frac{\mathbf{W}_q\boldsymbol{O}_{final}\boldsymbol{I}_{final}^T\mathbf{W}_k^T}{\sqrt{d}}\right)\mathbf{W}_v\boldsymbol{I}_{final}\end{aligned} \quad (4)$$

$$\boldsymbol{H} = \text{LN}\left(\text{Multi-Transformer}\left(\boldsymbol{H}_{cross}\right) + \boldsymbol{O}_{final}\right) \quad (5)$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d\times d}$ denote learnable weights multiplied by the queries $\boldsymbol{O}_{final}$, keys $\boldsymbol{I}_{final}$, and values $\boldsymbol{I}_{final}$, respectively. LN represents layer normalization. $\boldsymbol{H}$ are the final multimodal representations. Notably, there are two special representations in $\boldsymbol{H}$ corresponding to the [CLS-O] token and masked omics token, which are denoted as $\boldsymbol{h}_{cls}$ and $\boldsymbol{h}_{mask}$.

3.3 Pre-training Tasks

We pre-train POMP using three tasks: pathology-omics contrastive learning (POC), pathology-omics match (POM), and masked omics modeling (MOM).

Pathology-Omics Contrastive Learning. Given a training batch of N pathology-omics pairs, there are N positive and $N^2 - N$ negative pathology-omics pairs. Here, we introduce a pathology-omics contrastive learning method that aims to learn effective representation by pulling pathology-omics pairs from the same patients closer and pushing negative pairs farther away. Specifically, we use the representation of [CLS-P], $\boldsymbol{v}$, and the representation of [CLS-O], $\boldsymbol{w}$, as the aggregated representation of the pathology image and multi-omics data, respectively. Then, the pathology-to-omics and omics-to-pathology similarities are computed between $\boldsymbol{v}$ and $\boldsymbol{w}$ within a training batch. Formally, the softmax-normalized pathology-to-omics and omics-to-pathology similarities can be written as:

$$\boldsymbol{s}_i^{p2o} = \frac{\exp\left(\boldsymbol{v}_i^\top \boldsymbol{w}_i/\sigma\right)}{\sum_{j=1}^N \exp\left(\boldsymbol{v}_i^\top \boldsymbol{w}_j/\sigma\right)} \quad (6)$$

$$\boldsymbol{s}_i^{o2p} = \frac{\exp\left(\boldsymbol{w}_i^\top \boldsymbol{v}_i/\sigma\right)}{\sum_{j=1}^N \exp\left(\boldsymbol{w}_i^\top \boldsymbol{v}_j/\sigma\right)} \quad (7)$$

where σ is the temperature parameter to scale the similarities. Let $\boldsymbol{y}^{p2o}$ and $\boldsymbol{y}^{o2p}$ represent the ground-truth one-hot similarity, where the positive pair has a probability of 1 and the negative pair has a probability of 0. The pathology-omics contrastive loss is calculated using cross-entropy (CE) between $\boldsymbol{s}$ and $\boldsymbol{y}$:

$$\mathcal{L}_{\text{poc}} = \frac{1}{2}\mathbb{E}_{(p,o)\sim D}\left[\text{CE}\left(\boldsymbol{y}^{p2o}, \boldsymbol{s}^{p2o}\right) + \text{CE}\left(\boldsymbol{y}^{o2p}, \boldsymbol{s}^{o2p}\right)\right] \quad (8)$$

Pathology-Omics Matching. Inspired by ALBEF [Li *et al.*, 2021], we sample two hard negative pathology-omics pairs based on the max softmax-normalized pathology-to-omics and omics-to-pathology similarities for each pathology-omics pair in the training batch. As a result, we obtain N positive pairs and $2N$ hard negative samples. Pathology-omics matching aims to predict whether a pair of pathology and omics is positive or negative. Specifically, we use the final multimodal representation of the [CLS-O] token, $\boldsymbol{h}_{cls}$, as the integrated representation of the pathology-omics pair, and feed it into a linear classifier g with cross-entropy loss for binary classification. Let $\boldsymbol{y}^{pom}$ denote the ground-truth label, the POM loss is written as:

$$\mathcal{L}_{pom} = \mathbb{E}_{(p,o)\sim D}\text{CE}\left(g\left(\boldsymbol{h}_{cls}\right), \boldsymbol{y}^{pom}\right) \quad (9)$$

Masked Omics Modeling. We randomly mask one of three omics data and replace it with a zero vector. Masked omics modeling aims to reconstruct omics from a corrupted version based on other unmasked omics and pathology cues. Specifically, we feed the final representation of masked omics, $\boldsymbol{h}_{mask}$, into a classifier f over the dimension of omics data for predicting the probability of masked omics token $f\left(\boldsymbol{h}_{mask}\right)$. Let $\boldsymbol{y}^{mask}$ denote the ground-truth masked omics data, the MOM loss is written as:

$$\mathcal{L}_{mom} = \mathbb{E}_{(p,o)\sim D}\text{CE}\left(f\left(\boldsymbol{h}_{mask}\right), \boldsymbol{y}^{mask}\right) \quad (10)$$

3.4 Fine-tuning POMP on Survival Prediction

After pre-training POMP, we fine-tune it for the downstream survival prediction. Traditionally, the Cox proportional hazards model is expressed as:

$$h(t|X) = h_0(t)\exp(x_1\beta_1 + \cdots + x_p\beta_p) \quad (11)$$

where t represents the survival time, $h_0(t)$ is the baseline hazard, $h(t|X)$ is determined by a linear combination of covariates $(x_1, ..., x_p)$ and their corresponding coefficients $(\beta_1, ..., \beta_p)$.

In this work, we employ a neural network to fit the Cox proportional hazards function, where we consider the final pathology representation $\boldsymbol{v}$ as covariates and the final omics representation $\boldsymbol{w}$ as coefficients. We also standardize the linear relationship between covariates and coefficients, making the model less sensitive to the scale of the variables. Formally the Cox proportional hazards function can be written as:

$$h(t|\boldsymbol{v}, \boldsymbol{w}) = h_0(t)\exp(\frac{v_1w_1 + \cdots + v_pw_p}{\|\boldsymbol{v}\| \cdot \|\boldsymbol{w}\|}) \quad (12)$$

$$= h_0(t)\exp(c\left(\boldsymbol{v}, \boldsymbol{w}\right)) \quad (13)$$

Parameter estimates in the Cox proportional hazards model are often obtained by maximizing the partial likelihood:

$$l_{surv} = \prod_{i:e_i=1} \frac{\exp(c\left(\boldsymbol{v}_i, \boldsymbol{w}_i\right))}{\sum_{j:t_j>t_i}\exp(c\left(\boldsymbol{v}_j, \boldsymbol{w}_j\right))} \quad (14)$$

Finally, the survival loss function is set to be the negative log partial likelihood:

$$\mathcal{L} = -\frac{1}{n_e}\sum_{i:e_i=1}\left(c\left(\boldsymbol{v}_i, \boldsymbol{w}_i\right) - \log\sum_{j:t_j>t_i}\exp(c\left(\boldsymbol{v}_j, \boldsymbol{w}_j\right))\right) \quad (15)$$

Unlike vanilla Cox models that only deal with linear conditions in the hazard function, our model can better fit multimodal data and learn complex interactions. Particularly, since $h(T|\boldsymbol{v}, \boldsymbol{w})$ is an increasing function of $c(\boldsymbol{v}, \boldsymbol{w})$, we can take the cosine similarity $c(\boldsymbol{v}, \boldsymbol{w})$ as the survival risk score when leveraging the concordance index as the evaluation metric.

4 Experiments and Results Analysis

We implement our framework POMP using PyTorch and train it on 3 NVIDIA A100 GPUs. The Pathology-Transformer, Omics-Transformer, Multimodal-Transformer, and ViT in POMP all adopt the vanilla Transformer, each module has a 2-layer Transformer block with 384 hidden dimensions and 6 attention heads. POMP is trained for 500 epochs during pre-training and 80 epochs during fine-tuning. For both training phases, we use Adam optimization with a weight decay of 1e-2 and a learning rate of 5e-4. Since the pathological images have different sizes and are cropped into various sub-region numbers, we use a batch size of 1 with 50 forward accumulation steps (i.e., the actual batch size is equivalent to 50), then calculate the loss function and update the weights once.

4.1 Datasets and Evaluation Metrics

Datasets. We perform experiments using six cancer datasets obtained from the Cancer Genome Atlas (TCGA) data portal. These datasets cover various cancer types, namely colon adenocarcinoma (COAD), hepatocellular carcinoma (LIHC), stomach adenocarcinoma (STAD), breast invasive carcinoma (BRCA), lung adenocarcinoma (LUAD), and lower grade glioma (LGG). Each patient sample consists of a complete set of data types, including pathological images, as well as multi-omics data composed of RNA-Seq, miRNA, and DNA methylation. To better train the multimodal model, the patient samples missing any data type are removed. Additionally, we exclude patients with a survival time of fewer than 30 days or lacking follow-up records. As a result, the sample counts for COAD, LIHC, STAD, BRCA, LUAD, and LGG are 250, 323, 298, 724, 374, and 451, respectively. We assess all investigated methods using the same 5-fold cross-validation splits on each cancer dataset.

Evaluation Metrics. Following previous works [Yao *et al.*, 2017; Wu *et al.*, 2023a; Chen *et al.*, 2021], we leverage the concordance index (C-index) as the evaluation metric of survival prediction. The C-index is commonly used to calculate the concordance pairs between actual survival times and predicted survival risk scores, which is computed as follows:

$$\mathbf{c} = \frac{1}{n} \sum_{i \in \{1 \ldots N | \boldsymbol{e}_i = 1\}} \sum_{t_j > t_i} \mathrm{I}\left[c(\boldsymbol{v}_i, \boldsymbol{w}_i) > c(\boldsymbol{v}_j, \boldsymbol{w}_j)\right] \quad (16)$$

where $\mathrm{I}[\cdot]$ stands for the indicator function, n signifies the number of comparable pairs in which the survival status e of patients is dead, and t denotes the actual survival time. A higher C-index suggests superior predictive performance, whereas a C-index of 0.5 denotes that the model's predictions are similar to random chance outcomes.

4.2 Comparison With State-of-the-art Methods

We compare POMP with several state-of-the-art integration methods, as shown in Table 1. The first observation is that our proposed framework POMP consistently outperforms all existing methods across all cancer datasets, both with and without multi-modal pre-training. Specifically, POMP achieves C-index values of 71.5% on COAD, 75.5% on LIHC, 65.1% on STAD, 70.0% on BRCA, 69.5% on LUAD, and 85.6% on LGG, outperforming the previous best-performing methods by 1.2%, 1.8%, 2.0%, 0.5%, 0.4%, and 0.5%, respectively.

4.3 Survival Analysis

To further evaluate the effectiveness of POMP for survival prediction, we plot the Kaplan-Meier curves and calculate the log-rank test p-values across all six cancer datasets, which are shown in Figure 2. We utilize the median of the predictive survival risk scores as a risk indicator to stratify patients into high-risk (blue) and low-risk (orange) groups, and then perform Kaplan-Meier analysis to visualize the survival probability over time for different groups. Additionally, the log-rank test p-value is used to measure the statistical significance and compare the survival distributions between the high-risk group (blue) and the low-risk group (orange). In Figure 2, it can be observed that the p-values for all datasets are below the commonly used significance level threshold of 0.05. This suggests a statistically significant difference between the survival curves of all compared groups.

4.4 Ablation Study

To examine the contribution of each component of POMP, we conduct a set of ablation studies, including input modality, survival risk computing method, and pre-training tasks.

Unimodal and Multimodal. After pre-training, we add a linear layer to the pathology encoder and omics encoder, respectively, computing survival risk under single-modality. The survival prediction results of unimodal and multimodal are shown in Figure 3a. Multimodal modeling surpasses any single modality modeling in all six cancer datasets, demonstrating the effectiveness of our proposed framework in fusing multi-modality. Interestingly, we observe that the performance of the omics modality is better than the pathology modality in five of six datasets, indicating that omics can provide more information than pathology in our framework.

Survival Risk. To investigate the impacts of various survival risk computing methods in our framework, we conduct experiments on three kinds of survival risk computing methods, as illustrated in Figure 4. In this work we propose to compute the cosine similarity between two modalities as a risk score. Here we compare our method with two other methods, including (i) computing the fusion risk by feeding the final multimodal representation into a Cox layer implemented by a fully connected network; and (ii) computing the joint risk by averaging the scores of two modalities. Figure 3b shows the experimental results of the three survival risk computing methods. The cosine similarity method achieves superior survival prediction compared to other methods, indicating that the survival risk calculation method consistent with the pathology-omics contrastive learning in pre-training is more suitable for our proposed POMP framework.

Pre-training Tasks. We analyze the performance under individual pre-training tasks and different combinations to evaluate their impact on survival prediction performance. The

Methods	MP	COAD	LIHC	STAD	BRCA	LUAD	LGG
MCAT [Chen *et al.*, 2021]	×	0.691 ± 0.132	0.711 ± 0.029	0.622 ± 0.034	0.663 ± 0.041	0.664 ± 0.026	0.844 ± 0.032
GPDBN [Wang *et al.*, 2021b]	×	0.593 ± 0.081	0.643 ± 0.019	0.587 ± 0.025	0.636 ± 0.047	0.615 ± 0.053	0.844 ± 0.025
CAMR [Wu *et al.*, 2023a]	×	0.606 ± 0.074	0.691 ± 0.052	0.587 ± 0.029	0.656 ± 0.072	0.647 ± 0.059	0.803 ± 0.044
CMTA [Zhou and Chen, 2023]	×	0.641 ± 0.055	0.708 ± 0.034	0.631 ± 0.039	0.680 ± 0.063	0.670 ± 0.046	0.840 ± 0.038
HC-MAE [Wang *et al.*, 2023]	×	0.703 ± 0.083	0.737 ± 0.031	0.623 ± 0.045	0.695 ± 0.026	0.691 ± 0.044	0.851 ± 0.033
DeepCorrSurv [Yao *et al.*, 2017]	✓	0.549 ± 0.026	0.700 ± 0.048	0.609 ± 0.049	0.659 ± 0.018	0.662 ± 0.032	0.828 ± 0.034
PathOmics [Ding *et al.*, 2023]	✓	0.629 ± 0.042	0.690 ± 0.013	0.622 ± 0.034	0.694 ± 0.074	0.662 ± 0.027	0.848 ± 0.032
POMP (Ours)	✓	**0.715 ± 0.090**	**0.755 ± 0.047**	**0.651 ± 0.035**	**0.700 ± 0.030**	**0.695 ± 0.053**	**0.856 ± 0.046**

Table 1: Performance comparison of POMP and state-of-the-art methods on six cancer datasets. MP indicates whether it is a multi-modal pre-training method.

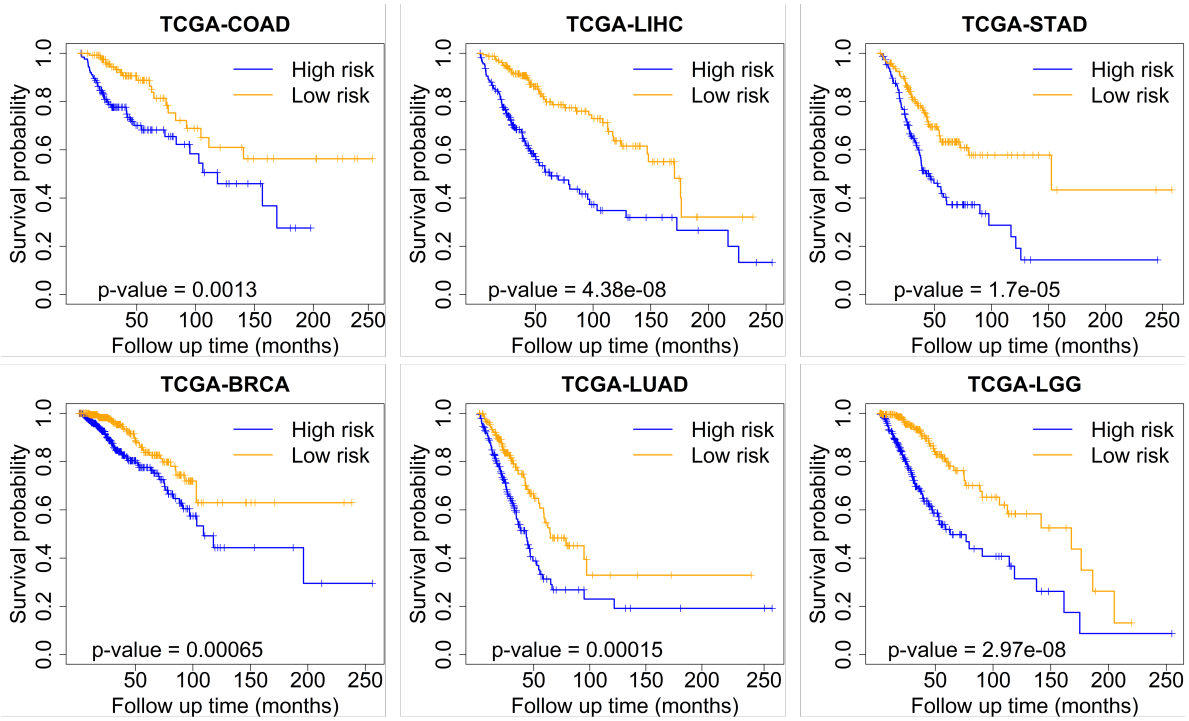
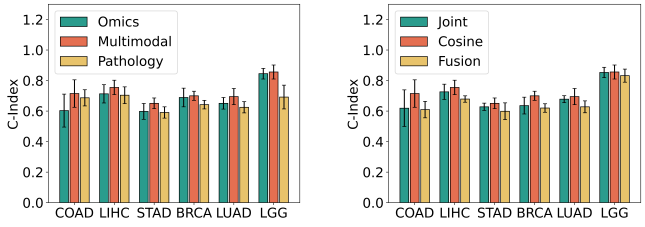


Figure 2: Survival analysis using Kaplan-Meier curves.



(a) Unimodal vs. multimodal (b) Risk computing methods

Figure 3: Ablation studies of modality and survival risk.

experimental results are presented in Table 2, where POMP-B0 denotes that POMP has not pre-training, $+\mathcal{L}_*$ represents the specific task is used to pre-train POMP. When looking at the individual pre-training tasks, we find that POMP-B1 with POC task obtains the superior C-index values on four out of six cancer datasets and achieves the best result in terms of macro average, demonstrating that the POC task contributes more than the POM and MOM tasks. Additionally, POMP-B3 with the MOM task obtains the worst performance among all individual pre-training tasks, even worse than POMP-B0 trained from scratch. However, in the scenario where two pre-training tasks are combined including POMP-B4, POMP-B5, and POMP-B6, we observe that POMP-B5 (which merges POC and MOM) achieves superior performance in terms of the macro average. Specifically, POMP-B5 surpasses POMP-B4 and POMP-B6 in four out of six datasets, suggesting that the MOM task together with the POM task can facilitate POMP learning to better semantic representation and fusion capabilities. Notably, the POMP ($\mathcal{L}_{total}$) framework, which joints all three pre-training tasks, achieves the best C-index value in terms of the macro average and outperforms other settings on five out of six datasets, demonstrating the effectiveness of our proposed POMP framework to joint three various pre-training tasks.

Pre-training Loss. To evaluate the impact of pre-training

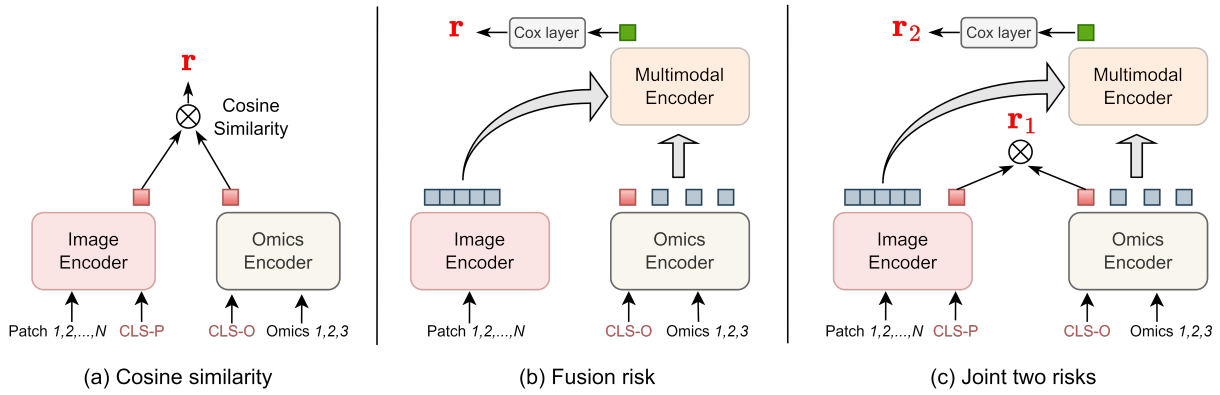
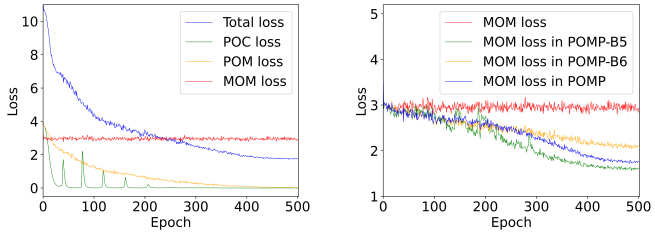


Figure 4: Survival risk computing methods.

Methods	COAD	LIHC	STAD	BRCA	LUAD	LGG	All (Macro-avg)
POMP-B0 (From scratch)	0.612 ± 0.030	0.706 ± 0.026	0.610 ± 0.041	0.648 ± 0.051	0.656 ± 0.022	0.847 ± 0.043	0.680 ± 0.036
POMP-B1 ($+\mathcal{L}_{poc}$)	0.638 ± 0.102	0.724 ± 0.033	0.583 ± 0.068	0.692 ± 0.048	0.692 ± 0.033	0.833 ± 0.059	0.694 ± 0.057
POMP-B2 ($+\mathcal{L}_{pom}$)	0.595 ± 0.076	0.738 ± 0.042	0.617 ± 0.025	0.670 ± 0.056	0.676 ± 0.032	0.823 ± 0.032	0.687 ± 0.044
POMP-B3 ($+\mathcal{L}_{mom}$)	0.618 ± 0.075	0.654 ± 0.039	0.594 ± 0.063	0.594 ± 0.055	0.625 ± 0.027	0.816 ± 0.041	0.650 ± 0.057
POMP-B4 ($+\mathcal{L}_{poc} + \mathcal{L}_{pom}$)	0.644 ± 0.049	0.714 ± 0.047	0.594 ± 0.005	0.679 ± 0.032	0.668 ± 0.023	0.815 ± 0.055	0.686 ± 0.035
POMP-B5 ($+\mathcal{L}_{poc} + \mathcal{L}_{mom}$)	0.660 ± 0.048	0.714 ± 0.016	0.596 ± 0.041	**0.706 ± 0.023**	0.679 ± 0.038	0.835 ± 0.062	0.698 ± 0.038
POMP-B6 ($+\mathcal{L}_{pom} + \mathcal{L}_{mom}$)	0.540 ± 0.074	0.715 ± 0.076	0.574 ± 0.030	0.644 ± 0.067	0.672 ± 0.040	0.836 ± 0.033	0.663 ± 0.053
POMP ($\mathcal{L}_{total}$)	**0.715 ± 0.090**	**0.755 ± 0.047**	**0.651 ± 0.035**	0.700 ± 0.030	**0.695 ± 0.053**	**0.856 ± 0.046**	**0.729 ± 0.050**

Table 2: Ablation experiments of different pre-training tasks.

tasks on the convergence of POMP training, we analyze the pre-training loss curves corresponding to different pre-training task settings. As shown in Figure 5a, we can see that (i) the total loss and POM loss decrease smoothly as the number of training papers increases; (ii) the POC loss decreases rapidly in the first 20 epochs, and then there are several fluctuations, but the fluctuations disappear after 300 epochs; (iii) it is interesting to see that MOM loss does not converge, which is consistent with POMP-B3 (Table 2) obtaining the worst performance in three individual pre-training tasks. We further verify whether the MOM loss will converge when the MOM task is combined with other tasks, as shown in Figure 5b. The MOM losses in POMP-B5 (which combines POC and MOM tasks), POMP-B6 (which combines POM and MOM tasks), and POMP (which combines three tasks) all decrease and converge normally, suggesting that the MOM task combined with other tasks contributes to the training of POMP.



(a) Pre-training loss under different tasks

(b) MOM loss under different tasks

Figure 5: Loss curves for different pre-training tasks.

5 Conclusion

We proposed a novel pathology-omics multimodal pre-training (POMP) framework for cancer survival prediction, by leveraging the principle of pre-trained models and exploring the benefit of aligning multimodal information of the same patient. POMP is jointly pre-trained on three tasks based on self-supervised learning and contrastive learning to align pathology image and omics data before fusion, allowing us to use their similarity as survival predictions. Experimental results show that POMP outperforms existing state-of-the-art methods on six cancer datasets.

While POMP shows promising results, there is still substantial room for improvement.

1. We will extend POMP to integrate other types of omics data associated with cancer survival, such as somatic mutations, copy number alterations, and clinical data.
2. We intend to pre-train the model based on a pancreatic cancer dataset to learn more prior knowledge about the potential relationship between multi-modalities.
3. We are going to explore alternative strategies for selecting the important genes in original multi-omics data, which will provide more valuable information in the omics modality.

Acknowledgments

This study was supported by the grants with No. 62162023 and No. KYQD(ZR)-21079.

References

[Bao *et al.*, 2022] Hangbo Bao, Wenhui Wang, Li Dong, et al. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022.

[Chen *et al.*, 2020] Richard J Chen, Ming Y Lu, Jingwen Wang, et al. 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.

[Chen *et al.*, 2021] Richard J Chen, Ming Y Lu, Wei-Hung Weng, et al. 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.

[Cheng *et al.*, 2021] Mingyue Cheng, Fajie Yuan, Qi Liu, Xin Xin, and Enhong Chen. Learning transferable user representations with sequential behaviors via contrastive pre-training. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 51–60. IEEE, 2021.

[Chicco and Jurman, 2020] Davide Chicco and Giuseppe Jurman. Survival prediction of patients with sepsis from age, sex, and septic episode number alone. *Scientific Reports*, 10(1):17156, 2020.

[Cui *et al.*, 2023] Can Cui, Haichun Yang, Yaohong Wang, et al. Deep multimodal fusion of image and non-image data in disease diagnosis and prognosis: a review. *Progress in Biomedical Engineering*, 5(2):022001, 2023.

[Deshmukh *et al.*, 2023] Soham Deshmukh, Benjamin Elizalde, Rita Singh, et al. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108, 2023.

[Ding *et al.*, 2023] Kexin Ding, Mu Zhou, Dimitris N Metaxas, et al. Pathology-and-genomics multimodal transformer for survival outcome prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 622–631. Springer, 2023.

[Dosovitskiy *et al.*, 2020] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[Elizalde *et al.*, 2023] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, et al. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[Gensheimer *et al.*, 2019] Michael F Gensheimer, A Solomon Henry, Douglas J Wood, et al. Automated survival prediction in metastatic cancer patients using high-dimensional electronic medical record data. *JNCI: Journal of the National Cancer Institute*, 111(6):568–574, 2019.

[Guzhov *et al.*, 2022] Andrey Guzhov, Federico Raue, Jörn Hees, et al. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980, 2022.

[Herrmann *et al.*, 2021] Moritz Herrmann, Philipp Probst, Roman Hornung, Vindi Jurinovic, and Anne-Laure Boulesteix. Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, 22(3):bbaa167, 2021.

[Jin *et al.*, 2023] Shan Jin, Zhikui Chen, Shuo Yu, Muhammad Altaf, and Zhenchao Ma. Self-augmentation graph contrastive learning for multi-view attribute graph clustering. In *Proceedings of the 2023 Workshop on Advanced Multimedia Computing for Smart Manufacturing and Engineering*, pages 51–56, 2023.

[Kim *et al.*, 2021] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594, 2021.

[Li *et al.*, 2021] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, et al. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34:9694–9705, 2021.

[Li *et al.*, 2022] Guoqiang Li, Qi Fang, Linlin Zha, Xin Gao, and Nenggan Zheng. Ham: Hybrid attention module in deep convolutional neural networks for image classification. *Pattern Recognition*, 129:108785, 2022.

[Lu *et al.*, 2021] Ming Y Lu, Drew FK Williamson, Tiffany Y Chen, et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

[McQuin *et al.*, 2018] Claire McQuin, Allen Goodman, Vasiliy Chernyshev, et al. Cellprofiler 3.0: Next-generation image processing for biology. *PLoS Biology*, 16(7):e2005970, 2018.

[Muzellec *et al.*, 2023] Boris Muzellec, Maria Teleńczuk, Vincent Cabeli, and Mathieu Andreux. Pydeseq2: a python package for bulk rna-seq differential expression analysis. *Bioinformatics*, 39(9):btad547, 2023.

[Pei *et al.*, 2024] Qizhi Pei, Lijun Wu, Zhenyu He, Jinhua Zhu, Yingce Xia, Shufang Xie, and Rui Yan. Exploiting pre-trained models for drug target affinity prediction with nearest neighbors. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 1856–1866, 2024.

[Pölsterl, 2020] Sebastian Pölsterl. 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.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, et al. Learning transferable visual models from

natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763, 2021.

[Srinidhi *et al.*, 2021] Chetan L Srinidhi, Ozan Ciga, and Anne L Martel. Deep neural network models for computational histopathology: A survey. *Medical image analysis*, 67:101813, 2021.

[Sung *et al.*, 2021] Hyuna Sung, Jacques Ferlay, Rebecca L Siegel, et al. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 71(3):209–249, 2021.

[Tomczak *et al.*, 2015] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

[Troyanskaya *et al.*, 2001] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, et al. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[Vale-Silva and Rohr, 2021] Luís A Vale-Silva and Karl Rohr. Long-term cancer survival prediction using multimodal deep learning. *Scientific Reports*, 11(1):13505, 2021.

[Wang *et al.*, 2019] Ping Wang, Yan Li, and Chandan K Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys*, 51(6):1–36, 2019.

[Wang *et al.*, 2021a] Tongxin Wang, Wei Shao, Zhi Huang, et al. Mogonet integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nature Communications*, 12(1):3445, 2021.

[Wang *et al.*, 2021b] Zhiqin Wang, Ruiqing Li, Minghui Wang, et al. Gpdbn: deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction. *Bioinformatics*, 37(18):2963–2970, 2021.

[Wang *et al.*, 2023] Suixue Wang, Xiangjun Hu, and Qingchen Zhang. Hc-mae: Hierarchical cross-attention masked autoencoder integrating histopathological images and multi-omics for cancer survival prediction. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 642–647, 2023.

[Wang *et al.*, 2024] Suixue Wang, Huiyuan Lai, Shuling Wang, and Qingchen Zhang. Contramae: Contrastive alignment masked autoencoder framework for cancer survival prediction. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2621–2626. IEEE, 2024.

[Wu *et al.*, 2022] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, et al. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567, 2022.

[Wu *et al.*, 2023a] Xingqi Wu, Yi Shi, Minghui Wang, et al. Camr: cross-aligned multimodal representation learning for cancer survival prediction. *Bioinformatics*, 39(1):btad025, 2023.

[Wu *et al.*, 2023b] Yusong Wu, Ke Chen, Tianyu Zhang, et al. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[Xu *et al.*, 2023a] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

[Xu *et al.*, 2023b] Xuenan Xu, Zhiling Zhang, Zelin Zhou, et al. Blat: Bootstrapping language-audio pre-training based on audioset tag-guided synthetic data. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2756–2764, 2023.

[Xu *et al.*, 2024] Bo Xu, Erchen Yu, Hongfei Lin, and Linlin Zong. Public opinion analysis in short videos of emergencies using pre-trained language models. In *2024 International Conference on Computational Linguistics and Natural Language Processing (CLNLP)*, pages 31–35. IEEE, 2024.

[Yao *et al.*, 2017] Jiawen Yao, Xinliang Zhu, Feiyun Zhu, et al. Deep correlational learning for survival prediction from multi-modality data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 406–414, 2017.

[Yin *et al.*, 2024] Ziyi Yin, Muchao Ye, Tianrong Zhang, Tianyu Du, Jinguo Zhu, Han Liu, et al. Vlattack: Multimodal adversarial attacks on vision-language tasks via pre-trained models. *Advances in Neural Information Processing Systems*, 36, 2024.

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