Distilled Prompt Learning for Incomplete Multimodal Survival Prediction

Supplementary Material

In this supplementary materials, we will include details of datasets and implementation, as well as more experimental results.

6. Datasets

The datasets involved in this paper are sourced from The Cancer Genome Atlas (TCGA) including Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Colon and Rectum Adenocarcinoma (COAD-READ), Lung Adenocarcinoma (LUAD) and Uterine Corpus Endometrial Carcinoma (UCEC). We used all paired 20× WSIs and RNA-Seq data to evaluate overall survival (OS) [10, 36]. After removing the absent genes, all genes are grouped into 330 pathways provided by the previous work [14]. The detailed list of genes will be present in the code repository upon acceptance.

7. Implementation

7.1. Model

Stage 1 - UniPro. In terms of UniPro for Pathology, a pathology foundation model, UNI [6], was used as the patch encoder to extract a 1024-d feature for each patch image. A linear layer followed by ReLU [2] was employed as the adapter to project each patch feature into a 768-d feature. Similarly, SNN [15] was used to encode each pathway into 768-d feature, leading to a sequence of tokens with the shape of (330, 768).

To learn the morphological descriptions in WSIs for every risk band of survival outcome, we set up 8 groups of prompts for the textual branch corresponding to the category of the pair of survival time (4 bins of time intervals, i.e., $\boldsymbol{I}_t = 4$) and censorship status (0 or 1). For every prompt of each category, three parts include learnable context tokens, prefix tokens and classname tokens. The sequence length of learnable context tokens is set as 255 (i.e., $k = 255$ in Eq. 3 and 4) and the embedding dimension of each token is 768, leading to a set of learnable context tokens with the shape of (8, 255, 768) concatenated into the [CLS] token of textual input of UniPro-P along the dimension of the sequence length.

Then, the pretrained tokenizer of BioBERT-v1.2 [16] was used to tokenize the prefix words of "*This is a pathology slide image from the patient with overall survival of*" into prefix tokens with the embedding dimension of 768, which are shared across different classes. Prefix tokens are put right after the context tokens along the sequence dimension.

The classname tokens were also obtained by embedding a group of classnames into 768-d tokens using the same tokenizer, including 1) *high risk, dead*, 2) *mid-high risk, dead*, 3) *mid-low risk, dead*, 4) *low risk, dead*, 5) *short observation, alive*, 6) *mid-short observation, alive*, 7) *mid-long observation, alive* and 8) *long observation, alive*. These classname tokens were put at the end of the textual input after prefix tokens, leading into the final textual input $\boldsymbol{T}_p$. By forwarding the constructed textual input $\boldsymbol{T}_p$ into a LLM, BioBERT-v1.2 [16], the [CLS] token of each class prompt in the output of the LLM was used as the representation $\boldsymbol{t}_p^{(j)}$ of a class j, which would be substituted into Eq. 5 for top-K MaxPooing (K=256).

Similarly, to capture the expression patterns in RNA-Seq data for every risk band, we set up prompts following the aforementioned steps. In particular, the sequence length of genomic context tokens is 256, and prefix words become "*These are gene expression profiles from the patient with overall survival of*". Other settings are the same as UniPro-P.

Stage 2 - MultiPro. In this stage, we used the same LLM, BioBERT-v1.2 [16] to encode the multimodal input, which consisted of 3 parts: a [CLS] token, a sequence of pathology feature tokens and a sequence of genomic feature tokens. The maximum length of input for BioBERT-v1.2 is 512, and thus we set the lengths of pathology and genomics by 255 and 256 (i.e., $K_p = 255$ and $K_g = 256$), respectively. The dimension of each token is 768 as well. Additionally, $K = 256$ is set in Top-K MaxPooling during inference of UniPro Scoring, unless otherwise specified. The coefficient factors α_1 and α_2 are simply set by 1.0 and 1.0, respectively.

7.2. Training

The setting for missing modalities is introduced in the main text. Here we present the details of the training procedure.

Following the previous setting [10, 36], we adopted Adam optimizer with the initial learning rate of 2×10^{-4} and weight decay of 1×10^{-5}, unless otherwise specified. Due to the large size and varying length of WSIs, the batch size is 1 following the common setting [5, 36]. All experiments are trained for 30 epochs by default to ensure the convergence of every model. Particularly, for prompt-based methods that employed BioBERT, the initial learning rate of genomic backbone becomes 1×10^{-5}, and these models are trained for 50 epochs to guarantee the full convergence of these models. For a fair comparison, all models are ensured to have fully converged.

Variants	Test		Modules			BLCA (372)	BRCA (1007)	COADREAD (533)	LUAD (443)	UCEC (478)	Avg
	P	G	Self	Cross	Uni						
w/o US	●	○				0.6213±0.010	0.6829±0.008	0.6740±0.022	0.6327±0.018	0.7152±0.011	0.6652
+ Self	●	○	✓			0.6095±0.007	0.6742±0.014	0.6748±0.016	0.6512±0.007	0.7187±0.011	0.6657
++ Cross	●	○	✓	✓		0.6208±0.010	0.6860±0.006	0.6800±0.016	0.6548±0.009	**0.7291±0.010**	0.6741
++ Uni	●	○	✓		✓	0.6214±0.008	0.6814±0.008	0.6795±0.025	0.6512±0.009	0.7196±0.016	0.6706
DisPro (full)	●	○	✓	✓	✓	**0.6319±0.014**	**0.6895±0.011**	**0.6880±0.010**	**0.6612±0.014**	0.7272±0.018	**0.6796**
w/o US	○	●				0.6473±0.017	0.6836±0.014	0.6707±0.022	0.6420±0.020	0.7226±0.013	0.6732
+ Self	○	●	✓			0.6514±0.013	0.6763±0.018	0.6662±0.026	0.6468±0.017	0.7191±0.020	0.6720
++ Cross	○	●	✓	✓		0.6532±0.020	0.6825±0.011	0.6754±0.032	0.6484±0.015	0.7237±0.023	0.6766
++ Uni	○	●	✓		✓	0.6498±0.014	0.6810±0.009	0.6637±0.026	0.6487±0.023	0.7199±0.013	0.6726
DisPro (full)	○	●	✓	✓	✓	**0.6547±0.012**	**0.6841±0.018**	**0.6804±0.024**	**0.6548±0.012**	**0.7271±0.017**	**0.6802**
w/o US	●	●				0.6585±0.007	0.7194±0.013	0.6963±0.009	0.6635±0.014	0.7250±0.012	0.6926
+ Self	●	●	✓			0.6499±0.009	0.7091±0.006	0.6959±0.014	0.6590±0.009	0.7315±0.009	0.6891
++ Cross	●	●	✓	✓		0.6608±0.007	0.7190±0.013	0.7015±0.025	0.6646±0.020	0.7324±0.009	0.6957
++ Uni	●	●	✓		✓	0.6582±0.003	0.7183±0.012	0.6992±0.015	0.6659±0.017	0.7263±0.019	0.6936
DisPro (full)	●	●	✓	✓	✓	**0.6638±0.006**	**0.7219±0.015**	**0.7029±0.016**	**0.6741±0.007**	**0.7476±0.020**	**0.7021**

Table 3. **Ablation Study (C-Index) on UniPro Scoring (US) under 60% training missing rate**. 'Self' refers to the *Self-Scoring* module in US. 'Uni' indicates considering the textual class representation of the query modality when calculating scores in Eq. 8, while 'Cross' indicates considering the textual class representation of non-query modalities.

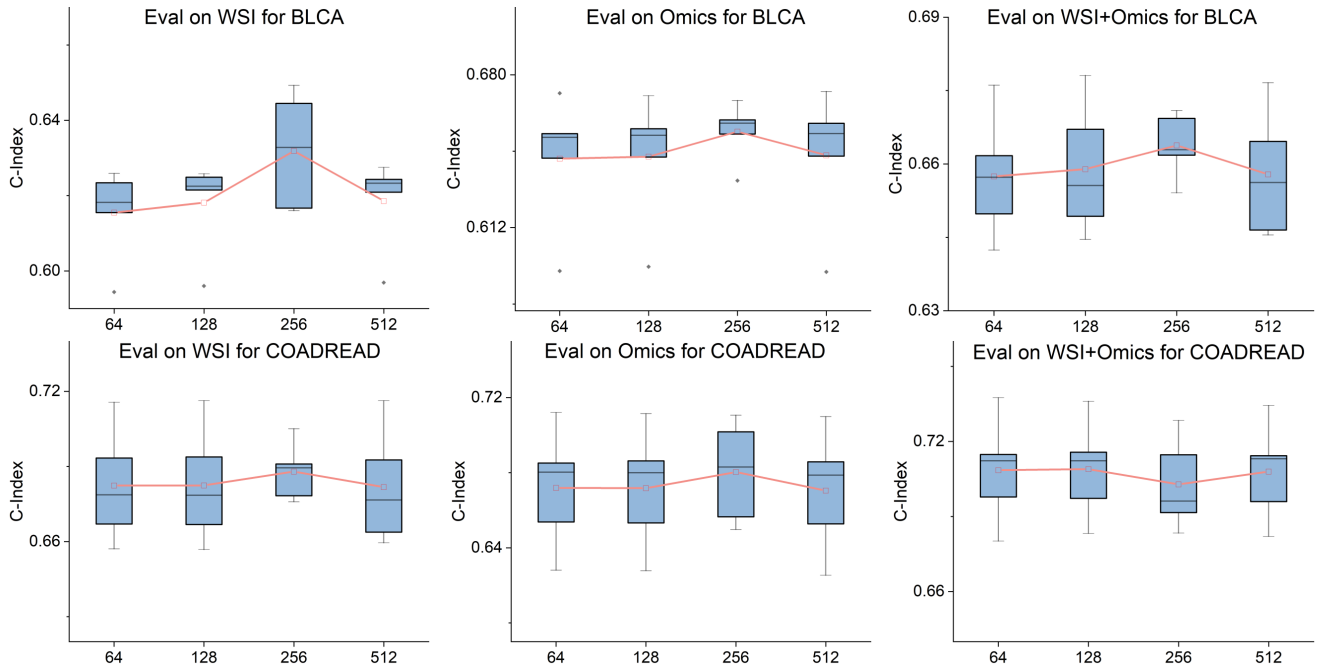

Figure 4. Performance on various K of Top-K MaxPooling in UniPro Scoring.

8. More Ablation Studies

8.1. Various Types of Tokens in UniPro Scoring

In this section, we investigate the roles of various tokens in UniPro Scoring, and results are shown in Tab. 3. Take the query modality for the UniPro Scoring module to be pathology as an example. The query tokens are a bag of features tokens $\{\mathbf{p}_n^i\}_{i=1}^{M_p}$. Then, 'Self' refers to the scores $\mathbf{a}_{n,p}$ computed by attention layers. 'Cross' suggests that the similarity between query features tokens and genomic

textual class representation $\boldsymbol{t}_g$ to get the scores $\mathbf{s}_{n,g}$ in Eq. 8. Similarly, 'Uni' suggests that the similarity between query features tokens and pathological textual class representation $\boldsymbol{t}_p$ to get the scores $\mathbf{s}_{n,p}$.

We observed that 1) when only incorporating Self-Scoring module, the performance become worse than the variant without US. The possible reason could be the information included in incomplete data is not enough for training a robust grader from scratch. 2) when additionally introducing either 'Cross' (++ Cross) or 'Uni' (++ Uni) scoring, the performance consistently surpasses the variant without US. This indicates the distilled knowledge can bring extra performance gains. In particularly, 'Cross' contributes to performance increases more significantly than 'Uni', which could be attributed to the assistance of 'Cross' in compensating for the modality-common knowledge. 3) The full version of DisPro achieves the best performance, suggesting the modal can benefit from their collaboration.

8.2. Top-K MaxPooling in UniPro Scoring

In this part, we explore the effect of the selection of K on DisPro. We set up a series of K including 64, 128, 256 and 512. Note that in other experiments, K is always 256. Results are shown in Fig. 4. In most cases, as K increases, the performance gradually rises until it reaches a peak, after which further increases in K will not result in significant performance gains. Therefore, considering the trade-off between performance and inference speed, we take K=256 as our default setting.

9. Visualization Interpretation

To intuitively validate if DisPro is robust to missing modalities, we visualize the attention signals of each token (512 in total for BioBERT) in LLM under the situations of missing modalities (WSI-only or Omics-only) and complete modality (WSI-Omics), and compare the prompt-based model, MAP [17]. We feed different combinations of missing and complete modalities of the same sample to the model and observe the differences of attention signals among them. If the model fed by incomplete data can predict similar attention signals to that of complete modality, the robustness to missing modalities has been learned by the model. The results are shown in Fig. 5, where we can see that attention signals for missing modalities in DisPro are more aligned with those of complete data, whereas MAP's predictions are chaotic across various modality combinations. This indicates that DisPro is more robust to missing modalities.

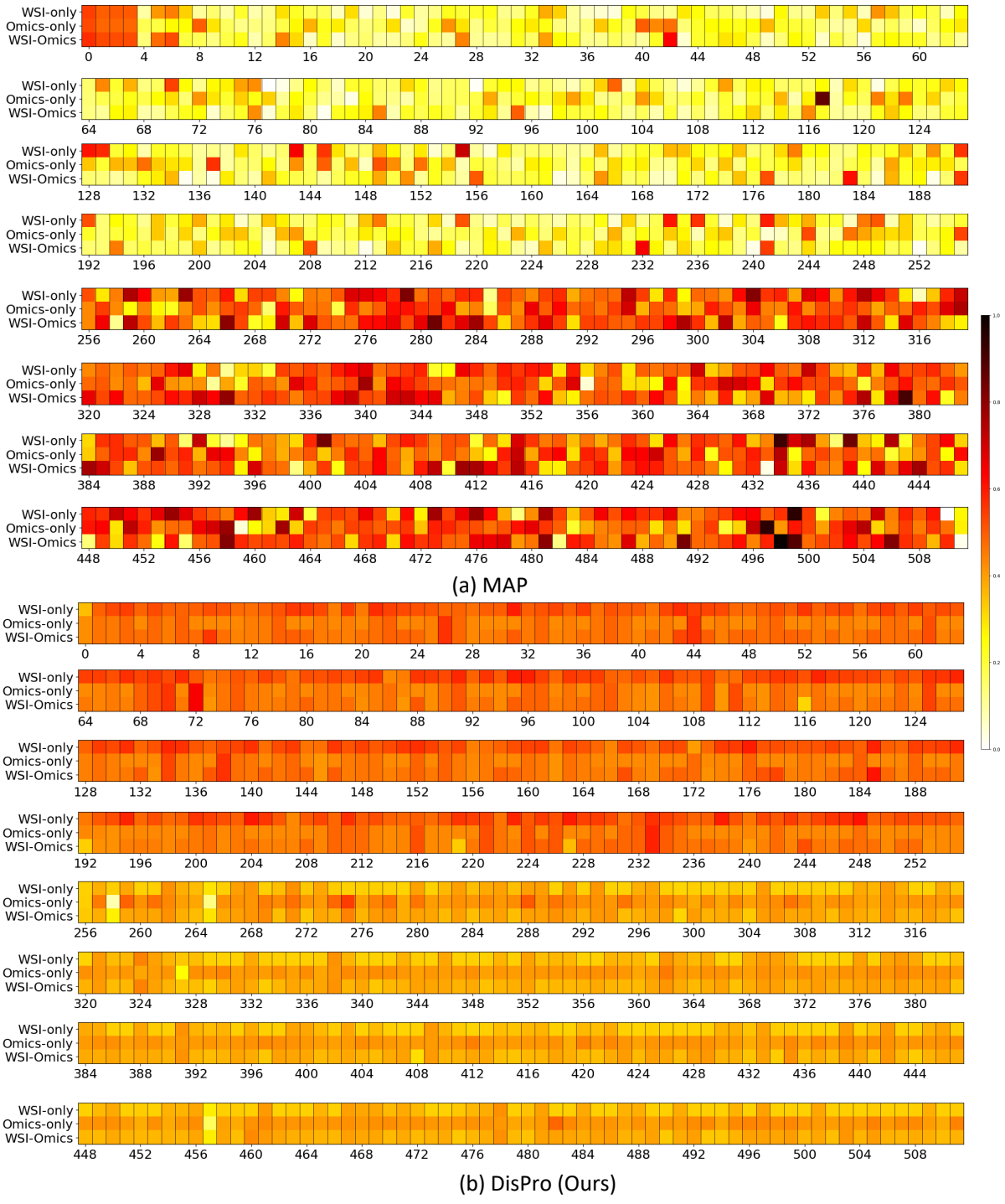

Figure 5. The visualization of attention signals for each token in LLM used in (a) MAP and (b) DisPro (Ours).