ORDINAL PRIVILEGED DISTILLATION FOR HISTOPATHOLOGY-ONLY SURVIVAL PREDICTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal learning can improve cancer survival prediction, but some modalities may be unavailable at deployment. Existing approaches often address this limitation through molecular reconstruction or representation matching, which can require learning variation beyond the survival task. We instead use a teacher trained on the auxiliary modality alone and take its survival outputs as the sole source of supervision from that modality. Transferring these outputs raises a further challenge. Existing survival distillation objectives typically encourage numerical agreement, although teachers and students observing different modalities need not share the same optimal predictions. We prove that matching the teacher's event probabilities at each time interval can shift the student's optimal predictions. We therefore introduce Ordinal Privileged Distillation (OPD), which transfers pairwise risk ordering derived from the teacher's survival curves without matching its absolute score scale or pairwise score gaps. A pairwise logistic loss supplies ordinal supervision alongside a survival loss that accounts for right censoring, leaving the student's inference architecture unchanged. We evaluate OPD using transcriptome data as the training-only modality and whole-slide images (WSIs) as the student's input. Extensive experiments on eight cancer benchmark datasets and four WSI backbones demonstrated our superiority over other methods.

1 INTRODUCTION

Survival prediction supports patient risk stratification and treatment planning in cancer care (Skrede et al., 2020; Sparano et al., 2018; Mobadersany et al., 2018). Multimodal data provide complementary information for this task: whole-slide images (WSIs) reveal tumor morphology and the surrounding tissue microenvironment (Bera et al., 2019; Kather et al., 2019), while transcriptomic data capture molecular information relevant to disease progression (Van't Veer et al., 2002; Perou et al., 2000). Combining these modalities can improve survival prediction (Chen et al., 2021; Jaume et al., 2024; Zhang et al., 2024; Yuan et al., 2012), but transcriptomic profiling requires additional assays and may be unavailable. We study this setting: paired modalities are observed during training, but inference uses only the deployment modality. The auxiliary modality thus provides privileged information for learning (Lopez-Paz et al., 2015). This raises two questions: (1) what should be learned from the training-only modality, and (2) how should that information supervise the deployed model?

Regarding the first question, existing methods often use molecular reconstruction, representation generation, or cross-modal feature translation (Wang et al., 2025; Zhou et al., 2025; Liu et al., 2025). Other approaches compensate for missing molecular representations or learn structured latent correspondences through cross-modal reconstruction (Xu et al., 2025; Zhang et al., 2026). These methods make recovery of auxiliary features part of the learning problem. Even with survival supervision, such targets can require recovering variation beyond what is needed to predict the endpoint. Molecular prediction studies also show that recoverability from histology varies across genes (Wang et al., 2021; Schmauch et al., 2020; Pizurica et al., 2024). We study an alternative: using the survival outputs of an auxiliary-modality teacher as the sole source of supervision from that modality. This places the transfer target at the task level, without requiring reconstruction or intermediate feature matching.

With the teacher's survival outputs as the transfer target, the second question is how to distill them across modalities. Existing survival-specific objectives encourage students to reproduce teacher pre-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

dictions by matching interval-wise event probabilities or probability distributions constructed from risk scores (Deng et al., 2026a;b). Such numerical matching constrains the values of the student's predictions (Huang et al., 2022). The teacher and student, however, condition their predictions on different observations, so their optimal predictions for the same endpoint need not coincide. Requiring numerical agreement can therefore impose overly restrictive constraints on cross-modal knowledge transfer (Xue et al., 2023; Huo et al., 2024).

To transfer the teacher's survival information with fewer numerical constraints, we introduce Ordinal Privileged Distillation (OPD), which distills pairwise risk ordering. Each comparison tells the student which patient has higher predicted risk, without specifying the teacher's absolute score scale or the size of the risk difference. This retains information relevant to risk stratification while avoiding a requirement to reproduce the teacher's survival probabilities or risk-score gaps. Specifically, we first train and freeze an RNA-only survival teacher and summarize its predicted survival curves as scalar risk scores. Their pairwise ordering supervises the WSI student through a pairwise logistic loss. The student also learns from the standard survival loss. Inference uses the student's existing WSI encoder and survival head.

Our contributions are: (1) We learn directly from an auxiliary-modality teacher's survival outputs and theoretically illustrate that learning task information need not require reconstructing the auxiliary modality. (2) We show that matching a teacher's predicted event probabilities in cross-modal survival distillation can shift the student's optimal survival predictions. Motivated by this finding, we propose OPD to transfer pairwise risk ordering without matching the teacher's absolute scores or score gaps. (3) Our experiments on eight cancer benchmark datasets and four WSI backbones show OPD's superiority over other methods and effectiveness across architectures.

2 Related Work

Survival Prediction with Missing Molecular Modalities. Learning from paired histopathology and transcriptomics while retaining WSI-only inference has motivated approaches that reconstruct or compensate for missing molecular information. G-HANet learns histo-genomic associations through functional gene reconstruction and incorporates these associations into WSI-based survival modeling (Wang et al., 2025). LD-CVAE generates molecular representations conditioned on WSIs using a variational framework with differentiated latent factors (Zhou et al., 2025). ProSurv uses survival-aware prototype banks to guide cross-modal feature translation (Liu et al., 2025), while DisPro combines unimodal prompt learning with multimodal prompting to compensate for missing representations (Xu et al., 2025). SlotSPE models patient-specific prognostic events with compact slots and uses cross-modal reconstruction to support prediction when molecular inputs are absent (Zhang et al., 2026). The survival-aware prompts, prototypes, and slots in these approaches already make representation learning task supervised. Their molecular supervision is mediated by reconstructed features, compensated representations, or structured latent correspondences. OPD studies whether this supervision can instead be supplied entirely through the teacher's survival outputs, without a molecular reconstruction or feature-alignment objective.

Knowledge Distillation for Survival Prediction. Classic knowledge distillation trains a student to match softened teacher predictions (Hinton et al., 2015), while generalized distillation extends this supervision to privileged inputs (Lopez-Paz et al., 2015). In histopathological survival prediction, AMKD distills a multimodal teacher into a WSI student, combining feature transfer with classification-style probability matching on discretized survival outputs (Xu et al., 2026). Survival-specific formulations further adapt distillation to censored time-to-event data. DiSKD augments the student's survival loss with time- and cause-specific Kullback–Leibler (KL) penalties between teacher and student conditional event probabilities over observed at-risk intervals (Deng et al., 2026a). Component-wise composite likelihood distillation aligns locally normalized risk-score distributions within comparability sets, such as Cox risk sets or sampled case–control sets (Deng et al., 2026b). These approaches use numerical probability matching to distill survival outputs. In our setting, however, the teacher and student use different modalities, so their optimal survival predictions need not coincide. Such matching therefore introduces cross-modal compatibility requirements that paired observations alone do not guarantee. Ranking-based knowledge transfer has also been explored in recommendation and learning-to-rank (Tang & Wang, 2018; Reddi et al., 2021). OPD uses this ordinal perspective for cross-modal survival learning: it derives pairwise risk comparisons from

teacher survival curves and combines an unweighted pairwise logistic loss with the student's censored survival likelihood. Its teacher supervision depends on comparison directions and is invariant to strictly increasing transformations of teacher risk scores.

3 METHOD

We first formulate our problem and prediction task (Sec. 3.1), then explain two design choices: using teacher's survival outputs as the source of supervision (Sec. 3.2) and distilling the pairwise risk ordering derived from them (Sec. 3.3). Finally, we describe the teacher–student training objective (Sec. 3.4).

3.1 PROBLEM FORMULATION AND SURVIVAL PREDICTION

The training cohort is $\mathcal{D}_{\text{train}} = \{(\mathbf{W}_n, \mathbf{G}_n, t_n, c_n)\}_{n=1}^{N}$, where $\mathbf{W}_n$ denotes a patient's WSI data, $\mathbf{G}_n$ the RNA expression profile, and t_n the observed follow-up time. We use $c_n = 1$ for right censoring and $c_n = 0$ for an observed event. Only $\mathbf{W}_n$ is available at inference.

We divide the prediction horizon into I_t intervals and let $T \in \{1, \dots, I_t\}$ denote the interval-coded endpoint. Each branch $b \in \{p, g\}$, for the WSI student and RNA teacher respectively, predicts hazards $\mathbf{h}_n^b = (h_n^{b,(1)}, \dots, h_n^{b,(I_t)})$. The hazard $h_n^{b,(k)}$ estimates the probability of an event in interval k, conditional on survival to its start and the branch's input. The survival probabilities are

$$S_n^{b,(k)} = \prod_{z=1}^{k} (1 - h_n^{b,(z)}), \qquad S_n^{b,(0)} = 1. \tag{1}$$

The predicted hazards and their associated survival curve form the model's survival outputs. Let τ_n be the interval assigned to the observed follow-up time. Under our interval-level observation convention, an event contributes $T = \tau_n$, while censoring contributes survival through τ_n. Each branch is trained with the negative log-likelihood (NLL) loss:

$$\mathcal{L}_{\text{surv}}^{b} = -\frac{1}{N} \sum_{n=1}^{N} \Big\{ (1 - c_n) \big[\log h_n^{b,(\tau_n)} + \log S_n^{b,(\tau_n - 1)} \big] + c_n \log S_n^{b,(\tau_n)} \Big\}. \tag{2}$$

We denote the corresponding per-patient loss by ℓ_{surv}^{b}. During minibatch training, $\mathcal{L}_{\text{surv}}^{b}$ is averaged over the patients in the current batch.

3.2 WHAT TO DISTILL: AUXILIARY-MODALITY SURVIVAL OUTPUTS

We examine the supervision target by treating $\mathbf{W}$ and $\mathbf{G}$ as the inputs observed by the student and teacher, respectively. The auxiliary modality can supply supervision through input reconstruction, representation transfer, or task outputs (Wang et al., 2025; Zhou et al., 2025; Liu et al., 2025; Xu et al., 2025; Zhang et al., 2026). We choose the survival outputs of a teacher trained on $\mathbf{G}$. The following analysis separates the information needed to predict the endpoint from the information needed to reconstruct the auxiliary input.

Prognostic information in the full survival curve. Define the ideal auxiliary-modality survival curve as

$$\mathbf{S}^{g,*}(\mathbf{G}) = \big(S^{g,*,(k)}(\mathbf{G})\big)_{k=1}^{I_t}, \qquad S^{g,*,(k)}(\mathbf{G}) = \mathbb{P}(T > k \mid \mathbf{G}). \tag{3}$$

The star denotes population conditional probabilities, which a learned auxiliary-modality teacher estimates. The complete curve has the following sufficiency property.

Proposition 1 (Sufficiency of the full auxiliary-modality survival curve)**.** *The full survival curve* $\mathbf{S}^{g,*}(\mathbf{G})$ *determines the auxiliary-modality conditional distribution of* T *and satisfies*

$$T \perp \mathbf{G} \mid \mathbf{S}^{g,*}(\mathbf{G}). \tag{4}$$

Successive differences of the survival probabilities recover the event probabilities. Thus, once the ideal full curve is known, observing the original auxiliary input does not further change the conditional distribution of T (Appendix A.1). This makes the survival curve a task-specific source of supervision.

Predicting survival need not require auxiliary-input reconstruction. A construction illustrates the distinction between these two learning targets. Let $\mathbf{G} = (V, \boldsymbol{\xi})$ and $\mathbf{W} = V$, where V has a finite second moment and determines the event distribution. The vector $\boldsymbol{\xi}$ contains d independent, zero-mean, unit-variance coordinates independent of V, and supplies no additional information about T given V. Then

$$\inf_f \mathbb{E}\|\mathbf{G} - f(\mathbf{W})\|_2^2 = d, \qquad \inf_u \mathbb{E}\|\mathbf{S}^{g,*}(\mathbf{G}) - u(\mathbf{W})\|_2^2 = 0. \tag{5}$$

In this example, the deployment input contains the information needed to recover the survival curve, while the additional auxiliary-input variation cannot be recovered from it (Appendix A.2).

Choice of supervision target. Together, Proposition 1 and this example support using auxiliary-modality survival outputs without making input reconstruction a prerequisite. We next consider how to transfer these auxiliary-modality outputs to a student model that observes a different modality.

3.3 How to distill: cross-modal risk ordering

Survival-specific objectives distill survival outputs through numerical matching (Deng et al., 2026a;b). The teacher and student condition on different inputs, $\mathbf{G}$ and $\mathbf{W}$, so their optimal predictions can differ for the same subject and endpoint. We first examine the condition under which hazard matching preserves the optimal student predictions, then motivate ordinal supervision based on pairwise risk ordering.

The prediction target induced by numerical matching. Consider interval-wise teacher-to-student Bernoulli KL, as in the single-event, untempered form of DiSKD (Deng et al., 2026a). For one subject, let $A_k = 1$ indicate that interval k is observed at risk. The distillation penalty is

$$\mathcal{L}_{\mathrm{KL}} = \sum_{k=1}^{I_t} A_k \ \mathrm{kl}\big(h^{g,(k)}(\mathbf{G}), h^{p,(k)}(\mathbf{W})\big), \tag{6}$$

where $\mathrm{kl}(a, b)$ is the KL divergence from $\mathrm{Bernoulli}(a)$ to $\mathrm{Bernoulli}(b)$. The survival loss and the KL penalty use the same observed at-risk intervals, as specified in Appendix A.3.

Two quantities determine the predictions that minimize the expected combined loss:

$$h_k^*(\mathbf{W}) = \mathbb{P}(T = k \mid T \geq k, \mathbf{W}), \qquad m_k(\mathbf{W}) = \mathbb{E}[h^{g,(k)}(\mathbf{G}) \mid \mathbf{W}, A_k = 1]. \tag{7}$$

Here h_k^* is the true hazard given the deployment input, while m_k averages teacher hazards conditional on that input among subjects observed at risk in interval k. Their difference determines whether adding distillation changes the predictions that minimize the expected survival loss.

Proposition 2 (Compatibility of cross-modal hazard matching)**.** *Assume independent censoring conditional on* $\mathbf{W}$ *and a fixed teacher. For* $\gamma > 0$*, minimize* $\mathbb{E}[\ell_{\mathrm{surv}}^p + \gamma \mathcal{L}_{\mathrm{KL}}]$ *over measurable student hazards. On intervals with* $\mathbb{P}(A_k = 1 \mid \mathbf{W}) > 0$*, the minimizing hazard is*

$$h_{\gamma,k}(\mathbf{W}) = \frac{h_k^*(\mathbf{W}) + \gamma m_k(\mathbf{W})}{1 + \gamma}. \tag{8}$$

It equals $h_k^*(\mathbf{W})$*, the minimizer of the expected survival loss, exactly when* $m_k(\mathbf{W}) = h_k^*(\mathbf{W})$ *almost surely on these intervals.*

The combined objective therefore pulls the optimal student hazard toward the average teacher hazard. When this average differs from the true deployment-modality hazard, numerical matching changes the optimal prediction. Thus, matching outputs across modalities requires a compatibility condition that paired observations alone do not ensure (Appendix A.3).

Pairwise risk ordering as ordinal supervision. This analysis motivates asking which information in the teacher's survival outputs can provide supervision with fewer numerical constraints. Following the principle of relational and ranking distillation (Tang & Wang, 2018; Park et al., 2019), we retain pairwise risk ordering. For two teacher risk scores r_i^g and r_j^g, the sign of $r_i^g - r_j^g$ specifies which subject has higher predicted risk, without specifying either subject's survival probabilities

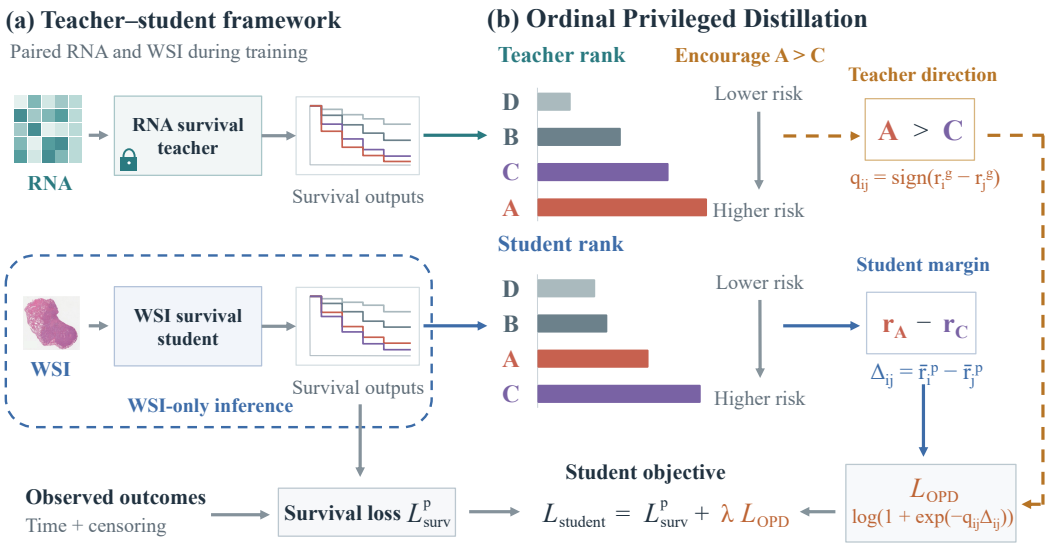

Figure 1: Overview of OPD. (a) A frozen RNA teacher provides ordinal supervision to a WSI student. Only the WSI student is used at inference. (b) Teacher risk scores supply comparison directions q_{ij}, while standardized student risk scores supply margins, abbreviated in the diagram as $\Delta_{ij} = \bar{r}_i^p - \bar{r}_j^p$. A pairwise logistic loss encourages agreement with the teacher's pairwise risk ordering and is added to the student's survival loss.

or the size of their risk difference. This comparison is invariant to any strictly increasing transformation of the teacher's scores. Such a transformation can change both the absolute scale and the gaps between subjects while preserving the comparison directions. These properties motivate using pairwise risk ordering as ordinal supervision across modalities. We evaluate its usefulness through comparisons of distillation objectives and teacher sources in Section 4.

3.4 Training with OPD

As shown in Figure 1, we implement the preceding choice with a frozen RNA teacher and a WSI student. The teacher supplies ordinal supervision, while observed outcomes directly supervise the student's survival outputs through the survival loss.

Train and freeze the RNA teacher. We first train an RNA-only teacher $\mathbf{h}_n^g = f_g(\mathbf{G}_n; \theta_g)$ by minimizing $\mathcal{L}^g_{\text{surv}}$ in Eq. (2), then freeze θ_g. The WSI student $\mathbf{h}_n^p = f_p(\mathbf{W}_n; \theta_p)$ has its own encoder and survival head, with no parameters shared with the teacher. Distillation operates on their survival outputs, so the two branches can use different architectures and feature dimensions.

Construct teacher comparison labels. For a batch of B paired patients, we compute each branch's survival curve using Eq. (1) and summarize it as a scalar risk score:

$$r_i^b = -\sum_{k=1}^{I_t} S_i^{b,(k)}, \qquad b \in \{p, g\}. \tag{9}$$

A larger score represents higher predicted risk. We use all ordered pairs of distinct patients whose teacher scores are non-tied:

$$\mathcal{P}_B = \{(i,j) : 1 \le i, j \le B,\ i \ne j,\ r_i^g \ne r_j^g\}, \qquad q_{ij} = \text{sign}(r_i^g - r_j^g). \tag{10}$$

Thus, $q_{ij} = 1$ indicates that the teacher assigns higher risk to patient i, and $q_{ij} = -1$ indicates the reverse. These labels are predictions supplied by the teacher and retain the teacher's comparison directions without transferring its score gaps.

Pairwise logistic loss. We standardize student risk scores within each batch:

$$\bar{r}_i^p = \frac{r_i^p - \mu_p}{s_p + \varepsilon}, \tag{11}$$

where $\mu_p = \frac{1}{B}\sum_{i=1}^{B} r_i^p$ and $s_p^2 = \frac{1}{B}\sum_{i=1}^{B}(r_i^p - \mu_p)^2$, and ε is a small constant for stability.

Gradients flow through the student batch statistics, while the frozen teacher's predictions are detached. For each valid pair, the signed risk difference $q_{ij}(\bar{r}_i^p - \bar{r}_j^p)$ is positive when the student's ordering agrees with the teacher. We encourage this agreement with the pairwise logistic loss

$$\mathcal{L}_{\mathrm{OPD}} = \frac{1}{|\mathcal{P}_B|} \sum_{(i,j)\in\mathcal{P}_B} \log\big(1 + \exp\big[-q_{ij}(\bar{r}_i^p - \bar{r}_j^p)\big]\big)\,. \tag{12}$$

The teacher supplies the comparison direction q_{ij}, while the student's standardized risk difference determines the margin. Teacher risk scores enter only through their ordering, without requiring the student to reproduce their absolute values or between-patient gaps. Each valid pair has equal weight. The comparison labels encode the teacher's predicted risk ordering rather than observed event order.

Combine survival and ordinal supervision. For each training batch, the student minimizes

$$\mathcal{L}_{\mathrm{student}} = \mathcal{L}_{\mathrm{surv}}^p + \lambda\mathcal{L}_{\mathrm{OPD}}, \tag{13}$$

where $\lambda = 0.3$ controls the strength of ordinal supervision. The survival loss supplies supervision from observed times and censoring indicators, and the pairwise logistic loss encourages agreement with the teacher's risk ordering. Both terms act on the same student model and update only θ_p.

At inference, the student predicts hazards, survival probabilities, and the risk score in Eq. (9) from WSI alone. Batch standardization is used only for training margins. Predicting a patient's risk score requires neither RNA nor other patients in a batch.

4 Experiments

Our experiments connect the design choices in Sections 3.2 and 3.3 to empirical comparisons. We first evaluate overall WSI-only survival performance, then examine the supervision target, the distillation objective, and the teacher source. Each comparison includes four WSI backbones, allowing us to assess whether the observed patterns depend on a particular student architecture.

4.1 Experimental setup

Datasets. We evaluate on eight publicly available cohorts from The Cancer Genome Atlas (TCGA) (Weinstein et al., 2013): breast invasive carcinoma (BRCA, n=1,047), colon and rectum adenocarcinoma (CRC, n=342), glioblastoma and lower-grade glioma (GBMLGG, n=585), head and neck squamous cell carcinoma (HNSC, n=441), kidney renal clear cell carcinoma (KIRC, n=510), lung adenocarcinoma (LUAD, n=453), skin cutaneous melanoma (SKCM, n=417), and uterine corpus endometrial carcinoma (UCEC, n=149). Models are trained to predict overall survival (OS). Following standard practice, we use 5-fold site-stratified cross-validation to mitigate batch effect (Howard et al., 2021). The performance of our model is assessed by concordance index (C-index). Log-2 transformed transcripts per million bulk RNA sequencing expression for all TCGA cohorts is accessed through UCSC Xena database (Goldman et al., 2020).

Implementation details. We use Trident (Zhang et al., 2025) to segment tissue and extract non-overlapping 256×256 patches at $20\times$ magnification, followed by feature extraction with the frozen UNI encoder (Chen et al., 2024). Transcriptomic genes are selected and grouped into 50 pathways according to the MSigDB Hallmark gene sets (Liberzon et al., 2015). The architecture of RNA teacher is the RNA-only branch of MMP (Song et al., 2024b): per-pathway SNN encoders (Klambauer et al., 2017) over 50 Hallmark pathways followed by one self-attention layer. All models are trained for 20 epochs using the AdamW optimizer with an initial learning rate of 1×10^{-4}, a weight decay of 1×10^{-5}, and a cosine learning rate decay schedule. We use an effective batch size of 64, achieved through gradient accumulation when necessary. During training, prototype-based methods aggregate all patches from each WSI into 16 prototypes (Song et al., 2024a), whereas the remaining methods randomly sample 4,096 patches per WSI to increase sampling diversity and reduce memory consumption. During inference, all methods use all patches from each WSI.

Table 1: WSI-only survival prediction results. The first block contains methods using paired modalities during training. Each subsequent backbone is shown with WSI NLL-only training and with OPD, and Δ is the gain of OPD over the backbone. Each cell reports the C-index with mean ± standard deviation. The best and second-best results are highlighted in **bold** and underline; OPD results marked with † outperform all existing methods with RNA during training on that cohort.

Method	BRCA	CRC	GBMLGG	HNSC	KIRC	LUAD	SKCM	UCEC
Existing methods with RNA during training								
G-HANet	$0.650_{\pm 0.079}$	$0.615_{\pm 0.070}$	$0.780_{\pm 0.049}$	$0.600_{\pm 0.079}$	$0.688_{\pm 0.077}$	$0.618_{\pm 0.045}$	$0.605_{\pm 0.054}$	$0.657_{\pm 0.149}$
ProSurv	$0.635_{\pm 0.041}$	$0.617_{\pm 0.035}$	$0.766_{\pm 0.055}$	$0.561_{\pm 0.049}$	$0.666_{\pm 0.084}$	$0.571_{\pm 0.037}$	$0.550_{\pm 0.031}$	$0.610_{\pm 0.191}$
LD-CVAE	$0.677_{\pm 0.029}$	$0.634_{\pm 0.034}$	$0.769_{\pm 0.059}$	$0.615_{\pm 0.058}$	$0.695_{\pm 0.098}$	$\mathbf{0.633}_{\pm 0.052}$	$0.628_{\pm 0.056}$	$0.660_{\pm 0.138}$
DisPro	$0.635_{\pm 0.049}$	$0.612_{\pm 0.037}$	$0.696_{\pm 0.083}$	$0.528_{\pm 0.042}$	$0.631_{\pm 0.079}$	$0.584_{\pm 0.053}$	$0.567_{\pm 0.035}$	$0.574_{\pm 0.233}$
SlotSPE	$0.646_{\pm 0.081}$	$0.688_{\pm 0.078}$	$0.777_{\pm 0.039}$	$0.611_{\pm 0.081}$	$0.678_{\pm 0.096}$	$0.587_{\pm 0.047}$	$0.592_{\pm 0.056}$	$0.591_{\pm 0.198}$
WSI backbones with NLL-only training or OPD								
ABMIL	$0.644_{\pm 0.058}$	$0.557_{\pm 0.103}$	$0.767_{\pm 0.044}$	$0.577_{\pm 0.058}$	$0.654_{\pm 0.073}$	$0.556_{\pm 0.040}$	$0.555_{\pm 0.043}$	$0.647_{\pm 0.191}$
+OPD	$0.665_{\pm 0.056}$	$0.640_{\pm 0.093}$	$0.780_{\pm 0.034}$	$\underline{0.636}^{\dagger}_{\pm 0.079}$	$0.676_{\pm 0.075}$	$0.591_{\pm 0.029}$	$0.623_{\pm 0.043}$	$0.713^{\dagger}_{\pm 0.176}$
Δ	+0.021	+0.083	+0.013	+0.059	+0.022	+0.035	+0.068	+0.066
TransMIL	$0.676_{\pm 0.060}$	$0.639_{\pm 0.035}$	$0.764_{\pm 0.051}$	$0.612_{\pm 0.039}$	$0.661_{\pm 0.091}$	$0.579_{\pm 0.034}$	$0.583_{\pm 0.052}$	$0.634_{\pm 0.187}$
+OPD	$\underline{0.683}^{\dagger}_{\pm 0.048}$	$0.664_{\pm 0.024}$	$0.777_{\pm 0.042}$	$0.622^{\dagger}_{\pm 0.041}$	$0.682_{\pm 0.076}$	$0.607_{\pm 0.039}$	$0.607_{\pm 0.043}$	$0.717^{\dagger}_{\pm 0.149}$
Δ	+0.007	+0.025	+0.013	+0.010	+0.021	+0.028	+0.024	+0.083
PANTHER	$\underline{0.683}_{\pm 0.110}$	$0.680_{\pm 0.104}$	$0.764_{\pm 0.031}$	$0.586_{\pm 0.075}$	$0.683_{\pm 0.069}$	$0.619_{\pm 0.048}$	$0.618_{\pm 0.064}$	$0.623_{\pm 0.109}$
+OPD	$\mathbf{0.688}^{\dagger}_{\pm 0.081}$	$\underline{0.690}^{\dagger}_{\pm 0.104}$	$\underline{0.782}^{\dagger}_{\pm 0.031}$	$0.631^{\dagger}_{\pm 0.048}$	$\underline{0.710}^{\dagger}_{\pm 0.054}$	$0.625_{\pm 0.054}$	$\underline{0.648}^{\dagger}_{\pm 0.026}$	$\underline{0.719}^{\dagger}_{\pm 0.213}$
Δ	+0.005	+0.010	+0.018	+0.045	+0.027	+0.006	+0.030	+0.096
MMP-Histo	$0.630_{\pm 0.060}$	$0.674_{\pm 0.104}$	$0.732_{\pm 0.103}$	$0.609_{\pm 0.055}$	$0.626_{\pm 0.057}$	$0.586_{\pm 0.055}$	$0.610_{\pm 0.078}$	$0.666_{\pm 0.158}$
+OPD	$\underline{0.683}^{\dagger}_{\pm 0.069}$	$\mathbf{0.699}^{\dagger}_{\pm 0.081}$	$\mathbf{0.783}^{\dagger}_{\pm 0.033}$	$\mathbf{0.640}^{\dagger}_{\pm 0.073}$	$\mathbf{0.713}^{\dagger}_{\pm 0.062}$	$\underline{0.630}_{\pm 0.057}$	$\mathbf{0.651}^{\dagger}_{\pm 0.079}$	$\mathbf{0.796}^{\dagger}_{\pm 0.109}$
Δ	+0.053	+0.025	+0.051	+0.031	+0.087	+0.044	+0.041	+0.130

4.2 Overall WSI-only survival performance

This experiment evaluates whether OPD improves WSI-only survival prediction and how the resulting models compare with approaches that use paired WSI–RNA data during training. We apply OPD to ABMIL (Ilse et al., 2018), TransMIL (Shao et al., 2021), PANTHER (Song et al., 2024a) and MMP-Histo (Song et al., 2024b). We also include G-HANet (Wang et al., 2025), ProSurv (Liu et al., 2025), LD-CVAE (Zhou et al., 2025), DisPro (Xu et al., 2025), and SlotSPE (Zhang et al., 2026) as existing approaches to learning with transcriptomics while retaining WSI-only inference. All baselines are trained with NLL loss. Table 1 shows that adding OPD improves the C-index of every backbone on each of the eight cohorts and the improvement extends across all four evaluated student architectures. Because OPD acts on their survival outputs and leaves their inference components unchanged, these results support ordinal distillation as a reusable form of training supervision for WSI survival models. An OPD-trained model achieves the highest reported C-index among other methods that use paired WSI–RNA data during training in seven of the eight cohorts. Specifically, MMP-Histo with OPD leads on six cohorts, while PANTHER with OPD leads on BRCA. The following sections examine the supervision target, the distillation objective, and the teacher source to identify which choices support these improvements.

4.3 What to distill: molecular features or survival outputs

Table 2 compares three forms of RNA-derived supervision for the WSI student: pathway features, learned representations, and survival outputs. Pathway matches 50 Hallmark pathway features derived from RNA expression, while RNA embedding matches an intermediate representation of the RNA teacher. The RNA teacher is fixed across comparisons that use its representations or survival outputs. Target definitions and implementation details are provided in Appendix B.1. In Table 2(a), pathway and embedding matching remain close to the no-distillation methods, with improvements in 18 and 15 of the 32 settings. OPD obtains the highest mean in each column, with improvements in all 32 settings. These results support teacher risk scores as effective supervision targets for the

Table 2: Ablations of the supervision target, objective, and teacher. Mean C-index over eight cohorts is reported. The last two columns give the mean C-index gain ($\times 100$) and wins/ties/losses against no distillation backbones. Detailed configurations and full tables are included in Appendix B.

	Mean C-index ↑				Compared with no distillation	
Variant	ABMIL	TransMIL	MMP-Histo	PANTHER	$\Delta C \times 100$	W/T/L
No distillation	0.620	0.644	0.642	0.657	–	–
(a) What to distill: supervision target						
Pathway	0.621	0.640	0.646	0.656	+0.03	18/2/12
RNA embedding	0.622	0.641	0.646	0.657	+0.09	15/1/16
(b) How to distill: distillation objective						
Survival-curve	0.622	0.646	0.644	0.659	+0.23	27/1/4
Hinton KL	0.626	0.642	0.638	0.656	-0.01	13/3/16
DiSKD	0.626	0.650	0.653	0.665	+0.82	25/1/6
Risk MSE	0.659	0.666	0.689	0.680	+3.31	30/0/2
(c) Which teacher: source of ordinal supervision						
WSI self-teacher	0.625	0.650	0.639	0.648	-0.01	16/3/13
Linear RNA	0.571	0.595	0.571	0.569	-6.38	5/0/27
Permuted RNA	0.564	0.611	0.587	0.579	-5.54	2/1/29
Multimodal MMP	0.656	0.662	0.678	0.667	+2.51	29/0/3
OPD (ours)	**0.666**	**0.670**	**0.699**	**0.687**	**+3.99**	**32/0/0**

evaluated WSI models and complement the analysis showing that molecular reconstruction need not be a prerequisite for distilling survival outputs.

4.4 How to distill: numerical matching or risk ordering

We next examine how the RNA teacher's survival outputs should supervise the WSI student, corresponding to the choice in Section 3.3. With the neural RNA teacher fixed, we compare survival-curve MSE, Hinton KL (Hinton et al., 2015), the single-event DiSKD adaptation (Deng et al., 2026a), Risk MSE, and OPD. Target definitions and implementation details are provided in Appendix B.1. Each auxiliary objective supplements the student's survival loss. Table 2(b) shows that the two scalar-risk objectives obtain higher mean C-index than the three probability-matching controls for every backbone. Risk MSE provides the closest control for ordinal transfer: it uses the same teacher and risk definition as OPD but matches separately standardized score values. OPD improves on Risk MSE by adding 0.68 C-index points on average. Its cohort mean is higher for every backbone. It directly connects the empirical gain to Section 3.3: teacher comparisons can provide useful supervision while leaving the student free from matching the teacher's numerical risk gaps.

4.5 Which teacher provides useful ordinal supervision

This experiment varies the teacher source while keeping the pairwise logistic loss fixed. We compare the default RNA teacher (OPD) with a WSI self-teacher using the student's architecture, a multimodal MMP teacher using WSI and RNA, and a linear RNA teacher mapping 50 pathway means to four hazard logits. Each teacher's survival curves are summarized as scalar risk scores using Eq. (9). A permuted-RNA control reassigns the default teacher's outputs among training patients, preserving the collection of teacher risk scores while disrupting patient correspondence. Teacher configurations are described in Appendix B.2. Table 2(c) shows that the default RNA teacher (OPD) yields the highest downstream student mean C-index for all four backbones. The multimodal teacher also improves on no-distillation backbones, while self-distillation has mixed effects. The linear RNA teacher performs below the no-distillation backbones, showing that access to RNA alone is insufficient to provide useful ordinal supervision in these runs. Permuting the neural teacher's outputs also reduces performance below the reference. Together, these controls support the use of patient-specific comparisons from a survival-trained neural RNA teacher and show how the teacher's modality, prediction model, and patient correspondence affect distillation.

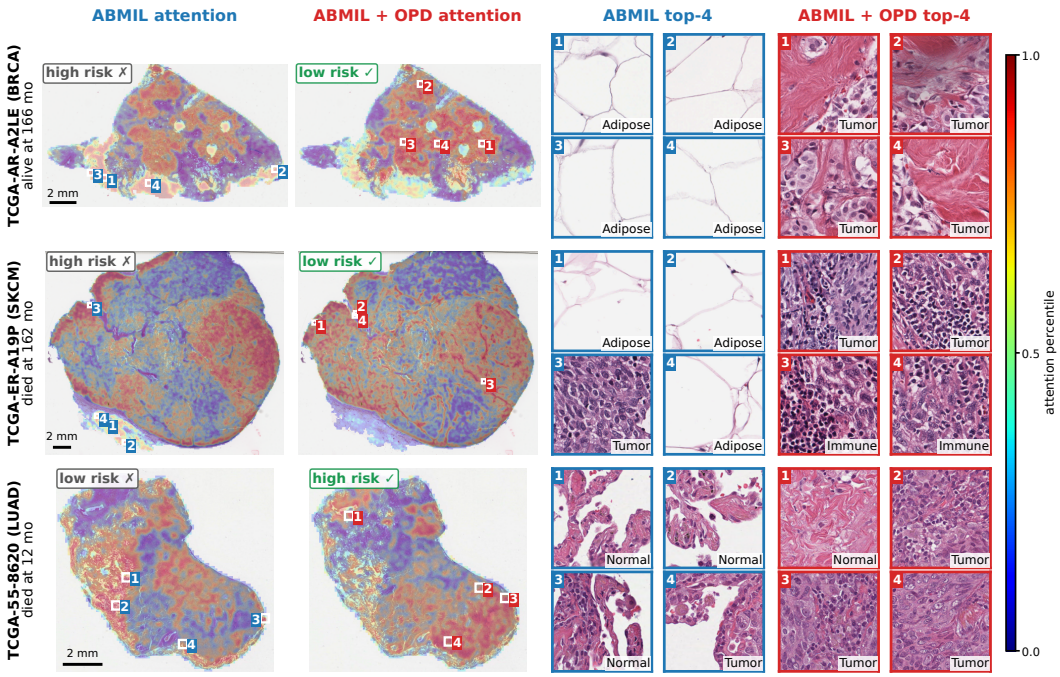

Figure 2: **OPD redirects attention to prognostically relevant tissue.** Heatmaps show attention percentiles. Numbered boxes locate each model's top-four patches with highest attention score. ✓/× indicate agreement/disagreement between the predicted risk group and the reference risk-group label. OPD shifts attention from adipose tissue to tumor (BRCA), toward lymphocyte-rich tumor regions (SKCM), and from lepidic-appearing regions to solid-appearing tumor (LUAD).

4.6 Qualitative analysis of attention

To examine how OPD affects the tissue regions emphasized by the WSI student, we visualize ABMIL and ABMIL + OPD for patients whose risk groups are misassigned by ABMIL but correctly assigned by both the RNA teacher and OPD. Figure 2 shows attention shifting from adipose tissue to tumor-containing stroma in BRCA, toward lymphocyte-rich tumor regions in SKCM, and from lepidic-appearing regions toward solid-appearing tumor in LUAD. The latter two patterns are consistent with the favorable prognosis associated with lymphocytic infiltration in melanoma (Azimi et al., 2012) and the poorer prognosis associated with solid growth in lung adenocarcinoma (Solis et al., 2012). These selected cases suggest that ordinal supervision can guide attention toward prognostically relevant morphology.

5 Conclusion

We introduced Ordinal Privileged Distillation (OPD) for survival prediction with an auxiliary modality available during training but absent at deployment. Our analysis motivates using auxiliary-modality survival outputs without reconstructing the auxiliary input and identifies when numerical hazard matching can shift the student's optimal predictions. OPD transfers pairwise risk ordering from teacher survival curves, allowing cross-modal supervision without matching absolute scores or score gaps. Combined with a censoring-aware survival loss, this supervision leaves the student's inference architecture unchanged. Experiments across eight TCGA cohorts and four WSI backbones show consistent improvements over other methods, while controlled comparisons examine the choice of supervision target, distillation objective, and teacher source. Selected attention visualizations further suggest a shift toward prognostically relevant tissue. These findings support OPD as an effective form of task-specific supervision from a modality unavailable at deployment.

6 AI USE STATEMENT

In this work, we used generative AI tools during manuscript preparation to refine the writing and improve clarity, conciseness, and readability. We have not used generative AI tools for: generate synthetic data sets, help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. We also used these tools for minor assistance with the presentation and formatting of figures and tables. All AI-assisted content was reviewed and revised by the authors, who take full responsibility for the final content, claims, and conclusions of the paper.

7 REPRODUCIBILITY STATEMENT

Section 3 specifies the survival formulation, teacher–student training procedure, and ordinal distillation objective. The assumptions and proofs supporting our theoretical analysis are provided in Appendix A. Section 4.1 describes the data sources, WSI and RNA preprocessing, cross-validation protocol, evaluation metric, and training hyperparameters. Appendix B.1 defines the supervision targets and loss functions used in the ablations, while Appendix B.2 describes the teacher variants.

REFERENCES

Farhad Azimi, Richard A Scolyer, Pavlina Rumcheva, Marc Moncrieff, Rajmohan Murali, Stanley W McCarthy, Robyn P Saw, and John F Thompson. Tumor-infiltrating lymphocyte grade is an independent predictor of sentinel lymph node status and survival in patients with cutaneous melanoma. *Journal of clinical oncology*, 30(21):2678–2683, 2012.

Kaustav Bera, Kurt A Schalper, David L Rimm, Vamsidhar Velcheti, and Anant Madabhushi. Artificial intelligence in digital pathology—new tools for diagnosis and precision oncology. *Nature reviews Clinical oncology*, 16(11):703–715, 2019.

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 *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3995–4005. IEEE, 2021.

Richard J Chen, Tong Ding, Ming Y Lu, Drew FK Williamson, Guillaume Jaume, Andrew H Song, Bowen Chen, Andrew Zhang, Daniel Shao, Muhammad Shaban, et al. Towards a general-purpose foundation model for computational pathology. *Nature medicine*, 30(3):850–862, 2024.

Feiyang Deng, Lingfeng Luo, Di Wang, Qinmengge Li, Lingxuan Kong, and Kevin He. Discrete survival knowledge distillation for competing risks analysis. In *Forty-third International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=wjsUR2vppX.

Feiyang Deng, Lingfeng Luo, Jiayu Zhou, and Kevin He. Component-wise composite likelihood distillation for censored time-to-event data. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=WgrRyHYbo5.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

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

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. *Advances in neural information processing systems*, 35:33716–33727, 2022.

Fushuo Huo, Wenchao Xu, Jingcai Guo, Haozhao Wang, and Song Guo. C 2 kd: Bridging the modality gap for cross-modal knowledge distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16006–16015. IEEE, 2024.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pp. 2127–2136. Pmlr, 2018.

Guillaume Jaume, Anurag Vaidya, Richard J Chen, Drew FK Williamson, Paul Pu Liang, and Faisal Mahmood. Modeling dense multimodal interactions between biological pathways and histology for survival prediction. In *2024 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 11579–11590. IEEE, 2024.

Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLoS medicine*, 16(1):e1002730, 2019.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. *Advances in neural information processing systems*, 30, 2017.

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

Fengchun Liu, Linghan Cai, Zhikang Wang, Zhiyuan Fan, Jin-gang Yu, Hao Chen, and Yongbing Zhang. Prototype-guided cross-modal knowledge enhancement for adaptive survival prediction. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 522–532. Springer, 2025.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. Unifying distillation and privileged information. *arXiv preprint arXiv:1511.03643*, 2015.

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.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3962–3971. IEEE, 2019.

Charles M Perou, Therese Sørlie, Michael B Eisen, Matt Van De Rijn, Stefanie S Jeffrey, Christian A Rees, Jonathan R Pollack, Douglas T Ross, Hilde Johnsen, Lars A Akslen, et al. Molecular portraits of human breast tumours. *nature*, 406(6797):747–752, 2000.

Marija Pizurica, Yuanning Zheng, Francisco Carrillo-Perez, Humaira Noor, Wei Yao, Christian Wohlfart, Antoaneta Vladimirova, Kathleen Marchal, and Olivier Gevaert. Digital profiling of gene expression from histology images with linearized attention. *Nature Communications*, 15(1): 9886, 2024.

Sashank Reddi, Rama Kumar Pasumarthi, Aditya Menon, Ankit Singh Rawat, Felix Yu, Seungyeon Kim, Andreas Veit, and Sanjiv Kumar. Rankdistil: Knowledge distillation for ranking. In *International Conference on Artificial Intelligence and Statistics*, pp. 2368–2376. PMLR, 2021.

Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, et al. A deep learning model to predict rna-seq expression of tumours from whole slide images. *Nature communications*, 11(1):3877, 2020.

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.

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

Luisa M Solis, Carmen Behrens, M Gabriela Raso, Heather Y Lin, Humam Kadara, Ping Yuan, Hector Galindo, Ximing Tang, J Jack Lee, Neda Kalhor, et al. Histologic patterns and molecular characteristics of lung adenocarcinoma associated with clinical outcome. *Cancer*, 118(11):2889–2899, 2012.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11566–11578. IEEE, 2024a.

Andrew H Song, Richard J Chen, Guillaume Jaume, Anurag J Vaidya, Alexander S Baras, and Faisal Mahmood. Multimodal prototyping for cancer survival prediction. *arXiv preprint arXiv:2407.00224*, 2024b.

Joseph A Sparano, Robert J Gray, Della F Makower, Kathleen I Pritchard, Kathy S Albain, Daniel F Hayes, Charles E Geyer Jr, Elizabeth C Dees, Matthew P Goetz, John A Olson Jr, et al. Adjuvant chemotherapy guided by a 21-gene expression assay in breast cancer. *New England Journal of Medicine*, 379(2):111–121, 2018.

Jiaxi Tang and Ke Wang. Ranking distillation: Learning compact ranking models with high performance for recommender system. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2289–2298, 2018.

Laura J Van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin Van Der Kooy, Matthew J Marton, Anke T Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536, 2002.

Yinxi Wang, Kimmo Kartasalo, Philippe Weitz, Balazs Acs, Masi Valkonen, Christer Larsson, Pekka Ruusuvuori, Johan Hartman, and Mattias Rantalainen. Predicting molecular phenotypes from histopathology images: a transcriptome-wide expression–morphology analysis in breast cancer. *Cancer research*, 81(19):5115–5126, 2021.

Zhikang Wang, Yumeng Zhang, Yingxue Xu, Seiya Imoto, Hao Chen, and Jiangning Song. Histogenomic knowledge association for cancer prognosis from histopathology whole slide images. *IEEE Transactions on Medical Imaging*, 44(5):2170–2181, 2025.

John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

Yangfan Xu, Linghan Cai, Yifeng Wang, Hailun Cheng, Fengchun Liu, Yuanjia Li, Rui Xue, Shuofeng Yuan, Yongbing Zhang, and Runming Wang. Distilling genomic knowledge into pathology slides for robust cancer survival prediction. *Computational Biomedicine*, 1(3):202527–202527, 2026.

Yingxue Xu, Fengtao Zhou, Chenyu Zhao, Yihui Wang, Can Yang, and Hao Chen. Distilled prompt learning for incomplete multimodal survival prediction. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5102–5111. IEEE, 2025.

Zihui Xue, Zhengqi Gao, Sucheng Ren, and Hang Zhao. The modality focusing hypothesis: Towards understanding crossmodal knowledge distillation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=w0QXrZ3N-s.

Yinyin Yuan, Henrik Failmezger, Oscar M Rueda, H Raza Ali, Stefan Gräf, Suet-Feung Chin, Roland F Schwarz, Christina Curtis, Mark J Dunning, Helen Bardwell, et al. Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. *Science translational medicine*, 4(157):157ra143–157ra143, 2012.

Andrew Zhang, Guillaume Jaume, Anurag Vaidya, Tong Ding, and Faisal Mahmood. Accelerating data processing and benchmarking of ai models for pathology. *arXiv preprint arXiv:2502.06750*, 2025.

Yilan Zhang, Yingxue Xu, Jianqi Chen, Fengying Xie, and Hao Chen. Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. In *International Conference on Learning Representations*, volume 2024, pp. 36854–36876, 2024.

Yilan Zhang, Li Nanbo, Changchun Yang, Jürgen Schmidhuber, and Xin Gao. Structural prognostic event modeling for multimodal cancer survival analysis. In *International Conference on Learning Representations*, volume 2026, pp. 27787–27822, 2026.

Junjie Zhou, Jiao Tang, Yingli Zuo, Peng Wan, Daoqiang Zhang, and Wei Shao. Robust multimodal survival prediction with conditional latent differentiation variational autoencoder. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10384–10393. IEEE, 2025.

APPENDIX CONTENTS

A Supporting Proofs 14
A.1 Sufficiency of the full RNA survival curve 14
A.2 Predicting survival without molecular reconstruction 14
A.3 The population target of hazard KL matching 15
B Experimental Details 16
B.1 Supervision targets and loss definitions 16
B.2 Teacher configurations 18
B.3 Sensitivity to the distillation weight 18
C Kaplan–Meier survival analysis 19

A SUPPORTING PROOFS

This appendix proves the two propositions and the reconstruction example that motivate the choices of supervision target and distillation objective. All conditional identities hold almost surely. The endpoint $T \in \{1, \dots, I_t + 1\}$ follows the interval coding in Section 3.1, including survival beyond the prediction horizon as the final category.

A.1 SUFFICIENCY OF THE FULL RNA SURVIVAL CURVE

Proof of Proposition 1. Let $U^* = \mathbf{S}^{g,*}(\mathbf{G})$ and set $S^{g,*,(0)}(\mathbf{G}) = 1$. For each $1 \le t \le I_t$, successive differences give

$$\begin{aligned} \mathbb{P}(T = t \mid \mathbf{G}) &= S^{g,*,(t-1)}(\mathbf{G}) - S^{g,*,(t)}(\mathbf{G}), \\ \mathbb{P}(T = I_t + 1 \mid \mathbf{G}) &= S^{g,*,(I_t)}(\mathbf{G}). \end{aligned} \tag{14}$$

Thus, there are measurable functions F_t such that $\mathbb{P}(T = t \mid \mathbf{G}) = F_t(U^*)$ for all possible t. Since U^* is a function of $\mathbf{G}$, the tower property gives

$$\begin{aligned} \mathbb{P}(T = t \mid U^*) &= \mathbb{E}[\mathbb{P}(T = t \mid \mathbf{G}) \mid U^*] \\ &= F_t(U^*) \\ &= \mathbb{P}(T = t \mid \mathbf{G}, U^*). \end{aligned} \tag{15}$$

These equalities establish $T \perp \mathbf{G} \mid U^*$. □

The sufficient object is the ideal full curve for the modeled endpoint. The result concerns what RNA alone predicts, and does not establish sufficiency after additionally conditioning on WSI. A learned teacher estimates this curve; neither exact recovery from WSI nor sufficiency of the scalar risk score or its pairwise ordering is assumed.

A.2 PREDICTING SURVIVAL WITHOUT MOLECULAR RECONSTRUCTION

We derive Eq. (5). Let Z have a finite second moment, let $\boldsymbol{\xi} \in \mathbb{R}^d$ have independent, zero-mean, unit-variance coordinates independent of Z, and set $\mathbf{G} = (Z, \boldsymbol{\xi})$ and $\mathbf{W} = Z$. Assume that $\mathbb{P}(T = t \mid Z, \boldsymbol{\xi}) = \mathbb{P}(T = t \mid Z)$ for every event category t.

The optimal squared-error molecular predictor is the conditional mean $\mathbb{E}[\mathbf{G} \mid \mathbf{W}] = (\mathbf{W}, \mathbf{0})$. For any measurable f with finite risk, orthogonality of the conditional-mean residual gives

$$\begin{aligned} \mathbb{E}\|\mathbf{G} - f(\mathbf{W})\|_2^2 &= \mathbb{E}\|\boldsymbol{\xi}\|_2^2 + \mathbb{E}\|(Z, \mathbf{0}) - f(Z)\|_2^2 \\ &\ge d. \end{aligned} \tag{16}$$

Equality is attained by $f(Z) = (Z, \mathbf{0})$. This is total squared reconstruction error, without averaging over molecular coordinates.

For the survival-curve target, define $u^*(z) = (\mathbb{P}(T > k \mid Z = z))_{k=1}^{I_t}$. The conditional event-law assumption and $\mathbf{W} = Z$ imply $\mathbf{S}^{g,*}(\mathbf{G}) = u^*(\mathbf{W})$ almost surely. Hence

$$\inf_u \mathbb{E}\|\mathbf{S}^{g,*}(\mathbf{G}) - u(\mathbf{W})\|_2^2 = 0. \tag{17}$$

Increasing d makes molecular reconstruction harder without changing recovery of the survival curve. This construction isolates the extra burden that irrelevant molecular variation can impose; it does not assert that every learned RNA representation retains such variation.

A.3 THE POPULATION TARGET OF HAZARD KL MATCHING

Observation convention and objective. Let C denote the discretized censoring horizon under the convention of Section 3.1. Define $A_k = \mathbf{1}\{T \geq k, C \geq k\}$ and $D_k = \mathbf{1}\{T = k\}$. When $A_k = 1$, the interval event indicator D_k is observed. With $\mathrm{bce}(d, u) = -d \log u - (1 - d) \log(1 - u)$, the per-patient survival NLL can be written as

$$\ell^p_{\mathrm{surv}} = \sum_{k=1}^{I_t} A_k \, \mathrm{bce}\Big(D_k, h^{p,(k)}(\mathbf{W})\Big). \tag{18}$$

This gives the event term and preceding survival terms for an uncensored observation, and survival terms through the last observed interval for a censored observation, as in Eq. (2). The forward Bernoulli KL is

$$\mathrm{kl}(a, u) = a \log \frac{a}{u} + (1 - a) \log \frac{1 - a}{1 - u}. \tag{19}$$

We analyze $\mathcal{R}_\gamma = \mathbb{E}[\ell^p_{\mathrm{surv}} + \gamma \mathcal{L}_{\mathrm{KL}}]$ with the same observation weights A_k on the survival loss and KL terms, a fixed teacher, and no temperature scaling.

Proof of Proposition 2. Assume $C \perp T \mid \mathbf{W}$. For an interval and input with $\rho_k(\mathbf{w}) = \mathbb{P}(A_k = 1 \mid \mathbf{W} = \mathbf{w}) > 0$, conditional independent censoring gives

$$\begin{aligned} \mathbb{E}[D_k \mid \mathbf{W} = \mathbf{w}, A_k = 1] &= \mathbb{P}(T = k \mid T \geq k, \mathbf{W} = \mathbf{w}) \\ &= h^*_k(\mathbf{w}). \end{aligned} \tag{20}$$

Fix k and $\mathbf{w}$, abbreviate $p = h^*_k(\mathbf{w})$ and $a = \mathbb{E}[h^{g,(k)}(\mathbf{G}) \mid \mathbf{W} = \mathbf{w}, A_k = 1]$, and let u be the candidate student hazard. Expanding the conditional contribution to $\mathcal{R}_\gamma$ yields

$$\begin{aligned} C_{g,k}(\mathbf{w}) + \rho_k(\mathbf{w}) \big[&- (p + \gamma a) \log u \\ &- \{1 - p + \gamma(1 - a)\} \log(1 - u) \big], \end{aligned} \tag{21}$$

where $C_{g,k}(\mathbf{w})$ collects the teacher-only entropy term and is independent of u. Set $m = (p + \gamma a)/(1 + \gamma)$. Apart from the additive constant, Eq. (21) is $\rho_k(\mathbf{w})(1 + \gamma)$ times the cross-entropy of a Bernoulli distribution with parameter m and one with parameter u. It is uniquely minimized at $u = m$, including $m = 0$ or $m = 1$ under the usual extended-value conventions.

Pointwise minimization over measurable hazards therefore gives Eq. (8). In particular,

$$h_{\gamma,k}(\mathbf{w}) - h^*_k(\mathbf{w}) = \frac{\gamma}{1 + \gamma} \big(m_k(\mathbf{w}) - h^*_k(\mathbf{w})\big). \tag{22}$$

For $\gamma > 0$, this difference vanishes almost surely on observed at-risk intervals exactly when the compatibility condition holds. Intervals with $\rho_k(\mathbf{w}) = 0$ are not identified by either term. □

Interpretation and scope. The teacher target is averaged conditional on both WSI and observed risk-set membership. These are population quantities used to characterize the objective; they are not estimated by the algorithm. An oracle RNA hazard conditions on RNA and survival to the interval, which does not generally make its observed-risk-set average equal the WSI hazard. Compatibility may nevertheless hold in a particular population, so different modalities alone do not prove incompatibility.

The proposition analyzes the untempered, single-event form of interval-wise forward KL with the same at-risk weights as the survival loss. It characterizes a population optimum over unrestricted hazards, and does not rule out finite-sample benefits from this objective or cover every variant of survival distillation. It motivates relaxing numerical agreement, without proving that teacher comparisons are always informative or that ranking must outperform KL. In the implemented method, the survival loss and pairwise logistic loss jointly affect the student's predictions. Including the survival loss therefore does not guarantee that the combined objective preserves the predictions that minimize the expected survival loss alone or the calibration of predicted probabilities.

Table 3: Configurations of the ablation variants in Table 2, listed in the same order. All variants produce WSI-only student predictions.

Variant	Configuration
No distillation	Survival NLL loss only; no auxiliary teacher supervision
(a) What to distill	
Pathway	50 Hallmark pathway features; normalized MSE
RNA embedding	Intermediate representation of the RNA teacher; normalized MSE
(b) How to distill	
Survival-curve	Four survival probabilities; MSE
Hinton KL	Conventional probability-based KL distillation
DiSKD	Observed at-risk hazard matching; single-event adaptation
Risk MSE	Separately standardized risk scores; MSE
(c) Which teacher	
WSI self-teacher	Frozen WSI-only checkpoint with the student's architecture
Linear RNA	50 pathway means $\rightarrow$ linear layer $\rightarrow$ 4 logits
Permuted RNA	Same RNA teacher as OPD; outputs permuted across training patients
Multimodal MMP	Joint WSI–RNA MMP teacher
OPD	Frozen RNA-only neural survival teacher; pairwise logistic loss

B Experimental Details

This appendix specifies the supervision targets and teacher variants in Tables 4 and 5. All variants retain WSI-only inference. Table entries are reproduced at the reported precision, and the NLL-only and OPD results provide shared references across comparisons.

B.1 Supervision targets and loss definitions

Notation and common training objective. For patient i and branch $b \in \{p, g\}$, let $\mathbf{a}_i^b \in \mathbb{R}^4$ be the hazard logits, $h_i^{b,(k)} = \sigma(a_i^{b,(k)})$, and $\mathbf{S}_i^b = (S_i^{b,(k)})_{k=1}^4$ the survival curve obtained from Eq. (1). Let $\mathbf{z}_i^p \in \mathbb{R}^{d_h}$ be the student representation preceding the survival head, $\mathbf{v}_i \in \mathbb{R}^{50}$ the Hallmark pathway features, and $\mathbf{e}_i^g \in \mathbb{R}^{144}$ the intermediate RNA teacher embedding. For batch tensors $U, V \in \mathbb{R}^{B\times d}$, define

$$\mathrm{MSE}(U, V) = \frac{1}{Bd} \sum_{i=1}^{B} \sum_{j=1}^{d} (U_{ij} - V_{ij})^2. \tag{23}$$

N_{rep} standardizes each dimension over the batch to zero mean and unit variance. Each auxiliary objective $\mathcal{D}_v$ is combined with the student's survival loss as

$$\mathcal{L}_v = \mathcal{L}_{\mathrm{surv}}^p + \lambda_v \mathcal{D}_v. \tag{24}$$

Teacher outputs and representations are detached during student optimization. Trainable projection heads are updated jointly with the student and are discarded at inference.

Pathway. A learned projection $P_{50} : \mathbb{R}^{d_h} \rightarrow \mathbb{R}^{50}$ predicts the Hallmark pathway features:

$$\mathcal{D}_{\mathrm{path}} = \mathrm{MSE}\big(\mathcal{N}_{\mathrm{rep}}(P_{50}\mathbf{z}^p), \mathcal{N}_{\mathrm{rep}}(\mathbf{v})\big). \tag{25}$$

This target is a 50-dimensional molecular summary. It does not require reconstructing the full gene-expression vector. The pathway definitions and gene-to-pathway aggregation are part of the preprocessing specification.

RNA embedding. A learned projection $P_{144} : \mathbb{R}^{d_h} \rightarrow \mathbb{R}^{144}$ matches the frozen teacher representation:

$$\mathcal{D}_{\mathrm{emb}} = \mathrm{MSE}\big(\mathcal{N}_{\mathrm{rep}}(P_{144}\mathbf{z}^p), \mathcal{N}_{\mathrm{rep}}(\mathbf{e}^g)\big). \tag{26}$$

The pathway and embedding projections have different output dimensions; neither changes the deployed WSI survival model.

Table 4: Supervision-target and objective ablation. Every auxiliary objective is added to the WSI student's survival loss. Pathway and embedding MSE use molecular or representation targets; the remaining objectives use teacher survival outputs. Risk MSE uses separately standardized risk scores. DiSKD denotes the single-event adaptation defined in Appendix B.1. NLL-only and OPD rows repeat the references in Table 1. Each cell reports the C-index with mean ± standard deviation.

Method / variant	BRCA	CRC	GBMLGG	HNSC	KIRC	LUAD	SKCM	UCEC
ABMIL								
No distillation	$0.644_{\pm 0.058}$	$0.557_{\pm 0.103}$	$0.767_{\pm 0.044}$	$0.577_{\pm 0.058}$	$0.654_{\pm 0.073}$	$0.556_{\pm 0.040}$	$0.555_{\pm 0.043}$	$0.647_{\pm 0.191}$
Pathway MSE	$0.649_{\pm 0.063}$	$0.549_{\pm 0.106}$	$0.775_{\pm 0.041}$	$0.578_{\pm 0.051}$	$0.661_{\pm 0.076}$	$0.561_{\pm 0.049}$	$0.546_{\pm 0.040}$	$0.651_{\pm 0.209}$
RNA embedding	$0.657_{\pm 0.055}$	$0.547_{\pm 0.107}$	$0.773_{\pm 0.042}$	$0.584_{\pm 0.061}$	$0.660_{\pm 0.079}$	$0.563_{\pm 0.040}$	$0.553_{\pm 0.044}$	$0.640_{\pm 0.206}$
Survival-curve	$0.651_{\pm 0.053}$	$0.556_{\pm 0.106}$	$0.772_{\pm 0.042}$	$0.579_{\pm 0.059}$	$0.655_{\pm 0.072}$	$0.559_{\pm 0.039}$	$0.557_{\pm 0.044}$	$0.650_{\pm 0.188}$
Hinton KL	$0.652_{\pm 0.050}$	$0.556_{\pm 0.102}$	$0.770_{\pm 0.041}$	$0.588_{\pm 0.063}$	$0.657_{\pm 0.080}$	$0.566_{\pm 0.036}$	$0.566_{\pm 0.053}$	$0.651_{\pm 0.187}$
DiSKD	$0.659_{\pm 0.052}$	$0.554_{\pm 0.111}$	$0.780_{\pm 0.036}$	$0.581_{\pm 0.059}$	$0.661_{\pm 0.070}$	$0.566_{\pm 0.038}$	$0.565_{\pm 0.055}$	$0.643_{\pm 0.186}$
Risk MSE	$0.662_{\pm 0.053}$	$0.639_{\pm 0.077}$	$0.780_{\pm 0.031}$	$0.630_{\pm 0.076}$	$0.671_{\pm 0.074}$	$0.585_{\pm 0.030}$	$0.612_{\pm 0.066}$	$0.693_{\pm 0.158}$
OPD	$0.665_{\pm 0.056}$	$0.640_{\pm 0.093}$	$0.780_{\pm 0.034}$	$0.636_{\pm 0.079}$	$0.676_{\pm 0.075}$	$0.591_{\pm 0.029}$	$0.623_{\pm 0.043}$	$0.713_{\pm 0.176}$
TransMIL								
No distillation	$0.676_{\pm 0.060}$	$0.639_{\pm 0.035}$	$0.764_{\pm 0.051}$	$0.612_{\pm 0.039}$	$0.661_{\pm 0.091}$	$0.579_{\pm 0.034}$	$0.583_{\pm 0.052}$	$0.634_{\pm 0.187}$
Pathway	$0.664_{\pm 0.049}$	$0.628_{\pm 0.047}$	$0.777_{\pm 0.043}$	$0.598_{\pm 0.047}$	$0.672_{\pm 0.088}$	$0.586_{\pm 0.057}$	$0.578_{\pm 0.048}$	$0.616_{\pm 0.147}$
RNA embedding	$0.683_{\pm 0.040}$	$0.616_{\pm 0.061}$	$0.769_{\pm 0.049}$	$0.590_{\pm 0.064}$	$0.669_{\pm 0.085}$	$0.575_{\pm 0.037}$	$0.578_{\pm 0.052}$	$0.644_{\pm 0.156}$
Survival-curve	$0.679_{\pm 0.061}$	$0.642_{\pm 0.033}$	$0.767_{\pm 0.048}$	$0.613_{\pm 0.036}$	$0.663_{\pm 0.089}$	$0.582_{\pm 0.035}$	$0.582_{\pm 0.051}$	$0.638_{\pm 0.187}$
Hinton KL	$0.686_{\pm 0.056}$	$0.635_{\pm 0.028}$	$0.770_{\pm 0.049}$	$0.604_{\pm 0.038}$	$0.659_{\pm 0.096}$	$0.577_{\pm 0.031}$	$0.581_{\pm 0.048}$	$0.626_{\pm 0.195}$
DiSKD	$0.694_{\pm 0.057}$	$0.634_{\pm 0.045}$	$0.783_{\pm 0.039}$	$0.612_{\pm 0.034}$	$0.671_{\pm 0.088}$	$0.589_{\pm 0.030}$	$0.581_{\pm 0.057}$	$0.636_{\pm 0.206}$
Risk MSE	$0.679_{\pm 0.043}$	$0.661_{\pm 0.026}$	$0.774_{\pm 0.041}$	$0.620_{\pm 0.047}$	$0.682_{\pm 0.074}$	$0.602_{\pm 0.050}$	$0.596_{\pm 0.048}$	$0.716_{\pm 0.129}$
OPD	$0.683_{\pm 0.048}$	$0.664_{\pm 0.024}$	$0.777_{\pm 0.042}$	$0.622_{\pm 0.041}$	$0.682_{\pm 0.076}$	$0.607_{\pm 0.039}$	$0.607_{\pm 0.043}$	$0.717_{\pm 0.149}$
MMP-Histo								
No distillation	$0.630_{\pm 0.060}$	$0.674_{\pm 0.104}$	$0.732_{\pm 0.103}$	$0.609_{\pm 0.055}$	$0.626_{\pm 0.057}$	$0.586_{\pm 0.055}$	$0.610_{\pm 0.078}$	$0.666_{\pm 0.158}$
Pathway	$0.644_{\pm 0.075}$	$0.677_{\pm 0.123}$	$0.735_{\pm 0.084}$	$0.604_{\pm 0.065}$	$0.644_{\pm 0.070}$	$0.586_{\pm 0.063}$	$0.608_{\pm 0.076}$	$0.667_{\pm 0.181}$
RNA embedding	$0.644_{\pm 0.072}$	$0.687_{\pm 0.088}$	$0.751_{\pm 0.071}$	$0.598_{\pm 0.057}$	$0.619_{\pm 0.053}$	$0.581_{\pm 0.062}$	$0.605_{\pm 0.079}$	$0.681_{\pm 0.171}$
Survival-curve	$0.632_{\pm 0.063}$	$0.672_{\pm 0.104}$	$0.736_{\pm 0.095}$	$0.612_{\pm 0.051}$	$0.631_{\pm 0.057}$	$0.591_{\pm 0.055}$	$0.611_{\pm 0.076}$	$0.663_{\pm 0.162}$
Hinton KL	$0.665_{\pm 0.050}$	$0.657_{\pm 0.094}$	$0.740_{\pm 0.083}$	$0.598_{\pm 0.060}$	$0.638_{\pm 0.071}$	$0.577_{\pm 0.028}$	$0.599_{\pm 0.089}$	$0.626_{\pm 0.146}$
DiSKD	$0.650_{\pm 0.061}$	$0.676_{\pm 0.104}$	$0.751_{\pm 0.072}$	$0.612_{\pm 0.044}$	$0.667_{\pm 0.074}$	$0.598_{\pm 0.053}$	$0.614_{\pm 0.074}$	$0.656_{\pm 0.135}$
Risk MSE	$0.666_{\pm 0.064}$	$0.689_{\pm 0.104}$	$0.780_{\pm 0.042}$	$0.637_{\pm 0.061}$	$0.710_{\pm 0.063}$	$0.610_{\pm 0.051}$	$0.643_{\pm 0.068}$	$0.778_{\pm 0.121}$
OPD	$0.683_{\pm 0.069}$	$0.699_{\pm 0.081}$	$0.783_{\pm 0.033}$	$0.640_{\pm 0.073}$	$0.713_{\pm 0.062}$	$0.630_{\pm 0.057}$	$0.651_{\pm 0.079}$	$0.796_{\pm 0.109}$
PANTHER								
No distillation	$0.683_{\pm 0.110}$	$0.680_{\pm 0.104}$	$0.764_{\pm 0.031}$	$0.586_{\pm 0.075}$	$0.683_{\pm 0.069}$	$0.619_{\pm 0.048}$	$0.618_{\pm 0.064}$	$0.623_{\pm 0.109}$
Pathway	$0.690_{\pm 0.115}$	$0.665_{\pm 0.101}$	$0.768_{\pm 0.030}$	$0.578_{\pm 0.076}$	$0.682_{\pm 0.066}$	$0.619_{\pm 0.044}$	$0.619_{\pm 0.063}$	$0.629_{\pm 0.118}$
RNA embedding	$0.696_{\pm 0.106}$	$0.676_{\pm 0.097}$	$0.767_{\pm 0.032}$	$0.586_{\pm 0.076}$	$0.682_{\pm 0.069}$	$0.617_{\pm 0.045}$	$0.617_{\pm 0.065}$	$0.614_{\pm 0.115}$
Survival-curve	$0.688_{\pm 0.111}$	$0.680_{\pm 0.102}$	$0.767_{\pm 0.031}$	$0.587_{\pm 0.075}$	$0.688_{\pm 0.067}$	$0.620_{\pm 0.046}$	$0.619_{\pm 0.061}$	$0.624_{\pm 0.117}$
Hinton KL	$0.683_{\pm 0.120}$	$0.676_{\pm 0.106}$	$0.764_{\pm 0.030}$	$0.584_{\pm 0.071}$	$0.683_{\pm 0.066}$	$0.613_{\pm 0.043}$	$0.623_{\pm 0.062}$	$0.620_{\pm 0.119}$
DiSKD	$0.699_{\pm 0.110}$	$0.681_{\pm 0.106}$	$0.776_{\pm 0.029}$	$0.585_{\pm 0.074}$	$0.697_{\pm 0.066}$	$0.623_{\pm 0.038}$	$0.622_{\pm 0.061}$	$0.640_{\pm 0.143}$
Risk MSE	$0.671_{\pm 0.073}$	$0.686_{\pm 0.092}$	$0.779_{\pm 0.037}$	$0.625_{\pm 0.059}$	$0.707_{\pm 0.064}$	$0.610_{\pm 0.039}$	$0.646_{\pm 0.032}$	$0.715_{\pm 0.219}$
OPD	$0.688_{\pm 0.081}$	$0.690_{\pm 0.104}$	$0.782_{\pm 0.031}$	$0.631_{\pm 0.048}$	$0.710_{\pm 0.054}$	$0.625_{\pm 0.054}$	$0.648_{\pm 0.026}$	$0.719_{\pm 0.213}$

Survival-curve. This objective matches the four time-point survival probabilities:

$$\mathcal{D}_{\text{curve}} = \frac{1}{4B} \sum_{i=1}^{B} \sum_{k=1}^{4} \big(S_i^{p,(k)} - S_i^{g,(k)}\big)^2. \tag{27}$$

It matches the survival curve before summarizing it as a risk score.

Risk MSE. We compute the scalar risk scores in Eq. (9) and standardize student and teacher scores separately over the batch. With standardized risk scores $\bar{r}_i^p$ and $\bar{r}_i^g$,

$$\mathcal{D}_{\text{risk}} = \frac{1}{B} \sum_{i=1}^{B} (\bar{r}_i^p - \bar{r}_i^g)^2. \tag{28}$$

This objective aligns positions and gaps on the standardized score scale. OPD uses only teacher comparison directions.

Hinton KL. This control uses conventional KL-based knowledge distillation (Hinton et al., 2015). Let $\mathbf{p}_i^{b,\tau}$ denote the normalized probability vector supplied to the KL loss after the implementation's

temperature transformation. Its generic discrepancy is

$$\mathcal{D}_{\text{Hinton}} = \text{Reduce}_i[\text{KL}(\mathbf{p}_i^{g,\tau} \| \mathbf{p}_i^{p,\tau})]\,. \tag{29}$$

Here Reduce denotes the implemented aggregation and scaling. This probability-matching control is reported separately from the survival-specific DiSKD adaptation below.

DiSKD. The single-event adaptation uses the frozen teacher hazards on observed at-risk intervals. Let A_{ik} be the interval observation indicator defined in Appendix A.3. The discrepancy is

$$\begin{aligned} \mathcal{D}_{\text{DiSKD}} &= \text{Reduce}_{i,k}\left[A_{ik}\,\text{ce}(h_i^{g,(k)}, h_i^{p,(k)})\right], \\ \text{ce}(a,b) &= -a\log b - (1-a)\log(1-b). \end{aligned} \tag{30}$$

For frozen teacher targets,

$$\text{ce}(a,b) = \text{kl}(a,b) + H_{\text{Bern}}(a), \tag{31}$$

where the teacher entropy term is constant with respect to the student. Consequently, the masked BCE objective has the same student gradient as forward Bernoulli KL with the same mask and reduction. This is the untempered, single-event adaptation of DiSKD's overall-risk objective (Deng et al., 2026a). The table label identifies this adaptation rather than a reproduction of every published DiSKD configuration.

B.2 Teacher configurations

The teacher comparison is repeated for ABMIL, TransMIL, MMP-Histo, and PANTHER. The pairwise logistic loss retains the form in Eq. (12); only the source of teacher comparisons is changed. NLL-only training provides the reference without distillation.

WSI self-distillation. The teacher uses the same WSI architecture as its student. It supplies teacher comparisons using WSI without molecular inputs.

Multimodal teacher. This variant freezes a jointly WSI–RNA MMP survival model. Its survival curve is summarized as the same scalar risk score used for all teacher comparisons. The student remains a WSI-only model; the multimodal teacher is used during training only.

Linear RNA teacher. Let $\mathbf{m}_i \in \mathbb{R}^{50}$ contain the pathway means. The teacher uses a single linear layer with bias:

$$\mathbf{a}_i^{g,\text{lin}} = M\mathbf{m}_i + \mathbf{b}, \qquad M \in \mathbb{R}^{4\times 50}, \quad \mathbf{b} \in \mathbb{R}^4. \tag{32}$$

It is trained for 20 epochs, after which its hazards and survival curve supply the teacher comparison labels. The experiment compares two RNA prediction models using the same form of ordinal supervision.

Permuted RNA outputs. A permutation π within the training set reassigns the teacher outputs across patients. Patient i retains its own WSI and outcome but receives risk $r^g_{\pi(i)}$. Its comparison labels are therefore

$$q_{ij}^{\text{perm}} = \text{sign}(r^g_{\pi(i)} - r^g_{\pi(j)}). \tag{33}$$

This operation preserves the multiset of teacher risk values in the permuted training set while changing patient-to-output correspondence. Only non-tied comparisons enter the loss.

B.3 Sensitivity to the distillation weight

We evaluate $\lambda \in \{0.1, 0.3, 1, 3, 10\}$ in the student objective $\mathcal{L}^p_{\text{surv}} + \lambda\mathcal{L}_{\text{OPD}}$. This coefficient controls the contribution of teacher-provided ordinal supervision relative to the survival loss. Table 6 reports the mean C-index over eight cohorts for each WSI backbone, with NLL-only training as the reference.

All five positive weights improve the cohort-averaged C-index of every backbone over NLL-only training. The shared setting $\lambda = 0.3$ used in the main experiments gives the highest reported mean

Table 5: Teacher-source ablation with ordinal supervision. Every numerical entry evaluates the resulting WSI student. The teacher-based rows use the pairwise logistic loss in Eq. (12); the reference without a teacher uses the survival loss alone. The linear RNA teacher maps 50 pathway means to four logits and is trained for 20 epochs. The permuted control reassigns teacher outputs within the training set. Each cell reports the C-index with mean $\pm$ standard deviation.

Method / variant	BRCA	CRC	GBMLGG	HNSC	KIRC	LUAD	SKCM	UCEC
ABMIL								
No distillation	$0.644_{\pm 0.058}$	$0.557_{\pm 0.103}$	$0.767_{\pm 0.044}$	$0.577_{\pm 0.058}$	$0.654_{\pm 0.073}$	$0.556_{\pm 0.040}$	$0.555_{\pm 0.043}$	$0.647_{\pm 0.191}$
WSI self-teacher	$0.653_{\pm 0.058}$	$0.564_{\pm 0.089}$	$0.767_{\pm 0.043}$	$0.572_{\pm 0.073}$	$0.661_{\pm 0.072}$	$0.576_{\pm 0.040}$	$0.563_{\pm 0.032}$	$0.645_{\pm 0.204}$
Multimodal MMP	$0.663_{\pm 0.068}$	$0.654_{\pm 0.095}$	$0.784_{\pm 0.036}$	$0.632_{\pm 0.071}$	$0.682_{\pm 0.076}$	$0.593_{\pm 0.028}$	$0.612_{\pm 0.054}$	$0.625_{\pm 0.136}$
Linear RNA	$0.523_{\pm 0.101}$	$0.506_{\pm 0.087}$	$0.790_{\pm 0.027}$	$0.568_{\pm 0.019}$	$0.643_{\pm 0.100}$	$0.514_{\pm 0.036}$	$0.439_{\pm 0.026}$	$0.588_{\pm 0.236}$
Permuted RNA	$0.563_{\pm 0.115}$	$0.532_{\pm 0.155}$	$0.668_{\pm 0.148}$	$0.591_{\pm 0.067}$	$0.563_{\pm 0.087}$	$0.548_{\pm 0.024}$	$0.498_{\pm 0.045}$	$0.545_{\pm 0.215}$
OPD	$0.665_{\pm 0.056}$	$0.640_{\pm 0.093}$	$0.780_{\pm 0.034}$	$0.636_{\pm 0.079}$	$0.676_{\pm 0.075}$	$0.591_{\pm 0.029}$	$0.623_{\pm 0.043}$	$0.713_{\pm 0.176}$
TransMIL								
No distillation	$0.676_{\pm 0.060}$	$0.639_{\pm 0.035}$	$0.764_{\pm 0.051}$	$0.612_{\pm 0.039}$	$0.661_{\pm 0.091}$	$0.579_{\pm 0.034}$	$0.583_{\pm 0.052}$	$0.634_{\pm 0.187}$
WSI self-teacher	$0.677_{\pm 0.057}$	$0.659_{\pm 0.031}$	$0.768_{\pm 0.049}$	$0.601_{\pm 0.041}$	$0.672_{\pm 0.083}$	$0.585_{\pm 0.046}$	$0.597_{\pm 0.040}$	$0.643_{\pm 0.165}$
Multimodal MMP	$0.685_{\pm 0.056}$	$0.675_{\pm 0.021}$	$0.777_{\pm 0.044}$	$0.628_{\pm 0.030}$	$0.682_{\pm 0.081}$	$0.601_{\pm 0.031}$	$0.592_{\pm 0.041}$	$0.655_{\pm 0.141}$
Linear RNA	$0.596_{\pm 0.067}$	$0.545_{\pm 0.076}$	$0.784_{\pm 0.033}$	$0.551_{\pm 0.049}$	$0.656_{\pm 0.083}$	$0.575_{\pm 0.049}$	$0.441_{\pm 0.018}$	$0.611_{\pm 0.194}$
Permuted RNA	$0.617_{\pm 0.078}$	$0.601_{\pm 0.145}$	$0.724_{\pm 0.080}$	$0.584_{\pm 0.063}$	$0.600_{\pm 0.101}$	$0.591_{\pm 0.075}$	$0.536_{\pm 0.039}$	$0.634_{\pm 0.094}$
OPD	$0.683_{\pm 0.048}$	$0.664_{\pm 0.024}$	$0.777_{\pm 0.042}$	$0.622_{\pm 0.041}$	$0.682_{\pm 0.076}$	$0.607_{\pm 0.039}$	$0.607_{\pm 0.043}$	$0.717_{\pm 0.149}$
MMP-Histo								
No distillation	$0.630_{\pm 0.060}$	$0.674_{\pm 0.104}$	$0.732_{\pm 0.103}$	$0.609_{\pm 0.055}$	$0.626_{\pm 0.057}$	$0.586_{\pm 0.055}$	$0.610_{\pm 0.078}$	$0.666_{\pm 0.158}$
WSI self-teacher	$0.617_{\pm 0.086}$	$0.652_{\pm 0.100}$	$0.758_{\pm 0.074}$	$0.595_{\pm 0.058}$	$0.677_{\pm 0.074}$	$0.585_{\pm 0.074}$	$0.596_{\pm 0.056}$	$0.628_{\pm 0.244}$
Multimodal MMP	$0.676_{\pm 0.079}$	$0.691_{\pm 0.110}$	$0.785_{\pm 0.050}$	$0.633_{\pm 0.061}$	$0.705_{\pm 0.074}$	$0.624_{\pm 0.049}$	$0.625_{\pm 0.055}$	$0.684_{\pm 0.146}$
Linear RNA	$0.512_{\pm 0.082}$	$0.523_{\pm 0.086}$	$0.778_{\pm 0.036}$	$0.514_{\pm 0.041}$	$0.637_{\pm 0.103}$	$0.517_{\pm 0.065}$	$0.482_{\pm 0.092}$	$0.606_{\pm 0.287}$
Permuted RNA	$0.556_{\pm 0.093}$	$0.627_{\pm 0.149}$	$0.629_{\pm 0.105}$	$0.576_{\pm 0.044}$	$0.558_{\pm 0.069}$	$0.566_{\pm 0.034}$	$0.519_{\pm 0.031}$	$0.664_{\pm 0.132}$
OPD	$0.683_{\pm 0.069}$	$0.699_{\pm 0.081}$	$0.783_{\pm 0.033}$	$0.640_{\pm 0.073}$	$0.713_{\pm 0.062}$	$0.630_{\pm 0.057}$	$0.651_{\pm 0.079}$	$0.796_{\pm 0.109}$
PANTHER								
No distillation	$0.683_{\pm 0.110}$	$0.680_{\pm 0.104}$	$0.764_{\pm 0.031}$	$0.586_{\pm 0.075}$	$0.683_{\pm 0.069}$	$0.619_{\pm 0.048}$	$0.618_{\pm 0.064}$	$0.623_{\pm 0.109}$
WSI self-teacher	$0.683_{\pm 0.106}$	$0.663_{\pm 0.117}$	$0.767_{\pm 0.037}$	$0.582_{\pm 0.068}$	$0.683_{\pm 0.076}$	$0.625_{\pm 0.057}$	$0.598_{\pm 0.069}$	$0.579_{\pm 0.171}$
Multimodal MMP	$0.692_{\pm 0.092}$	$0.684_{\pm 0.109}$	$0.790_{\pm 0.034}$	$0.608_{\pm 0.042}$	$0.715_{\pm 0.056}$	$0.611_{\pm 0.041}$	$0.633_{\pm 0.027}$	$0.602_{\pm 0.157}$
Linear RNA	$0.514_{\pm 0.077}$	$0.620_{\pm 0.076}$	$0.786_{\pm 0.036}$	$0.535_{\pm 0.057}$	$0.615_{\pm 0.089}$	$0.519_{\pm 0.044}$	$0.462_{\pm 0.035}$	$0.502_{\pm 0.150}$
Permuted RNA	$0.589_{\pm 0.068}$	$0.623_{\pm 0.139}$	$0.649_{\pm 0.101}$	$0.570_{\pm 0.071}$	$0.543_{\pm 0.054}$	$0.534_{\pm 0.030}$	$0.518_{\pm 0.027}$	$0.606_{\pm 0.093}$
OPD	$0.688_{\pm 0.081}$	$0.690_{\pm 0.104}$	$0.782_{\pm 0.031}$	$0.631_{\pm 0.048}$	$0.710_{\pm 0.054}$	$0.625_{\pm 0.054}$	$0.648_{\pm 0.026}$	$0.719_{\pm 0.213}$

Table 6: Sensitivity to the distillation weight λ. Each entry is the mean C-index over eight TCGA cohorts. The NLL-only reference ($\lambda = 0$) is reused from Table 1. Bold indicates the highest reported mean for each backbone, including ties at three-decimal precision.

λ	ABMIL	TransMIL	MMP-Histo	PANTHER
0 (NLL only)	0.620	0.644	0.642	0.657
0.1	0.654	0.660	0.692	0.677
0.3	**0.666**	**0.670**	**0.699**	**0.687**
1	**0.666**	0.667	0.693	0.685
3	0.664	0.667	0.695	0.685
10	0.663	0.666	0.692	0.686

for TransMIL, MMP-Histo, and PANTHER; for ABMIL, it ties with $\lambda = 1$ at three-decimal precision. Across $\lambda \in \{0.3, 1, 3, 10\}$, the mean C-index varies by less than 0.01 for each backbone. Increasing the weight beyond 0.3 yields no consistent additional improvement across backbones. These results show that the benefits of ordinal supervision persist across a broad range of weights, with 0.3 providing strong average performance across the evaluated students.

C Kaplan–Meier survival analysis

Figures 3–10 show Kaplan–Meier curves for the eight TCGA cohorts. In these figures, MMP denotes the WSI-only MMP-Histo backbone.

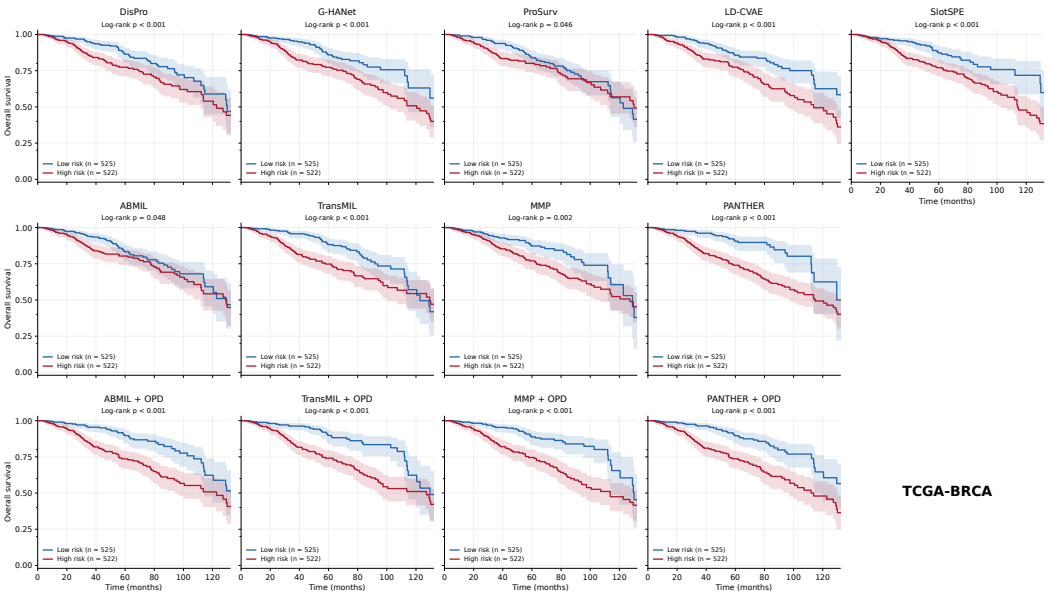

Figure 3: Kaplan–Meier risk stratification on TCGA-BRCA. Blue and red indicate low and high predicted risk, respectively. Legends report group sizes; each panel displays the log-rank p-value.

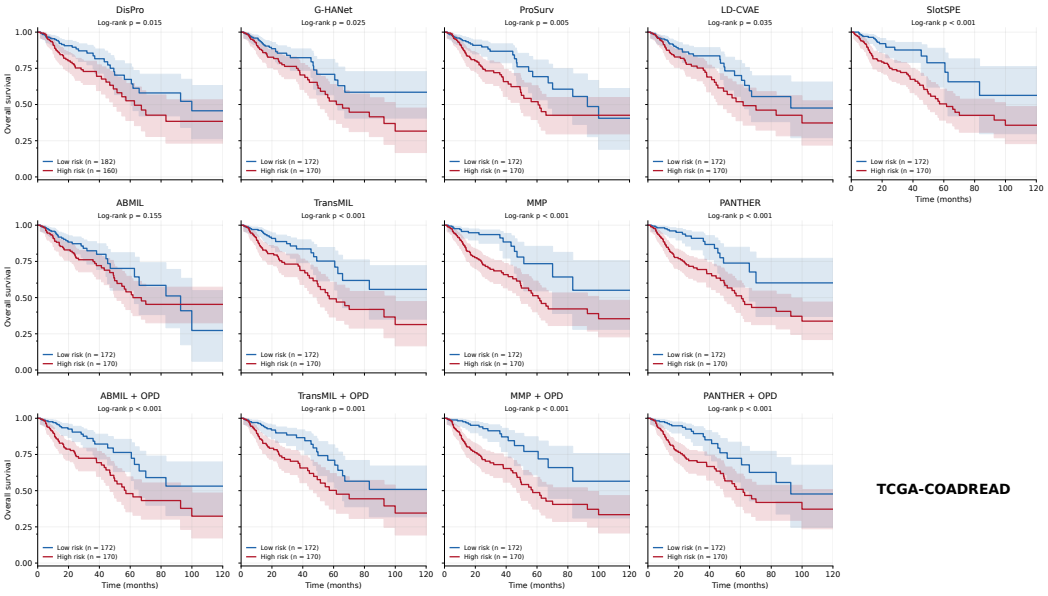

Figure 4: Kaplan–Meier risk stratification on TCGA-CRC. Blue and red indicate low and high predicted risk, respectively. Legends report group sizes; each panel displays the log-rank p-value.

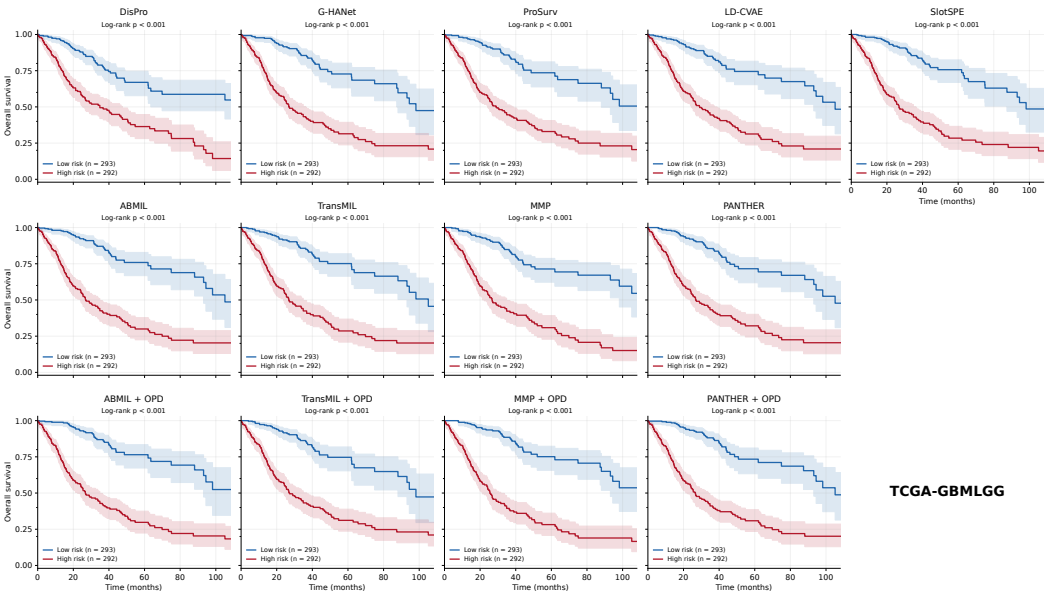

Figure 5: Kaplan–Meier risk stratification on TCGA-GBMLGG. Blue and red indicate low and high predicted risk, respectively. Legends report group sizes; each panel displays the log-rank p-value.

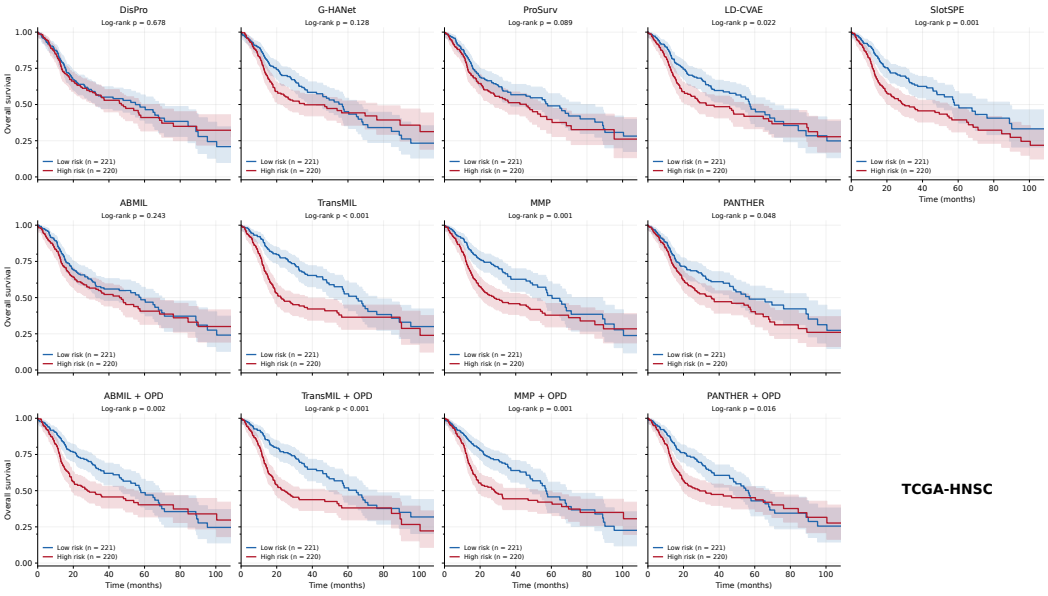

Figure 6: Kaplan–Meier risk stratification on TCGA-HNSC. Blue and red indicate low and high predicted risk, respectively. Legends report group sizes; each panel displays the log-rank p-value.

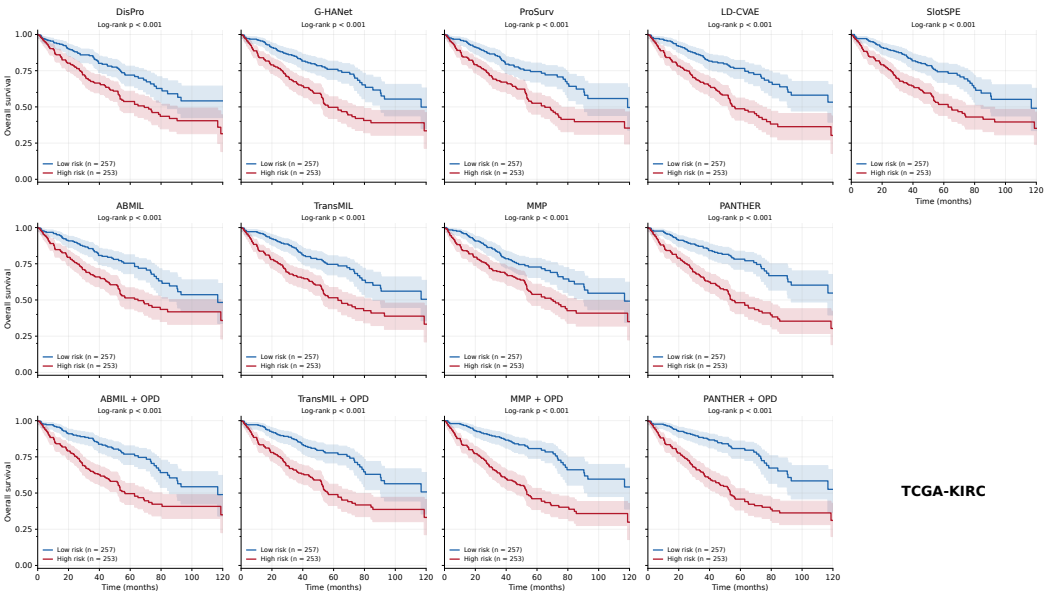

Figure 7: Kaplan–Meier risk stratification on TCGA-KIRC. Blue and red indicate low and high predicted risk, respectively. Legends report group sizes; each panel displays the log-rank p-value.

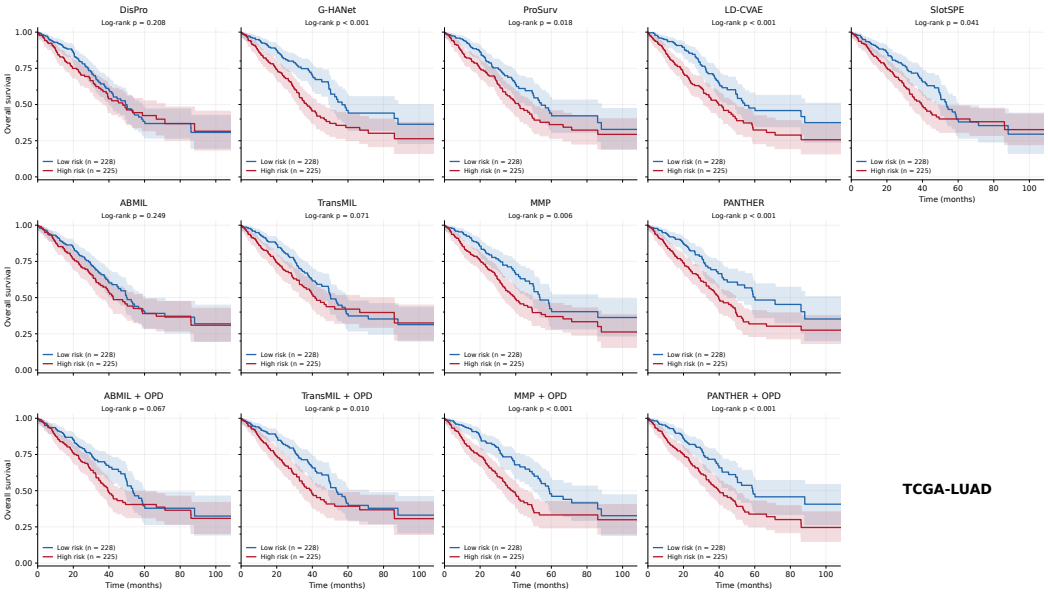

Figure 8: Kaplan–Meier risk stratification on TCGA-LUAD. Blue and red indicate low and high predicted risk, respectively. Legends report group sizes; each panel displays the log-rank p-value.

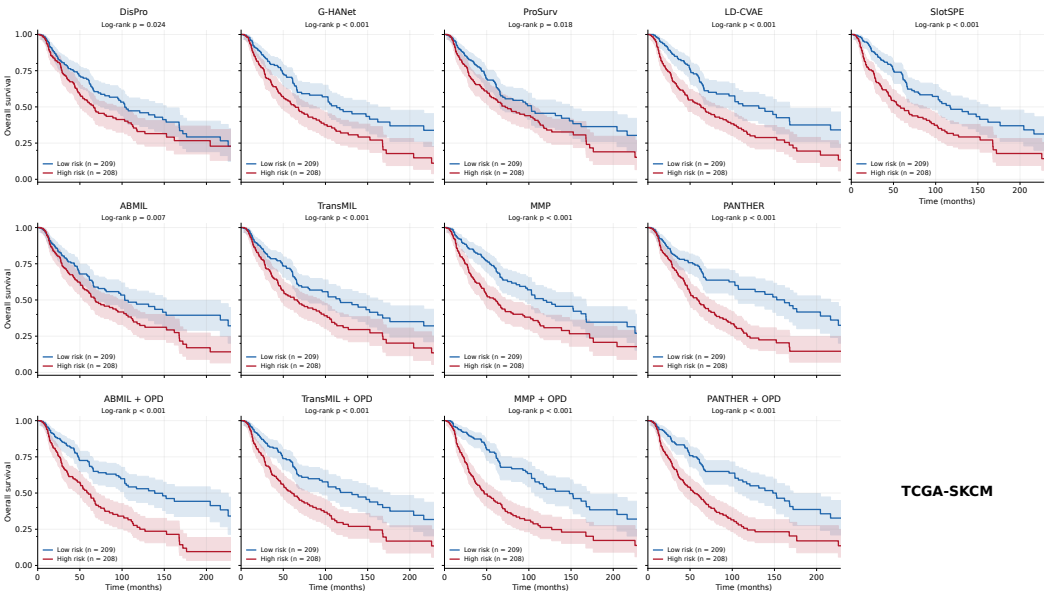

Figure 9: Kaplan–Meier risk stratification on TCGA-SKCM. Blue and red indicate low and high predicted risk, respectively. Legends report group sizes; each panel displays the log-rank p-value.

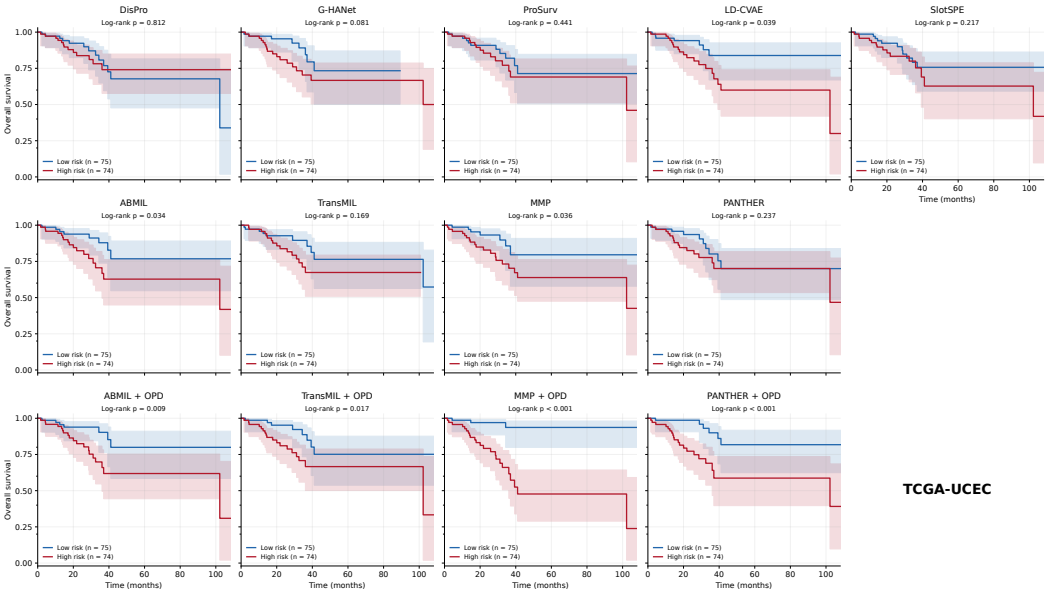

Figure 10: Kaplan–Meier risk stratification on TCGA-UCEC. Blue and red indicate low and high predicted risk, respectively. Legends report group sizes; each panel displays the log-rank p-value.