Advancing Cancer Prognosis with Hierarchical Fusion of Genomic, Proteomic and Pathology Imaging Data from a Systems Biology Perspective

Supplementary Material

6. Survival Analysis

6.1. Hazard and Survival Functions

For the k-th patient with data $I^{(k)} = (\mathbf{Y}^{(k)}, \mathbf{X}_g^{(k)}, \mathbf{X}_p^{(k)}, c^{(k)}, t^{(k)})$, where $\mathbf{Y}^{(k)}$ denotes WSI patch features, $\mathbf{X}_g^{(k)}$ and $\mathbf{X}_p^{(k)}$ represent genomic and proteomic embeddings, $c^{(k)} \in \{0, 1\}$ indicates the censorship status, and $t^{(k)} \in \mathbb{R}^+$ is the survival time in months.

The hazard function $h^{(k)}(t|I^{(k)})$ quantifies the instantaneous risk of death at time t, defined as:

$$h^{(k)}(t|I^{(k)}) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t, I^{(k)})}{\Delta t}. \tag{15}$$

The survival function $S^{(k)}(t|I^{(k)})$ represents the probability of surviving beyond time t, which can be expressed in terms of the cumulative hazard:

$$\begin{aligned} S^{(k)}(t|I^{(k)}) &= \prod_{u=1}^{t} \left(1 - h^{(k)}(u|I^{(k)})\right) \\ &= \exp\left(-\sum_{u=1}^{t} h^{(k)}(u|I^{(k)})\right). \end{aligned} \tag{16}$$

6.2. Cox Proportional Hazards Model

The Cox Proportional Hazards (CoxPH) model is a widely used approach for estimating the hazard function, in which $h(t|I^{(k)})$ is parameterized as:

$$h(t|I^{(k)}) = h_0(t) \exp(\theta^\top f(I^{(k)})), \tag{17}$$

where $h_0(t)$ is the baseline hazard function, θ is a vector of coefficients, and $f(I^{(k)})$ represents the learned multimodal representation.

6.3. Negative Log-Likelihood Loss

Following [65], we optimize our model using the negative log-likelihood (NLL) loss for discrete-time survival analysis:

$$\begin{aligned} \mathcal{L}_{\text{surv}} = -\frac{1}{N_D} \sum_{k=1}^{N_D} \Big[& c^{(k)} \log S^{(k)}(t^{(k)}|f(I^{(k)})) \\ & + (1 - c^{(k)}) \log S^{(k)}(t^{(k)} - 1|f(I^{(k)})) \\ & + (1 - c^{(k)}) \log h^{(k)}(t^{(k)}|f(I^{(k)})) \Big], \end{aligned} \tag{18}$$

where the first term accounts for observed events (uncensored cases), and the second and third terms handle censored observations by maximizing the probability of surviving up to the censoring time while penalizing the hazard at that time.

7. Additional Detailed Implementation

7.1. Molecular Feature Extraction

Genomic Data Processing. For genomic data, we employ Gene2Vec [13] to extract 200-dimensional gene identity embeddings that capture functional relationships between genes. Given the high dimensionality of gene expression profiles (approximately 30,000 genes per sample), we perform feature selection by identifying the top $N_g = 2,000$ genes with the highest variance across the entire cohort. This variance-based selection ensures that we retain the most informative genes that exhibit significant biological variability relevant to patient outcomes.

The gene expression values are first normalized using log-transformation: $\tilde{x}_i = \log_2(x_i + 1)$, where x_i represents the raw expression count for gene i. Subsequently, we apply z-score normalization across samples to standardize the expression profiles:

$$x_i^{\text{norm}} = \frac{\tilde{x}_i - \mu_i}{\sigma_i} \tag{19}$$

where μ_i and σ_i denote the mean and standard deviation of gene i across all samples.

Proteomic Data Processing. For proteomic data, we leverage GPT-5 to generate comprehensive semantic descriptions of each protein's biological function, subcellular localization, and morphological correlates in tissue histology. The prompt template used for description generation is:

> *"Describe the biological function and typical histopathological appearance of [protein name] in cancer tissue. Include information about its visual characteristics that would be observable in H&E stained tissue sections."*

These generated descriptions are then encoded using the CONCH [35] text encoder, which produces 512-dimensional protein identity embeddings. Crucially, this encoding naturally aligns the protein feature space with the histopathology image space, as CONCH is trained contrastively on paired pathology images and text descriptions.

Example Protein Descriptions:

- **4E-BP1:** A translational repressor protein that regulates mRNA translation by binding eIF4E. Its elevated expression is often associated with increased cellular proliferation and altered nuclear morphology detectable in H&E-stained tissue sections.

- **Akt:** A serine/threonine kinase involved in regulating cell survival, growth, and metabolism. Elevated Akt expression often correlates with increased cellular density and atypical morphology observable in H&E-stained tissue sections.
- **Bax:** A pro-apoptotic member of the Bcl-2 protein family that promotes programmed cell death. Higher Bax expression typically correlates with increased cellular shrinkage, nuclear condensation, and tissue architecture disruption observable in H&E-stained sections.
- **CD31:** An endothelial cell adhesion molecule (PECAM-1) that facilitates leukocyte transmigration. Higher expression correlates with increased vascular density and prominent endothelial structures visible in H&E-stained tissue sections.
- **CDK1:** A serine/threonine kinase that regulates cell cycle progression. Its elevated expression is often associated with increased mitotic activity and higher nuclear atypia observable in H&E-stained tissue sections.

7.2. Histopathological Feature Extraction

Whole Slide Image Processing. All whole slide images (WSIs) are processed at 20× magnification, corresponding to a resolution of approximately 0.5 μm/pixel. We apply an automated tissue segmentation pipeline to distinguish tissue regions from background.

Patch Extraction and Encoding. The remaining tissue regions are cropped into non-overlapping patches of size 512×512 pixels. For each WSI, this typically yields between 1,000 and 10,000 patches depending on tissue size. We employ the pre-trained CONCH [35] image encoder to extract 512-dimensional patch-level features. The CONCH encoder ensures semantic alignment with protein embeddings through its shared architecture and contrastive pretraining on pathology-specific image-text pairs.

8. Baseline Methods

8.1. Uni-modal Baselines

SNN [29]: Self-Normalizing Networks employ scaled exponential linear units (SELUs) to enable self-normalization properties, which we apply to gene and protein expression profiles for survival prediction.

ABMIL [21]: ABMIL uses an attention mechanism to aggregate patch-level features from whole slide images, enabling the model to focus on diagnostically relevant regions for survival prediction.

CLAM [34]: CLAM incorporates instance-level clustering to identify distinct morphological patterns and uses attention-based aggregation for slide-level survival prediction.

TransMIL [47]: TransMIL employs self-attention mechanisms to capture long-range dependencies among image patches, enabling more effective context-aware feature aggregation for histopathology-based survival analysis.

WiKG [31]: WiKG constructs dynamic instance graphs to model spatial relationships among tissue regions and employs graph neural networks for survival prediction from histopathological images.

8.2. Two-modality Baselines

MCAT [9]: MCAT employs cross-modal attention to learn interpretable, dense co-attention mappings between WSI features and genomic profiles within a shared embedding space.

MOTCat [60]: MOTCat employs optimal transport to establish soft correspondences between genomic data and image patches, providing global awareness to capture structural interactions within the tumor microenvironment.

CMTA [70]: CMTA introduces two parallel encoder-decoder structures to integrate intra-modal information and generate cross-modal representation.

SurvPath [23]: SurvPath proposes a memory-efficient, resolution-agnostic multimodal Transformer that integrates transcriptomic pathway tokens and histology patch tokens for patient survival prediction.

MMP [48]: MMP learns prototypical representations for each modality, with the resulting multimodal tokens processed by a fusion network using either Transformers or optimal transport-based cross-alignment.

PIBD [68]: PIBD designs a Prototypical Information Bottleneck (PIB) that models prototypes for selecting discriminative information to reduce intra-modal redundancy, while Prototypical Information Disentanglement (PID) addresses inter-modal redundancy by decoupling multimodal data into distinct components with the guidance of joint prototypical distribution.

MoME [58]: MoME designs a mixture of multimodal experts layer which enables the network to selectively focus on the information from a specific modality and utilizes the reference information in different forms across encoding stages.

8.3. Three-modality Baselines

PS3 [43]: PS3 proposes a prototype-based multimodal fusion framework that integrates cancer aggressiveness-related signals from whole slide images, pathology reports, and transcriptomic data for improved survival prediction.

ICFNet [66]: ICFNet integrates histopathology whole slide images, genomic expression profiles, and clinical textual data. It employs three distinct encoders to extract modality-specific features and leverages optimal transport algorithms to align and fuse interrelated features across modalities.

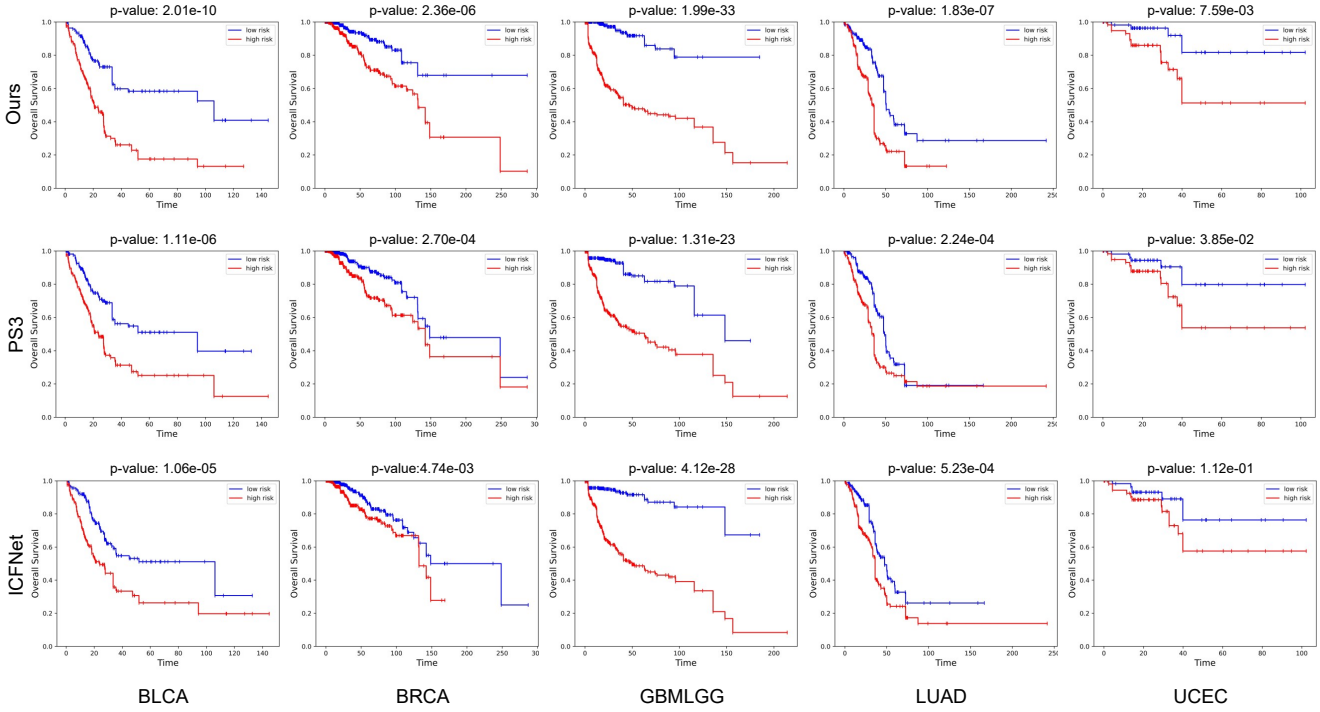

Figure 3. Kaplan-Meier Analysis of predicted high-risk (red) and low-risk (blue) groups on five cancer datasets.

Table 4. Impact of the number of selected genes (N_g) on model performance across five TCGA datasets.

N_g	BLCA	BRCA	GBMLGG	LUAD	UCEC	Overall
1000	0.711±0.033	0.709±0.049	0.862±0.056	0.650±0.042	0.779±0.074	0.742
2000	0.717±0.022	0.715±0.043	0.873±0.064	0.680±0.039	0.782±0.062	0.753
3000	0.710±0.066	0.698±0.038	0.847±0.071	0.662±0.042	0.751±0.071	0.734

To enable fair comparison on our three-modality setting (genomic, proteomic and pathology imaging data), we maintain the original architectures of PS3 and ICFNet but replace their textual modality (pathology reports or clinical text) with proteomic features.

8.4. Patient Stratification

Beyond prediction accuracy, effective patient stratification into distinct risk groups is crucial for personalized treatment planning. We evaluate stratification performance using Kaplan-Meier survival analysis, where patients are divided into high-risk and low-risk groups based on median predicted risk scores, with statistical significance assessed via the log-rank test. As presented in Fig. 3, our approach demonstrates significantly improved discrimination between the two groups and achieves the lowest p-values when compared to baseline methods. These results demonstrate that our hierarchical fusion of genomic, proteomic, and pathology imaging data yields biologically meaningful risk stratification with strong statistical power, essential for clinical decision-making.

8.5. Additional Visualization Results

To further demonstrate the hierarchical interpretability of HFGPI, we provide additional case studies in the BLCA (EGFR protein) and GBMLGG (CD31 protein) datasets, as shown in Figs. 4 and 5.

Case Study 1: EGFR in BLCA Dataset. EGFR (Epidermal Growth Factor Receptor) is a receptor tyrosine kinase that is frequently overexpressed in bladder cancer and plays a critical role in tumorigenesis through activation of downstream signaling pathways [7, 11, 44]. In Fig. 4, we identify genes associated with EGFR, including *ARAF*, *NRAS*, *FGF7* and *STAT5A*. These genes are involved in key signaling pathways that promote cell growth and tumor progression, which aligns with EGFR's role in driving cancer cell proliferation [15, 49, 52]. At the protein-to-phenotype level, representative image patches from the PGHL mod-

Table 5. Impact of the number of neighbors (k_g) in gene graph construction on model performance across five TCGA datasets.

k_g	BLCA	BRCA	GBMLGG	LUAD	UCEC	Overall
50	0.703±0.037	0.715±0.027	0.864±0.087	0.674±0.041	0.778±0.062	0.747
100	0.717±0.022	0.715±0.043	0.873±0.064	0.680±0.039	0.782±0.062	0.753
150	0.711±0.055	0.705±0.069	0.865±0.090	0.668±0.051	0.763±0.073	0.742

Table 6. Impact of the number of neighbors (k_p) in protein graph construction on model performance across five TCGA datasets.

k_p	BLCA	BRCA	GBMLGG	LUAD	UCEC	Overall
10	0.698±0.047	0.694±0.050	0.842±0.131	0.670±0.065	0.762±0.081	0.733
20	0.717±0.022	0.715±0.043	0.873±0.064	0.680±0.039	0.782±0.062	0.753
30	0.699±0.042	0.717±0.037	0.854±0.103	0.673±0.034	0.781±0.031	0.745

ule exhibit typical aggressive tumor features, including high nuclear-to-cytoplasmic ratio, prominent nucleoli, nuclear pleomorphism, solid tumor architecture with loss of normal urothelial differentiation, and increased mitotic activity. These histopathological patterns reflect the proliferative and invasive phenotype characteristic of EGFR-driven bladder cancer.

Case Study 2: CD31 in GBMLGG Dataset. CD31 (PECAM-1, Platelet Endothelial Cell Adhesion Molecule-1) is a transmembrane glycoprotein predominantly expressed on endothelial cells and serves as a critical marker for tumor-associated vasculature. Microvascular proliferation is a histopathological hallmark of glioblastomas, and CD31 expression correlates with tumor neovascularization, disease progression, and patient prognosis [39, 51]. In Fig. 5, we identify genes associated with CD31, including *ANGPT1*, *ANGPT2*, *FLT1*, and *PDGFB*. These genes are involved in angiogenesis and vascular remodeling, which aligns with CD31's role in mediating endothelial cell interactions and tumor neovascularization [8, 10, 18]. Representative image patches from the PGHL module display characteristic features of glioblastoma vasculature, including microvascular proliferation with glomeruloid tufts, multilayered endothelial cell networks, and regions of palisading necrosis surrounded by hypercellular areas. These morphological patterns are consistent with the aggressive angiogenic phenotype associated with CD31 expression in high-grade gliomas.

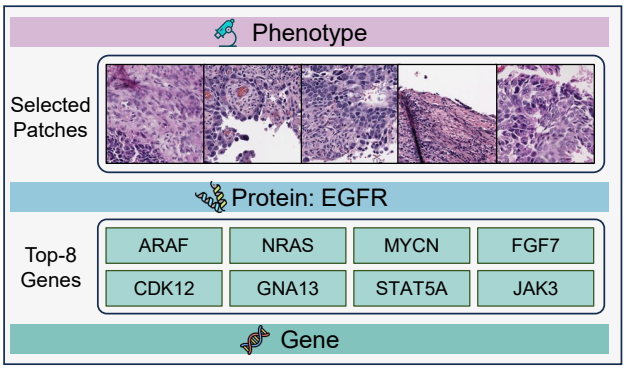

Figure 4. Visualization results on BLCA dataset using EGFR as an exemplar protein.

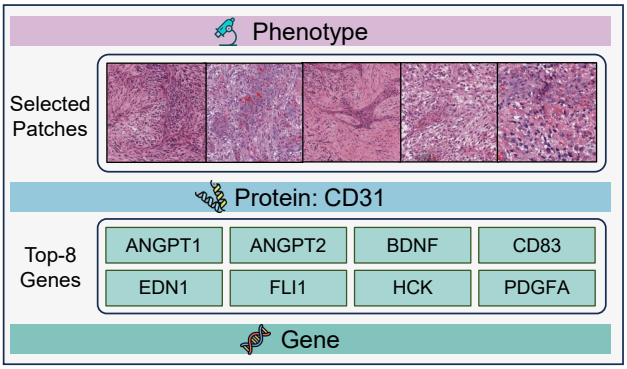

Figure 5. Visualization results on GBMLGG dataset using CD31 as an exemplar protein.

9. Additional Discussions

9.1. Number of Selected Genes

Highly variable genes capture the most informative transcriptional variations across samples. We investigate the impact of gene selection by varying $N_g \in \{1000, 2000, 3000\}$. As shown in Tab. 4, the model achieves optimal performance with $N_g = 2000$. Using fewer genes ($N_g = 1000$) leads to information loss, reducing the C-index to 0.742, while including more genes ($N_g = 3000$) introduces noisy features that do not contribute to survival prediction and slightly degrades performance. These results suggest that $N_g = 2000$ provides an

Table 7. Impact of the number of top-k patches selected per protein in PGHL on model performance across five TCGA datasets.

k	BLCA	BRCA	GBMLGG	LUAD	UCEC	Overall
8	0.694±0.031	0.684±0.039	0.871±0.047	0.665±0.038	0.730±0.069	0.729
16	0.707±0.061	0.697±0.051	0.878±0.049	0.672±0.048	0.770±0.024	0.745
32	0.717±0.022	0.715±0.043	0.873±0.064	0.680±0.039	0.782±0.062	0.753
64	0.707±0.053	0.708±0.041	0.866±0.080	0.678±0.025	0.775±0.122	0.747
128	0.703±0.055	0.704±0.051	0.864±0.071	0.683±0.024	0.760±0.110	0.743

Table 8. Impact of the regularization parameter (λ) on overall model performance (average C-index) across five TCGA datasets.

Metric	0	0.1	0.3	0.5	0.7	0.9	1.0
Average C-index	0.738	0.741	**0.753**	0.747	0.741	0.735	0.730

appropriate balance between capturing biologically relevant variations and avoiding noise.

9.2. Number of Neighbors in Gene Graph

We evaluate neighborhood sizes $k_g \in \{50, 100, 150\}$ for graph construction. As shown in Tab. 5, $k_g = 100$ yields the best performance. When $k_g = 50$, the graph is too sparse to capture sufficient functional relationships, while $k_g = 150$ introduces excessive connections that include biologically irrelevant interactions. Thus, $k_g = 100$ provides an optimal balance between capturing meaningful gene interactions and maintaining graph sparsity.

9.3. Number of Neighbors in Protein Graph

We similarly evaluate $k_p \in \{10, 20, 30\}$ for protein graph construction. Tab. 6 shows that $k_p = 20$ achieves optimal performance, with smaller values ($k_p = 10$) providing insufficient connectivity and larger values ($k_p = 30$) introducing weakly related proteins.

9.4. Number of Top-k Patches in PGHL

The PGHL module constructs hyperedges by selecting the top-k most relevant patches for each protein. We investigate the effect of hyperedge size with $k \in \{8, 16, 32, 64, 128\}$. As shown in Tab. 7, $k = 32$ achieves the best performance. Smaller values ($k \leq 16$) provide insufficient spatial coverage of protein-related morphological patterns, while larger values ($k \geq 64$) incorporate patches with weak protein-morphology associations, introducing noise. Thus, $k = 32$ optimally balances coverage of protein-relevant patches with association strength.

9.5. Effect of λ

Tab. 8 summarizes the influence of the structure-preserving regularization on model performance. Incorporating the regularization can improve the overall performance, with $\lambda = 0.3$ achieving the best balance between structural coherence and predictive accuracy. These results indicate that insufficient regularization fails to constrain the attention map, leading to noisy gene–protein associations, whereas excessive regularization forces the learned matrix $\mathbf{T}$ to overfit the network topology, thereby weakening the survival prediction.

Table 9. Comparisons of model complexity and efficiency. We report the number of parameters (MB), inference time per slide (ms), and training time (s) on the BRCA dataset.

Methods	#Params (M)	Inference Time (s / slide)	Training Time (s)
PS3	1.08	0.013	1876
ICFNet	6.18	0.104	11232
Ours	2.03	0.017	2228

9.6. Computational Complexity

To evaluate the practical efficiency of HFGPI, we compare its computational cost with state-of-the-art three-modality methods on the BRCA dataset. As shown in Tab. 9, HFGPI achieves superior efficiency compared to ICFNet while remaining comparable to PS3. Specifically, HFGPI requires only 2.03 M parameters with an inference time of 0.017 s/slide and training time of 2228 s. In contrast, ICFNet demands 6.18 M parameters with significantly higher computational costs: 0.104 s/slide for inference and 11232 s for training. These results demonstrate that HFGPI achieves superior prognostic performance while maintaining high computational efficiency, making it practical for clinical deployment.