SCOPE: SPECTRAL-CONDITIONED OMIC-PATHOLOGY ENCODER FOR INTERPRETABLE MULTIMODAL CANCER SURVIVAL ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Predicting patient survival from cancer biopsies is a multimodal learning problem in which whole-slide images (WSI) and gene expression must be fused effectively. The state of the art relies almost exclusively on flattened bag of patch tokens, discarding the 2D structure of tissue and yielding fusion mechanisms that are difficult to interpret biologically. We introduce **SCOPE** (**S**pectral-**C**onditioned **O**mic-**P**athology **E**ncoder), the first spectral-conditioned multimodal architecture for joint pathology-genomics survival modeling. The core primitive is the Gene-Modulated Convolution: a global 2D convolution whose kernel is generated with FiLM modulations driven by pathway-level gene embeddings. Stacking these modules across a multi-scale Haar wavelet pyramid delivers a global receptive field at each spatial-frequency band at log-linear cost. A mask-aware differentiable Top-K predictor then selects the most prognostic spatial positions for the discrete-time negative-log-likelihood survival head. On 10 TCGA cohorts, SCOPE raises the mean C-index to 0.744 ($+3.3\%$ relative) over the strongest baseline while using $1.9\times$ lower peak memory and running $1.6\times$ faster, and it remains the top method in both the pathology-only and missing-omics settings. SCOPE further exposes three complementary axes of intrinsic interpretability (Top-K spatial saliency, wavelet sub-band allocation, and pathway attention) that bridge molecular biology and observable tissue architecture. Our code/models will be publicly available.

1 INTRODUCTION

Predicting patient survival from cancer biopsies has emerged as an important multimodal learning problem in computational pathology (Lu et al., 2021; Chen et al., 2022b; Jaume et al., 2024). Each patient is characterised by two complementary views: a gigapixel whole-slide image (WSI) capturing tissue morphology, and high-throughput omics data, which includes, for instance, transcriptomic profiles that summarize gene expression at the pathway level. Strong empirical evidence shows that the two modalities carry complementary information. Morphology reflects emergent tissue organisation, while omics data reveals the underlying molecular programs that drive it, and joint modeling has been shown to improve prognostic accuracy (Chen et al., 2021; Xu & Chen, 2023).

The state of the art (SOTA) in omic-pathology survival modeling is built almost exclusively around self-attention between a flattened bag of WSI patch tokens and a sequence of pathway tokens (Song et al., 2024; Zhang et al., 2026). This paradigm has three structural blind spots that limit further progress: (i) The global 2D structure of tissue is discarded. A WSI is not a permutation-invariant set of patches; tumour-stromal interfaces, infiltrating-lymphocyte gradients, and glandular architecture are spatial signals that the bag-of-tokens treatment cannot exploit. (ii) The fusion mechanism is implicit and weakly interpretable. A self-attention map shows which patches a given pathway attends to, but does not reveal what morphological computations the genome induces; the visual processing path itself remains gene-agnostic. (iii) Missing-modality robustness is fragile. Self-attention architectures collapse when one modality is absent at test time, requiring auxiliary reconstruction losses, dual encoders, or distinct inference graphs (Cui et al., 2022; Zhou et al., 2025).

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.

A deeper gap is more conceptual: tissue morphology has an intrinsic multi-scale spectral character: fine-grained nuclear texture, mid-scale glandular structure, and slide-level architectural patterns coexist in the same image, and yet **no** existing omic-pathology model exploits spectral-conditioned priors. Spectral-domain operators have recently dominated long-context modeling in language and DNA via implicit-filter parameterisations such as Hyena (Poli et al., 2023; Nguyen et al., 2023) and SPU (Liu et al., 2026a), but the generalisation of these ideas to pathology has not been attempted.

In this work we propose **SCOPE** (**S**pectral-**C**onditioned **O**mic-**P**athology **E**ncoder), a multimodal architecture that fuses WSI and gene expression by letting the latter directly parameterise the convolutional filters that process the former. SCOPE is built based on the Gene-Modulated Convolution (GM-Conv), whose kernel is generated by an implicit MLP from a learnable positional embedding, with Gene-FiLM (FiLM (Perez et al., 2018) modulation predicted from a global pathway embedding) injected at every hidden layer of that MLP. Because the FiLM heads are zero-initialised, the same forward pass cleanly degrades to a pathway-agnostic global convolution when omics is unavailable, yielding native missing-modality robustness without any auxiliary objective. The 2D WSI feature grid is decomposed by a J-level Haar wavelet transform, each subband is processed by its own GM-Conv evaluated in the Fourier domain, and the outputs are recombined via inverse DWT plus a full-resolution residual. This grants the model a global receptive field at every scale while keeping cost at log-linear per layer; Appendix A proves this cost bound and shows it is uniform in J. A mask-aware differentiable Top-K gate selects the most prognostic spatial positions per slide, and the survival head is trained directly on this sparse selection.

SCOPE exposes three complementary axes of interpretability. *(a) Spatial: Top-K Aware Predictor.* The hard-mask emitted by the differentiable Top-K gate is a per-slide saliency map that tells the clinician precisely which tissue regions drove the risk score. *(b) Spectral: wavelet sub-band allocation.* Because the operator filters each spatial-frequency band separately, the energy a patient's representation places in the coarse versus the detail bands is directly readable, and we find that this allocation shifts systematically with predicted risk. *(c) Molecular: pathway attention.* The gated pooling weights of the pathway encoder rank which gene programmes a given patient's conditioning rests on, and how much each programme varies across a cohort, providing a concrete bridge between molecular biology and observable tissue architecture.

Our main contributions are threefold: (1) We introduce SCOPE, the first spectral-conditioned architecture for joint omic-pathology survival analysis. It delivers global receptive fields across scales at log-linear cost. (2) SCOPE provides intrinsic spatial, spectral and molecular interpretability, providing a concrete bridge between molecular biology and observable tissue architecture. (3) On 10 TCGA cohorts, SCOPE beats SOTAs under multi-modal, pathology-only, and missing-omics settings, while running $1.9\times$ leaner and $1.6\times$ faster than the strongest baseline.

2 Related Work

Multimodal Omic-Pathology Survival Modeling. Earlier multimodal survival models concatenated patch-level WSI features with bulk gene expression and trained an MLP head on the union (Mobadersany et al., 2018; Chen et al., 2022a). The current generation is dominated by self-attention fusion: MCAT (Chen et al., 2021) introduced co-attention between WSI patches and gene-pathway tokens. MMP (Song et al., 2024) extended this to pathway hierarchies. Porpoise (Chen et al., 2022b) and MOTCat (Xu & Chen, 2023) refined the architecture along the same axis. Slot-SPE (Zhang et al., 2026) replaces dense cross-attention with sparse slot attention (Locatello et al., 2020), modeling the multimodal interaction as a small set of patient-specific structural prognostic events. All of these methods, however, share the same structural choices: the WSI is treated as a permutation-invariant bag of patch tokens, the fusion mechanism is some flavour of soft attention, and missing-modality inference is an ad hoc extension of the architecture. SCOPE departs from this paradigm on all three axes simultaneously.

Long-Range and Spectral-Domain Sequence Operators. Implicit-filter architectures (Poli et al., 2023; Romero et al., 2022) have emerged as a competitive alternative to attention for long sequences, parameterising the convolutional kernel as the output of a position-MLP and realising the convolution via FFT for $O(N \log N)$ complexity (Poli et al., 2023; Massaroli et al., 2023). State-space models such as S4 (Gu et al., 2022) and Mamba (Gu & Dao, 2024) share the long-range philosophy. Wavelet-augmented variants such as SPU (Liu et al., 2026a) additionally decompose the input across

spatial-frequency scales before applying implicit filters at each band. A common limitation across this entire family is that the operators have been studied almost exclusively for 1D, unimodal signals (DNA, text, audio, time series). To our knowledge, SCOPE is the first to generalise this design to a 2D multimodal setting, and the first to introduce external-modality conditioning into the implicit filter itself.

Interpretability in Computational Pathology. Saliency in WSI analysis is typically obtained post hoc by visualising attention weights of an attention-based MIL model (Ilse et al., 2018). While useful, attention maps are known to be unfaithful surrogates for feature importance (Wiegreffe & Pinter, 2019). Differentiable Top-K gates (Jang et al., 2017; Xie et al., 2020) provide a trained sparse selection that is committed to at inference time. SCOPE places such a gate at the survival head level, yielding spatial saliency that is consistent with what the model actually uses to make a prediction. On the molecular side, prior works have visualised attention-weighted pathway importance scores (Jaume et al., 2024). SCOPE provides additional spectral interpretability: because each spatial-frequency band is filtered separately, the energy a patient's representation allocates across the wavelet sub-bands is directly measurable. SCOPE's three complementary axes of interpretability provide a concrete bridge between molecular biology and observable tissue architecture.

3 METHOD

3.1 PROBLEM SETUP

We formulate survival analysis as a discrete-time prediction problem. For each patient i, the dataset consists of a tuple $(\mathcal{X}_i, \mathcal{G}_i, t_i, \delta_i)$, where $\mathcal{X}_i$ is the set of whole-slide images (WSIs), $\mathcal{G}_i$ is the bulk gene expression data, t_i is the observed event time, and $\delta_i \in \{0, 1\}$ is the event indicator (1 if observed, 0 if right-censored). Following standard practice (Zadeh & Schmid, 2021), we discretize the continuous time into $C = 4$ bins based on the quantiles of the uncensored event distribution, transforming the target into a discrete label $y_i \in \{1, \ldots, C\}$.

WSI Representation. We process each WSI into a sparse 2D feature grid. Initially, patches are extracted at $20\times$ magnification (256×256 pixels) using the CLAM pipeline (Lu et al., 2021). Each patch is then embedded into a feature vector using a frozen, pretrained patch encoder; our primary encoder is CTransPath (Wang et al., 2022). To create a fixed-size grid representation, we map the patch features to a uniform $H \times W$ grid. Each grid cell's feature is computed via adaptive average pooling of all patch embeddings that fall within its spatial extent. The fraction of tissue-containing patches within each cell is retained as a density channel. For a patient with S_i slides, this process yields a set of feature tensors. After batching and padding, the final WSI input has the shape $\mathbf{X} \in \mathbb{R}^{B\times S\times(d_{\text{in}}+1)\times H\times W}$, where B is the batch size, S is the maximum number of slides in the batch and d_{in} is the patch feature dimension.

Omics Representation. We process bulk RNA-seq data by grouping gene expression values into M pre-defined pathway signatures. Following Jaume et al. (2024), we extract pathways from two resources: Reactome (Gillespie et al., 2022) and the MSigDB Hallmarks collection (Liberzon et al., 2015), comprising 1,281 and 50 human biological pathways respectively, and retain only pathways for which at least 90% of the constituent genes are available in the transcriptomics. This yields $M = 331$ pathways covering 4,999 unique genes (281 Reactome pathways from 1,577 genes and 50 Hallmark pathways from 4,241 genes). Each patient is thus represented as a set of variable-length vectors $\{\mathbf{g}_q \in \mathbb{R}^{G_q}\}_{q=1}^{M}$, where G_q is the number of genes in pathway q.

3.2 OVERVIEW OF SCOPE

The SCOPE model is composed of three main components, as illustrated in Figure 1. First, a Pathway Encoder processes pathway-level gene expression signatures into a single, patient-level conditioning vector $\mathbf{c} \in \mathbb{R}^{D_c}$. Second, the SCOPE Encoder, a stack of L custom transformer-style blocks, processes the 2D WSI feature grid. The core of each block is a novel SCOPE Operator that performs multi-scale, globally-aware filtering. Third, a Top-K Aware Predictor identifies the most salient spatial regions within each slide, computes slide-level predictions, and aggregates them into a final patient-level survival prediction via an inter-slide attention mechanism. A central design

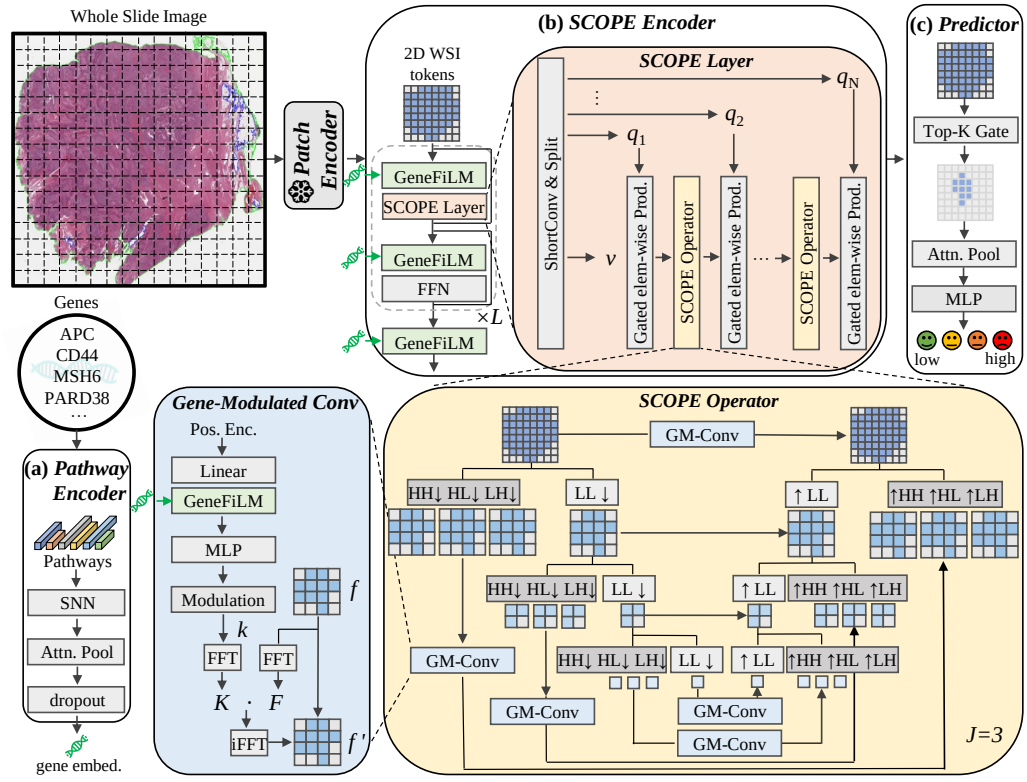


Figure 1: Overview of SCOPE. The core of the encoder is a multi-scale global convolution operator whose kernels are dynamically generated based on the patient's omics profile. The resulting deep features are then processed by a Top-K aware predictor to produce the final survival prediction.

principle of SCOPE is that omics information exclusively influences the model by modulating the WSI feature extraction process via a conditional normalization mechanism we term Gene-FiLM.

Pathway Encoding and Conditioning. For each patient, we first encode the M pathway expression vectors $\{\mathbf{g}_q \in \mathbb{R}^{G_q}\}_{q=1}^{M}$ into a sequence of fixed-dimensional embeddings $\{\mathbf{e}_q \in \mathbb{R}^{D_e}\}_{q=1}^{M}$ using independent Self-Normalizing Networks (SNNs) (Klambauer et al., 2017). These embeddings are then aggregated into the single patient-level conditioning vector $\mathbf{c}$ using a standard gated attention mechanism:

$$\mathbf{c} = \sum_{q=1}^{M} \alpha_q \mathbf{e}_q, \quad \text{where} \quad \alpha_q \propto \exp\left(\mathbf{w}_a^\top (\tanh(\mathbf{W}_a \mathbf{e}_q) \odot \sigma(\mathbf{V}_a \mathbf{e}_q))\right). \tag{1}$$

Here, $\mathbf{W}_a, \mathbf{V}_a$, and $\mathbf{w}_a$ are learnable parameters of the attention network. We set the embedding dimension D_e and conditioning dimension D_c to 256 in all experiments.

Gene-Modulated Convolution (GM-Conv). The fundamental building block of our architecture is the GM-Conv. Given a WSI feature grid $\mathbf{U} \in \mathbb{R}^{B\times D\times H\times W}$ and the conditioning vector $\mathbf{c}$, it computes $\mathbf{U}' = (\mathbf{k_c} * \mathbf{U}) + \boldsymbol{\rho} \odot \mathbf{U}$, where $\boldsymbol{\rho} \in \mathbb{R}^D$ is a learnable per-channel skip gain. The convolutional kernel $\mathbf{k_c} \in \mathbb{R}^{B\times D\times H\times W}$ is dynamically generated for each patient. Specifically, $\mathbf{k_c}$ is the output of a small implicit MLP whose inputs are a learnable 2D positional embedding and the conditioning vector $\mathbf{c}$. This modulation is achieved via Gene-FiLM, a conditional normalization layer that acts on the MLP's hidden activations $\mathbf{a}$:

$$\text{GeneFiLM}_\mathbf{c}(\mathbf{a}) = \big(\mathbf{1} + \boldsymbol{\gamma}(\mathbf{c})\big) \odot \frac{\mathbf{a} - \mu(\mathbf{a})}{\sqrt{\sigma^2(\mathbf{a}) + \epsilon}} + \boldsymbol{\beta}(\mathbf{c}), \tag{2}$$

The affine parameters $\boldsymbol{\gamma}, \boldsymbol{\beta}$ are predicted from $\mathbf{c}$ and are zero-initialized. This structural choice is critical: when $\mathbf{c} = \mathbf{0}$ (e.g., during conditional dropout or at test-time with missing omics), Gene-FiLM defaults to a parameter-free standardisation, providing inherent robustness without auxiliary

losses. The convolution $\mathbf{k_c} * \mathbf{U}$ is efficiently computed in the frequency domain using the Fast Fourier Transform (FFT) to achieve a global receptive field.

The SCOPE Encoder. The full encoder stacks L residual blocks. Each block sequentially processes the input $\mathbf{U}$ through a SCOPE operator layer and a SwiGLU feed-forward network (FFN), with both sub-layers preceded by Gene-FiLM normalization. The SCOPE operator itself, detailed in Section 3.3, is wrapped in a gated recurrence structure, enabling it to model complex, long-range dependencies within the feature map.

Prediction and Aggregation. The encoder's output $\mathbf{Z} \in \mathbb{R}^{BS\times D\times H\times W}$ is passed to the Top-K Aware Predictor. For each slide, a scoring MLP assigns an importance score to every spatial position. A differentiable Top-K operator (Xie et al., 2020) then retains the K highest-scoring positions. These features are weighted-averaged to produce a slide-level feature $\mathbf{z}_s$, from which slide-level logits $\boldsymbol{\ell}_s$ are predicted. Finally, for patients with multiple slides, we apply a masked attention mechanism over the set of slide-level logits $\{\boldsymbol{\ell}_s\}$ to compute the final patient-level logits $\boldsymbol{\ell}_i$, which are then converted to survival probabilities via a softmax function.

3.3 The SCOPE Operator

A single GM-Conv mixes information globally but operates at a single scale. To capture morphology across a range of spatial frequencies, we embed the operator within a J-level 2D Haar wavelet pyramid (Daubechies, 1990). Let DWT and IDWT be the 2D Haar analysis and synthesis operators, respectively, with subbands $\{\mathrm{LL}, \mathrm{LH}, \mathrm{HL}, \mathrm{HH}\}$. The SCOPE operator processes an input feature map $\mathbf{U} \in \mathbb{R}^{B\times D\times H\times W}$ conditioned on an omics embedding $\mathbf{c}$.

Decomposition and Filtering. The operator first computes a full-resolution residual $\mathbf{U}_b = \boldsymbol{\eta} \odot \mathrm{GMConv}_b(\mathbf{U}, \mathbf{c})$, where $\boldsymbol{\eta} \in \mathbb{R}^D$ is a learnable channel-wise scale. Concurrently, it performs a multi-level decomposition by setting $\mathrm{LL}_0 = \mathbf{U}$ and iteratively computing $(\mathrm{LL}_j, \mathrm{LH}_j, \mathrm{HL}_j, \mathrm{HH}_j) = \mathrm{DWT}(\mathrm{LL}_{j-1})$ for levels $j = 1, \ldots, J$. Every subband is then filtered by its *own* gene-modulated convolution: writing $\mathrm{GMConv}_j^{\mathrm{B}}$ for the filter assigned to band $\mathrm{B} \in \{\mathrm{LH}, \mathrm{HL}, \mathrm{HH}\}$ at level j,

$$\widehat{\mathrm{B}}_j = \mathrm{GMConv}_j^{\mathrm{B}}\big(\mathrm{B}_j, \mathbf{c}\big) \quad \text{for } \mathrm{B} \in \{\mathrm{LH}, \mathrm{HL}, \mathrm{HH}\}, \qquad \widehat{\mathrm{LL}}_J = \mathrm{GMConv}_J^{\mathrm{LL}}(\mathrm{LL}_J, \mathbf{c}), \quad (3)$$

where the low-frequency convolution is applied only at the deepest level J; the intermediate LL_j for $j < J$ are not filtered themselves, since they are further decomposed at the next level.

Recombination with Learned Mixing. The filtered subbands are reconstructed level-by-level from $j = J$ down to 1. Writing $\widetilde{\mathrm{LL}}_J = \mathbf{0}$ and $\mathrm{LL}_j^\star = \widehat{\mathrm{LL}}_J$ for $j = J$ but $\mathrm{LL}_j^\star = \mathrm{LL}_j$ for $j < J$, then

$$\widetilde{\mathrm{LL}}_{j-1} = \mathrm{IDWT}\big(\mathrm{LL}_j^\star + \boldsymbol{\nu}_j \odot \widetilde{\mathrm{LL}}_j,\ \widehat{\mathrm{LH}}_j, \widehat{\mathrm{HL}}_j, \widehat{\mathrm{HH}}_j\big), \quad (4)$$

where $\boldsymbol{\nu}_j \in \mathbb{R}^D$ is a learnable cross-scale mixing parameter. The final output is the sum of the recombined wavelet path and the base residual path: $\widetilde{\mathrm{LL}}_0 + \mathbf{U}_b$.

Design Rationale. Each subband possesses an independent GM-Conv, allowing the same gene embedding to shape filtering differently at each scale, a property we leverage for interpretability in Section 4.4. The zero initializations of $\boldsymbol{\eta}$ and $\{\boldsymbol{\nu}_j\}$ ensure the operator begins as the identity mapping passed through a lossless wavelet pipeline, for every depth J and independently of the omics condition (Proposition 3), with cross-scale and cross-band interactions learned gradually. This spectral-domain approach uniquely connects three key properties: (1) a *global receptive field* per layer; (2) a *natural multi-scale decomposition* via the wavelet pyramid; and (3) a *clean interface for cross-modal conditioning*. Appendix A makes these properties precise: the wavelet analysis is an exact isometry, so depth never discards signal (Proposition 1), and the whole J-level pyramid costs the same as two single-scale global convolutions regardless of J (Proposition 2).

3.4 Training Objective

Our primary training objective is to minimize the discrete-time negative log-likelihood (NLL) for survival analysis, following Zadeh & Schmid (2021):

$$\mathcal{L}_{\mathrm{NLL}} = -\frac{1}{N_{\mathrm{pat}}} \sum_{i=1}^{N_{\mathrm{pat}}} \Big[\delta_i \log P(Y_i = y_i \mid \boldsymbol{\ell}_i) + (1 - \delta_i) \log P(Y_i > y_i \mid \boldsymbol{\ell}_i)\Big], \quad (5)$$

Table 1: Main results on 10 TCGA cohorts of C-index (mean ± std) under multi-modal, pathology-only, and missing-omics settings. Best per cohort in **bold**; second-best underline.

Method	BLCA	BRCA	COADREAD	HNSC	KIRC	LUAD	LUSC	SKCM	STAD	UCEC
Pathology + Transcriptome										
MCAT	0.688±.024	0.709±.052	0.737±.033	0.622±.049	0.795±.018	0.659±.054	0.583±.051	0.667±.049	0.639±.032	0.802±.060
Porpoise	0.690±.025	0.651±.046	0.699±.050	0.599±.048	0.788±.028	0.647±.091	0.597±.050	0.682±.021	0.619±.022	0.788±.092
CMTA	0.702±.033	0.731±.066	0.720±.008	0.619±.051	0.794±.018	0.662±.053	0.599±.071	0.681±.036	0.639±.038	0.764±.069
MOTCat	0.700±.010	0.717±.064	0.742±.042	0.627±.049	0.799±.013	0.663±.052	0.573±.026	0.688±.054	0.642±.019	0.773±.060
SurvPath	0.666±.034	0.728±.040	0.727±.024	0.611±.024	0.787±.025	0.661±.050	0.577±.024	0.666±.054	0.647±.027	0.751±.084
PIBD	0.693±.032	0.662±.054	0.716±.054	0.627±.069	0.772±.007	0.634±.065	0.575±.098	0.649±.048	0.644±.038	0.783±.053
MMP	0.686±.004	0.714±.050	0.686±.016	0.580±.049	0.770±.011	0.634±.048	0.580±.043	0.682±.029	0.602±.020	0.755±.073
LD-CVAE	0.635±.037	0.705±.050	0.753±.029	0.635±.051	0.792±.025	0.651±.055	0.620±.045	0.682±.037	0.653±.036	0.788±.068
SlotSPE	0.708±.025	**0.779±.061**	**0.773±.042**	0.642±.035	0.815±.017	0.683±.058	0.634±.079	0.688±.051	0.671±.028	**0.813±.059**
SCOPE	**0.747±.025**	0.778±.066	0.754±.047	**0.690±.058**	**0.845±.029**	**0.709±.052**	**0.641±.059**	**0.722±.025**	**0.745±.067**	0.809±.062
Pathology-Only										
ABMIL	**0.663±.072**	0.656±.037	0.715±.029	0.664±.043	0.782±.026	0.645±.061	0.593±.055	**0.668±.042**	0.642±.044	0.779±.057
TransMIL	0.647±.032	0.647±.032	0.723±.027	0.644±.020	0.767±.035	0.655±.052	0.582±.039	0.632±.050	0.656±.038	0.759±.059
CLAM-SB	0.645±.052	0.696±.049	0.742±.046	0.649±.042	0.784±.033	0.639±.059	0.584±.065	0.643±.049	0.637±.022	0.757±.072
CLAM-MB	0.657±.029	0.715±.034	0.721±.041	0.636±.038	0.782±.014	0.632±.070	0.608±.056	0.640±.032	0.641±.032	**0.788±.074**
SlotSPE	0.642±.035	0.738±.019	0.729±.038	0.635±.038	0.790±.023	0.656±.038	0.598±.054	0.659±.063	0.676±.033	0.779±.057
SCOPE	0.660±.049	**0.742±.023**	**0.745±.036**	**0.666±.036**	**0.799±.019**	**0.668±.042**	**0.612±.049**	**0.668±.043**	**0.678±.028**	0.786±.056
Missing-Omics										
CMTA	0.558±.054	0.604±.112	0.683±.065	0.564±.074	0.662±.045	0.593±.051	0.541±.045	0.541±.048	0.608±.065	0.676±.124
MOTCat	0.565±.098	0.600±.077	0.645±.028	0.578±.030	0.738±.034	0.643±.023	0.540±.024	0.642±.050	0.585±.071	0.659±.144
LD-CVAE	0.644±.048	0.715±.056	0.736±.042	0.649±.049	0.798±.014	0.638±.016	0.596±.042	0.676±.038	0.647±.049	0.775±.073
SlotSPE	0.678±.041	0.734±.056	0.741±.040	0.649±.031	0.789±.029	0.663±.054	0.634±.032	**0.678±.068**	0.677±.029	0.802±.047
SCOPE	**0.708±.050**	**0.767±.054**	**0.753±.041**	**0.690±.067**	**0.838±.023**	**0.665±.031**	**0.647±.047**	0.676±.056	**0.743±.065**	**0.807±.096**

where N_{pat} is the number of patients and, for patient i, y_i is the discrete event time bin, $\delta_i \in \{0, 1\}$ is the event indicator (1 if the event was observed, 0 if censored), and $\boldsymbol{\ell}_i$ are the model logits. The survival probabilities $P(Y_i = y_i \mid \boldsymbol{\ell}_i)$ and $P(Y_i > y_i \mid \boldsymbol{\ell}_i)$ are derived from the logits.

To enable robust inference when omics data is unavailable, we employ a conditional dropout strategy during training. For each sample in a batch, we replace its omics conditioning vector $\mathbf{c}_i$ with a zero vector with a fixed probability p_{drop}. This forces the model to learn meaningful representations from the WSI data alone. We use a dropout probability of $p_{\text{drop}} = 0.25$ in all experiments.

4 Experiments

4.1 Experimental Setup

Datasets. We evaluate SCOPE on ten TCGA cohorts: BLCA (N_{pat}=381), BRCA (1046), COADREAD (573), HNSC (438), KIRC (488), LUAD (467), LUSC (460), SKCM (409), STAD (366), and UCEC (488). We use patient-level 5-fold splits inherited verbatim from Jaume et al. (2024) for direct comparability, and predict disease-specific survival (DSS) discretized into four intervals. Each WSI is tiled into 256×256 patches at 20× magnification and encoded to 768-d features with CTransPath (Wang et al., 2022), then rasterized onto a 64×64 tissue grid (with a density channel). RNA-seq is aggregated into pathway signatures drawn from Reactome (Gillespie et al., 2022) and the MSigDB Hallmarks (Liberzon et al., 2015).

Baselines. Under the multi-modal (WSI + RNA-seq) setting we compare against MCAT (Chen et al., 2021), Porpoise (Chen et al., 2022b), CMTA (Zhou & Chen, 2023), MOTCat (Xu & Chen, 2023), MMP (Song et al., 2024), SurvPath (Jaume et al., 2024), PIBD (Zhang et al., 2024), LD-CVAE (Zhou et al., 2025), and SlotSPE (Zhang et al., 2026). Under the pathology-only and missing-omics setting we also compare against ABMIL (Ilse et al., 2018), TransMIL (Shao et al., 2021), and CLAM-SB/MB (Lu et al., 2021).

Implementation and metrics. All results use 5-fold cross-validation with five random seeds; we report the concordance index (C-index) as mean ± std across folds. For risk stratification we addi-

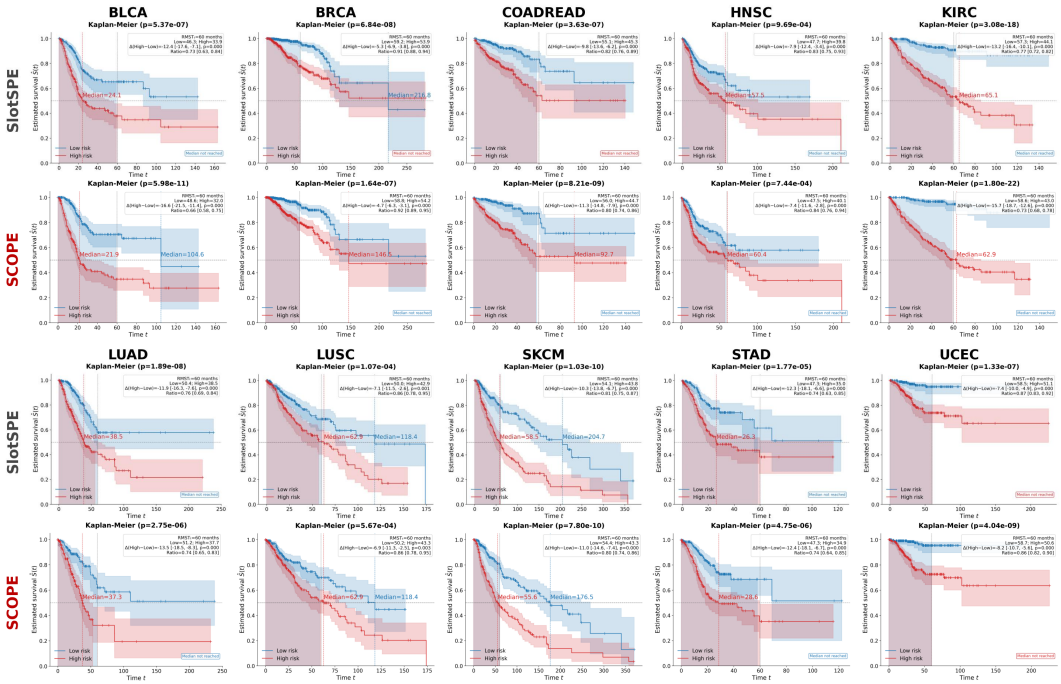


Figure 2: Kaplan-Meier survival curves for low- (blue) and high-risk (red) groups on TCGA cohorts.

tionally report the log-rank test and restricted mean survival time (RMST) up to τ=60 months on median-split risk groups. Models are trained with AdamW under a cosine schedule, the discrete-time NLL survival loss, batch size 8, and 30 epochs. The remaining hyperparameters are selected by hyperparameter search (further details are provided in Appendix E).

4.2 Comparisons with SOTAs

Table 1 reports the multi-modal C-index across the ten TCGA cohorts. SCOPE attains the best C-index on 7 of 10 cohorts and the second-best on the remaining three (BRCA, COADREAD, UCEC), where it is statistically on par with the strongest published baseline SlotSPE (the gaps are within a single standard deviation). Averaged over the ten cohorts, SCOPE reaches a mean C-index of 0.744 versus SlotSPE's 0.721 (a relative gain of $+3.3\%$); treating each cohort as a paired observation, SCOPE's improvement over every baseline is significant (Appendix B). The improvements are largest on the cohorts where spatial tissue architecture is most informative (STAD $+0.074$, HNSC $+0.048$, and BLCA $+0.039$) confirming that SCOPE's spectral-genomic decomposition captures prognostic structure that the bag-of-tokens attention baselines miss. Crucially, SCOPE delivers this accuracy while training in a fraction of the memory and wall-clock time of the strongest baselines (Appendix D).

The middle and lower blocks of Table 1 probe the omics-absent regimes. Trained and tested on WSI alone (pathology-only), SCOPE tops 8 of 10 cohorts with a cohort-mean of 0.702, ahead of the image-only SlotSPE (0.690), CLAM-MB (0.682) and ABMIL (0.681), so its spectral operator alone already matches or beats dedicated pathology-only models. When omics is present for training but withheld at inference (missing-omics), the conditional-dropout scheme lets SCOPE top 9 of 10 cohorts, its mean falling by only 0.015 (0.744$\rightarrow$0.729) versus SlotSPE's drop to 0.704; the degradation concentrates on the omics-informative cohorts (SKCM, LUAD, BLCA), showing omics acts as a targeted enhancement.

Beyond ranking accuracy, a clinically useful prognostic model must translate its continuous risk scores into well-separated survival strata. We pool the out-of-fold risk predictions across all five folds, split patients at the median into low- and high-risk groups, and assess separation with the log-rank test and the restricted mean survival time (RMST) up to $\tau = 60$ months. Fig. 2 contrasts SCOPE against SlotSPE on all ten cohorts, and Appendix C reports the full statistics. SCOPE separates the two strata highly significantly on every cohort, and relative to SlotSPE it attains the smaller log-rank p-value on 6 of 10 cohorts, the larger absolute RMST gap $|\Delta|$ on 7 of 10, and the largest

Table 2: Ablation on wavelet decomposition depth J, per cohort. Entries are C-index (mean $\pm$ std).

Structure	BLCA	BRCA	COADREAD	HNSC	KIRC	LUAD	LUSC	SKCM	STAD	UCEC
SelfAttn	0.695±.029	0.695±.067	0.712±.050	0.626±.061	0.794±.006	0.649±.056	0.572±.054	0.674±.039	0.597±.060	0.797±.075
J=0	0.724±.034	0.696±.048	0.698±.052	0.623±.046	0.780±.017	0.678±.056	0.598±.050	0.693±.023	0.627±.071	0.782±.073
J=1	0.722±.025	0.752±.075	0.724±.024	0.665±.048	0.810±.034	0.692±.046	0.580±.062	0.708±.022	0.696±.035	0.808±.065
J=2	0.744±.030	0.760±.081	0.726±.060	0.676±.061	0.824±.028	0.689±.047	0.600±.068	0.704±.017	0.741±.050	0.802±.053
J=3	0.747±.025	0.778±.066	0.754±.047	0.690±.058	0.845±.029	0.709±.052	0.641±.059	0.722±.025	0.745±.067	0.809±.062
J=4	0.742±.028	0.755±.075	0.757±.046	0.695±.048	0.847±.034	0.696±.057	0.644±.075	0.711±.015	0.716±.049	0.777±.044

individual gaps overall (BLCA -16.6, KIRC -15.7 months). Even on BRCA, where SCOPE's p-value is not the smaller one, SCOPE's high-risk (red) curve stays below its low-risk (blue) curve over the entire follow-up, whereas the SlotSPE curves cross near the end of the observation window.

4.3 Ablation Study

The core inductive bias of SCOPE is the multi-level 2D wavelet decomposition inside the SCOPE Operator, governed by the number of levels J. We isolate its contribution by sweeping J together with two non-wavelet controls: a standard self-attention block (*SelfAttn*) and disabling the wavelet path (J=0, a single full-resolution gene-modulated convolution). We report mean $\pm$ std C-index over the five folds for each of the ten cohorts individually. Appendix A gives structural reasons for the gap we observe against the *SelfAttn* control, including a spatial transfer characteristic that softmax attention's token mixing cannot vary independently of the rest of the spectrum (Proposition 5). The complementary ablation study is deferred to Appendix E.

The wavelet decomposition is essential and its benefit is consistent across cohorts: on all ten cohorts the best wavelet setting ($J\geq1$) outperforms both non-wavelet controls, with cohort-mean C-index rising from 0.681 (SelfAttn) and 0.690 (J=0) to 0.716-0.744 for $J\geq1$ (Table 2). The gain is large where spatial structure matters most, e.g. STAD improves from 0.597/0.627 to 0.745 and HNSC from 0.626/0.623 to 0.695, confirming that hierarchical spatial-spectral decomposition, not merely the gene-modulated convolution, drives SCOPE's accuracy. Going deeper than one level also helps, and the cohort-mean increases monotonically up to J=3 (0.744), after which it saturates. Crucially, depth is nearly free: each additional level operates on a feature map a quarter of the previous area while carrying three high-frequency bands, and these two factors cancel exactly, so the whole J-level pyramid costs the same as two single-scale global convolutions for *every* $J \geq 1$ rather than growing with depth (Proposition 2). Peak training memory is essentially flat from J=1 to J=4 ($221 \rightarrow 222$ MB), per-step time grows only mildly ($21 \rightarrow 26$ ms), and parameters rise modestly ($2.65 \rightarrow 3.58$M). In all other experiments in this work, we consistently applied J=3.

4.4 Interpretability Analysis

SCOPE affords three complementary, architecturally intrinsic views of its own reasoning: *where* it looks (the differentiable Top-K gate), *at what spatial scale* it looks (the wavelet sub-bands of the gene-modulated feature map), and *which molecular programmes* drive its conditioning (the pathway attention of Eq. 1). We illustrate all three on COADREAD (Fig. 3), using an uncensored patient (TCGA-AA-A02H) who died at 2.0 months and is among the highest-risk cases in the validation fold; further exemplars are given in Appendix F.

Spatial interpretability. Ranking tissue patches by their gate weight and displaying the top 25% yields a sparse, spatially structured map (Fig. 3b) whose strongest responses form a few well-separated foci rather than diffuse coverage. Cropping the raw H&E at the three highest-scoring foci (Fig. 3c) shows densely packed, architecturally distorted glandular tissue at each site, while large homogeneous regions are down-weighted.

Spectral interpretability. Applying the block-0 Haar DWT to the gene-modulated feature map decomposes it into one coarse (LL) and three detail (LH/HL/HH) sub-bands (Fig. 3d). For this patient 94.7% of the energy sits in LL, consistent with the learned GM-Conv kernel whose radial magnitude spectrum falls by three orders of magnitude from low to high frequency (Fig. 3e). This allocation varies with the patient: across all patients the LL energy fraction correlates positively with predicted risk (Pearson r=$+0.112$, p=0.007) while all three detail bands correlate negatively

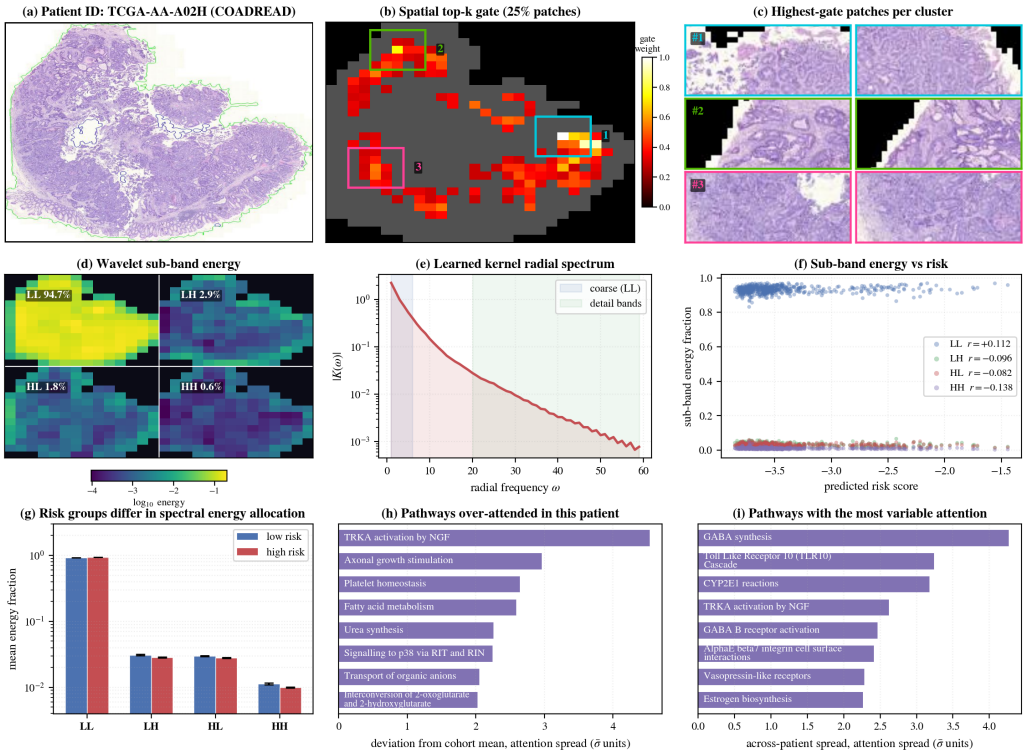


Figure 3: Interpretability analysis on COADREAD for an uncensored patient (TCGA-AA-A02H). More exemplars are given in Appendix F.

(-0.096, -0.082, -0.138; Fig. 3f), and the risk-median split (Fig. 3g) shows the same shift. On this cohort, then, a rising risk estimate moves the representation towards large-scale architectural structure and away from fine-grained texture, a spectral signature that a purely spatial attention model cannot express.

Omics interpretability. The pathway attention α_q of Eq. 1 states how much each of the M pathways contributes to the conditioning vector, so it is the natural place to read off which molecular programmes a prediction rests on. Fig. 3h ranks the pathways at which this patient departs furthest above its cohort, led by TRKA activation by NGF ($+4.5\,\bar{\sigma}_\alpha$) and axonal growth stimulation ($+3.0$). That the two strongest programmes are both neurotrophic is notable for this case: the colon and rectum are densely innervated, and neurotrophin (NGF/TRK) signalling drives the perineural invasion that is an established marker of tumour progression and unfavourable survival in colorectal cancer (Wang et al., 2024). The conditioning of this rapidly fatal patient therefore concentrates on a programme whose clinical correlate is precisely early, aggressive spread. Fig. 3i widens the view from the individual to the population, ranking the pathways whose attention varies most between patients: GABA synthesis ($4.3\,\bar{\sigma}_\alpha$), TLR10 signalling and CYP2E1 reactions head that list, so the encoder reserves most of its patient-to-patient dynamic range for neuroendocrine, innate-immune and xenobiotic-metabolic programmes.

5 Conclusion

We presented **SCOPE**, a spectral-conditioned multimodal architecture for pathology-genomics survival prediction, which replaces the prevailing self-attention fusion paradigm with gene-modulated convolution: gene expression directly parameterises the spectral filters that process the 2D WSI grid, evaluated across multiple wavelet scales. Three properties follow from the design: (a) a mean C-index of 0.744 over 10 TCGA cohorts, $+3.3\%$ above the strongest published baseline; (b) it remains the leading method without omics at inference, topping the pathology-only setting (mean 0.702) and retaining 0.729 under missing omics; (c) intrinsic spatial, spectral and molecular interpretability via Top-K selection, wavelet sub-band allocation and pathway attention. The design choices behind these results also open several natural extensions, which are discussed in Appendix G.

AI USE STATEMENT

In this work, generative AI tools were used solely to polish the manuscript: LLM-based assistants helped edit the authors' own draft prose for grammar, clarity, and readability. All substantive work was carried out entirely by the human authors without generative AI assistance. The architecture, the code, the proofs, the experimental protocol, and every reported number and its interpretation are the authors' own.

We have reviewed all AI-assisted prose. All code was written, executed, and verified by the authors; the optimisations described in Appendix D were checked for numerical equivalence against the original implementations before being adopted. Every quantitative claim in the paper was computed from the raw per-fold result files, and every passage polished with AI assistance was checked by the authors against those results and against the intended meaning of the original draft. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work uses only the publicly available TCGA cohorts, which consist of de-identified whole-slide images, bulk transcriptomic profiles, and clinical outcome annotations collected under the consent and governance procedures of the original TCGA programme. No new human-subject data were collected and no re-identification was attempted, so separate IRB approval was not required; we accessed and used the data in accordance with the NIH genomic data-sharing policy and the TCGA publication guidelines.

We note two concerns that bear on responsible use. First, TCGA cohorts are retrospective and demographically skewed, over-representing patients of European ancestry and particular care settings, so prognostic accuracy reported here may not transfer equitably across populations; the per-cohort variance we observe should not be read as evidence of generalisation to under-represented groups. Second, SCOPE is a research prototype evaluated by cross-validation on retrospective data, not a clinically validated device. Its risk scores and the saliency and spectral visualisations of Section 4.4 are intended to support model understanding and hypothesis generation, and must not be used to guide patient management without prospective, externally validated clinical trials. We declare no competing interests or sponsorship that could have influenced the reported results.

REPRODUCIBILITY STATEMENT

We have taken the following steps to make this work reproducible. The model is specified in full in Section 3.2 and Section 3.3, including the initialisation choices that the architecture depends on, and the training objective and conditional-dropout scheme in Section 3.4. Section 4.1 states the cohorts and their sizes, the patch-extraction and grid-rasterisation pipeline, the pathway construction, the evaluation protocol, and the selected hyperparameters; we reuse the patient-level 5-fold splits of Jaume et al. (2024) verbatim, so our numbers are directly comparable to published results on those splits. Appendix E reports the hyperparameter search behind those settings, and Appendix D documents the implementation optimisations together with the profiling protocol used for the efficiency claims. Appendix C gives the complete statistics underlying the risk-stratification figure, including the bootstrap procedure and resample count. All data used are public: the TCGA slides and transcriptomes are obtainable from the GDC portal, and the pathway signatures from Reactome and MSigDB. Source code for the model, the data-preparation pipeline, and the 5-fold training and evaluation harness will be released publicly.

REFERENCES

Richard J. Chen, Ming Y. Lu, Wei-Hung Weng, and et al. 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, 2021.

Richard J. Chen, Ming Y. Lu, Jingwen Wang, and et al. Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Transactions on Medical Imaging*, 41(4):757–770, 2022a.

Richard J. Chen, Ming Y. Lu, Drew F.K. Williamson, and et al. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. *Cancer Cell*, 40(8):865–878, 2022b.

Can Cui, Han Liu, Quan Liu, and et al. Survival prediction of brain cancer with incomplete radiology, pathology, genomic, and demographic data. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pp. 626–635, 2022.

I. Daubechies. The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, 36(5):961–1005, 1990.

Marc Gillespie, Bijay Jassal, Ralf Stephan, and et al. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D687–D692, 2022.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling (COLM)*, 2024.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations (ICLR)*, 2022.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, 2018.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. In *International Conference on Learning Representations (ICLR)*, 2017.

Guillaume Jaume, Anurag Vaidya, Richard J. Chen, and et al. 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, 2024.

Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing neural networks. In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pp. 972–981, 2017.

Arthur Liberzon, Chet Birger, Helga Thorvaldsdóttir, and et al. The molecular signatures database hallmark gene set collection. *Cell Systems*, 1(6):417–425, 2015.

Binghao Liu, Wenzheng Zhao, Zhijie Zheng, and Fei Gu. Adaptive dna sequence modeling via synergistic plasticity units. In *Proceedings of the 43th International Conference on Machine Learning (ICML)*, 2026a.

Xiao Liu, Chenxu Zhang, Fuxiang Huang, and et al. Vision mamba: A comprehensive survey and taxonomy. *IEEE Transactions on Neural Networks and Learning Systems*, 37(2):505–525, 2026b.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, and et al. Object-centric learning with slot attention. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

M.Y. Lu, D.F.K. Williamson, T.Y. Chen, and et al. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nature Biomedical Engineering*, 5:555–570, 2021.

S. G. Mallat. A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(7):674–693, 1989.

Stefano Massaroli, Michael Poli, Daniel Y. Fu, and et al. Laughing hyena distillery: extracting compact recurrences from convolutions. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, and et al. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proceedings of the National Academy of Sciences*, 115(13):E2970–E2979, 2018.

Eric Nguyen, Michael Poli, Marjan Faizi, and et al. Hyenadna: long-range genomic sequence modeling at single nucleotide resolution. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 2023.

Ethan Perez, Florian Strub, Harm de Vries, and et al. Film: visual reasoning with a general conditioning layer. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI) and Thirtieth Innovative Applications of Artificial Intelligence Conference (IAAI) and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI)*, 2018.

Michael Poli, Stefano Massaroli, Eric Nguyen, and et al. Hyena hierarchy: towards larger convolutional language models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 2023.

David W. Romero, Anna Kuzina, Erik J. Bekkers, and et al. Ckconv: Continuous kernel convolution for sequential data. In *International Conference on Learning Representations (ICLR)*, 2022.

Zhuchen Shao, Hao Bian, Yang Chen, and et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. In *Proceedings of the International Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pp. 2136–2147, 2021.

Andrew H. Song, Richard J. Chen, Guillaume Jaume, and et al. Multimodal prototyping for cancer survival prediction. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

H. Wang, R. Huo, K. He, and et al. Perineural invasion in colorectal cancer: mechanisms of action and clinical relevance. *Cellular oncology (Dordrecht, Netherlands)*, 47(1):1—-17, 2024.

Xiyue Wang, Sen Yang, Jun Zhang, and et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Medical Image Analysis*, 81:102559, 2022.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

Yujia Xie, Hanjun Dai, Minshuo Chen, and et al. Differentiable top-k with optimal transport. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Yingxue Xu and Hao Chen. Multimodal optimal transport-based co-attention transformer with global structure consistency for survival prediction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21184–21194, 2023.

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

Yilan Zhang, Yingxue Xu, Jianqi Chen, and et al. Prototypical information bottlenecking and disentangling for multimodal cancer survival prediction. In *International Conference on Learning Representations (ICLR)*, 2024.

Yilan Zhang, Nanbo Li, Changchun Yang, and et al. Structural prognostic event modeling for multimodal cancer survival analysis. In *International Conference on Learning Representations (ICLR)*, 2026.

F. Zhou and H. Chen. Cross-modal translation and alignment for survival analysis. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21485–21494, 2023.

Junjie Zhou, Jiao Tang, Yingli Zuo, and et al. 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, 2025.

A THEORETICAL ANALYSIS OF THE SCOPE OPERATOR

This appendix formalises the SCOPE Operator and establishes the properties that distinguish it from the two dominant families of token mixers: softmax self-attention ($\Theta(N^2)$) and raster-scanned linear state-space models ($\Theta(N)$). Propositions 1–4 are exact statements about the operator as implemented: exact reconstruction, an exact cost accounting, the identity at initialisation, and the nesting of the depth hierarchy. Propositions 5–7 are structural contrasts: they identify a degree of freedom that attention's spatial mixing cannot exercise independently of the rest of the spectrum, and an axis asymmetry that any fixed raster scan necessarily imposes. We state each precisely and flag the regime in which it holds.

A.1 SETUP AND NOTATION

Let $\mathbf{U} \in \mathbb{R}^{D\times H\times W}$ be a feature grid with $N = HW$ spatial positions, and let $\mathbf{c} \in \mathbb{R}^{D_c}$ be the patient conditioning vector. Write $\mathrm{vec}(\mathbf{U})$ for the flattening of $\mathbf{U}$ into a sequence of N tokens of width D. All three families are token mixers: maps $\mathbb{R}^{D\times H\times W} \to \mathbb{R}^{D\times H\times W}$ that move information between spatial positions.

The gene-modulated convolution. For a kernel field $\mathbf{k_c} \in \mathbb{R}^{D\times H\times W}$ and a per-channel skip gain $\boldsymbol{\rho} \in \mathbb{R}^D$, the primitive of Section 3.2 is

$$\big[\mathrm{GMConv}(\mathbf{U},\mathbf{c})\big]_d \;=\; \mathcal{C}\big[\,\mathbf{k}_{\mathbf{c},d} * \mathbf{U}_d\,\big] \;+\; \rho_d\,\mathbf{U}_d, \tag{6}$$

where $*$ is the linear (zero-padded) 2D convolution on the $2H \times 2W$ grid and $\mathcal{C}$ is the centre crop back to $H \times W$. We stress that Eq. 6 is a cropped Toeplitz operator: the $2\times$ zero-padding is exactly what breaks periodic wrap-around, so the FFT is used only as an $O(N \log N)$ evaluation device and does not diagonalise the operator. Each output channel is a scalar 2D convolution of the corresponding input channel (a depthwise operator), plus a residual scaled by ρ_d.

The kernel field is generated implicitly. With $\boldsymbol{\pi}_n \in \mathbb{R}^E$ the (shared, patient-independent) positional embedding of spatial position n, $\phi_{\mathbf{c}} : \mathbb{R}^E \to \mathbb{R}^{D_k}$ the FiLM-conditioned hidden stack of the kernel generator, of width $D_k = 64$ throughout, $\mathbf{A} \in \mathbb{R}^{D\times D_k}$ the output projection, and $\boldsymbol{\Omega} \in \mathbb{R}^{D\times N}$ the radial-decay window,

$$\mathbf{K_c} \;=\; \boldsymbol{\Omega} \odot \big(\mathbf{A}\,\boldsymbol{\Phi}_{\mathbf{c}}\big), \qquad \boldsymbol{\Phi}_{\mathbf{c}} := \big[\phi_{\mathbf{c}}(\boldsymbol{\pi}_1), \ldots, \phi_{\mathbf{c}}(\boldsymbol{\pi}_N)\big] \in \mathbb{R}^{D_k\times N}, \tag{7}$$

where $\mathbf{K_c} \in \mathbb{R}^{D\times N}$ is $\mathbf{k_c}$ reshaped. Only $\phi_{\mathbf{c}}$ depends on $\mathbf{c}$, and it does so solely through the FiLM parameters $(\boldsymbol{\gamma}(\mathbf{c}), \boldsymbol{\beta}(\mathbf{c}))$ of Eq. 2.

The wavelet pyramid. Let $\mathcal{B}$ denote one level of the depthwise 2D Haar analysis operator, mapping $\mathbb{R}^{D\times H\times W}$ to four subbands $(\mathrm{LL}, \mathrm{LH}, \mathrm{HL}, \mathrm{HH})$ each of size $D \times \frac{H}{2} \times \frac{W}{2}$, and $\mathcal{B}^*$ its synthesis. The J-level SCOPE Operator $\mathcal{F}_J(\cdot\,;\mathbf{c})$ is the cascade of Section 3.3: analysis to depth J, a per-band GMConv on each packed high-frequency triple and on the deepest LL, then synthesis with cross-scale mixing coefficients $\{\boldsymbol{\nu}_j\}$, added to the base path $\boldsymbol{\eta} \odot \mathrm{GMConv}_b(\mathbf{U},\mathbf{c})$.

A.2 EXACT STRUCTURAL PROPERTIES

Proposition 1 (Orthogonality and perfect reconstruction)**.** *The single-level Haar analysis operator satisfies $\mathcal{B}^*\mathcal{B} = \mathrm{Id}$ and $\|\mathcal{B}\mathbf{U}\|_2 = \|\mathbf{U}\|_2$ for all $\mathbf{U}$. Consequently the J-level analysis is an isometry of $\mathbb{R}^{D\times H\times W}$ onto its coefficient space, and the pyramid loses no information at any depth.*

Proof. Per channel and per 2×2 spatial block, $\mathcal{B}$ acts as the matrix $\mathbf{Q} = \frac{1}{2}\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$ on the vectorised block, which is the choice implemented by the filter bank. Direct computation gives $\mathbf{Q}\mathbf{Q}^\top = \mathbf{I}_4$, so $\mathbf{Q}$ is orthogonal; since the transform is applied blockwise with stride 2 and the blocks partition the grid, the full operator is a permutation of a direct sum of copies of $\mathbf{Q}$, hence orthogonal. Isometry and $\mathcal{B}^*\mathcal{B} = \mathrm{Id}$ follow, and composing J isometries yields an isometry. Numerically, $\|\mathbf{Q}\mathbf{Q}^\top - \mathbf{I}\|_{\max} = 0$ exactly. □

Proposition 1 is what makes depth safe: because analysis is lossless, adding a level can only enlarge the hypothesis class, never discard signal. It also gives the Parseval identity $\|\mathbf{U}\|_2^2 = \sum_j \|\mathrm{HF}_j\|_2^2 + \|\mathrm{LL}_J\|_2^2$ used implicitly by the energy-fraction diagnostics of Section 4.4 (Mallat, 1989).

Proposition 2 (Log-linear cost with a depth-independent bound)**.** *Let $\mathcal{T}_1(N)$ denote the cost of one full-resolution global convolution (Eq. 6) over N positions and D channels, so $\mathcal{T}_1(N) = \Theta(DN \log N)$. Counting the base path, the packed high-frequency convolution at every level, and the low-frequency convolution at the deepest level, the total mixing cost of the J-level operator is*

$$\mathcal{T}_{\mathrm{SCOPE}}(N, J) \;=\; \underbrace{\mathcal{T}_1(N)}_{\text{base path}} \;+\; \underbrace{\sum_{j=1}^{J} \frac{3}{4^j}\, \mathcal{T}_1(N)}_{\text{HF triples}} \;+\; \underbrace{\frac{1}{4^J}\, \mathcal{T}_1(N)}_{\text{deepest LL}} \;=\; 2\,\mathcal{T}_1(N) \tag{8}$$

up to the log *terms,* exactly and for every $J \geq 1$. *The operator is therefore $\Theta(DN \log N)$ per layer with a constant factor that does not grow with depth.*

Proof. Level j operates on a grid of $N/4^j$ positions. Its high-frequency convolution is applied to the packed triple $(\mathrm{LH}_j, \mathrm{HL}_j, \mathrm{HH}_j)$, i.e. to $3D$ channels, contributing $3 \cdot 4^{-j}\mathcal{T}_1(N)$; the deepest level additionally convolves the single D-channel LL_J at $N/4^J$ positions, contributing $4^{-J}\mathcal{T}_1(N)$; and the base path is one full-resolution convolution, contributing $\mathcal{T}_1(N)$. Summing the finite geometric series, $\sum_{j=1}^{J} 3 \cdot 4^{-j} = 1 - 4^{-J}$, so the wavelet branch costs exactly $(1 - 4^{-J}) + 4^{-J} = 1$ in units of $\mathcal{T}_1(N)$ and the total is $2\,\mathcal{T}_1(N)$ for every J. The cancellation is exact because the 3 channels of each high-frequency triple compensate precisely for the $4\times$ reduction in spatial extent; each extra level redistributes work between the HF and LL terms without adding any. Ignoring the $\log$ factors is the standard convention here and is conservative, since $\log(N/4^j) < \log N$. □

Proposition 2 is the formal counterpart of the empirical observation in Section 4.3 that peak memory and per-step time are nearly flat in J: depth is not merely asymptotically cheap but carries no marginal cost in this accounting, because each additional level trades an equal amount of work between its high- and low-frequency terms.

Proposition 3 (Exact identity at initialisation)**.** *The base-path scale is initialised depth-dependently, $\boldsymbol{\eta} = \mathbf{0}$ when $J \geq 1$ and $\boldsymbol{\eta} = \mathbf{1}$ when $J = 0$, and the remaining parameters are initialised as shipped ($\boldsymbol{\nu}_j = \mathbf{0}$, zero-initialised GeneFiLM, zero-initialised kernel output layer $\mathbf{A} = \mathbf{0}$, skip gain $\boldsymbol{\rho} = \mathbf{1}$). Then for every depth $J \geq 0$ and every $\mathbf{c}$,*

$$\mathcal{F}_J(\mathbf{U}; \mathbf{c}) \;=\; \mathcal{F}_J(\mathbf{U}; \varnothing) \;=\; \mathbf{U} \qquad \textit{for all } \mathbf{U}, \tag{9}$$

where $\varnothing$ denotes the omics-missing path.

Proof. $\mathbf{A} = \mathbf{0}$ forces $\mathbf{K}_\mathbf{c} = \mathbf{0}$ by Eq. 7, so every GMConv collapses to its skip term $\boldsymbol{\rho} \odot \mathbf{U} = \mathbf{U}$, independently of $\mathbf{c}$. For $J = 0$ the operator is $\boldsymbol{\eta} \odot \mathrm{GMConv}_b(\mathbf{U}, \mathbf{c}) = \mathbf{1} \odot \mathbf{U} = \mathbf{U}$. For $J \geq 1$ each band passes through unchanged, so by Proposition 1 the analysis/synthesis cascade reconstructs $\mathbf{U}$ exactly, while $\boldsymbol{\nu}_j = \mathbf{0}$ removes the cross-scale terms and $\boldsymbol{\eta} = \mathbf{0}$ removes the base path; the output is again $\mathbf{U}$. Independence of $\mathbf{c}$ holds because $\boldsymbol{\gamma}(\mathbf{c}) = \boldsymbol{\beta}(\mathbf{c}) = \mathbf{0}$ at zero-init, so $\phi_\mathbf{c}$ is the unconditional map, and in any case $\mathbf{A} = \mathbf{0}$ annihilates its output. □

Two consequences deserve emphasis. First, the identity at init means training begins from a well-conditioned residual stream regardless of J, which is why deeper pyramids do not destabilise optimisation. Second, the $\mathbf{c}$-independence is the precise sense in which missing-modality inference is architectural: the conditioned and unconditioned forward passes are bit-identical at initialisation, so no auxiliary reconstruction loss is needed to keep them aligned.

Proposition 4 (Nesting of the depth hierarchy)**.** *Let $\mathcal{H}_J = \{\mathcal{F}_J(\cdot\,; \mathbf{c})\}$ denote the family of maps realisable at depth J as the parameters range over their admissible values. Then $\mathcal{H}_0 \subseteq \mathcal{H}_1$ and, more generally, $\mathcal{H}_J \subseteq \mathcal{H}_{J+1}$ for every $J \geq 0$.*

Proof. For $\mathcal{H}_0 \subseteq \mathcal{H}_1$: given any $\mathcal{F}_0(\cdot\,; \mathbf{c}) = \boldsymbol{\eta} \odot \mathrm{GMConv}_b(\cdot\,, \mathbf{c})$, instantiate $\mathcal{F}_1$ with the same base-path parameters and with every band kernel and band skip gain set to zero. All four subbands then vanish, so $\mathcal{B}^*(\mathbf{0}) = \mathbf{0}$, the wavelet branch contributes nothing, and only the base path remains, reproducing $\mathcal{F}_0$ pointwise.

For $\mathcal{H}_J \subseteq \mathcal{H}_{J+1}$: given any depth-$J$ instance, build a depth-$(J{+}1)$ instance that copies the base path and the level-1 through level-J band parameters, and at the new deepest level $J{+}1$ sets the

high-frequency kernels and skip gains to zero and the cross-scale coefficient $\boldsymbol{\nu}_{J+1} = \mathbf{0}$. The extra analysis/synthesis stage is then applied to LL_J and inverted with unmodified coefficients, which by Proposition 1 returns LL_J exactly; the level-J low-frequency convolution of the shallower instance is recovered by placing it on LL_J before that stage. Every remaining term is identical, so the two maps agree pointwise. □

Because each depth is contained in the next, increasing J can only enlarge the hypothesis class, and the zero-initialised $\{\boldsymbol{\nu}_j\}$ and $\boldsymbol{\eta}$ mean the added capacity starts inactive and is recruited only if training finds it useful. This is what makes the depth sweep of Section 4.3 a clean capacity ablation.

A.3 Contrast with self-attention

The next result isolates a spatial degree of freedom, the gain on the constant mode, that SCOPE can set with a single per-channel scalar independently of its oscillatory response, whereas attention's spatial mixing can only rescale it together with the rest of the spectrum, at any width or head count.

Proposition 5 (Spatial DC gain is pinned for attention and free for SCOPE)**.** *Let* $\mathbf{P}_{\mathrm{attn}} = \mathrm{softmax}(\mathbf{S}_{\mathrm{attn}}) \in \mathbb{R}^{N\times N}$ *be any single-head attention matrix, where* $\mathbf{S}_{\mathrm{attn}}$ *is the scaled query–key score matrix. Then* $\mathbf{P}_{\mathrm{attn}}\mathbf{1} = \mathbf{1}$*: the spatial token-mixing operator fixes the constant (zero-frequency) spatial mode with gain exactly* 1*, independently of* $\mathbf{S}_{\mathrm{attn}}$*. The value and output projections act identically at every spatial position, so they can rescale the channel vector but cannot make attention's spatial gain on the DC mode differ from its gain on any oscillatory mode. By contrast, let* $\mathbf{T}_{\mathbf{c},d} \in \mathbb{R}^{N\times N}$ *be the matrix of the spatial operator (Eq. 6) on channel* d*. For any position* n *whose support lies in the grid interior, the corresponding row sum equals*

$$\big(\mathbf{T}_{\mathbf{c},d}\mathbf{1}\big)_n \;=\; \sum_m k_{\mathbf{c},d}(m) \;+\; \rho_d, \tag{10}$$

a single learnable scalar per channel that can be set independently of the operator's response to any oscillatory input.

Proof. Each row of $\mathbf{P}_{\mathrm{attn}}$ is a softmax, hence non-negative and summing to 1, so $\mathbf{P}_{\mathrm{attn}}$ is row-stochastic and $\mathbf{P}_{\mathrm{attn}}\mathbf{1} = \mathbf{1}$; therefore $\mathbf{P}_{\mathrm{attn}}(v\mathbf{1}) = v\mathbf{1}$ for any scalar v, independently of $\mathbf{S}_{\mathrm{attn}}$. The value/output projections and multi-head concatenation apply the same linear map at every spatial position, i.e. they act as $\mathbf{I}_N \otimes \mathbf{R}$ on the (position $\times$ channel) tensor; such a map commutes with any spatial-frequency decomposition and therefore multiplies every spatial mode by the same channel operator $\mathbf{R}$. It cannot give the DC spatial mode a gain that differs from the gain applied to an oscillatory mode, which is the sense in which attention's spatial DC response is pinned relative to the rest of the spectrum.

For SCOPE, apply Eq. 6 to the constant input $\mathbf{U}_d = \mathbf{1}$. By definition the convolution term at position n is $\sum_m k_{\mathbf{c},d}(m)\,\mathbf{U}_d(n-m)$ restricted to the indices retained by the centre crop $\mathcal{C}$; when the whole kernel support fits inside the grid at n, every tap is multiplied by 1, giving $\sum_m k_{\mathbf{c},d}(m)$, and the skip term adds ρ_d, which is Eq. 10. The kernel taps enter this sum only through their total, whereas the response to an oscillatory input weights them by the corresponding oscillation, so the interior DC gain can be moved without constraining the oscillatory response; ρ_d alone already suffices, since it shifts Eq. 10 while leaving every zero-mean input's image unchanged. □

Two remarks on the scope of Proposition 5. First, the operator of Eq. 6 is a cropped Toeplitz map, so it is not diagonalised by the DFT and the constant vector is not an eigenvector: near the boundary the crop truncates part of the kernel support and the row sums in Eq. 10 shrink accordingly. This does not affect the contrast being drawn, because attention's row sums are pinned to 1 at every position, interior and boundary alike, whereas SCOPE's are a free scalar in the interior and a determined function of the same parameters at the edge. Second, the practical consequence for attention is not that it leaves a global brightness or stain-intensity offset untouched (the channel projections can rescale it), but that any rescaling it applies to that offset is applied equally to the informative oscillatory structure at the same channels, because the projection is a single position-independent map. Attention therefore cannot spatially suppress such a nuisance offset relative to the prognostic content it wants to keep; the removal must happen downstream. SCOPE can attenuate the constant mode inside the mixer while preserving oscillatory content (Proposition 1 guarantees the detail bands

survive), a genuine gain in expressible band-pass behaviour that is consistent with the strongly low-pass learned kernels reported in Section 4.4.

Remark 6 (A low-dimensional parameterisation of the reachable kernel family)**.** *By equation 7, before the radial window the reachable kernel matrices* $\mathbf{A}\mathbf{\Phi}_\mathbf{c}$ *have rank at most* $\min(D_k, D)$ *regardless of* $\mathbf{c}$*, with* $D_k = 64$*: conditioning steers a kernel inside an at-most-*D_k*-dimensional subspace rather than choosing it freely. The Hadamard window only bounds the rank of the product from above,* $\mathrm{rank}(\mathbf{\Omega} \odot \mathbf{A}\mathbf{\Phi}_\mathbf{c}) \le \mathrm{rank}(\mathbf{\Omega}) \cdot \mathrm{rank}(\mathbf{A}\mathbf{\Phi}_\mathbf{c})$*, and does not by itself increase it; a separable (rank-one) window would leave the rank untouched. The radial window used here is not separable across the channel and position axes, so in practice it does lift the final kernel to full rank, and the bound should therefore be read as constraining the* generator *rather than the realised kernel. The substantive point for Section 4.4 is the parameterisation itself: patient-to-patient variation enters through* $\phi_\mathbf{c}$ *alone and is thus confined to a low-dimensional, inspectable family, which is also why a near-identity FiLM produces the observed near-degenerate per-pathway spectra.*

A.4 Contrast with raster-scanned SSMs

Proposition 7 (Raster anisotropy of 1D SSMs)**.** *Let a diagonal linear SSM with spectral radius* $\lambda \in (0, 1)$ *be applied to* $\mathrm{vec}(\mathbf{U})$ *under row-major flattening. For a single causal pass the influence of token* m *on token* n *decays as* λ^{n-m} *for* $n \ge m$ *and vanishes for* $n < m$*. Consequently, comparing two pairs of grid positions at the same relative orientation, a horizontally adjacent pair is separated by raster distance* 1 *whereas a vertically adjacent pair is separated by* W*, so the ratio of vertical to horizontal influence is* λ^{W-1}*, exponentially small in the grid width. The SCOPE operator has no such anisotropy: its kernel is indexed by the 2D offset* (Δ_y, Δ_x) *and treats both axes symmetrically by construction.*

Proof. A diagonal SSM with transition λ has impulse response $h[n] \propto \lambda^n$ supported on $n \ge 0$, so the contribution of token m to token n scales as λ^{n-m} for $n \ge m$ and is zero otherwise. Row-major flattening sends grid position (y, x) to index $yW + x$, whence horizontal neighbours differ by 1 and vertical neighbours by W; taking the ratio of the two influences at matched orientation gives λ^{W-1}. For Eq. 6 the weight linking (y, x) to (y', x') is $k_\mathbf{c}(y - y', x - x')$, a free function of the 2D offset, so no axis is privileged. □

Bidirectional or multi-directional scans mitigate but do not remove this: each scan direction still imposes a 1-vs-W asymmetry, and averaging a fixed number of scans cannot recover a genuinely isotropic 2D kernel (Liu et al., 2026b). Tissue architecture (glandular orientation, invasive fronts, and stromal banding) has no preferred raster axis, which is the structural reason we adopt a 2D operator rather than porting a 1D long-context model.

A.5 Summary

Table 3 collects the comparison. SCOPE's advantage is that it attains a global receptive field at $\Theta(DN \log N)$ with an isotropic 2D kernel, a constant-mode response that can be set independently of the oscillatory one, exact identity at initialisation for every depth, and a multi-scale decomposition whose constant factor does not grow with depth, none of which attention or a raster SSM provides simultaneously.

Table 3: Structural comparison of token mixers on a $H \times W$ grid, $N = HW$.

Property	Self-attention	Raster SSM	SCOPE Operator
Cost per layer	$\Theta(DN^2)$	$\Theta(DN)$	$\Theta(DN \log N)$
Global receptive field	yes	yes (decaying)	yes (per band)
Isotropic in 2D	yes (permutation-inv.)	no, ratio λ^{W-1}	yes, by construction
Spatial DC gain	pinned rel. to spectrum	free	free (Prop. 5)
Identity at init, any depth	no	no	yes (Prop. 3)
Multi-scale by construction	no	no	yes, isometric (Prop. 1)
Cross-modal conditioning	via tokens	via input	in the kernel (Rem. 6)

Table 4: Paired significance of SCOPE against each baseline, with the ten cohorts of Table 1 as paired observations. $\overline{\Delta}$ is the mean per-cohort C-index difference (SCOPE minus baseline), Wins the number of cohorts on which SCOPE is higher, p_t the paired t-test p-value, and d_z Cohen's paired effect size.

Setting	Baseline	$\overline{\Delta}$	Wins	p_t	d_z
Pathology + Transcriptome	MCAT	$+0.0539$	$10/10$	1.6×10^{-4}	1.96
	Porpoise	$+0.0680$	$10/10$	1.9×10^{-4}	1.91
	CMTA	$+0.0529$	$10/10$	2.3×10^{-5}	2.52
	MOTCat	$+0.0516$	$10/10$	8.8×10^{-5}	2.12
	SurvPath	$+0.0619$	$10/10$	4.2×10^{-6}	3.10
	PIBD	$+0.0685$	$10/10$	1.9×10^{-5}	2.57
	MMP	$+0.0751$	$10/10$	2.4×10^{-5}	2.51
	LD-CVAE	$+0.0526$	$10/10$	8.4×10^{-4}	1.55
	SlotSPE	$+0.0234$	$7/10$	2.6×10^{-2}	0.84
Pathology-Only	ABMIL	$+0.0217$	$8/10$	2.7×10^{-2}	0.83
	TransMIL	$+0.0312$	$10/10$	2.4×10^{-3}	1.32
	CLAM-SB	$+0.0248$	$10/10$	1.8×10^{-4}	1.93
	CLAM-MB	$+0.0204$	$9/10$	1.4×10^{-3}	1.44
	SlotSPE	$+0.0122$	$10/10$	1.2×10^{-3}	1.47
Missing-Omics	CMTA	$+0.1264$	$10/10$	1.2×10^{-6}	3.60
	MOTCat	$+0.1099$	$10/10$	5.8×10^{-5}	2.24
	LD-CVAE	$+0.0420$	$9/10$	7.2×10^{-4}	1.59
	SlotSPE	$+0.0249$	$9/10$	6.8×10^{-3}	1.10

B Statistical Significance of the Main Comparisons

Table 1 reports a 5-fold mean C-index per cohort with an inter-fold standard deviation of 0.03–0.06, so any single per-cohort difference is individually noisy and a per-cohort win should not be read as a significant one. We therefore assess significance at the level at which the claim is actually made by treating each of the ten TCGA cohorts as one paired observation and comparing SCOPE against every baseline in the same evaluation setting. Our primary test is a paired t-test. Alongside the p-values we give the mean paired difference $\overline{\Delta}$, the number of cohorts on which SCOPE wins, and Cohen's $d_z = \overline{\Delta}/\mathrm{sd}(\Delta)$ as an effect size.

Every comparison in Table 4 is significant at the 0.05 level under the test. The margins are large and unambiguous against the older fusion architectures ($\overline{\Delta}$ between $+0.05$ and $+0.08$ multi-modally, and $+0.11$ to $+0.13$ against the omics-dropout variants of CMTA and MOTCat, whose $d_z > 2$ reflects that these models degrade sharply without omics). As expected, the tightest comparison in each setting is against the strongest published baseline: $+0.0234$ (p_t=2.6×10^{-2}) multi-modally, $+0.0122$ (p_t=1.2×10^{-3}) pathology-only, and $+0.0249$ (p_t=6.8×10^{-3}) under missing omics.

C Kaplan-Meier Risk-Stratification Statistics

Table 5 reports the full log-rank and RMST-based statistics underlying Fig. 2. For each cohort we pool the out-of-fold risk scores across the five folds, split patients at the median risk into low- and high-risk groups, and quantify separation by (i) the log-rank p-value, (ii) the restricted mean survival time (RMST) of each group at horizon $\tau = 60$ months, (iii) the RMST difference $\Delta = \mathrm{RMST}_{\mathrm{High}} - \mathrm{RMST}_{\mathrm{Low}}$ with a bootstrap 95% confidence interval (2,000 resamples), and (iv) the RMST ratio (High/Low) with its bootstrap CI. A more negative Δ and a smaller ratio indicate stronger, clinically more meaningful stratification.

SCOPE achieves a highly significant separation on all ten cohorts ($p < 10^{-3}$ throughout), with the largest absolute RMST gaps on BLCA (-16.6 months), KIRC (-15.7) and LUAD (-13.5). Against SlotSPE, SCOPE attains the smaller log-rank p-value on 6 of 10 cohorts, the larger $|\Delta|$ on 7 of 10, and the lower RMST ratio on 6 of 10 with two exact ties.

Table 5: Summary of log-rank p-values and RMST-based statistics (RMST values, differences, and ratios) for median-split group separation across ten TCGA cohorts, corresponding to Figure 2. RMST is computed at $\tau = 60$ months; confidence intervals are bootstrap 95% intervals.

Dataset	Method	Log-rank p-value	RMST (Low)	RMST (High)	Δ (High–Low) [95% CI]	Ratio (High/Low) [95% CI]
BLCA	SlotSPE	5.37×10^{-7}	46.3	33.9	-12.4 [-17.6, -7.1]	0.73 [0.63, 0.84]
	SCOPE	5.98×10^{-11}	48.6	32.0	-16.6 [-21.5, -11.4]	0.66 [0.58, 0.75]
BRCA	SlotSPE	6.84×10^{-8}	59.2	53.9	-5.3 [-6.9, -3.8]	0.91 [0.88, 0.94]
	SCOPE	1.64×10^{-7}	58.8	54.2	-4.7 [-6.3, -3.1]	0.92 [0.89, 0.95]
COADREAD	SlotSPE	3.63×10^{-7}	55.1	45.3	-9.8 [-13.6, -6.2]	0.82 [0.76, 0.89]
	SCOPE	8.21×10^{-9}	56.0	44.7	-11.3 [-14.8, -7.9]	0.80 [0.74, 0.86]
HNSC	SlotSPE	9.69×10^{-4}	47.7	39.8	-7.9 [-12.4, -3.4]	0.83 [0.75, 0.93]
	SCOPE	7.44×10^{-4}	47.5	40.1	-7.4 [-11.6, -2.8]	0.84 [0.76, 0.94]
KIRC	SlotSPE	3.08×10^{-18}	57.3	44.1	-13.2 [-16.4, -10.1]	0.77 [0.72, 0.82]
	SCOPE	1.80×10^{-22}	58.6	43.0	-15.7 [-18.7, -12.6]	0.73 [0.68, 0.78]
LUAD	SlotSPE	1.89×10^{-8}	50.4	38.5	-11.9 [-16.3, -7.6]	0.76 [0.69, 0.84]
	SCOPE	2.75×10^{-6}	51.2	37.7	-13.5 [-18.5, -8.3]	0.74 [0.65, 0.83]
LUSC	SlotSPE	1.07×10^{-4}	50.0	42.9	-7.1 [-11.5, -2.6]	0.86 [0.78, 0.95]
	SCOPE	5.67×10^{-4}	50.2	43.3	-6.9 [-11.3, -2.5]	0.86 [0.78, 0.95]
SKCM	SlotSPE	1.03×10^{-10}	54.1	43.8	-10.3 [-13.8, -6.7]	0.81 [0.75, 0.87]
	SCOPE	7.80×10^{-10}	54.4	43.3	-11.0 [-14.6, -7.4]	0.80 [0.74, 0.86]
STAD	SlotSPE	1.77×10^{-5}	47.3	35.0	-12.3 [-18.1, -6.6]	0.74 [0.63, 0.85]
	SCOPE	4.75×10^{-6}	47.3	34.9	-12.4 [-18.1, -6.7]	0.74 [0.64, 0.85]
UCEC	SlotSPE	1.33×10^{-7}	58.5	51.1	-7.4 [-10.0, -4.9]	0.87 [0.83, 0.92]
	SCOPE	4.04×10^{-9}	58.7	50.6	-8.2 [-10.7, -5.6]	0.86 [0.82, 0.90]

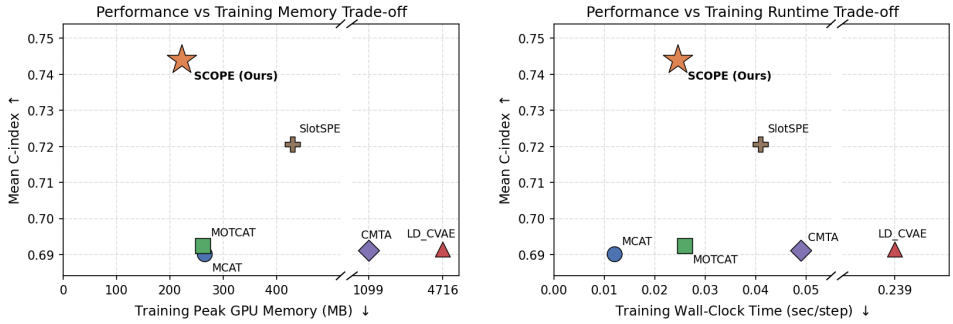
D Efficiency Analysis

We profile SCOPE against the strongest multi-modal baselines from Table 1 (MCAT, MOTCat, CMTA, LD-CVAE, and SlotSPE). All measurements use the full multi-modal setting with no missing modality. The y-axis in Figure 4(a) is the mean C-index across all ten TCGA cohorts (Table 1).

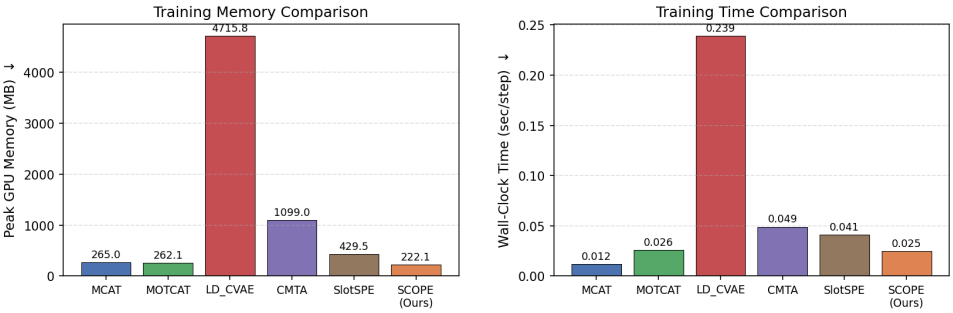
Three implementation-level optimizations are decisive for SCOPE's efficiency, and none of them alters the model's outputs. First, the pathway-omics encoder originally applied one small MLP per pathway in a Python loop ($\sim$330 pathways), issuing thousands of tiny GPU kernels per step; we replace it with a small set of batched (`einsum`) operations that are numerically identical but launch only a handful of kernels. Second, because pathways vary enormously in gene count, naively batching them would pad every pathway's weight matrix to the global maximum width; we instead bucket pathways by size and pad only within each bucket, which shrinks the backbone parameters (and their optimizer state) at bit-identical outputs. Third, the spectral encoder (a stack of FFT-based Gene-Modulated convolution blocks) is launch-overhead bound (its actual GPU compute is only a fraction of the observed wall-clock), so we capture it as a CUDA graph. Since CUDA graphs require a fixed input shape, we lock onto the dominant single-slide shape (which covers $\approx 99\%$ of patients) and use the captured graph for both training and inference, transparently falling back to eager execution on the rare multi-slide batch, so correctness never depends on the batch shape.

Inference runtime and memory. At inference SCOPE processes a patient with a peak GPU memory of 85 MB and a wall-clock latency of 7.4 ms per patient. Because the model rasterises every slide onto a fixed grid before the spectral encoder, this cost is independent of the raw patch count, giving SCOPE a bounded, predictable inference footprint.

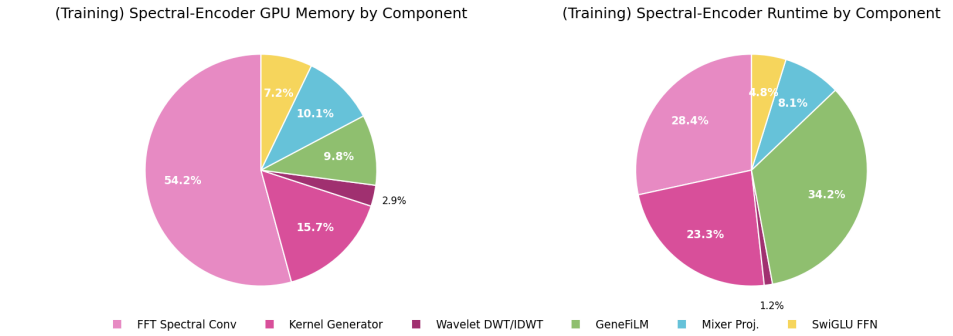
Training runtime and memory. Under the unified batch-size setting (Figure 4(b)), SCOPE trains with a peak GPU memory of 222 MB and a per-step wall-clock time of 0.025 s, lower than the strongest published baseline, SlotSPE (429 MB, 0.041 s) on both axes, and far below the heaviest baselines LD-CVAE (4716 MB, 0.239 s) and CMTA (1099 MB, 0.049 s). Since SCOPE also attains the highest mean C-index (0.744 vs SlotSPE's 0.721), it dominates SlotSPE on all three axes simultaneously, accuracy, memory, and runtime, and occupies the top-left corner of the accuracy-



(a) Mean C-index versus training peak GPU memory / wall-clock trade-off



(b) Training peak memory and per-step wall-clock time



(c) Attributing SCOPE's spectral encoder training GPU memory / runtime to its sub-components

Figure 4: Efficiency analysis. (a) Mean C-index (averaged over the ten TCGA cohorts) versus training peak GPU memory / wall-clock trade-off across compared models. (b) Training peak memory and per-step wall-clock time under a unified batch-size setting. (c) Internal breakdown of SCOPE's spectral encoder, attributing its training GPU memory / runtime to its own sub-components. The SCOPE Operator itself split into the FFT spectral convolution, the implicit kernel generator, and the wavelet transform (DWT/IDWT), together with GeneFiLM conditioning, the mixer projections, and the SwiGLU FFN.

efficiency plane in Figure 4(a). In fact SCOPE has the lowest peak training memory of all six compared methods (222 MB, below MOTCat's 262 MB and MCAT's 265 MB), and only MCAT is faster per step (0.012 vs 0.025 s) at the cost of a 0.054 lower mean C-index.

Module-level cost breakdown. Figure 4(c) opens up the spectral encoder (which itself accounts for essentially all of SCOPE's training memory and runtime) and attributes its cost to its own sub-components. Within the encoder, the SCOPE Operator (FFT spectral convolution + implicit kernel generator + wavelet transform) collectively dominates activation memory ($\approx 73\%$). Broken down further, the FFT spectral convolution is the single largest consumer (54.2% of encoder memory and 28.4% of runtime), the implicit kernel-generator MLP is second (15.7% memory, 23.3% runtime), and the Haar wavelet transform (DWT/IDWT) is negligible (2.9% memory, 1.2% runtime). The

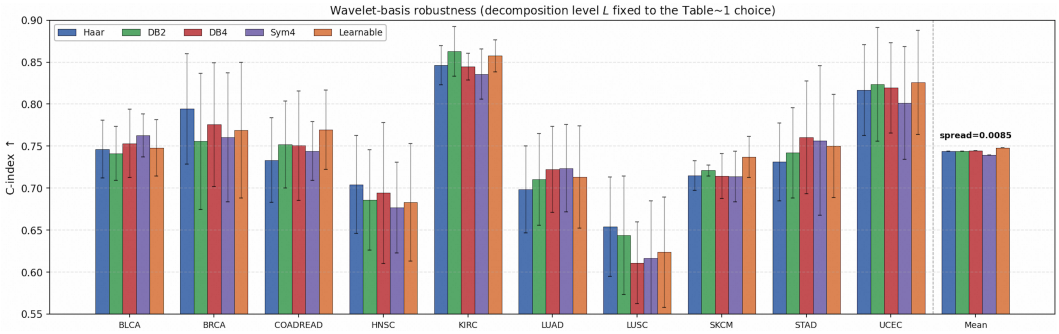


Figure 5: Wavelet-basis robustness. Grouped bars show the mean C-index of the five wavelet bases (Haar, DB2, DB4, Sym4, Learnable) with $\pm$ std error bars over five folds. The rightmost Mean group gives the cohort-averaged C-index per basis: its total spread is only 0.0085, well within the inter-fold noise, showing SCOPE is insensitive to the choice of wavelet basis.

remaining encoder cost is the GeneFiLM conditioning (9.8% memory but 34.2% runtime, reflecting its many small elementwise modulations), the 1×1 mixer projections ($\sim 10\%$ memory, $\sim 8\%$ runtime), and the SwiGLU FFN ($\sim 7\%$ memory, $\sim 5\%$ runtime). This confirms that SCOPE's compute is concentrated in the FFT-based spectral-genomic mixing, precisely the static, control-flow-free computation that the CUDA-graph capture above exploits, removing its launch overhead without altering its numerics. The six-way split also cleanly separates SCOPE's two efficiency levers: peak memory is bounded almost entirely by the FFT spectral convolution, whereas per-step runtime is spread across the FFT convolution, the implicit kernel generator, and the many small GeneFiLM modulations. This separation is exactly what our targeted optimizations address: the CUDA-graph capture fuses the launch-heavy GeneFiLM and kernel-generator kernels that would otherwise dominate the wall-clock, while the fixed tissue grid caps the FFT working set that dominates memory. Notably, the wavelet transform itself costs under 3% of either budget; this is the micro-level explanation for why SCOPE's peak memory and per-step time stay nearly flat as the decomposition level grows (Section 4.3, Table 2), making deeper multi-scale decomposition effectively free.

E More Ablations

This section complements the main-text ablation study (Section 4.3).

E.1 Ablation on Wavelet Basis

Varying only the wavelet basis over {Haar, DB2, DB4, Sym4, Learnable} leaves performance essentially unchanged as shown in Fig. 5. The cohort-mean C-index is 0.7440 (Haar), 0.7438 (DB2), 0.7445 (DB4), 0.7391 (Sym4), and 0.7477 (Learnable), a total spread of only 0.0085, far below the typical inter-fold std of 0.03–0.06. Within every individual cohort the five bases likewise cluster within one standard deviation of one another. SCOPE is therefore robust to the choice of wavelet basis: the benefit comes from performing a multi-scale spectral decomposition at all, not from any particular filter. We default to Haar for its simplicity, while the learned lifting basis offers a marginal average edge for practitioners willing to add a handful of parameters.

E.2 Ablation on the Remaining Hyperparameters

Beyond the wavelet design, SCOPE exposes seven training/architecture hyperparameters. We run a separate one-factor-at-a-time grid for hyperparameter searching. Because cohorts differ substantially in baseline C-index, we center every trial by its own cohort's trial mean and report the average centered ΔC-index per hyperparameter value (Table 6).

Two effects stand out. First, omics dropout is the most influential hyperparameter (spread 0.0092): disabling it ($p_{\text{drop}}{=}0$) is consistently worst, and enabling it improves the cohort-centered mean. This matches its dual role: a modality-dropout regularizer that prevents the model from over-relying

Table 6: Hyperparameter sensitivity aggregated over 5-fold trials on the ten cohorts. Each trial is centered by its cohort's trial mean; entries are the average centered ΔC-index for each value. Spread is the max-min across values of that hyperparameter.

Hyperparameter	Values	Centered ΔC-index			Spread
Omics dropout p_{drop}	0.0 / 0.25 / 0.5	-0.0092	0.0	-0.0022	0.0092
Spectral blocks	2 / 4 / 6	0.0	-0.0019	-0.0071	0.0071
Learning rate	1e-4 / 5e-4 / 1e-3	-0.0052	-0.0050	0.0	0.0052
Top-K ratio	0.1 / 0.25 / 0.5	-0.0003	0.0	-0.0048	0.0048
Gate temperature	0.01 / 0.05 / 0.1	-0.0039	-0.0015	0.0	0.0039
Weight decay	1e-3 / 5e-3 / 1e-2	0.0	-0.0011	-0.0032	0.0032
WSI projection dim	128 / 256 / 384	-0.0022	0.0	-0.0015	0.0022

on the omics branch, while simultaneously training the omics-missing forward path of Section 3.4, in which Gene-FiLM reduces to a parameter-free standardisation, and which is what makes the missing-omics regime of Table 1 work without architectural change. Second, encoder depth mildly favors shallower stacks (2–4 blocks over 6, spread 0.0071), consistent with the modest sample sizes of survival cohorts; higher learning rates and lower weight decay also help slightly. The remaining hyperparameters shift the centered mean by less than 0.005, an order of magnitude below the typical inter-fold std of 0.03-0.06, so SCOPE is effectively insensitive to them.

F Interpretability: Further Exemplars and Cross-Cohort Analysis

Figs. 6 and 7 reproduce the full analysis of the main text (Fig. 3) on two further cohorts (KIRC and STAD). The two cohorts deliberately span the two branches of the selection protocol: KIRC is a correctly-ranked high-risk uncensored case, and STAD a correctly-ranked long-surviving censored case, so the panels are not all drawn from a single prognostic regime.

Two properties recur in both renders. First, the learned kernel is strongly low-pass: panel (e) falls by roughly three orders of magnitude from $\omega{=}1$ to $\omega{\approx}60$ in each case ($2.8{\times}10^3$-fold for KIRC, $2.0{\times}10^3$-fold for STAD). Second, and as a direct consequence, the coarse band dominates the energy budget, carrying 93.4% (KIRC) and 85.8% (STAD) of tissue-masked energy in panel (d), so the low-pass character reported in the main text is a property of the operator.

The two cohorts differ in how strongly their internals couple to risk. Panel (f) shows KIRC's coarse-band coupling is statistically indistinguishable from zero ($r_{\mathrm{LL}}{=}{+}0.046$, $p{=}0.31$) with the diagonal-detail band the only individually significant one ($r_{\mathrm{HH}}{=}{-}0.115$, $p{=}0.011$), whereas STAD couples positively in the coarse band ($r_{\mathrm{LL}}{=}{+}0.147$, $p{=}0.005$) and negatively in two of the three detail bands ($r_{\mathrm{LH}}{=}{-}0.176$, $r_{\mathrm{HL}}{=}{-}0.106$). Panels (h) and (i) separate the two cohorts on the omics side as well. For the high-risk KIRC case, panel (h) is led by SHC activation ($+4.4\,\bar{\sigma}_\alpha$) and the p53-independent DNA damage response ($+3.9$), while the cohort-level ranking in panel (i) is dominated by Hallmark programmes, MYC targets far ahead of the rest at $11.1\,\bar{\sigma}_\alpha$, followed by IL2/STAT5 signalling, allograft rejection and complement. The long-surviving STAD case instead over-attends GABA synthesis ($+7.6\,\bar{\sigma}_\alpha$) and the ATM-mediated double-strand-break response ($+6.8$), and its cohort spreads widest over gap-junction assembly, TASK channels and connexon trafficking.

For each cohort we correlated every patient's LL (coarse-band) energy fraction with the model's predicted risk, pooling all five validation folds so each patient appears exactly once, in the fold where it was held out. These are the coefficients quoted throughout the paper, including in Section 4.4 and in panel (f) of every exemplar figure. The Pearson correlation runs from -0.259 (BLCA, $p{<}10^{-6}$) through -0.227 (LUSC), -0.126 (UCEC), -0.103 (SKCM) and -0.056 (HNSC) to $+0.018$ (LUAD), $+0.046$ (KIRC), $+0.112$ (COADREAD), $+0.147$ (STAD) and $+0.155$ (BRCA): positive in five cohorts, negative in five, and $|r|$ significant ($p{<}0.05$) in seven. SCOPE consistently re-allocates energy across scales as its risk estimate changes, but the *direction* of that association is cancer-type-specific.

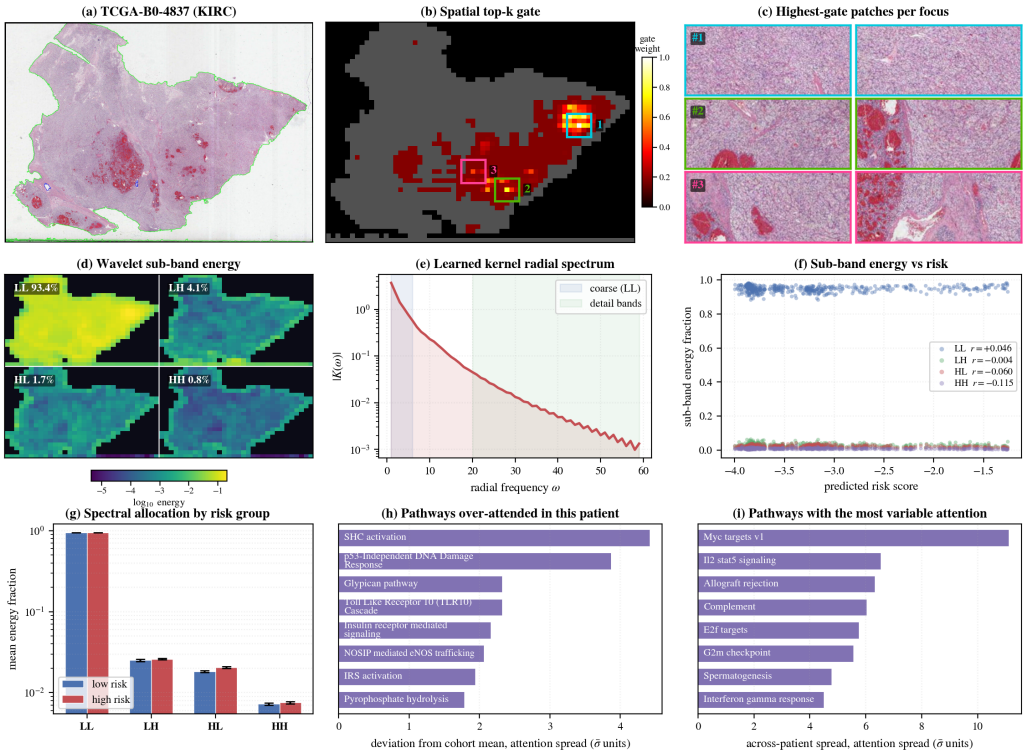


Figure 6: Per-cohort interpretability composite for KIRC (TCGA-B0-4837). Panels follow Fig. 3: (a) tissue mask; (b) the Top-K gate with the three strongest spatially separated foci boxed; (c) raw H&E at each focus; (d) Haar sub-band energies of the block-0 gene-modulated feature map; (e) radial magnitude spectrum of the learned kernel; (f) sub-band energy fraction against predicted risk, pooled over all five validation folds of this cohort so each patient appears exactly once; (g) risk-median split of the same pooled quantities; (h) pathways this patient attends to most strongly relative to its cohort, in units of the mean across-patient attention spread $\bar{\sigma}_\alpha$; (i) pathways whose attention varies most across the patients of this cohort, in the same $\bar{\sigma}_\alpha$ units.

Fig. 8 resolves this coupling by sub-band and shows that the coarse and detail bands carry the same information with opposite sign. Across the ten cohorts r_{LL} is almost perfectly anti-correlated with each detail-band coefficient (-0.97 for LH, -0.94 for HL, -0.88 for HH), and the opposition holds within cohorts as well: in 8 of 10 cohorts all three detail bands take the sign opposite to LL, with LUAD and BRCA the only partial exceptions (one detail band each falls on the same side of zero, in both cases with a confidence interval spanning it). What the decomposition adds is where the compensating mass goes and how strong the effect is in each band. Magnitudes are comparable across bands (mean $|r|$ of 0.125, 0.124, 0.106 and 0.117 for LL, LH, HL and HH), and the largest single coefficient in the whole grid is BLCA's HH at $+0.322$ ($p<10^{-9}$), i.e. the diagonal-detail band, not the coarse band, gives the sharpest spectral read on risk in that cohort.

Read this way the cohorts fall into two regimes: a *coarse-risk* group in which rising risk shifts energy towards large-scale architecture (BRCA, STAD, COADREAD), a *detail-risk* group in which it shifts towards fine texture (BLCA, LUSC, UCEC, SKCM), and three cohorts whose coarse-band coupling is statistically indistinguishable from zero (KIRC, LUAD, HNSC). We stress that these are all modest effect sizes ($|r|\leq 0.32$) that characterise how the operator internally distributes energy across scales as its risk estimate varies; consistent with the main text's observation that the LL fraction alone ranks survival far worse than the full model.

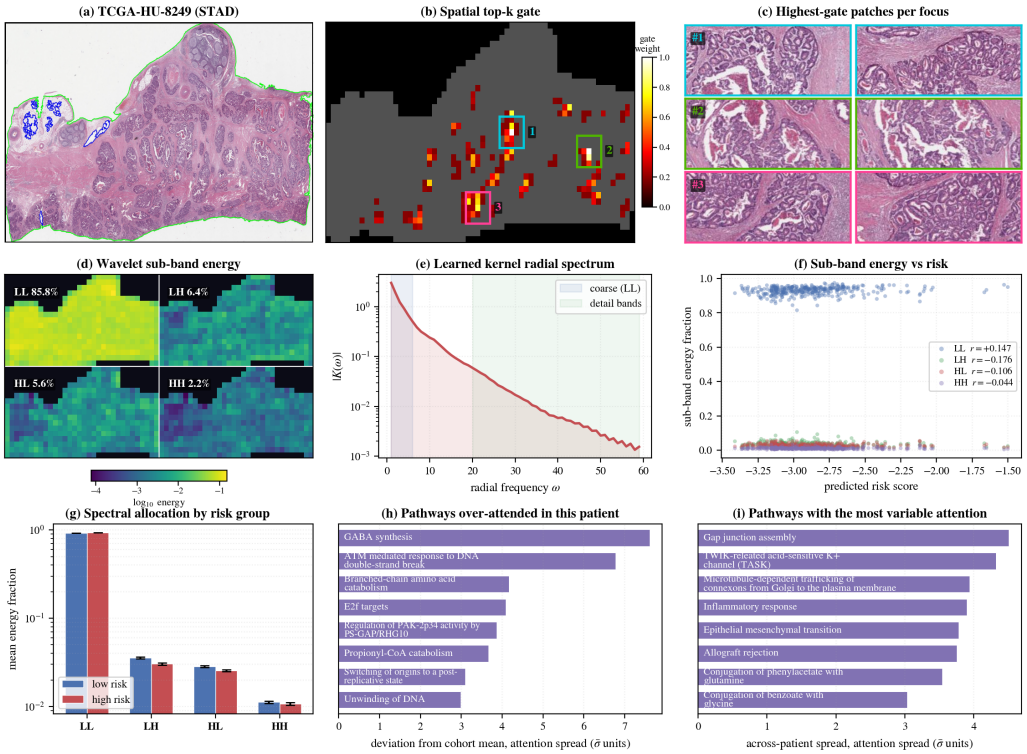


Figure 7: Per-cohort interpretability composite for STAD (TCGA-HU-8249). Panels as in Fig. 6.

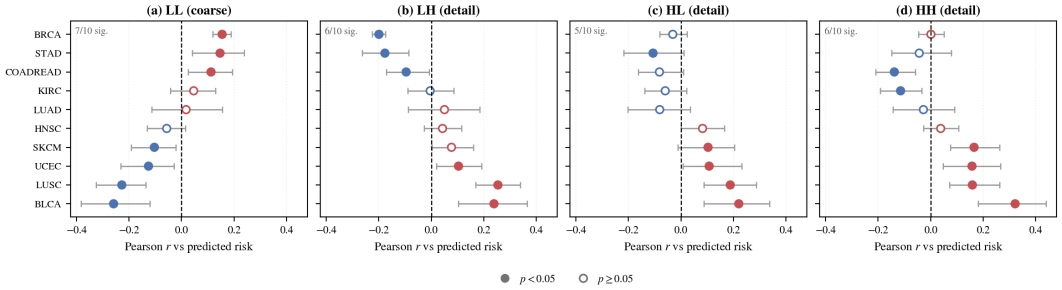


Figure 8: Per-cohort spectral–risk coupling, resolved by Haar sub-band. Each panel plots the Pearson correlation between a patient's energy fraction in one sub-band and the model's predicted risk, pooled over all five validation folds of that cohort; whiskers are bootstrap 95% confidence intervals, filled markers are significant at $p<0.05$ and hollow markers are not, and red/blue encode the sign. Cohorts keep the same ordering (by r_{LL}) in all four panels and all panels share one x-range, so the sign reversal between the coarse band (a) and the three detail bands (b–d) is readable across each row.

G Scope and Future Directions

The results in this paper are established under a deliberately controlled protocol, held-out patch encoder, fixed wavelet basis, and bulk transcriptomic conditioning, chosen so that every comparison in Table 1 is like-for-like against the published baselines on the same 5-fold splits. The design choices that make this scope tractable are also the ones that leave the most room to grow, and we outline three complementary directions below.

Richer patch encoders. All experiments use a single frozen patch encoder, CTransPath (Wang et al., 2022), applied identically to SCOPE and to every baseline, so the reported ranking isolates

the fusion architecture. Because SCOPE consumes patch features through a generic rasterised grid, it is agnostic to this choice: swapping in a larger pathology foundation model changes only the input channel dimension and could raise absolute accuracy for all methods simultaneously. We therefore view the encoder as an orthogonal axis of improvement.

Beyond the Haar basis. We default to a Haar wavelet for its simplicity and exact-reconstruction guarantees (Proposition 1). Appendix E shows that DB2, DB4, Sym4 and a fully learnable lifting basis all land within inter-fold noise of Haar (total spread 0.0086, well below the 0.03–0.06 inter-fold std), so the multi-scale decomposition itself is what carries the benefit. This makes the framework a natural testbed for task-adapted or steerable bases; we expect such extensions to refine the present findings, and the learnable basis already offers a small average edge for practitioners willing to add a handful of parameters.

Higher-resolution omics. Our conditioning is bulk RNA-seq aggregated into pathway signatures, which is exactly the modality for which large, outcome-annotated cohorts such as TCGA exist, and it is what enables the ten-cohort evaluation reported here. Spatial transcriptomics and single-cell profiling would let the same Gene-FiLM mechanism modulate filters locally, a strict generalisation of the current patient-level conditioning, and are a promising route to stronger and more spatially resolved omics signal as such cohorts reach comparable scale. The zero-initialised, missing-modality-robust conditioning path we already use means these richer inputs can be incorporated without changing the core operator.