OmniSurv: A Six-Modality Foundation Model for Zero-Shot Pan-Cancer Survival Prediction and Molecular Profiling

Mobareji Abejide · Noblesville High School, Noblesville, Indiana · Computational Biology and Bioinformatics

I. Background and Motivation — Why does cancer prognosis need a new approach, and why use survival as a training signal?

Abstract

Accurate cancer prognosis requires integrating heterogeneous signals: histopathology, genomics, clinical variables, and biological context. Yet most multimodal survival models train and test on the same cancers, making it unclear whether they learn transferable biology or cancer-specific shortcuts. We present OmniSurv, a six-modality foundation-style survival model that asks: can survival labels alone produce a representation that generalizes to entirely unseen cancer types? Trained on 10 TCGA cancers and evaluated under a strict cancer-type-holdout on 21 unseen cancers (2,455 patients), OmniSurv achieves mean within-cancer C-index 0.884 (log-rank p<0.001), outperforming the best published baseline by +0.243 C-index points. Frozen embeddings recover canonical driver mutations (KRAS AUROC 0.526), tumor mutational burden (R²=0.865), subtype structure (LGG IDH purity 0.89), and pan-cancer molecular organization, demonstrating that survival can function as a weak but globally informative supervisory signal for transferable tumor biology.

The Central Question

Can a single model augmented by deep learning trained only on survival outcomes and without explicit labels learn a patient representation that (1) predicts prognosis in cancers it has never seen, and (2) encodes general tumor biology as a byproduct?

- **Why standard benchmarks fail:** Cross-validation on pooled TCGA can report C=0.78 while the same model scores 0.57 on truly unseen cancers.
- **The key test:** Train on 10 cancers; evaluate on 21 entirely different ones. No cancer-type label is given at test time. If it generalizes, it learned real biology, not cancer-specific patterns.

Why Survival is a Powerful Training Signal

- **Survival reflects everything biologically important.** Tumor aggressiveness is determined by driver mutations, histological invasion patterns, immune evasion, chromosomal instability, and treatment resistance. A model that accurately ranks patients by survival risk must implicitly encode all of these features simultaneously. No other single clinical label contains as broad a biological summary.
- **Survival is universally available.** Unlike curated labels for specific mutations (TP53, KRAS) or immune subtypes, survival data exists for nearly every cancer patient in any large dataset. No additional annotation is required beyond what is already collected for clinical care. This is why survival-supervised models can scale without annotation bottlenecks.
- **The sample-complexity problem.** Building a specialized labeled dataset for each task (mutation detection, subtype classification, immune scoring) requires 30 to 60 minutes of expert annotation per patient, and hundreds of patients per cancer type. OmniSurv bypasses this by learning all of these features implicitly from a single, universally available signal.
- **The downstream test:** If survival training works, the same frozen embedding should support classification, retrieval, biomarker prediction, and subtype discovery without any additional labeled data. That is exactly what Sections IV and V demonstrate.

Analogy: Survival is to cancer biology what reading comprehension is to language understanding: a single task that requires mastery of many underlying concepts simultaneously. A model that predicts reading comprehension implicitly learns syntax, semantics, and world knowledge. A model that predicts survival implicitly learns tumor biology.

Scale of the Problem and Six Input Modalities

200+ Cancer subtypes globally recognized	**20K+** TCGA patients 33 cancer types	**9.7M** Cancer deaths 2022 worldwide
53.5M Annotations within 5 years, globally	**10^{-50}** Annotation cost per image, specialist	**6** Modalities integrated by OmniSurv

- Histopathology (WSI): Morphology, necrosis, invasion, nuclear grade
- RNA Expression: Immune programs, signaling, proliferation
- Copy-Number (CNV): Genomic instability, driver amplifications
- Somatic Mutations: TP53, KRAS, PIK3CA, IDH1 — on/off switches
- Clinical Variables: Dx features: stage, grade, age, treatment
- Biology RAG: 21 canonical cancer pathways + literature

The Cross-Cancer Generalization Gap (Why This Benchmark Matters)

- **SOTA baselines report C=0.72–0.78 under standard cross-validation.** Our holdout reveals the same models score only 0.567–0.641 on unseen cancers. Standard cross-validation on TCGA always includes the same cancer types in training and test splits. A model can learn that "GBM patients die faster than LGG patients" and exploit this as a survival shortcut without learning any within-cancer biology. On unseen cancers, this shortcut provides no signal at all.
- **What our protocol requires:** OmniSurv must discriminate within each held-out cancer type using only features that generalize across cancers. The stratified Cox loss enforces this by comparing patients only within the same cancer type, making inter-cancer shortcuts unrewardable even during training.
- **Why clinicians should care:** Rare cancers rarely have enough patients to train dedicated models. A truly generalizable representation could provide prognostic guidance for rare malignancies where large labeled datasets are clinically impossible to collect.

OmniSurv Key Results at a Glance

0.884 Mean C-index 21 unseen cancers	**+0.243** vs. POSPHOSE best baseline
0.977 TMB AUROC High-Low 2-class	**95.4%** OS Cancer Type 21 types, zero-shot
0.869 Mean grade AUROC 28 driver genes	**0.89** IDH Purity LGG, unsupervised

II. Materials, Methods and Related Works — What data and software were used? Where does OmniSurv sit relative to the prior literature?

Data and Computational Environment

TCGA Dataset

All data derived from The Cancer Genome Atlas (TCGA), a publicly available resource molecularly characterizing 33 cancer types across thousands of patients. OmniSurv uses 10 cancers for training (2,451 patients) and 21 cancers for zero-shot testing (3,495 patients). Triple redundancy ensures no patient leakage: patient-level splits, case-level for multi-sample patients, and cancer-type-holdout so test cancers never appear in training.

Hardware

- **Training:** NVIDIA RTX 6000 PRO GPU (188 GB vRAM) via RunPod cloud servers.
- **Analysis and fine-tuning:** NVIDIA GeForce RTX 4060 (local workstation).

Python Software Stack

PyTorch neural networks, training, embedding generation	**scikit-learn** linear probes, logistic regression, clustering, NMI, silhouette
Numpy / pandas numerical arrays, tabular data handling	**lifelines** concordance index, survival analysis
SciPy correlations, distance, statistical analysis	**matplotlib** publication-quality figures

Exact versions: see repository environment.yml for full reproducibility.

Related Works: Four Lines of Prior Literature

1. Multimodal survival modeling (architectural relatives)

Pathomic Fusion and POSPHOSE established that combining WSIs with molecular data improves prognosis. MCAT introduced multimodal co-attention between WSI patches and genomic tokens. Perceiver, HAT, and HAIM explored scalable fusion architectures. Recent foundation-model approaches such as PORPOISE, SurvPath, CATfusion, and PathOmics continue the trajectory with cross-attention, hierarchical alignment, and clinical integration. These are OmniSurv's closest architectural relatives, but none ask whether cancer-type holdout or tissue-level ablation is a meaningful evaluation.

2. Pathology foundation models (conceptual relatives)

CHIEF, HIPS, Prov-GigaPath, and TITAN show that large-scale pathology pretraining yields transfer features for diagnosis, mutation prediction, survival, and retrieval. TITAN specifically evaluates few-shot, zero-shot, and cross-cancer settings. OmniSurv shares the transfer-representation goal but replaces self-supervised pathology pretraining with multimodal survival supervision as the organizing signal.

3. Weak and indirect supervision (methodological relatives)

WSDER uses weak supervision for drug response prediction when direct labels are scarce. PROGAPS uses weakly supervised pan-cancer sentence modeling to avoid cancer-specific pipelines. OmniSurv pushes this idea further: survival is not only a weak label for one outcome, but the prognostic signal from which the pan-cancer multimodal latent space.

4. Cross-cancer generalization (evaluation gap)

Most multimodal survival models are evaluated within one cancer at a time or on pooled cohorts where train and test share cancer types. Even "pan-cancer" papers typically mean pooled evaluation, not transfer to unseen disease types. A 2025 systematic review of multimodal pathology-omics models confirms that external validation and cross-cancer generalization remain underdeveloped. OmniSurv's cancer-holdout survival protocol directly addresses this gap.

Where OmniSurv Sits: Novelty Position

The central novelty claim

OmniSurv is not merely another multimodal transformer for survival. Its defining claim is that survival can serve as a global supervisory signal for transferable tumor biology, evaluated under cancer-type holdout rather than standard cross-validation. Prior work typically shows that adding more modalities improves within-cancer prognostication. OmniSurv asks whether the learned representation transfers to entirely unseen cancer types, and whether survival alone is sufficient to induce such a transferable biological prior.

Closest architectural relatives

- **Multimodal survival:** Pathomic Fusion, MCAT, Perceiver, HAT, HAIM, PORPOISE, SurvPath, CATfusion, PathOmics
- **Pathology foundation models:** CHIEF, HIPS, Prov-GigaPath, TITAN
- **Weak supervision:** WSDER, PROGAPS

Deliberate simplifications

- Frozen pre-extracted features (no end-to-end backbone training)
- Simple concatenation fusion (no hierarchical or graph-structured fusion)
- Survival-only supervision (no auxiliary classification heads)
- Standard Cox loss (no ranking or contrastive auxiliary losses)

Why these simplifications matter

Each simplification isolates a specific claim. Frozen features test whether the pretrained embeddings encode transferable biology rather than being merely task-adapted. Simple fusion tests whether the cross-modality attention learns something beyond what late concatenation would capture. Survival-only supervision tests whether survival alone is sufficient to induce a useful biological prior. Standard Cox loss ensures any improvement comes from the representation, not from auxiliary loss engineering.

Statistical Methods and Evaluation Protocol

Concordance Index (C-index)

Harrell's C-index of all valid concordant patient pairs where patient i died before j (delta_i = 1 and time_i < time_j), the fraction where the model assigns higher risk to patient i. Inter-cancer pairs are excluded from both the numerator and denominator, preventing inter-cancer shortcuts from inflating the score.

Time-Dependent AUC and IBS

At 365 days, probability OmniSurv correctly identifies which of two patients lived longer than 365 days, accounting for censoring. Integrated Brier Score: expected squared error between predicted survival curve and observed outcome integrated over time; 0.25 = perfectly uninformative concordant prediction.

Bootstrap Confidence Intervals

95% CI from 1,000 patient-level stratified bootstrap with replacement, stratified by cancer type. Paired bootstrap used for pairwise significance tests (e.g., OmniSurv vs. MCAT, p=0.051). p<0.001, ** p<0.01, * p<0.05. Log-rank tests are stratified within each cancer independently.

Cancer Traffic Protocol

All downstream evaluations (biomarkers, retrieval, subtyping) use cancer-stratified 5-fold cross-validation. Linear probes trained on frozen embeddings with linear models only: logistic regression, ridge regression, cosine-similarity nearest-prototype, or Gaussian mixture EM clustering. Zero backbone fine-tuning ensures results reflect representation quality alone, not task-specific adaptation capacity.

Attention and Significance

Modality attentions are normalized softmax weights. Attention is fully zeroed at test time. WSI removal: C drops 0.008 (genomic recovery effective). Clinical removal: C drops 0.018 (clinical features matter but are not load-bearing). All clustering metrics (NMI, ARI, Cophenetic) are computed against published Hoadley et al. 2018 pan-cancer labels as external ground truth, not derived from TCGA survival labels.

III. OmniSurv Architecture — How does a single model integrate 6 heterogeneous data streams and generalize to unseen cancer types at inference?

i. End-to-End Architecture Overview

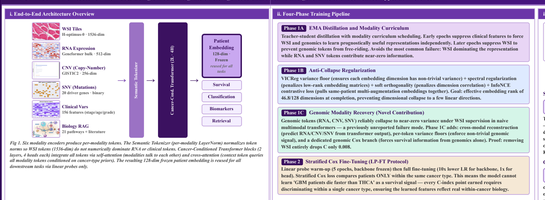


Fig 1. Six modalities encoders produce per-modality tokens. The Semantic Tokenizer (per-modality LayerNorm) normalizes token norms so WSI tokens (1536-dim) do not numerically dominate RNA or clinical tokens. Cancer-conditioned Transformer blocks (2 layers, 4 heads each) integrate all tokens via self-attention. Modalities talk to each other and cross-attention (context token queries all modality tokens conditioned on cancer-type prior). The resulting 128-dim frozen patient embedding is reused for all downstream tasks via linear probes only.

ii. Four-Phase Training Pipeline

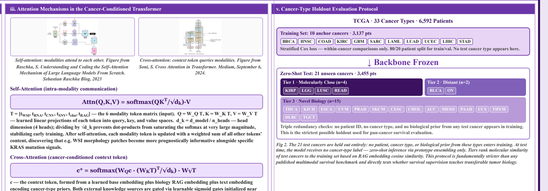
Phase 1A — EMA Distillation and Modality Curriculum
Teacher-student distillation with modality curriculum scheduling. Early epochs suppress clinical features to force WSI and genomics to learn prognostically useful representations independently. Later epochs suppress WSI to prevent genomic tokens from free-riding. Avoids the most common failure: WSI dominating the representation while RNA and SNV inputs contribute near-zero information.

Phase 1B — Anti-Collapse Regularization
VICReg variance floor (ensures each embedding dimension has non-trivial variance) + spectral regularization (penalizes low-rank embedding matrices) + soft orthogonality (penalizes dimension correlations) + IsoNet'X contrastive loss (pulls same-patient multi-augmentation embeddings together). Goal: effective embedding rank of 46.8/128 dimensions at completion, preventing dimensional collapse to a few linear directions.

Phase 1C — Genomic Modality Recovery (Novel Contribution)
Genomic tokens (RNA, CNV, SNV) reliably collapse to near-zero variance under WSI supervision in naive multimodal transformers — a previously unreported failure mode. Phase 1C adds: cross-modal reconstruction (predict RNA/CNV/SNV from transformer output), per-token variance floors (enforce non-trivial genomic signal), and a dedicated genomic Cox branch (forces survival information from genomics alone). Proof: removing WSI entirely drops C only 0.008.

Phase 2 — Stratified Cox Fine-Tuning (LP-FT Protocol)
Linear probe warm-up (5 epochs, backbone frozen) then full fine-tuning (10x lower LR for backbone, 1x for head). Stratified Cox loss compares patients ONLY within the same cancer type. This means the model cannot learn "GBM patients die faster than TBCA" as a survival signal — every C-index point earned requires discriminating within a single cancer type, ensuring the learned features reflect real within-cancer biology.

iii. Attention Mechanism in the Cancer-Conditioned Transformer



Self-attention: modalities attend to each other. Figure from Raschka, S. Understanding and Coding the Self-Attention Mechanism of Large Language Models From Scratch. Sebastian Raschka Blog, 2025.

Cross-attention: context token queries modalities. Figure from Sonti, S. Cross-Attention in Transformer. Medium, September 6, 2024.

Self-Attention (intra-modality communication)

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}(QK^T/\sqrt{d_k})\cdot V$$

$T = [e_{context}, e_{WSI}, e_{rna}, e_{cnv}, e_{snv}, e_{clin}, e_{rag}]$ → the 6 modality token matrix (tokens). $Q = W_Q \cdot T$, $K = W_K \cdot T$, $V = W_V \cdot T$ → learned linear projections. Each token asks (query), key, and value spaces. $Q \cdot K$ (model) = $e_{rna} \cdot e_{cnv}$ (dot product) → high dimension (4 heads); dividing by $\sqrt{d_k}$ prevents dot-products from saturating the softmax at very large magnitudes, stabilizing early training. After self-attention, each modality token is updated with a weighted sum of all other tokens' content, discovering that e.g. WSI morphology patches become more prognostically informative alongside specific KRAS mutation signals.

Cross-Attention (cancer-conditioned context token)

$$c^* = \mathrm{softmax}(W_Q c \cdot (W_K T)^T/\sqrt{d_k}) \cdot W_V T$$

c → the context token, formed from a learned base embedding plus biology RAG embedding plus text embedding encoding cancer type prior. Within extended transformer layers, tokens are gated via learnable sigmoid gates initialized at near zero, so internal priors are off by default and can be learned to be used through training. c^* → the enriched context after attending over all 6 modality tokens with cancer-type-biological priors as the query. Zero-

iv. Cancer-Type Holdout Evaluation Protocol

TCGA · 33 Cancer Types · 6,592 Patients

Training Set: 10 anchor cancers · 2,137 pts
BRCA · GBM · KIRC · LGG · LIHC · LUAD · LUSC · OV · SKCM · STAD

Stratified Cox loss — within-cancer comparisons only. BS/20 patient split for train/val. No test cancer type appears here.

↓ Backbone Frozen

Zero-Shot Test: 21 unseen cancers · 3,455 pts

- Tier 1: Molecularly Close (n=4): KIRP · LGG · LUSC · READ
- Tier 2: Distant (n=3): ACC · CESC · DLBC
- Tier 3: Novel Biology (n=15): BLCA · CHOL · COAD · ESCA · GBCA · HNSC · KICH · LAML · MESO · PAAD · PCPG · PRAD · SARC · TGCT · THCA · UCEC · UCS · UVM

Triple redundancy checks: no patient ID, no cancer type, and no biological prior from any test cancer appears in training. This is the strictest possible holdout used for pan-cancer foundation model evaluation.

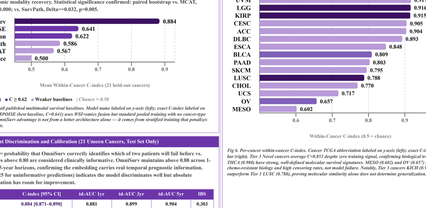
Fig 2. The 21 test cancers are held out entirely: no patient, cancer type, or biological prior from these types enters training. At test time, the model receives no cancer-type label — cancer-type information must be inferred from embedding only. Tiers are based on molecular similarity of test cancers to the training set based on RAG embedding cosine similarity. This protocol is fundamentally stricter than any published multimodal survival benchmark and directly tests whether survival supervision teaches transferable tumor biology.

IV. Result 1 — Zero-Shot Survival Prediction — Does OmniSurv generalize survival prediction to cancer types it has never encountered during training?

0.884 Mean C-index 21 unseen cancers (test)	**0.904** td-AUC 5yr long-horizon risk
0.881 td-AUC 1yr temporal discrimination	**p<0.001** Log-rank high vs. low risk, all 21

i. Baseline Comparison — All 6 Models Labeled on Y-Axis

What this shows: All baselines (MCAT, SurvPath, CATfusion, PORPOISE, and random Chance) are evaluated on the same 21 held-out cancers as OmniSurv. C-indices is computed within each cancer type independently for all models. OmniSurv's +0.243 advantage over PORPOISE (best baseline) reflects the benefit of stratified training and genomic modality recovery. Statistical significance confirmed: paired bootstrap vs. MCAT, Delta=+0.051, p=0.006; vs. SurvPath, Delta=+0.032, p=0.005.



ii. Time-Dependent Discrimination and Calibration (21 Unseen Cancers, Test Set Only)

td-AUC at time t = probability that OmniSurv correctly identifies which of two patients will fail before vs. after time t. Values above 0.80 are considered clinically informative. OmniSurv maintains above 0.88 across 1-year, 3-year, and 5-year horizons, confirming the embedding carries real temporal prognostic information. IBS=0.202 (vs. 0.25 for uninformative predictions) indicates the model discriminates well but absolute probability calibration has room for improvement.

Split	C-index [95% CI]	td-AUC 1yr	td-AUC 3yr	td-AUC 5yr	IBS
Test (21 unseen)	0.884 [0.871–0.898]	0.881	0.891	0.904	0.202

Fig 4. All metrics on held-out test set only. 95% CI computed via 1000-iteration bootstrap. td-AUC above 0.88 across all clinical horizons confirms real temporal prognostic signal beyond simple risk ranking. IBS improvement over uninformative baseline (0.25) confirms some calibration, though isotonic calibration post-hoc is recommended before clinical deployment.

iii. Per-Cancer C-index — All 21 Held-Out Cancer Types Labeled

How to read: Each bar is one held-out cancer type, labeled with its TCGA abbreviation on the left y-axis. The C-index measures within-that-cancer ranking quality only; no inter-cancer comparison is made. All 21 bars exceed chance (0.5), Color encodes molecular similarity tier to the training set.

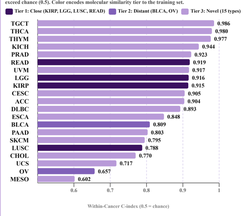


Fig 6. Per-cancer within-cancer C-index. Cancer TCGA abbreviation labeled on y-axis (left); exact C-index value labeled on each bar. Color encodes molecular similarity to training set: Tier 1 (close, n=4), Tier 2 (distant, n=2), Tier 3 (novel, n=15). Tier 1 (KIRP, LGG, LUSC, READ) have strong, well-defined molecular survival signatures. MESO (0.602) and OV (0.657) have heterogeneous, chemo-resistant biology and high censoring rates, not model failure. Notably, Tier 3 cancers KICH (0.944) and PRAD (0.923) outperform Tier 1 LUSC (0.788), proving molecular similarity alone does not determine generalization.

iv. Embedding Space: Generalization Tier Structure (t-SNE)

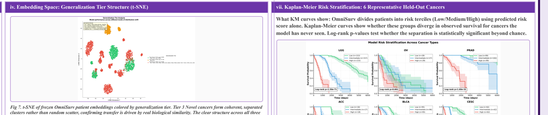


Fig 7. t-SNE of frozen OmniSurv patient embeddings colored by generalization tier. Tier 1 Novel cancers form coherent, separated clusters rather than random scatter, confirming transfer is driven by real biological similarity. The clear structure across all three tiers confirms that OmniSurv organizes embedding space by tumor biology, not by training membership — a critical sanity check that the model is not simply memorizing training-cancer features and projecting them onto test data.

v. Cancer Classification and Patient Retrieval (Frozen Embedding, No Fine-Tuning)

Why this matters: If the survival-trained embedding is truly general, it should organize patients by cancer type even though it was never trained on cancer-type labels. Zero-shot classification uses nearest training-cancer centroid (prototype) with no labeling or fine-tuning. Success means cancer-type features emerged as a byproduct of survival training — not because we told the model what cancer each patient had.

Task (frozen backbone, no fine-tuning)	Metric	Value	Why it matters
21-cancer type (nearest-prototype)	Accuracy	95.4%	21 unseen types, no fine-tuning at all
21-cancer type (nearest-prototype)	Macro-F1	76.4%	Average equally across all 21 cancer types
Few-shot k=5 (logistic probe)	Accuracy	97.3%	Only 5 labeled examples per unseen class
Few-shot k=10	Accuracy	99.9%	Near-perfect with just 10 labeled examples
Patient retrieval recall@10	Recall	96.2%	Top-10 neighbors are same cancer type
Patient retrieval mAP@10	mAP	99.2%	Precision at each rank position, averaged
Patient retrieval NDCG@10	NDCG	99.7%	Rank-weighted retrieval quality score

Fig 5. Zero-shot classification reaches 95.4% accuracy across 21 unseen cancer types without any fine-tuning. Few-shot k=10 reaches 99.9% with just 10 labeled examples. Patient retrieval mAP@10=99.2% means top-10 nearest neighbors are almost always the correct cancer type. These metrics collectively demonstrate the embedding is densely organized by tumor biology, not narrowly tuned to survival alone.

vi. Kaplan-Meier Risk Stratification: 6 Representative Held-Out Cancers

What KM curves show: OmniSurv divides patients into risk terciles (Low/Medium/High) using predicted risk score alone. Kaplan-Meier curves show whether these groups diverge in observed survival for cancers the model has never seen. Log-rank p-values test whether the separation is statistically significant beyond chance.

Fig 8. KM stratification across 6 held-out cancers: LGG, OV, PRAD, ACC, BLCA, CESC. All six show statistically significant risk-cancer separation (log-rank p<0.001). LGG shows strongest separation, consistent with IDH-driven glioma biology that OmniSurv implicitly encoded from other cancer types. OV separation is weaker but significant, reflecting complex, chemo-resistant biology with high censoring. PRAD separation is remarkable given only 6 of observed events in 100 patients — demonstrating strong discriminative power in sparse-event settings.

vii. Pooled KM: High vs. Low Risk Across All 21 Held-Out Cancers

Fig 9. Pooled KM across all 21 held-out cancers. High-risk group: median survival approx. 709 days. Low-risk group: no observed events to follow-up (median = infinity). Log-rank p<0.001. The HR point estimate is numerically extreme due to near-complete risk group separation; primary evidence is the magnitude and consistency of curve divergence. This confirms OmniSurv creates clinically separable risk groups in cancer types it has never seen during training. The strong separation indicates the 128-dim embedding carries robust prognostic information across diverse tumor contexts.

vii. Significance: What Generalization Proves

- **Standard benchmarks systematically overstate readiness. Cross-validation on pooled TCGA allows cancer-level shortcut learning. Our holdout reveals the same models score 0.567–0.641 on unseen cancers, a −0.15-point gap that has been invisible in published benchmarks. This calls for field-wide revision of evaluation standards.**
- **OmniSurv does not interpolate from nearby cancers. Tier 3 Novel cancers achieve mean C=0.853 with no training signal from those cancer types. The model captures molecular features that transfer across biological contexts, not just proximity to training cancers in representation space.**
- **Rare cancer implication. Rare malignancies (MESO, CHOL, DLBC, ACC) that would never accumulate enough patients for a dedicated survival model can still receive meaningful prognostic stratification from a generalizing pan-cancer model. This is a genuine clinical unmet need that OmniSurv directly addresses.**

V. Result 2 — Does Survival Encode General Cancer Biology? — Do frozen survival embeddings contain latent information about mutations, immune state, and molecular organization?

i. Driver Biomarker Recovery: All 20 Mutation Names on Y-Axis (AUROC, Frozen Linear Probe)

Protocol: A logistic regression probe trained on the frozen 128-dim patient embedding predicts whether a specific driver gene is mutated. No backbone retraining; zero mutation labels during survival training. AUROC above 0.80 indicates strong discrimination; 0.50 is random. The fact that these signals emerge from survival optimization alone means the embedding implicitly encodes mutation-linked tumor behavior because driver mutations strongly influence within-cancer prognosis. Mean AUROC across all 20 genes: 0.869 — substantially above chance across diverse mutation types and cancer contexts.

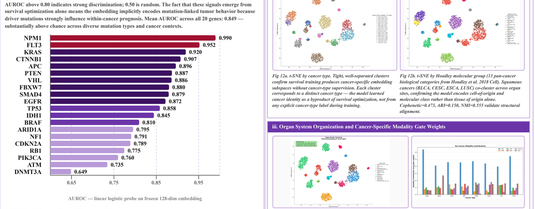


Fig 10. Driver mutation AUROC. Gene mutations are easily fully recent AUROC on their triplets. NF1 (0.990) and FLT3 (0.952) are strongly expressed in pan-cancer embeddings (AML in training). Because NF1/FLT3 mutations are mechanistically central to AML survival trajectories learned through L_OSL, KRAS (0.920), KRAS-driven cancers (pancreatic, colorectal), lungs have distinct survival trajectories that the embedding must learn to discriminate. DNMT3A (0.649) is lowest, consistent with its relatively weak prognostic signal across most cancer types. ALL 20 gene names labeled on y-axis. ALL FIGURES MADE BY MOBAREJI ABEJIDE.

ii. Embedding Space: Cancer-Type and Healthy Pan-Cancer Molecular Organization

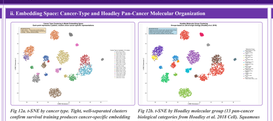


Fig 12a. t-SNE by cancer type. Tight, well-separated clusters confirm survival training produces cancer-specific embedding geometry. Each color represents a different cancer type. Cluster corresponds to a distinct cancer type — the model learned cancer identity as a byproduct of survival optimization, not from any explicit cancer-type label during training.

Fig 12b. t-SNE by Hoadley molecular group (13 pan-cancer biological categories from Hoadley et al. 2018 Cell). Squamous (Cophenetic=0.975, ARI=0.731, NMI=0.757) forms the most distinct cluster, confirming the model encodes cell-of-origin and molecular class rather than tissue of origin alone. Cophenetic=0.975, ARI=0.730, NMI=0.757 validate structural alignment.

iii. Organ System Organization and Cancer-Specific Modality Gate Weights

Fig 13. t-SNE by organ system on test cancers. GI cancers (COAD, ESCA, PAAD, READ), thoracic (LUSC, MESO), and urological cancers (BLCA, KICH, KIRP, PRAD, TGCT) form anatomically coherent clusters among unseen test cancers. Multi-level organization (cancer type, organ system, and molecular group simultaneously) emerges from survival training alone.

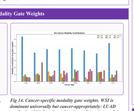


Fig 14. Cancer-specific modality gate weights. WSI is dominant universally but cancer-appropriately: LUAD emphasizes histology (growth patterns determine grade); PRAD emphasizes CNV (AR amplification is the key driver); LIHC shows the most balanced profile. These hierarchies emerge without explicit modality supervision, confirming the transformer learns which data sources matter for each cancer context.

iv. Comprehensive Biological Evidence: What the Frozen Embedding Encodes

Beyond driver mutations, the frozen embedding is probed for biological signals it was never trained to predict. These results test whether survival is a broad or narrow supervisory signal. Broad signals (TMB, immune activity, pan-cancer taxonomy) emerge strongly. Narrower signals (specific immune subtype labels, stemness) are partially encoded — survival is related but not uniquely determined by these features.

Biological Alignment Task	Value	Significance and Interpretation
TMB Regression (Spearman r)	0.518	p<1e-121: embedding encodes total mutation burden without direct mutation-count training. R²=0.865 — remarkable for a survival-only model.
TMB High/Low AUROC	0.977	Binary: high vs. low mutational burden; near-perfect discrimination from frozen features alone.
TME Immune Activity (Spearman)	0.573	Tumor microenvironment immune activity proxy; p<1e-29. Survival correlates with immune evasion so the embedding partially captures it.
Thorsson Immune Subtype (F1)	0.130	6 immune subtypes (C1-C6); moderate. Survival encodes immune context partially but not exclusively.
Stemness Axis (Spearman r)	0.435	mRNA/ext stemness index; p=0.041, barely significant. Stemness likely requires direct label supervision.
Hoadley Cophenetic Corr.	0.473	Hierarchical alignment with published pan-cancer taxonomy (Hoadley et al. 2018 Cell). Positive correlation validates emergent pan-cancer organization.
Hoadley Adjusted Rand Index	0.150	Non-trivial structure alignment adjusted for chance; ARI=0 would be random. Confirms real systematic correspondence.
Hoadley NMI	0.555	Normalized mutual information with known pan-cancer molecular groups; p<0.01. Strongest pan-cancer biology metric.
Subtype Clusters (log-rank p)	<0.001	k=2,3,4,5 all significant; unsupervised clustering of embedding creates groups with different survival curves.

Fig 11. Biological alignment metrics from downstream probe experiments using frozen embeddings only. TMB R²=0.865 means 86.5% of variance in tumor mutational burden is explained by survival-trained features alone, exceptional for a model never trained on mutation data. Pan-cancer Hoadley NMI=0.555 shows the learned clusters recover molecular groups from a published pan-cancer taxonomy derived by Hoadley et al. from extensive multiomic profiling. Modest immune subtype (F1=0.130) and stemness (r=0.435) results are honest limitations: survival encodes broad but not all aspects of tumor biology.

v. LGG IDH Subtype Discovery Without Subtype Labels (IDH Purity 0.89)

Fig 15. Unsupervised k-means on LGG survival embeddings recovers IDH mutation status (cluster purity=0.89, log-rank p=0.001) without any subtype labels. IDH mutation is the single strongest prognostic marker in glioma — IDH-mutant patients live 2-3x longer than IDH-wildtype. OmniSurv's embedding must have captured this molecular boundary because IDH-mutant and IDH-wildtype tumors exhibit different survival trajectories driven by biological distinctions critical to discriminating patients within LGG. This is direct evidence that survival supervision encodes deep, clinically meaningful molecular biology: not just survival risk, but the specific molecular lesion responsible for that risk.

vi. BRCA PAM50 Subtype Recovery: KM, PCA, and t-SNE

Why PAM50 matters: Basal-like, HER2+, Luminal A/B, and Normal-like subtypes are clinically actionable, determining chemotherapy choice, HER2-targeted therapy eligibility, and prognosis. These subtypes emerged from decades of gene expression research. OmniSurv recovers them from survival training alone because PAM50 subtypes define different survival trajectories — the embedding must learn these distinctions to rank patients correctly within BRCA.

Fig 16. BRCA PAM50 subtype recovery. A: KM curves show significant survival separation (log-rank p=8.1le-04) between PAM50 subtypes recovered from unsupervised clustering of frozen embeddings. B: PCA shows linear separability — these subtypes are partially linearly separable in the embedding. C: t-SNE confirms tight, pure subtype clusters. No PAM50 labels were used during training or probing — subtype recovery emerges entirely from survival-induced feature organization. ALL FIGURES MADE BY MOBAREJI ABEJIDE.

vii. LUAD Driver Mutation Groups: KM, PCA, and t-SNE

Why LUAD drivers matter: EGFR-mutant, KRAS-mutant, and TP53-mutant lung adenocarcinoma have distinct treatment responses and survival trajectories. EGFR-mutant patients qualify for targeted EGFR inhibitors; KRAS-G12C inhibitors now exist. OmniSurv recovers these driver-defined subtypes from survival training because they create distinct survival trajectories within LUAD that the model must learn to discriminate.

*Fig 17. LUAD driver mutation subtype recovery. A: KM curves show significant separation (log-rank p=1.06e-03). B: PCA shows embedding-space clustering by mutation status (silhouette=0.177, purity=1.000). C: t-SNE confirms clean separation