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

Mobareji Abejide, Noblesville High School, Noblesville, Indiana

**All figures generated by Mobareji Abejide unless stated otherwise*

Problem

- Modern cancer AI often depends on **task-specific labeled datasets** for mutations, subtypes, or treatment response, which makes it expensive and difficult to scale across oncology.
- This is especially challenging for **rare cancers**, which the National Cancer Institute defines as cancers with incidence below 15 cases per 100,000 people per year; although individually uncommon, they account for **about 25% of new adult cancer cases** in the United States.
- At the same time, many models are evaluated on cancers seen during training, so strong performance may reflect memorizing disease-specific patterns rather than learning shared tumor biology.

I asked:

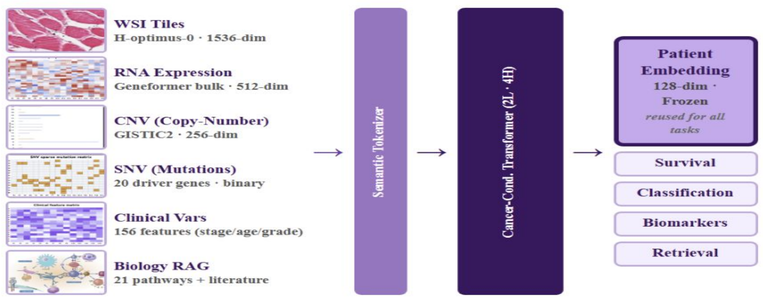
Can we use a global supervisory signal (survival) for learning transferable tumor biology?

Can a model trained on some cancers generalize to entirely unseen cancers without cancer-type labels at inference?

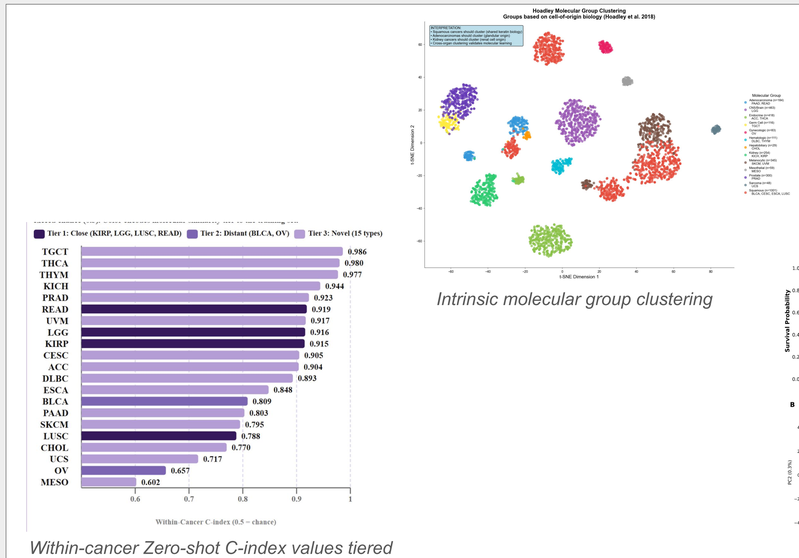
Does a survival-trained embedding recover broader biology, such as biomarkers, immune state, and latent subtypes?

Methodology

I built OmniSurv, a six-modality multimodal model that integrates histopathology, RNA expression, copy-number variation, somatic mutations, clinical variables, and retrieval-augmented biology text/pathway context into a shared patient embedding. The system uses norm-balanced multimodal tokenization, cancer-conditioned attention-based fusion, a dedicated genomic-recovery path to prevent omics collapse, and a **stratified Cox survival objective** so patients are compared **within cancer type** during training rather than across cancers. Additionally, the model is trained on **10 anchor cancers** and evaluated on **22 entirely unseen cancers** under strict cancer-type holdout, rather than conventional pooled cross-validation.This makes shortcut learning harder and encourages the model to learn transferable aggressiveness biology instead of simply memorizing which cancers are globally worse. At inference, the model remains cancer-agnostic and can use prototype neighbor information for safer transfer.

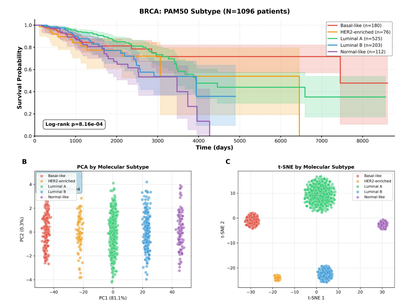


Data analysis and results



Within-cancer Zero-shot C-index values tiered by molecular distance from training data.

Intrinsic molecular group clustering



Unsupervised subtype recovery for Breast Cancer

Interpretation & Conclusions

- OmniSurv achieved strong zero-shot survival prediction on unseen cancers (C-index 0.8836, mean td-AUC 0.8957), beating state-of the art multimodal cancer survival baselines and showing that the model can rank patient risk well even without seeing those cancers during training.
- The embedding also captured structure beyond prognosis. It showed strong **retrieval and ranking performance**, supported **subtype discovery and meaningful clustering**, and aligned with known biological organization through **Hoadley cancer structure analyses**. Additional probe tasks showed recovery of external biology such as **immune state, stemness, and pathway / mutation-related signals** from the frozen embedding.
- Overall, these findings support the conclusion that OmniSurv learned **shared biological signals of tumor aggressiveness**, not just cancer-specific shortcuts.
- **Impact:** a single survival-trained multimodal backbone could reduce dependence on many separate labeled datasets and make AI more scalable for **rare, low-resource, and previously unseen cancers**.