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

Mobareji Abejide

Noblesville High School, Noblesville, Indiana

Motivation / Problem

Modern cancer AI usually depends on many disparate task-specific labeled datasets, supervised models, and validation pipelines for different problems (eg. mutation prediction, subtype classification)

Many published models are evaluated on cancers that also appear in training -> Strong performance may reflect memorizing cancer-specific patterns rather than learning the shared biological mechanisms that make tumors aggressive.

Survival offers a different possibility. Unlike specialized labels, survival outcomes are relatively common in research datasets and reflect the combined effects of many biological processes, including driver mutations, genomic instability, immune state, and therapy resistance.

Research Questions and Hypotheses

Q1: Can survival act as a global supervision signal for learning transferable tumor biology leveraging deep learning?

Q2: Can a model trained on some cancers generalize to entirely unseen cancers without being told the cancer type at inference?

Q3: Does the learned embedding capture broader biology, such as biomarkers, immune state, and latent subtypes, even without direct supervision on those tasks?

Hypothesis

Because survival reflects many underlying biological processes related to tumor aggressiveness, a multimodal model trained on survival should learn a general biological embedding that:

- supports zero-shot survival prediction in unseen cancers,
- preserves meaningful molecular and clinical structure,
- and can be reused for downstream oncology tasks with simple probes.

# OmniSurv: Multimodal Survival-Supervised Representation Learning

To test this idea, I built OmniSurv, a multimodal model that integrates six patient data modalities into a shared patient embedding:

- Pathology images to capture tumor morphology
- RNA expression to measure transcriptional activity
- Copy-number variation (CNV) to reflect genomic instability
- Somatic mutation (SNV) data to capture driver alterations
- Clinical variables to provide patient-level context
- Biology/text context from curated pathways and literature

Each modality is first encoded into a common token space. These tokens are then fused by cancer-conditioned transformer blocks, where:

- modality tokens first interact through self-attention,
- a context token then uses cross-attention to pool information across modalities,
- and the final context token becomes the patient embedding used for survival prediction and downstream tasks.

Training uses a stratified Cox survival loss, which compares patients within the same cancer during training so the model cannot rely on simple between-cancer survival shortcuts.
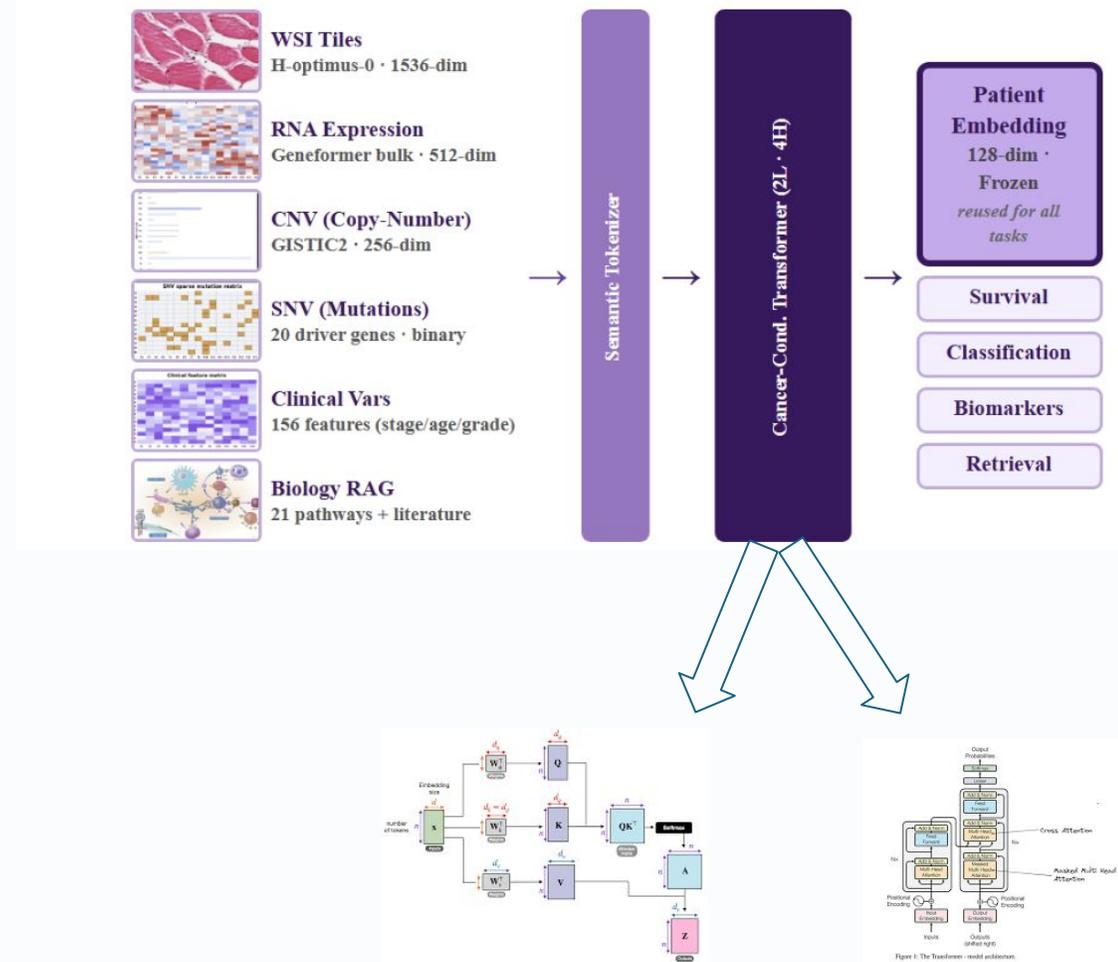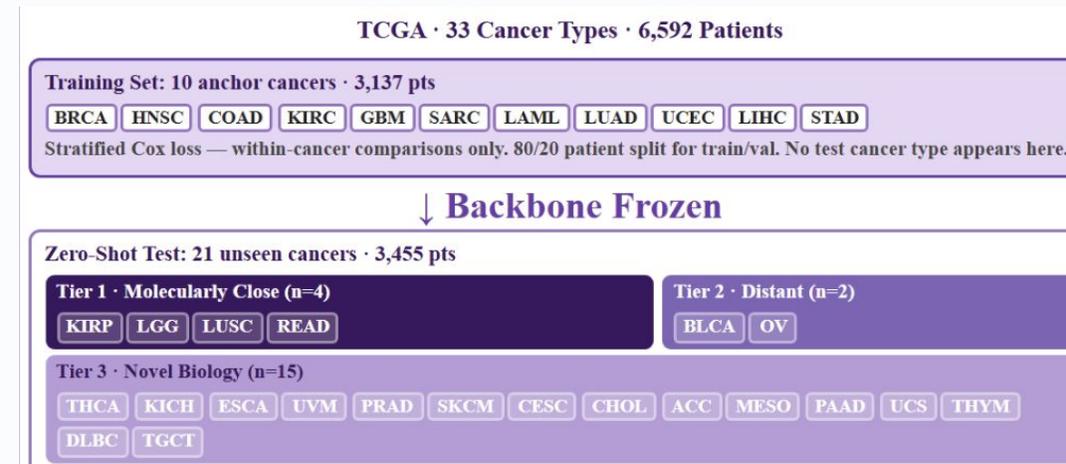
Evaluation Protocol

The central goal of this project is not just to predict survival on familiar data, but to test whether the model learns transferable tumor biology.

To do that, I used a cancer-type holdout protocol instead of standard pooled cross-validation:

- The model was trained on only 10 anchor cancers
- It was then evaluated on 22 completely unseen cancers
- At inference, the model was not given the cancer type

This creates a zero-shot setting where the model cannot rely on memorized disease-specific patterns. Instead, it must generalize using biological signals that are shared across cancers

To make this evaluation more honest, the primary metric is within-cancer C-index, which measures whether the model can correctly rank patient risk inside each unseen cancer, rather than benefiting from broad survival differences between cancer types.

TCGA · 33 Cancer Types · 6,592 Patients

Training Set: 10 anchor cancers · 3,137 pts

BRCA | HNSC | COAD | KIRC | GBM | SARC | LAML | LUAD | UCEC | LIHC | STAD

Stratified Cox loss — within-cancer comparisons only. 80/20 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=2)

BLCA | OV

Tier 3 · Novel Biology (n=15)

THCA | KICH | ESCA | UVM | PRAD | SKCM | CESC | CHOL | ACC | MESO | PAAD | UCS | THYM
DLBC | TGCT

$$C\text{-index} = \frac{\sum_{i,j} 1_{T_j < T_i} \cdot 1_{\eta_j > \eta_i} \cdot \delta_j}{\sum_{i,j} 1_{T_j < T_i} \cdot \delta_j}$$

Retrieved from Deotte, C. (2024). C-index calculation formula [Image]. Kaggle.
https://www.kaggle.com/competitions/equity-post-HCT-survival-predictions/discussion/554225

Hypothesis 2

We hypothesize that [independent variable] will [expected effect] on [dependent variable] in [specific population or condition].

Research Objectives

| Research objective 1. For example, "To determine the impact of cytokine blockers in the prevention of graft-versus-host disease (GVHD) following bone marrow transplant." **Tip: Highlight** or **bold** keywords to emphasize your point.

|| Research objective 2. For example, "To determine the impact of cytokine blockers in the prevention of graft-versus-host disease (GVHD) following bone marrow transplant."
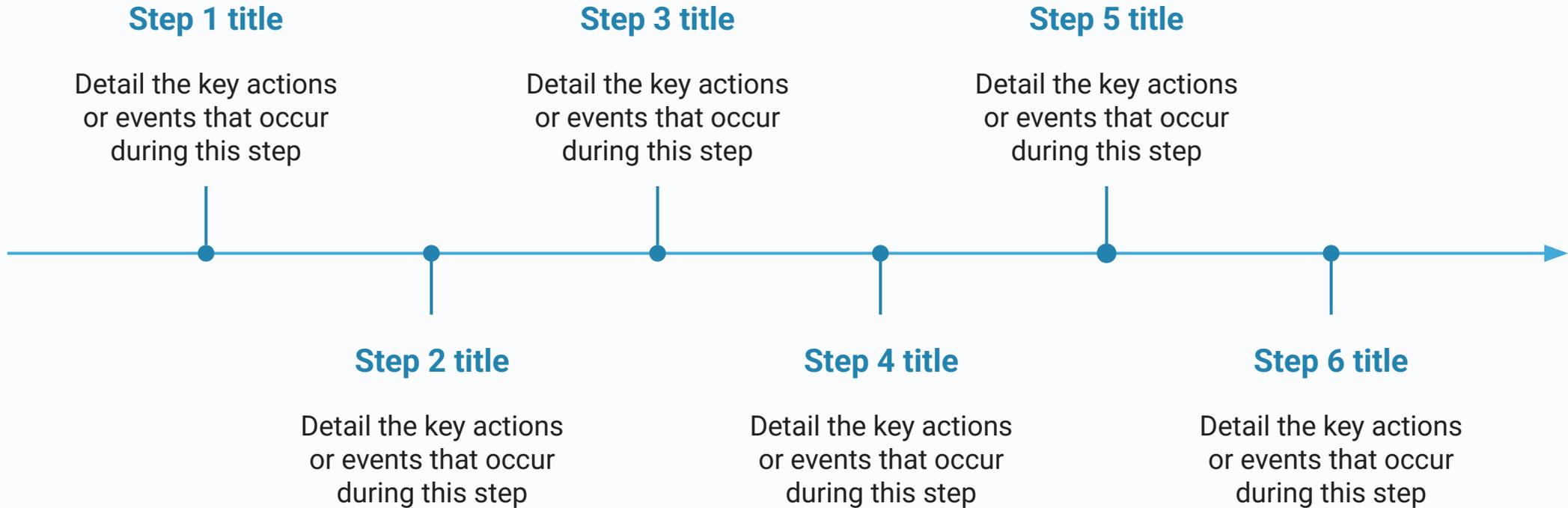
Research Methodology

Provide a brief overview of your research approach or the goal of the methodology.

• Overview of your experimental design

• **Tip:** Create visual protocols and/or timeline to demonstrate your methodology using BioRender

Use the Resource section to swap this figure

Experiment Timeline

Step 1 title

Detail the key actions
or events that occur
during this step

Step 3 title

Detail the key actions
or events that occur
during this step

Step 5 title

Detail the key actions
or events that occur
during this step

Step 2 title

Detail the key actions
or events that occur
during this step

Step 4 title

Detail the key actions
or events that occur
during this step

Step 6 title

Detail the key actions
or events that occur
during this step

Detailed Result 1 (for example, "Effect of Treatment A on Survival")

- Provide a brief explanation of your interpretation of the data

- Use less words and more visuals

- **Tip:** Ensure each slide focuses on one key result or set of related results to avoid overcrowding

- **Tip:** Create clear, beautiful visualizations of your research data with BioRender
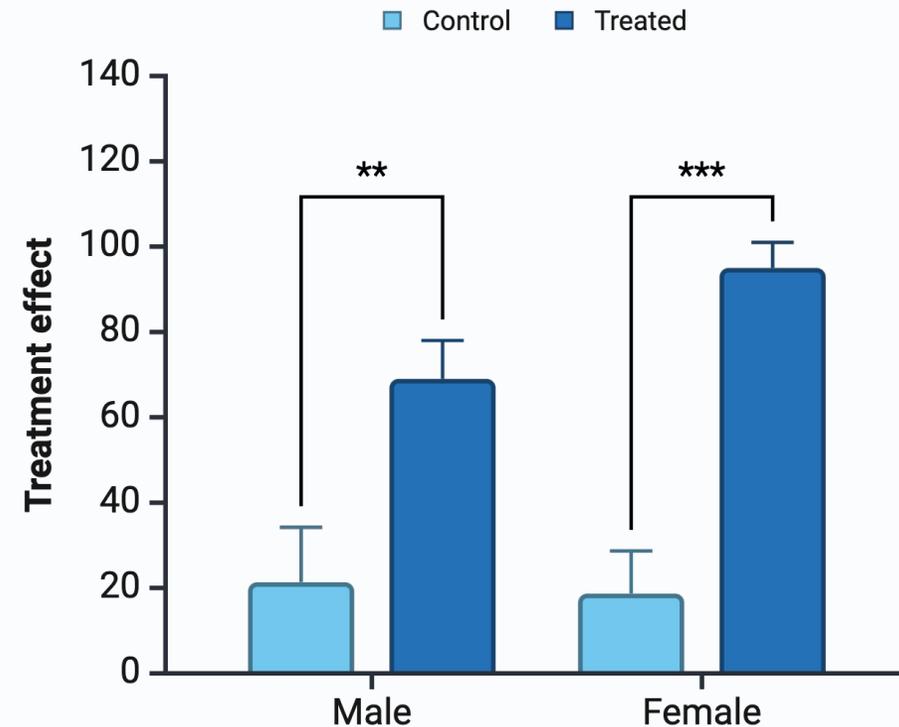
References

1. Author, A. A., Author, B. B., & Author, C. C. (Year). Title of the article. Title of the Journal, volume number(issue number), page range. https://doi.org/xx.xxx/yyyy

2. Author, A. A., Author, B. B., & Author, C. C. (Year). Title of the article. Title of the Journal, volume number(issue number), page range. https://doi.org/xx.xxx/yyyy

3. Author, A. A., Author, B. B., & Author, C. C. (Year). Title of the article. Title of the Journal, volume number(issue number), page range. https://doi.org/xx.xxx/yyyy

4. Author, A. A., Author, B. B., & Author, C. C. (Year). Title of the article. Title of the Journal, volume number(issue number), page range. https://doi.org/xx.xxx/yyyy

5. Author, A. A., Author, B. B., & Author, C. C. (Year). Title of the article. Title of the Journal, volume number(issue number), page range. https://doi.org/xx.xxx/yyyy

6. Author, A. A., Author, B. B., & Author, C. C. (Year). Title of the article. Title of the Journal, volume number(issue number), page range. https://doi.org/xx.xxx/yyyy

7. Author, A. A., Author, B. B., & Author, C. C. (Year). Title of the article. Title of the Journal, volume number(issue number), page range. https://doi.org/xx.xxx/yyyy

8. Author, A. A., Author, B. B., & Author, C. C. (Year). Title of the article. Title of the Journal, volume number(issue number), page range. https://doi.org/xx.xxx/yyyy