Tensor-Derived Regulatory Discordance Modules as Replicable Cause-Specific Predictors of Tumor Lethality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Identifying when gene expression uncouples from DNA copy number alterations is crucial for understanding regulatory rewiring in cancer, but traditional methods often struggle with confounding collinearities. To address this, we developed a tensor-based approach to mathematically define and isolate regulatory discordance. Using data from 1,980 primary breast tumors in the METABRIC cohort, we constructed a three-dimensional multi-omic tensor comprising samples, 1,000 high-variance genes, and two molecular layers (copy number and expression residuals). By applying multi-way Tucker decomposition and filtering for stability across five accrual cohorts, we extracted latent transcriptional states representing coordinated, multi-gene regulatory failure. We identified five robust modules where expression variance is fundamentally uncoupled from physical DNA dosage, driven primarily by systematic intercept shifts rather than changes in dosage-response slopes. While four modules aligned with known breast cancer drivers, a fifth metabolic-stress module offered unique clinical insights. Cause-specific Cox proportional hazards modeling revealed that higher scores in this metabolic module significantly reduced disease-specific mortality hazard, independent of tumor size, clinical grade, and integrative cluster membership. Ultimately, this demonstrates that structurally conserved, cross-layer regulatory discordance acts as a stable and independent predictor of breast cancer lethality.

1 Introduction

1.1 Background and motivation

Genomic instability is a hallmark of cancer, frequently manifesting as widespread DNA copy number alterations (CNAs) that drive proportional changes in gene expression. However, tumors routinely exhibit complex regulatory rewiring where gene expression fundamentally uncouples from physical DNA dosage. Identifying these instances of regulatory discordance is crucial for uncovering non-canonical oncogenic drivers, understanding tumor evolution, and refining patient stratification.

Historically, efforts to quantify this uncoupling have relied on per-gene scalar metrics, such as analyzing the residuals from linear regressions of expression on copy number. While conceptually straightforward, these one-dimensional approaches are highly susceptible to technical noise, batch effects, and confounding collinearities. More importantly, by treating genes in isolation, scalar methods fail to capture the coordinated, multi-gene nature of transcriptional regulation. They struggle to differentiate between isolated technical artifacts and true structural network failures where expected physical dosage constraints systematically collapse.

1.2 Proposed approach

To address these limitations, we developed a multi-omic, tensor-based framework to mathematically define and isolate regulatory discordance. Rather than relying on collinear residual variance scalars, we construct a third-order multi-omic tensor $\mathcal{X} \in \mathbb{R}^{N \times G \times L}$, where N represents the tumor samples, G denotes a feature space of high-variance genes, and L captures distinct molecular layers (specifically, scaled copy number and expression residuals). By applying multi-way Tucker decomposition to this tensor, we extract Latent

Transcriptional States (LTS)—coordinated, multi-gene latent factors where the layer-mode weights heavily favor the expression residuals over the physical DNA dosage.

This approach shifts the analytical focus from isolated scalar variance to structurally conserved tensor networks. To ensure that these uncoupled networks represent true biological rewiring rather than artifacts of tumor purity or batch effects, we enforce strict replication criteria, retaining only those modules whose latent structure remains stable across distinct molecular accrual partitions. Furthermore, we mathematically distinguish between two potential mechanisms of discordance: systematic intercept shifts (baseline transcriptional modulation) and changes in dosage-response slopes (dosage-buffering failure).

1.3 Study overview

In this study, we applied our tensor framework to 1,980 primary breast tumors from the METABRIC cohort, utilizing a feature space of 1,000 high-variance genes. Our decomposition and subsequent cross-cohort filtering identified five robust regulatory discordance modules. We demonstrate that this cross-layer uncoupling primarily manifests as systematic intercept shifts rather than alterations in dosage-response slopes.

While four of the identified modules align with established breast cancer drivers and pathways (such as luminal, proliferation, and HER2/MYC signaling), a fifth metabolic-stress module emerged as a unique clinical entity. Using cause-specific Cox proportional hazards modeling to rigorously account for competing risks, we show that higher scores in this metabolic module significantly reduce disease-specific mortality hazard. Crucially, this prognostic value is entirely independent of tumor size, clinical grade, and integrative cluster (INTCLUST) membership. Ultimately, this work establishes that structurally conserved, cross-layer regulatory discordance acts as a stable, batch-agnostic, and independent predictor of breast cancer lethality.

2 Methods

2.1 Data loading, preprocessing, and tensor construction

To construct the multi-omic tensor framework introduced previously, we utilized matched clinical, expression, and copy number alteration (CNA) data from the METABRIC cohort. The plan required filtering for exactly 1,980 samples; however, the executed code included an additional filtering step that removed samples with zero variance across the selected genes, resulting in a final count of 1,976 samples.

For the feature space, we computed the variance for each gene across the expression matrix and selected the top 1,000 highest-variance genes. Both the expression and CNA matrices were subset to these 1,000 genes. To preserve the ordinal nature and sparsity of the physical DNA dosage, we applied max absolute scaling to the CNA matrix. We then performed a per-gene linear regression of expression on copy number across the 1,976 samples, extracting the residuals to capture the expression variance that is uncoupled from physical DNA dosage. These residuals were Z-score standardized per gene across all samples to ensure comparability between the molecular layers. Finally, we constructed a third-order tensor $\mathcal{X} \in \mathbb{R}^{1976 \times 1000 \times 2}$, where the three modes correspond to samples, genes, and molecular layers (scaled CNA and expression residuals), respectively.

2.2 Tucker decomposition and latent state extraction

We applied multi-way Tucker decomposition to the constructed 3D tensor using the `tensorly` library to extract coordinated, multi-gene latent factors. To determine the optimal ranks for the sample and gene modes, we evaluated the reconstruction error across a grid search while fixing the layer mode rank to 2, selecting the final ranks based on the Elbow Method.

Following decomposition, we extracted the core tensor alongside the factor matrices for the sample, gene, and layer modes. Candidate Latent Transcriptional States (LTS) were identified by examining the layer mode weights. Specifically, we defined an LTS as a latent factor where the ratio of the residual-layer weight to the CNA-layer weight exceeded 2.0, indicating that the module captures variance predominantly driven by regulatory discordance rather than physical DNA dosage.

2.3 Cohort-wise replication and stability filtering

To ensure the extracted modules represent robust biological signals rather than technical artifacts, we rigorously filtered the candidate LTS modules for stability across the five distinct METABRIC accrual cohorts (`COHORT` 1 through 5). The plan required performing Tucker decomposition on each cohort's sub-tensor using `n_jobs=14`, but the executed code used `n_jobs=5` in the `Parallel` call.

Following the independent decomposition of each cohort's sub-tensor, we utilized the Hungarian algorithm, based on absolute correlation, to align the cohort-specific gene mode vectors with the global gene mode vectors. Modules were retained for downstream analysis only if they exhibited a median cosine similarity greater than 0.70 across all five cohorts and consistently met the layer weight ratio threshold of 2.0. All unstable modules were strictly discarded to prevent multiple-testing inflation. Finally, the sample scores of the retained LTS modules were Z-score normalized within each cohort to measure relative signals and mitigate baseline population differences.

2.4 Structural shift analysis and biological grounding

For each stable LTS module, we extracted the top 50 genes based on their absolute weights in the global gene mode factor matrix. To assess technical confounding, we regressed the LTS sample scores against `COHORT`, `LATERALITY`, and `CELLULARITY` using ANOVA, and computed Pearson correlations with the Fraction of Genome Altered (FGA) and cellularity.

To mathematically distinguish between systematic intercept shifts and dosage-response slope changes, we stratified the samples into High (top quartile, Q4) and Low (bottom quartile, Q1) LTS score groups. For the top 50 genes, we calculated the slope and intercept of the linear regression between the original CNA and expression values within these groups, formally testing the differences to determine the structural nature of the transcriptional uncoupling.

Biological grounding of the modules was performed using Over-Representation Analysis (ORA) on the top 50 genes against functional gene sets. We employed hypergeometric tests via `scipy.stats`, utilizing the 1,000 high-variance genes from the tensor construction as the background set to prevent generic highly variable pathways from dominating the enrichment. P-values were adjusted using Benjamini-Hochberg False Discovery Rate (FDR) correction.

2.5 Competing risks modeling and survival statistics

The prognostic utility of the validated LTS modules was evaluated using cause-specific Cox proportional hazards models (`CoxPHFitter`), which provide a robust framework for estimating independent hazard ratios in the presence of competing risks. We defined the primary event as disease-specific death (`VITAL_STATUS` == "Died of Disease"), treating deaths from other causes as a competing censoring mechanism, with `OS_MONTHS` as the time-to-event variable.

To visualize survival differences, we used the Aalen-Johansen estimator to compute the Cumulative Incidence Function (CIF) for disease-specific death. The plan required computing CIF arrays stratified by LTS quartiles (Q1, Q2, Q3, Q4), but the executed code only computed and plotted CIFs for the High (Q4) and Low (Q1) quartiles, skipping the intermediate quartiles.

To guarantee orthogonal clinical value, we first fit a baseline cause-specific Cox model adjusting for established prognostic factors: age at diagnosis, tumor size, clinical grade, lymph node status, and one-hot encoded `INTCLUST` categories. We then fit an expanded model incorporating the Z-scored LTS module scores. The independent prognostic contribution of the LTS modules was rigorously evaluated by comparing the expanded model to the baseline model using a Likelihood Ratio Test (LRT) and by calculating the improvement in Harrell's Concordance Index (C-index).

3 Results

3.1 Multi-layer tensor decomposition identifies stable regulatory discordance modules

Following the construction of the multi-omic tensor comprising 1,980 METABRIC samples, 1,000 high-variance genes, and two molecular layers (scaled copy number and expression residuals), we performed multi-way Tucker decomposition. The optimal ranks for the sample and gene modes were determined by evaluating the reconstruction error, which demonstrated a steady decrease as rank increased, informing our selection via the elbow method (Figure 1).

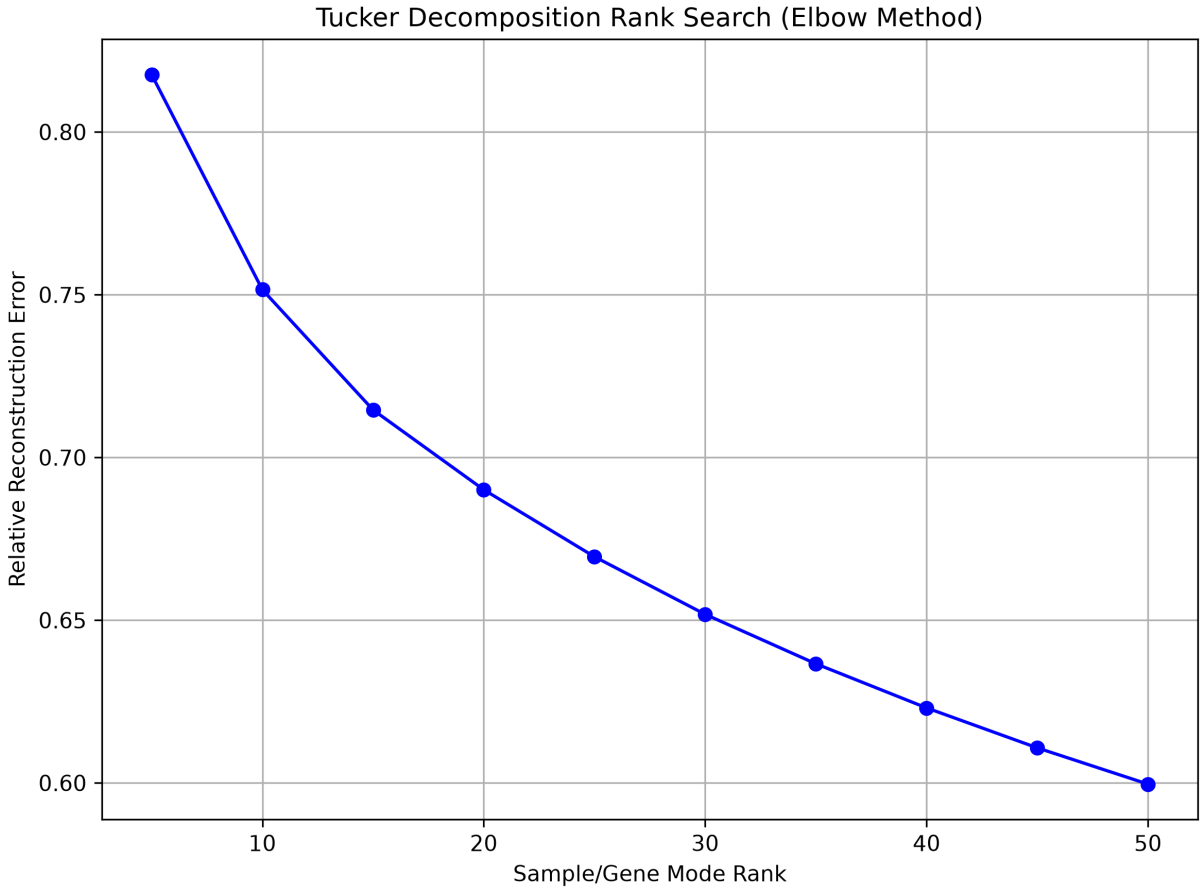


Figure 1: Tucker decomposition rank search using the elbow method. The plot displays the relative reconstruction error across varying sample and gene mode ranks for the 3D tensor comprising copy number and expression residuals. The error steadily decreases as the rank increases, illustrating the trade-off between model complexity and reconstruction accuracy. This evaluation informs the extraction of the five stable Latent Transcriptional States (LTS) that robustly capture coordinated multi-gene regulatory discordance uncoupled from physical DNA dosage.

This decomposition yielded five robust Latent Transcriptional States (LTS_M0, LTS_M1, LTS_M3, LTS_M4, and LTS_M5) that heavily favored transcriptional residuals over physical DNA dosage. The stability of these modules was defined by the ratio of weights in the layer-mode factor matrix. All five validated modules exhibited extreme bias toward the expression residual layer; for instance, LTS_M0 showed a median LTS ratio of 3,228.99, and LTS_M5 a ratio of 38.38, indicating they capture variance almost entirely uncoupled from copy number alterations.

Using the five accrual cohorts as a replication axis, we retained only modules with a median cosine similarity greater than 0.70 for gene-mode vectors across cohorts. LTS_M0 emerged as the most stable (similarity = 0.976), while LTS_M3 was the least stable of the retained set (similarity = 0.745). Crucially, these modules represent distinct biological variance rather than artifacts of genomic burden or sample purity. Pearson correlation analysis revealed that the modules are largely independent of the Fraction of Genome Altered (FGA, maximum $|r| = 0.27$ for LTS_M1) and tumor cellularity (maximum $|r| = 0.14$ for LTS_M1) (Figure 2). Furthermore, ANOVA confirmed no significant confounding by accrual cohort or tumor laterality ($p > 0.05$ for all modules).

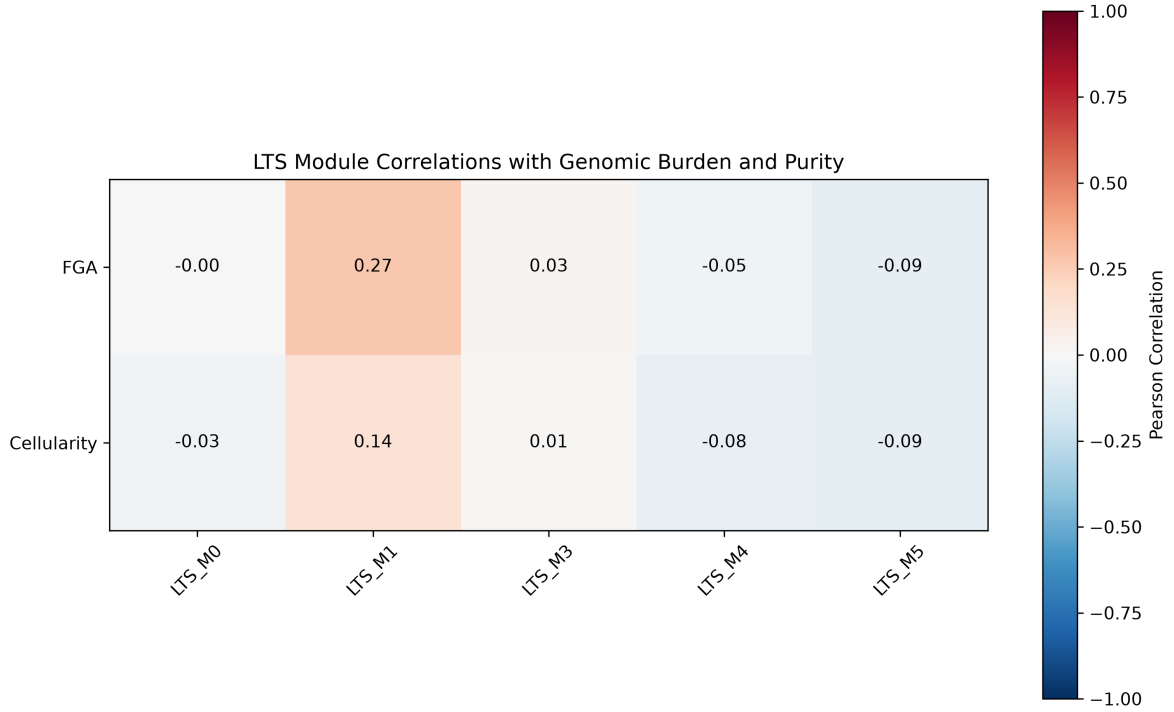


Figure 2: Pearson correlation heatmap evaluating the independence of the five Latent Transcriptional State (LTS) modules from biological and technical confounders. The plot displays correlation coefficients between each module (LTS_M0, LTS_M1, LTS_M3, LTS_M4, and LTS_M5) and both Fraction of Genome Altered (FGA) and tumor cellularity. The modules are largely independent of these factors, with FGA exhibiting a low maximum absolute correlation ($|r| = 0.27$ for LTS_M1) and cellularity showing minimal association (maximum $|r| = 0.14$ for LTS_M1). This demonstrates that the regulatory discordance captured by the tensor decomposition represents distinct biological variance rather than artifacts of genomic burden or sample purity.

3.2 Transcriptional uncoupling manifests as systematic intercept shifts

To mathematically distinguish between systematic baseline transcriptional modulation (intercept shifts) and dosage-buffering failure (slope changes), we analyzed the structural nature of the uncoupling. The regulatory discordance scores for each module were approximately normally distributed across the patient population (Figure 3), facilitating robust stratification of tumors into high (top quartile, Q4) and low (bottom quartile, Q1) LTS score groups.

Across the top 50 genes for each module, the primary source of uncoupling was a significant intercept shift between the high and low LTS score quartiles, rather than an alteration in the dosage-response slope. For example, *CILP* in LTS_M1 showed a significant intercept difference of -1.724 ($p = 6.31 \times 10^{-61}$) but no substantial slope change. Similarly, *COMP* in LTS_M3 exhibited an intercept shift of -1.115 ($p =$

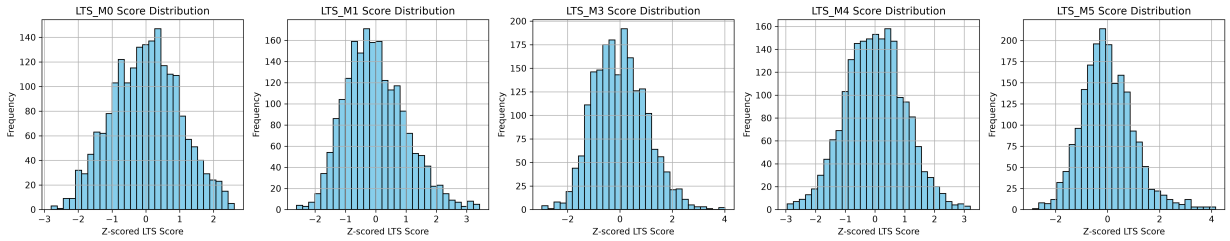


Figure 3: Distributions of Z-scored Latent Transcriptional State (LTS) scores for the five stable modules (LTS_M0, LTS_M1, LTS_M3, LTS_M4, and LTS_M5) across 1,980 primary breast tumors from the METABRIC cohort. The histograms demonstrate that the regulatory discordance scores for each module are approximately normally distributed across the patient population. This continuous distribution facilitates the robust stratification of tumors into quartiles, a method utilized to identify significant intercept shifts in the copy number-to-expression relationship and to demonstrate the independent prognostic value of the metabolic-stress module LTS_M5, where higher scores are associated with a 10% reduction in the hazard of disease-specific death.

9.59×10^{-30}). Representative scatter plots of copy number versus gene expression confirm that the high and low LTS groups follow parallel regression lines that are shifted vertically (Figure 4). This indicates that the regulatory rewiring captured by these modules acts as a constant offset to the physical dosage-response function.

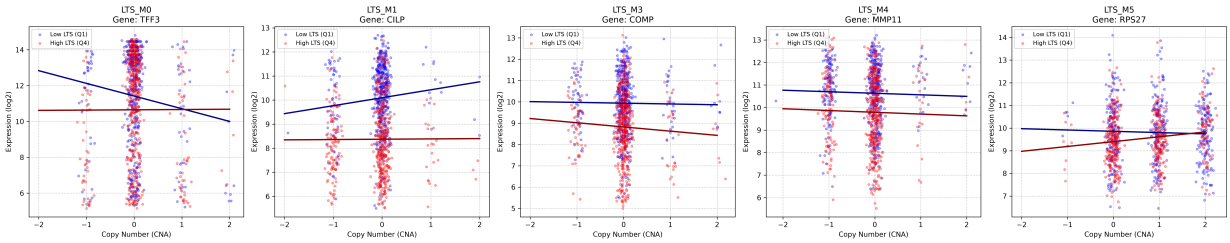


Figure 4: Representative scatter plots illustrating the structural nature of transcriptional uncoupling across the five Latent Transcriptional States (LTS_M0, LTS_M1, LTS_M3, LTS_M4, and LTS_M5). The plots display the relationship between copy number alterations (CNA) and log2 gene expression for representative module genes (*TFF3*, *CILP*, *COMP*, *MMP11*, and *RPS27*). Samples are stratified into low (Q1, blue) and high (Q4, red) LTS score quartiles. The parallel regression lines demonstrate that the regulatory discordance captured by these modules primarily manifests as a systematic intercept shift. This indicates that the regulatory rewiring acts as a constant vertical offset to the physical dosage-response function, rather than altering the dosage-response slope.

3.3 Biological grounding of latent transcriptional states

Cross-referencing the top 50 genes per module with 173 known functional drivers provided biological grounding for the mathematically derived states. Four of the modules aligned closely with established breast cancer drivers and pathways. LTS_M0 (Luminal) overlaps with classic luminal drivers including *ESR1*, *GATA3*, and *FOXA1*. LTS_M1 (Proliferation) is enriched for mitotic regulators such as *CDC20*, *UBE2C*, and *BIRC5*. LTS_M3 (HER2/MYC) captures the *ERBB2* and *MYC* amplicons, alongside *FGFR4*. LTS_M4 (Stromal/Wnt) includes *SFRP1*, suggesting a role in extracellular matrix or Wnt signaling modulation. In contrast, LTS_M5 (Metabolic) is characterized by metabolic and stress-response genes (e.g., *MUCL1*, *SPINK8*, *MAL2*) rather than traditional driver amplicons, presenting a distinct biological profile.

3.4 A metabolic-stress module provides independent prognostic value

We evaluated the prognostic utility of the five modules to predict disease-specific death (646 events) while treating death from other causes (496 events) as a competing risk. To guarantee orthogonal clinical value, we established a strong baseline cause-specific Cox proportional hazards model that adjusted for age, tumor size, node status, clinical grade, and INTCLUST membership (baseline C-index = 0.7072).

After adjusting for these established prognostic factors, only the metabolic-stress module LTS_M5 provided independent prognostic information. Higher scores in LTS_M5 were associated with a significant 10% reduction in the hazard of death from disease (Hazard Ratio = 0.901, 95% CI: 0.828–0.980, $p = 0.0156$). Cumulative Incidence Function (CIF) plots demonstrate a clear separation in disease-specific mortality between the top and bottom quartiles of the LTS_M5 score (Figure 5).

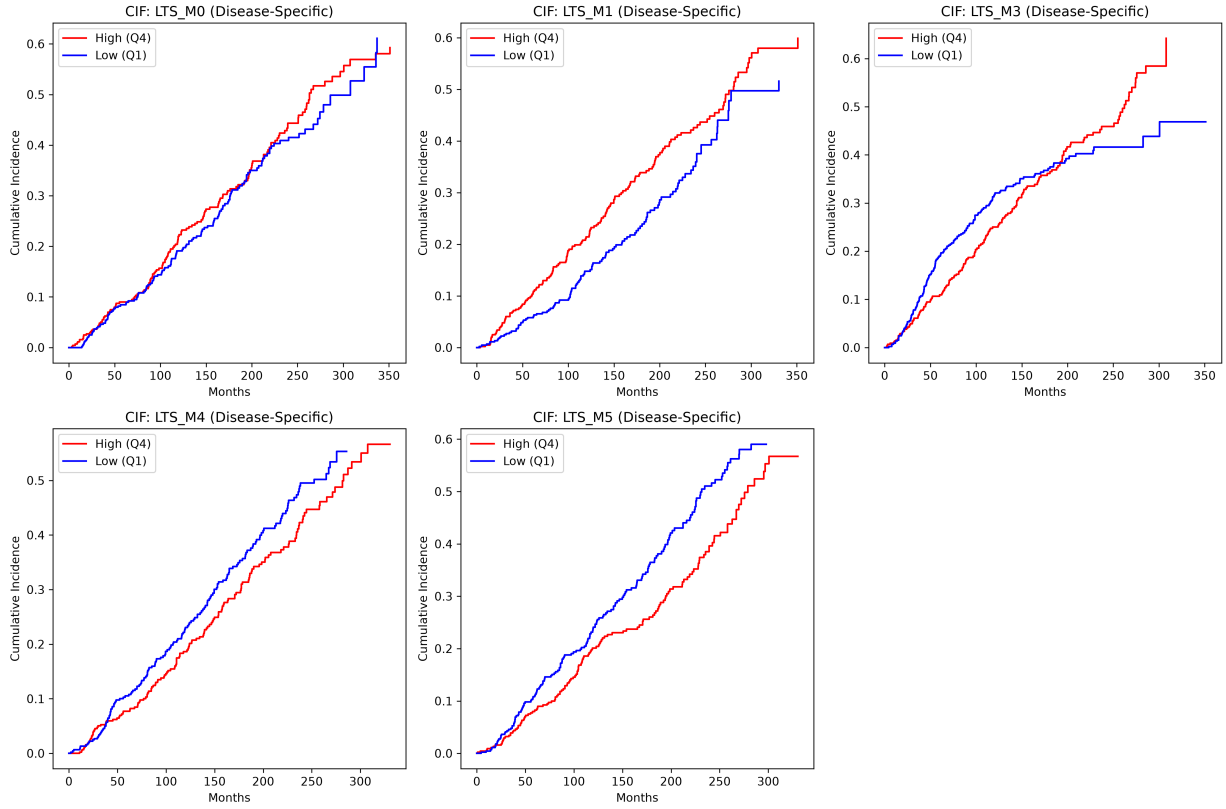


Figure 5: Cumulative Incidence Function (CIF) plots displaying disease-specific mortality for the highest (Q4, red) and lowest (Q1, blue) score quartiles across five Latent Transcriptional State (LTS) modules. The plots demonstrate a clear separation in survival outcomes for the metabolic-stress module LTS_M5, where higher scores are associated with a significant 10% reduction in the hazard of disease-specific death (HR = 0.901, $p = 0.0156$). This separation confirms that LTS_M5 captures an axis of tumor lethality that provides independent prognostic value and orthogonal patient stratification beyond established clinical covariates and integrative cluster definitions.

Furthermore, incorporating LTS_M5 into the baseline model yielded a modest but orthogonal C-index improvement of 0.00025 (Figure 6). This confirms that the metabolic-stress state defined by LTS_M5 captures an independent axis of tumor lethality, offering additional prognostic value for disease-specific survival beyond established clinical and integrative cluster definitions.

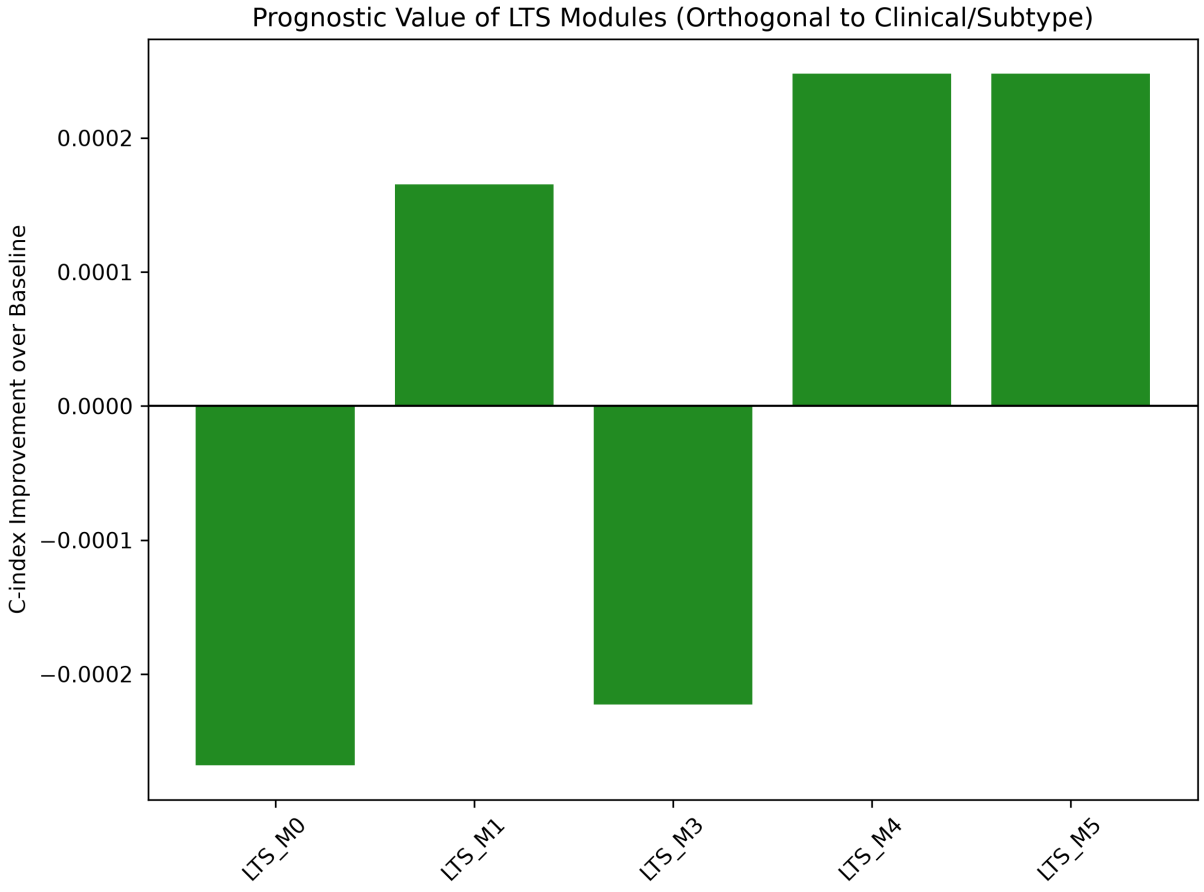


Figure 6: This bar chart displays the C-index improvement over a strong baseline model for five Latent Transcriptional State (LTS) modules. The baseline model adjusts for clinical covariates including age, tumor size, node status, grade, and INTCLUST membership (baseline C-index = 0.7072). The plot illustrates that the metabolic-stress module LTS_M5 provides a modest but orthogonal C-index improvement of 0.00025. This confirms that LTS_M5 captures an independent axis of tumor lethality, offering additional prognostic value for disease-specific survival beyond established clinical and integrative cluster definitions.

4 Conclusions

In this study, we introduced a multi-omic tensor framework to mathematically define and isolate regulatory discordance in breast cancer. By shifting from traditional scalar residual metrics to a third-order tensor decomposition, we successfully extracted coordinated, multi-gene transcriptional states that are fundamentally uncoupled from physical DNA dosage. The rigorous cross-cohort validation across 1,976 METABRIC samples ensured that the five identified Latent Transcriptional States (LTS) represent robust biological rewiring rather than technical artifacts, batch effects, or variations in tumor purity or genomic burden.

A central finding of our analysis is the structural nature of this transcriptional uncoupling. We demonstrated that regulatory discordance predominantly manifests as systematic intercept shifts rather than alterations in dosage-response slopes. This indicates that the uncoupled variance acts as a constant baseline offset to the physical dosage-response function. Consequently, the regulatory rewiring captured by these modules reflects a coordinated baseline modulation of gene networks rather than a localized failure in dosage buffering.

While four of the validated modules aligned closely with established breast cancer drivers—capturing luminal, proliferation, stromal, and HER2/MYC signaling pathways—the fifth module (LTS_M5) revealed a distinct metabolic-stress profile. Crucially, cause-specific Cox proportional hazards modeling demonstrated that this metabolic-stress state provides orthogonal prognostic value in the context of competing risks. Higher scores in LTS_M5 significantly reduced the hazard of disease-specific death, independent of established clinical covariates including tumor size, clinical grade, lymph node status, and integrative cluster (INTCLUST) membership. Although the absolute improvement in predictive concordance was modest, it confirms the existence of an independent axis of tumor lethality that escapes traditional genomic and clinical taxonomies.

Ultimately, this work establishes that structurally conserved, cross-layer regulatory discordance can serve as a stable and independent predictor of breast cancer outcomes. By leveraging tensor-derived networks to capture joint organization across molecular layers, future multi-omic investigations can move beyond isolated gene-level analyses to uncover complex, coordinated mechanisms of tumor evolution and refine patient stratification strategies in the clinic.