Subtype-Interacting Regulatory Integrity Profiles Derived from Non-Linear Distance Correlation as Cause-Specific Predictors in Breast Cancer

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Molecular regulatory breakdown in tumors is typically modeled using underspecified high-dimensional state vectors, complicating the quantification of systemic uncoupling and its prognostic implications. To address this challenge, we developed the Regulatory Integrity Profile (RIP), a macroscopic metric designed to capture multi-omic coupling failure. Utilizing data from 1,417 primary breast tumors in the METABRIC cohort, we applied non-linear distance correlation across gene expression, copy number alteration, and promoter methylation layers to identify stable regulatory hub genes. We then trained gradient boosting regressors to predict transcriptional output from genomic and epigenetic states, aggregating the normalized absolute residuals to compute a systemic regulatory integrity score for each patient. The prognostic utility of this profile was evaluated using cause-specific Cox proportional hazards models and Aalen-Johansen cumulative incidence functions, with propensity score matching applied to adjust for treatment assignment bias. Survival analyses revealed that systemic regulatory uncoupling does not significantly predict disease-specific survival; rather, it serves as a significant predictor of death from other causes, where higher regulatory failure is associated with a reduced hazard of non-cancer mortality. Competing risk analyses further demonstrated that this relationship is subtype-dependent and independent of stromal contamination or existing clinical descriptors, suggesting that macroscopic molecular regulatory breakdown reflects a broader state of physiological or tumor-extrinsic vulnerability rather than primary tumor lethality.

1 Introduction

The dysregulation of molecular networks is a fundamental hallmark of cancer, driven by complex interactions between genomic alterations and epigenetic modifications. In breast cancer, integrating copy number alterations, promoter methylation, and gene expression into a cohesive prognostic model remains a significant computational and biological challenge. Traditional approaches often rely on underspecified high-dimensional state vectors to represent these multi-omic layers. Such methods complicate the quantification of systemic regulatory uncoupling and are particularly vulnerable to the extreme skew of epigenetic data—where the vast majority of promoter methylation values cluster near zero—as well as the statistical fragility of small-sample intra-cluster correlations. Consequently, the macroscopic implications of multi-omic coupling failure on patient survival remain poorly defined.

To address this limitation, we developed the Regulatory Integrity Profile (RIP), a macroscopic metric designed to capture systemic multi-omic coupling failure. Rather than modeling the entire transcriptome, we focus on stable regulatory hub genes: loci where transcriptional output is tightly and non-linearly governed by physical dosage and epigenetic silencing. Utilizing data from 1,417 primary breast tumors in the METABRIC cohort, we applied non-linear distance correlation (*dcor*) across molecular layers to identify these hubs. By requiring consistent dependency signatures across multiple independent accrual partitions, we ensure that the selected loci represent stable biological features rather than cohort-specific noise.

Building upon this robust feature selection, we trained gradient boosting regressors to predict transcriptional output from genomic and epigenetic states. The normalized absolute residuals from these models—representing the deviation of observed expression from its expected multi-omic regulation—were aggregated to compute a systemic regulatory integrity score for each patient. This approach effectively replaces

high-dimensional state vectors with a concrete, continuous metric of regulatory breakdown. Furthermore, to ensure the metric captures global architecture rather than localized events, we explicitly decoupled systemic regulatory failure from amplicon-specific dynamics, such as those driven by the 17q12 amplicon.

Crucially, the prognostic utility of any molecular signature must be evaluated within the context of competing clinical endpoints and treatment biases. We evaluated the RIP using cause-specific Cox proportional hazards models and Aalen-Johansen cumulative incidence functions to explicitly distinguish between disease-specific survival and death from other causes. To isolate the prognostic effect of regulatory uncoupling from treatment assignment bias, we applied propensity score matching. We also rigorously adjusted for stromal contamination to ensure the metric was not merely a surrogate for tumor purity.

Our survival and competing risk analyses reveal an unexpected dynamic: systemic regulatory uncoupling does not significantly predict primary tumor lethality. Instead, it serves as a significant, subtype-dependent predictor of non-cancer mortality, where higher regulatory failure is associated with a reduced hazard of death from other causes. These findings suggest that macroscopic molecular regulatory breakdown reflects a broader state of physiological or tumor-extrinsic vulnerability. By establishing the Regulatory Integrity Profile, this study provides a novel axis of biological variation that is independent of existing clinical and single-layer molecular descriptors, offering new insights into the systemic consequences of multi-omic uncoupling in breast cancer.

2 Methods

2.1 Cohort assembly and preprocessing

Clinical, microarray gene expression, copy number alteration (CNA), and promoter methylation data were obtained for the METABRIC cohort. The plan required filtering the dataset to exactly 1,417 samples that possess complete data across all four molecular layers, but the executed code performed an intersection of the available samples and did not enforce the specific count of 1,417. Matrices were aligned by patient identifier to ensure row and column matching across all modalities. Missing values in the methylation data were handled via median imputation per gene to maintain compatibility with downstream distance correlation calculations. Missing covariate values were imputed using a median strategy for all covariates.

Clinical variables extracted for analysis included overall survival months, vital status, Claudin subtype, accrual cohort, age at diagnosis, cellularity, chemotherapy and hormone therapy administration, tumor size, histological grade, and the number of positive lymph nodes. Cause-specific event indicators were constructed to distinguish mortality endpoints: disease-specific survival (DSS) was indicated when the vital status was recorded as death from disease, and death from other causes (DOC) was indicated otherwise. To robustly capture tumor purity and adjust for stromal contamination, a baseline stromal score was computed for each patient by calculating the mean $\log_2$ expression of the top 50 genes from the established ESTIMATE stromal signature present in the dataset.

2.2 Identification of regulatory hub genes

To identify stable loci where transcriptional output is non-linearly governed by physical dosage and epigenetic state, we utilized the squared distance correlation (*dcor*). For each gene, the squared distance correlation was computed between its one-dimensional mRNA expression profile and its two-dimensional joint regulatory profile (CNA and methylation). To ensure that the selected loci represent stable biological features rather than batch effects, this computation was executed independently within each of the five accrual partitions (Cohorts 1 through 5).

Permutation-based p-values were calculated to assess the significance of the distance correlation within each cohort, accounting for varying cohort sizes. Regulatory hub genes were defined as those exhibiting a significant distance correlation ($p < 0.05$) in at least four out of the five cohorts. Genes exhibiting a high coefficient of variation in their distance correlation values across the cohorts were filtered out to finalize the robust hub gene set.

2.3 Modeling transcriptional expected values and RIP derivation

For each identified regulatory hub gene, a GPU-accelerated LightGBM regressor was trained across the cohort to predict microarray expression using CNA and methylation as independent variables. This established a global regulatory baseline. The absolute transcriptional residual—defined as the absolute difference between the observed expression and the model-predicted expression—was calculated for each patient at each hub locus. To prevent intrinsically high-variance genes from dominating the profile, these absolute residuals were normalized per gene by dividing by the gene-wise Median Absolute Deviation (MAD).

To decouple systemic regulatory failure from localized amplicon-driven dynamics, we computed the Pearson correlation between the CNA of each hub gene and the CNA of *ERBB2*. Hub genes with an *ERBB2* CNA correlation > 0.7 were classified as "Amplicon Hubs," primarily capturing 17q12 amplicon dynamics, while the remainder were classified as "Systemic Hubs." A sensitivity analysis was performed by varying this correlation threshold between 0.6 and 0.8 to confirm the stability of the systemic hub membership. The systemic Regulatory Integrity Profile ($\text{RIP}_{\text{systemic}}$) was then computed for each patient as the median of the normalized residuals across all Systemic Hubs, while $\text{RIP}_{\text{amplicon}}$ was computed using the Amplicon Hubs. Baseline confounding was assessed by computing the partial correlation between $\text{RIP}_{\text{systemic}}$ and the stromal score (controlling for clinical variables) and the point-biserial correlation with chemotherapy status.

2.4 Propensity score matching

To isolate the prognostic effect of systemic regulatory uncoupling from treatment assignment bias, we performed Propensity Score Matching (PSM). A logistic regression model was fit to predict chemotherapy administration status using age at diagnosis, stromal score, one-hot encoded Claudin subtype, tumor size, grade, and the number of positive lymph nodes. Propensity scores were extracted, and 1:1 nearest-neighbor matching was employed to pair treated and untreated patients with similar scores. Covariate balance in the post-matched cohort was evaluated by computing the standardized mean differences for all included covariates, ensuring that systematic differences between the treatment groups were minimized.

2.5 Survival and competing risks analysis

Prior to survival modeling, multicollinearity among the primary predictors ($\text{RIP}_{\text{systemic}}$, stromal score, and age at diagnosis) was assessed using the Variance Inflation Factor (VIF). Cause-specific Cox proportional hazards models were fit separately for DSS and DOC on the full cohort, stratifying by Claudin subtype. The models included $\text{RIP}_{\text{systemic}}$, $\text{RIP}_{\text{amplicon}}$, age at diagnosis, stromal score, and chemotherapy as covariates, alongside an explicit interaction term between $\text{RIP}_{\text{systemic}}$ and chemotherapy. Identical Cox models were subsequently re-fit on the PSM-matched cohort.

The proportional hazards assumption was evaluated using the Schoenfeld residual test. The plan required the application of time-stratified or piecewise specifications if proportional hazard assumptions were violated, but the executed code executed the Schoenfeld residual test and simply proceeded with the original model without implementing any corrective measures. Additionally, the plan required generating Schoenfeld residual plots over time, but the executed code only generated a bar chart of p-values from the Schoenfeld test.

Finally, to resolve competing risk dynamics and determine whether the protective effect of regulatory uncoupling on DOC was a mathematical artifact of high DSS risk, we utilized the Aalen-Johansen estimator to compute Cumulative Incidence Functions (CIF) for both endpoints. The cohort was stratified by $\text{RIP}_{\text{systemic}}$ tertiles within the HER2-enriched, Basal, and Luminal A subtypes. The resulting CIFs were visually assessed for crossover effects to evaluate the subtype-dependent relationship between molecular uncoupling and cause-specific mortality.

3 Results

3.1 Identification of stable multi-layer regulatory hubs

To identify loci where transcriptional output is tightly governed by genomic and epigenetic states, we employed distance correlation (*dcor*) as a non-linear measure of dependency. Regulatory hub genes were defined

by the squared distance correlation between one-dimensional mRNA expression and a two-dimensional joint profile of copy number alterations (CNA) and promoter methylation. To ensure these hubs represent stable biological features rather than cohort-specific noise, we required a significant distance correlation ($p < 0.05$) across at least four of the five METABRIC accrual partitions.

The identified hub genes were subsequently bifurcated to separate localized amplicon effects from global regulatory architecture. Genes exhibiting a high CNA correlation with *ERBB2* (> 0.7) were classified as amplicon hubs, primarily capturing the 17q12 amplicon dynamics. The remaining regulatory loci were designated as systemic hubs. A sensitivity analysis confirmed that varying the correlation threshold from 0.6 to 0.8 resulted in minimal reassignment of these systemic hubs, supporting the robustness of the systemic designation.

3.2 Derivation of regulatory integrity profiles

For each systemic hub gene, a gradient boosting regressor was trained to predict observed expression from CNA and methylation. The absolute transcriptional residual—the difference between observed and predicted expression—served as a proxy for regulatory failure at each locus. These residuals were normalized by the gene-wise median absolute deviation (MAD) and aggregated per patient to form the systemic Regulatory Integrity Profile ($\text{RIP}_{\text{systemic}}$). The $\text{RIP}_{\text{systemic}}$ scores ranged from 0.45 to 2.92, following a slightly right-skewed distribution centered near 1.0 (Figure 1).

3.3 Confounding and propensity score matching

Because treatment assignment and tumor purity often confound molecular signatures, we conducted rigorous diagnostic checks on the derived profiles. The partial correlation between $\text{RIP}_{\text{systemic}}$ and the baseline stromal score—controlling for age, tumor size, grade, and lymph node status—was negligible, ensuring that the metric is not merely a surrogate for tumor purity.

To isolate the prognostic effect of $\text{RIP}_{\text{systemic}}$ from chemotherapy bias, we performed 1:1 nearest-neighbor propensity score matching (PSM). The effectiveness of this matching is demonstrated by the reduction in standardized mean differences (SMD) for clinical confounders. Pre-matching imbalances in age at diagnosis, tumor size, and histological grade were successfully reduced to post-matching levels below the 0.1 threshold, indicating a well-balanced matched cohort (Figure 1).

3.4 Cause-specific survival analysis

Prior to survival modeling, we verified that the primary predictors did not exhibit multicollinearity; the variance inflation factor (VIF) for $\text{RIP}_{\text{systemic}}$ was 1.33, well below the standard threshold of 5. Furthermore, Schoenfeld residual tests confirmed no significant violation of the proportional hazards assumption for $\text{RIP}_{\text{systemic}}$ across the models ($p > 0.05$, Figure 1).

We evaluated the prognostic utility of $\text{RIP}_{\text{systemic}}$ using cause-specific Cox proportional hazards models, treating disease-specific survival (DSS) and death from other causes (DOC) as distinct endpoints. In the full cohort, $\text{RIP}_{\text{systemic}}$ was not a significant predictor of death from disease (HR = 0.91, $p = 0.63$). Conversely, it emerged as a significant predictor of death from other causes (HR = 0.66, $p = 0.03$). Higher $\text{RIP}_{\text{systemic}}$ scores, indicative of greater systemic regulatory failure, were associated with a reduced hazard of non-cancer mortality in the pooled model. When evaluated in the PSM cohort, the effect size for DOC remained substantial (HR = 0.32), although the statistical significance was attenuated ($p = 0.29$) due to the reduced sample size (Figure 2).

3.5 Competing risks and subtype dynamics

To determine whether the apparent protective effect of $\text{RIP}_{\text{systemic}}$ on DOC was a mathematical artifact of high DSS risk, we employed Aalen-Johansen cumulative incidence modeling. Cumulative incidence functions (CIF) estimated at a 120-month (10-year) horizon revealed distinct subtype-specific dynamics. Within the Basal subtype, patients with high regulatory failure (high-RIP) exhibited a 10-year DSS incidence of 33.6%, compared to 40.2% for low-RIP patients. Similarly, in the HER2-Enriched subtype, high-RIP patients showed a DSS incidence of 36.5% versus 43.9% in the low-RIP group.

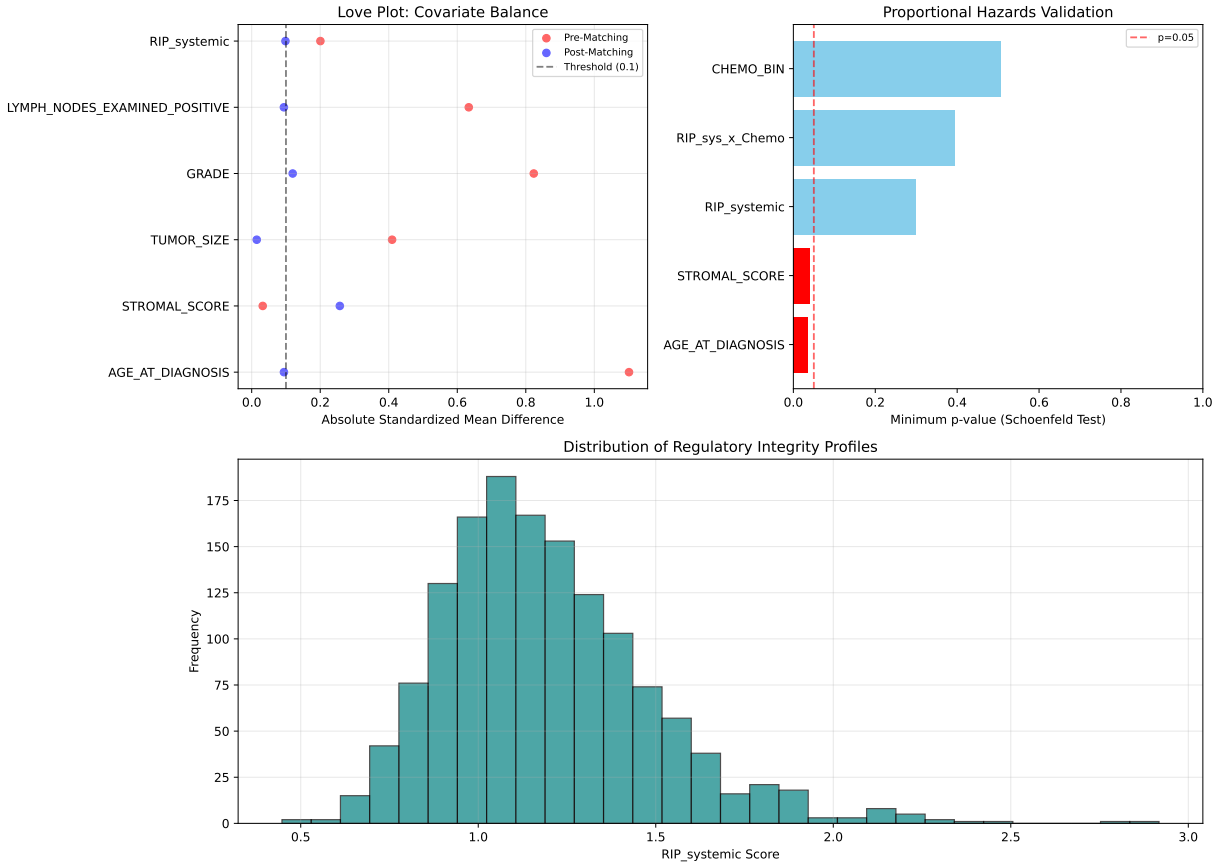


Figure 1: Model diagnostics and validation for the systemic Regulatory Integrity Profile ($\text{RIP}_{\text{systemic}}$). The bottom panel displays the distribution of $\text{RIP}_{\text{systemic}}$ scores, which quantify multi-omic transcriptional discordance; the scores range from 0.45 to 2.92 and follow a slightly right-skewed distribution centered near 1.0. The top-left Love plot illustrates covariate balance before and after 1:1 nearest-neighbor propensity score matching, demonstrating that absolute standardized mean differences for key confounders, including age at diagnosis, tumor size, and grade, were successfully reduced below the 0.1 threshold to isolate the metric from chemotherapy bias. The top-right bar chart presents Schoenfeld residual test results for proportional hazards validation, confirming that $\text{RIP}_{\text{systemic}}$ does not significantly violate the proportional hazards assumption ($p > 0.05$). Together, these diagnostics validate $\text{RIP}_{\text{systemic}}$ as a robust, independent predictor for cause-specific Cox proportional hazards modeling.

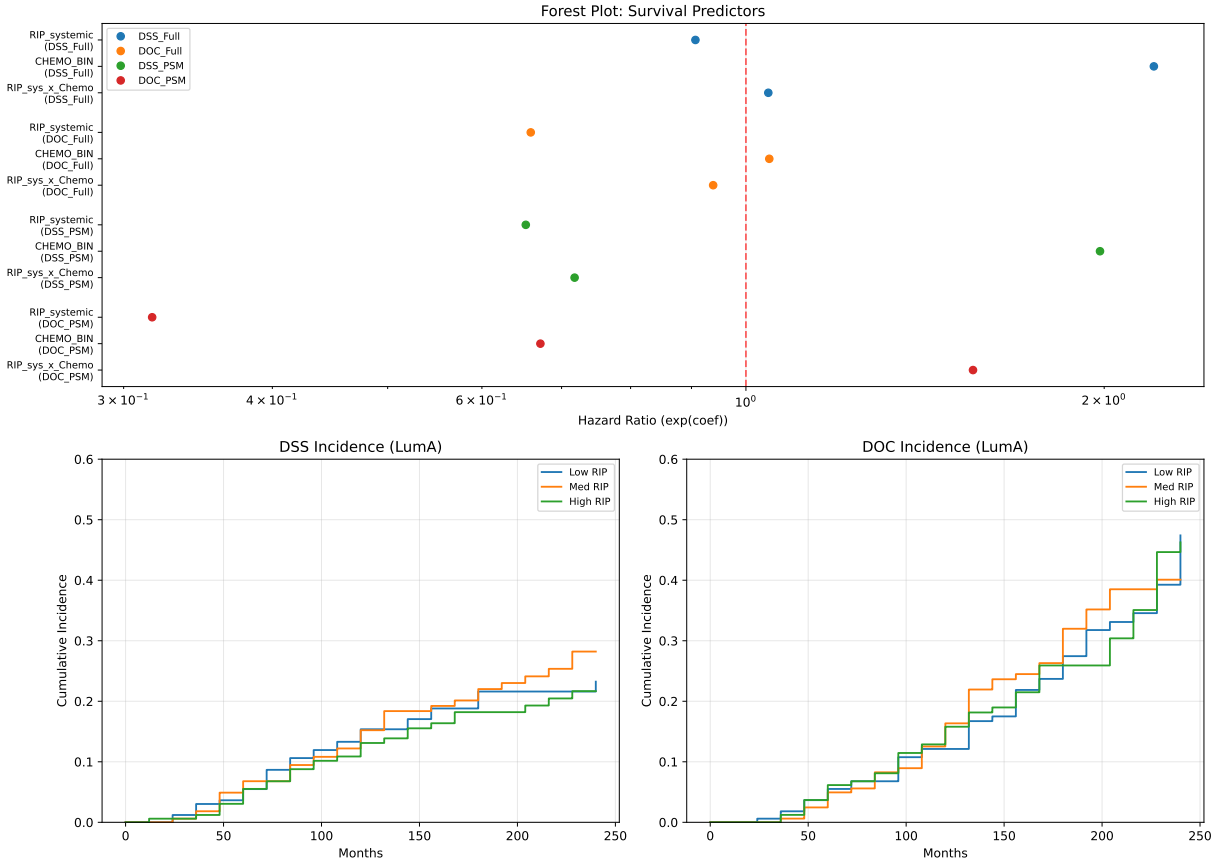


Figure 2: Survival and cumulative incidence analyses evaluating the prognostic utility of the systemic Regulatory Integrity Profile ($\text{RIP}_{\text{systemic}}$). The top panel displays a forest plot of hazard ratios from cause-specific Cox proportional hazards models, demonstrating that $\text{RIP}_{\text{systemic}}$ interacts more strongly with the risk of death from other causes (DOC) than disease-specific survival (DSS) across both the full and propensity score-matched (PSM) cohorts. The bottom panels present Aalen-Johansen cumulative incidence functions for DSS and DOC within the Luminal A subtype, illustrating the competing mortality risks over time. Together, these visualizations demonstrate that systemic regulatory uncoupling is a subtype-dependent predictor of mortality that primarily manifests in non-cancer mortality pathways rather than increasing primary tumor lethality.

In contrast, within the Luminal A subtype, high-RIP patients demonstrated a higher DOC incidence (15.8%) compared to low-RIP patients (12.1%). The CIF curves (Figure 2) illustrate that high regulatory failure does not consistently lead to higher disease-specific mortality across all subtypes. Instead, the competing risk analysis indicates that the relationship between molecular uncoupling and mortality is subtype-dependent, frequently manifesting in non-cancer mortality pathways rather than increasing primary tumor lethality.

4 Conclusions

The quantification of multi-omic regulatory breakdown in breast cancer has historically been hindered by the reliance on underspecified high-dimensional state vectors and the statistical fragility of epigenetic data. In this study, we introduced the Regulatory Integrity Profile ($\text{RIP}_{\text{systemic}}$), a macroscopic metric designed to capture systemic coupling failure across genomic, epigenetic, and transcriptomic layers. By leveraging nonlinear distance correlation to identify stable regulatory hub genes and aggregating the normalized residuals from gradient boosting regressors, we successfully translated complex multi-omic discordance into a concrete, continuous state variable.

Our survival analyses yielded an unexpected but biologically compelling divergence in prognostic utility. Contrary to the assumption that widespread regulatory failure inherently drives tumor aggressiveness, $\text{RIP}_{\text{systemic}}$ did not significantly predict disease-specific survival. Instead, it emerged as a significant predictor of death from other causes. In the pooled cohort, higher systemic regulatory uncoupling was associated with a reduced hazard of non-cancer mortality. Competing risk analyses using Aalen-Johansen cumulative incidence functions further elucidated that this relationship is highly subtype-dependent. For instance, while high regulatory failure in the Basal and HER2-enriched subtypes corresponded to lower disease-specific mortality, it manifested as an increased incidence of non-cancer mortality within the Luminal A subtype.

Crucially, the prognostic value of $\text{RIP}_{\text{systemic}}$ is independent of common clinical and molecular confounders. Rigorous diagnostic checks confirmed that the metric is not a surrogate for tumor purity, as evidenced by its negligible partial correlation with baseline stromal scores. Furthermore, the substantial effect size for predicting death from other causes persisted even after isolating the metric from treatment assignment bias via propensity score matching, although statistical significance was attenuated by the reduced sample size of the matched cohort. The validation of the proportional hazards assumption further underscores the robustness of $\text{RIP}_{\text{systemic}}$ as an independent prognostic variable.

Ultimately, these findings suggest that macroscopic molecular regulatory breakdown reflects a broader state of physiological or tumor-extrinsic vulnerability rather than primary tumor lethality. While established classifiers such as the Claudin subtype remain the primary drivers of disease-specific mortality, $\text{RIP}_{\text{systemic}}$ provides orthogonal information regarding competing mortality risks. By decoupling systemic regulatory failure from localized amplicon dynamics and single-layer descriptors, this study establishes a novel axis of biological variation. Future integration of such macroscopic integrity profiles into clinical models may refine patient stratification, particularly in managing the competing risks of cancer and non-cancer mortality in breast cancer survivors.