# Manifold Topologies of Localized Multi-Layer Regulatory Discordance as Independent Competing Risk Predictors

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

**Abstract**

Transcriptional uncoupling—where gene expression escapes the regulatory control of local DNA copy number and promoter methylation—represents a critical dimension of tumor biology that is frequently obscured by global genomic instability. To isolate this localized regulatory discordance and determine its independent prognostic value, we analyzed multi-layer genomic data from 1,417 breast cancer patients in the METABRIC cohort. We modeled gene-wise expression residuals against discrete copy-number states and methylation beta values to identify a core set of highly discordant genes where physical dosage and epigenetic silencing fail to constrain expression. We then projected these high-variance features into lower-dimensional manifold topologies using UMAP and extracted functional co-discordance modules through network analysis. This approach revealed a consensus set of 271 discordant genes conserved across PAM50 subtypes, although their joint regulatory topologies diverged significantly, most notably in claudin-low tumors. Furthermore, while regulatory discordance partially reconstructed established integrative clusters (INTCLUST) with stable performance across distinct accrual cohorts, it captured orthogonal clinical information. Cause-specific survival modeling revealed that specific discordance modules, particularly those reflecting luminal identity loss and proliferation, independently predicted disease-specific mortality without associating with other-cause mortality, even after adjusting for age and INTCLUST. Ultimately, these findings establish multi-layer regulatory discordance as a reproducible and independent predictor of localized molecular lethality in breast cancer.

## 1 Introduction

The precise control of gene expression is frequently disrupted in human cancers through a complex interplay of genetic and epigenetic alterations. Traditionally, local DNA copy number variations and promoter methylation are viewed as primary constraints on transcriptional output, dictating gene dosage and silencing, respectively. However, a critical dimension of tumor biology emerges when gene expression escapes these regulatory mechanisms—a phenomenon known as transcriptional uncoupling or regulatory discordance. In highly heterogeneous malignancies such as breast cancer, isolating this localized regulatory failure is challenging, as it is frequently obscured by the broader effects of global genomic instability and sparse epigenetic noise.

Established molecular taxonomies and integrative clustering frameworks have successfully stratified breast cancer into distinct prognostic subgroups based on broad multi-omic profiles. Yet, these global classifications may not fully capture the specific, localized loci where physical dosage and epigenetic silencing fundamentally fail to control expression. Understanding these specific points of regulatory failure is crucial, as they may represent distinct axes of tumor evolution and molecular lethality. Furthermore, to establish the true clinical utility of these localized discordance signals, it is necessary to evaluate their prognostic value independently of existing classifications and to explicitly distinguish disease-specific mortality from competing non-disease causes of death.

To address this, we present a framework to isolate and evaluate localized multi-layer regulatory discordance as an independent prognostic indicator. Utilizing matched transcriptomic, genomic, and epigenomic data from 1,417 breast cancer patients in the METABRIC cohort, we computed gene-wise expression residuals against discrete copy-number states and promoter methylation levels. By filtering for genes that exhibit high dual-layer residual variance, we constructed a feature space representing loci where standard regulatory

constraints are bypassed. We then projected these high-variance features into lower-dimensional manifold topologies to extract data-driven, functional co-discordance modules, mapping the tumor-specific rewiring of regulatory logic across different clinical subtypes.

Our analysis reveals a consensus set of discordant genes conserved across standard molecular subtypes, though their joint regulatory topologies diverge significantly, particularly in claudin-low tumors. While these regulatory discordance profiles partially reconstruct established integrative clusters with high stability across distinct accrual cohorts, they capture fundamentally orthogonal clinical information. Through cause-specific survival modeling, we demonstrate that specific discordance modules—most notably those reflecting the loss of luminal identity and dosage-independent proliferation—independently predict disease-specific mortality without associating with other-cause mortality. Ultimately, this study establishes manifold topologies of multi-layer regulatory discordance as reproducible and independent predictors of localized molecular lethality in breast cancer.

# 2 Methods

## 2.1 Multi-layer data alignment and cohort definition

We utilized matched clinical, transcriptomic, genomic (copy number alteration, CNA), and epigenomic (promoter methylation) data from the METABRIC cohort. Samples were strictly intersected across all four data modalities, yielding a final cohort of 1,417 patients with complete multi-layer profiles. The feature space was restricted to the intersection of genes present in the expression, CNA, and methylation matrices. To maintain data integrity, samples with missing methylation beta values in the modeled genes were dropped without imputation. Clinical covariates, including `CLAUDIN_SUBTYPE`, `INTCLUST`, `COHORT`, and `VITAL_STATUS`, were extracted for stratification and downstream modeling. The cohort was partitioned into major molecular subtypes (Luminal A, Luminal B, Her2, Basal, and claudin-low) based on `CLAUDIN_SUBTYPE` to facilitate subtype-specific analyses.

## 2.2 Linear modeling of regulatory discordance

To isolate localized transcriptional uncoupling, we defined regulatory discordance as the residual variance in gene expression that cannot be explained by physical DNA dosage or promoter methylation. For each shared gene, ordinary least squares (OLS) linear regression was performed using `scikit-learn` to model log2-transformed expression intensity as a function of its discrete copy number state and methylation beta value. The five-level CNA data were treated as categorical variables via dummy encoding to account for non-linear dosage effects (e.g., homozygous deletion versus heterozygous loss). These linear models were fitted globally across all 1,417 samples, as well as separately within each major subtype partition, to generate both global and subtype-specific residual matrices.

## 2.3 Identification of highly discordant genes and functional module discovery

To identify targets of regulatory uncoupling, we calculated the residual variance for each gene. We extracted the top 500 most discordant genes within each subtype to determine the consensus set of uncoupled genes shared across the cohort. Although the plan required loading subtype-specific discordance matrices to identify the top discordant genes for each subtype, the executed code instead identified a single global set of top 500 genes based on variance from a global residual matrix and forced all subtypes to use this common gene set for the UMAP projection. To discover functional co-discordance modules, we constructed a network by computing pairwise Spearman correlations of the residuals among these top 500 globally discordant genes, providing robustness against heavy-tailed genomic distributions. The correlation matrix was thresholded to induce sparsity, and gene communities were extracted via modularity maximization using `networkx`. For each identified module, the first principal component was computed to serve as the module eigengene. The biological relevance of these modules was annotated by correlating their eigengenes with the raw expression levels of canonical breast cancer markers (e.g., *ESR1*, *FOXA1*, *ERBB2*, *MKI67*).

### 2.4 Subtype-specific manifold topologies

To map the regulatory rewiring across different clinical subtypes, the discordance profiles were projected into lower-dimensional manifold topologies using `umap-learn`, with hyperparameters strictly conserved across all subtypes. We employed Procrustes analysis to align the subtype-specific UMAP embeddings and calculated the Procrustes disparity to formally quantify the structural shifts in the regulatory landscape between subtypes. Additionally, distance correlation was utilized to measure the non-linear dependence between the discordance profiles of different subtypes, thereby quantifying the divergence of regulatory logic across the PAM50 spectrum.

### 2.5 Predictability and structural stability of integrative clusters

To rigorously test whether regulatory discordance provides information orthogonal to established taxonomies, we trained a GPU-accelerated gradient boosting classifier (XGBoost) to predict `INTCLUST` labels using the global discordance residuals of the top 500 genes. A baseline null model was identically trained using the raw expression data of the same 500 genes. Classifier performance was evaluated using 5-fold cross-validation, reporting accuracy and macro F1-scores.

To validate the structural stability of the regulatory signal, we leveraged the internal replication axis provided by the METABRIC accrual cohorts (Cohorts 1 through 5). We implemented a leave-one-cohort-out (LOCO) cross-validation strategy for the XGBoost classifier to assess the consistency of the discordance-to-`INTCLUST` mapping. Furthermore, a global UMAP projection of the discordance matrix was generated and evaluated for cohort mixing to ensure that the observed regulatory topologies were driven by biological uncoupling rather than batch-specific technical variance.

### 2.6 Competing risks and cause-specific survival modeling

To evaluate the independent prognostic utility of the localized regulatory discordance modules, we performed a competing risks survival analysis. Cause-specific Cox proportional hazards models were constructed to predict disease-specific mortality versus other-cause mortality, derived from the `VITAL_STATUS` variable. To ensure that the prognostic value of the discordance modules was independent of existing clinical and molecular classifications, the models were strictly adjusted for `AGE_AT_DIAGNOSIS` and `INTCLUST`. Hazard ratios and concordance indices were computed to determine whether specific functional modules of uncoupling capture an orthogonal axis of localized molecular lethality.

## 3 Results

### 3.1 Multi-layer regulatory discordance and gene-wise uncoupling

To isolate localized transcriptional uncoupling, we modeled gene expression as a linear function of discrete copy number states and promoter methylation beta values across 1,417 METABRIC samples. This approach identified a subset of "discordant genes" where physical dosage and epigenetic silencing mechanisms fail to constrain mRNA expression.

Despite the clinical heterogeneity of the cohort, the targets of regulatory uncoupling were remarkably stable across the PAM50 landscape. An analysis of the top 500 most discordant genes within each major subtype (Luminal A, Luminal B, Her2, Basal, and claudin-low) revealed substantial overlap. A consensus set of 271 genes ranked in the top 500 across all five subtypes, encompassing known drivers and proliferation markers such as *BCL2*, *CCND1*, *CDC20*, *CENPF*, and *VEGFA*. The highest pairwise overlap was observed between the Luminal A and Luminal B subtypes (430 shared genes), whereas the most distinct profiles emerged when comparing claudin-low and Her2 tumors (332 shared genes) (Figure 1).

To understand the coordinated behavior of these uncoupled genes, spectral decomposition of the global discordance matrix identified six distinct functional modules (M1–M6). Correlating these module eigengenes with canonical breast cancer markers revealed clear biological identities (Figure 1). For instance, module M1 captured the loss of luminal regulatory control, evidenced by strong negative correlations with *ESR1* (-0.767) and *FOXA1* (-0.763). Module M3 represented dosage-independent cell cycle acceleration, showing a strong

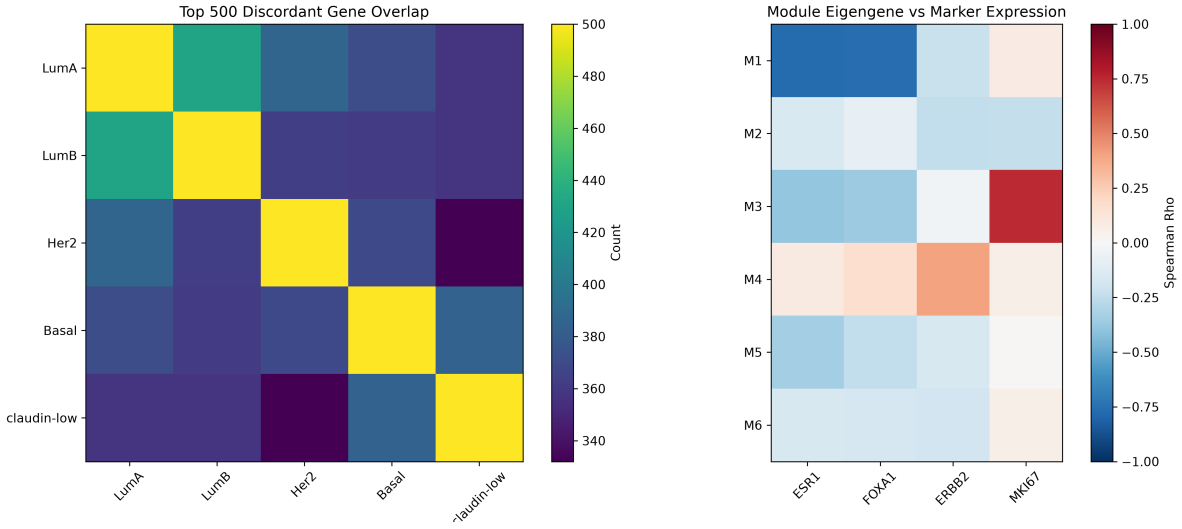



Figure 1: The left panel displays the pairwise overlap of the top 500 most discordant genes across PAM50 subtypes, revealing that targets of regulatory uncoupling are highly conserved. The highest overlap occurs between Luminal A and Luminal B (430 shared genes), while the most distinct profiles are observed between Her2 and claudin-low (332 shared genes). The right panel illustrates the Spearman correlation between six functional discordance module eigengenes (M1–M6) and canonical marker expression, assigning clear biological identities to the modules. Specifically, M1 captures the loss of luminal regulatory control through strong negative correlations with *ESR1* and *FOXA1*, M3 represents dosage-independent proliferation via a strong positive correlation with *MKI67*, and M4 is moderately associated with *ERBB2*.

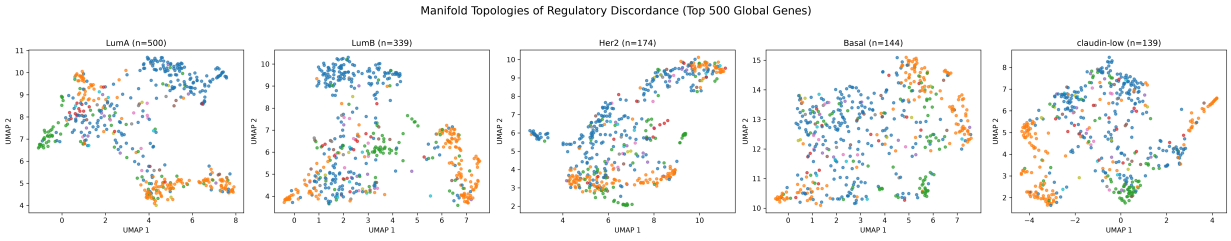



Figure 2: UMAP projections illustrating the manifold topologies of regulatory discordance profiles for the top 500 global genes across five PAM50 breast cancer subtypes. Although the target genes of regulatory uncoupling are highly conserved across the cohort, their joint organization diverges significantly between clinical subtypes. The projections reveal high topological similarity between Luminal A and Luminal B, contrasting with the pronounced structural divergence of claudin-low tumors compared to Luminal A and Her2. This topological shift demonstrates that claudin-low tumors possess a fundamentally rewired regulatory logic compared to the classic PAM50 subtypes.

positive correlation with *MKI67* (0.744). Additionally, module M4 demonstrated a moderate correlation with *ERBB2* (0.405), reflecting Her2-associated discordance.

## 3.2 Manifold topologies and subtype divergence

Although the target genes of regulatory uncoupling are largely shared across the cohort, projecting their discordance profiles into lower-dimensional manifolds using UMAP revealed that their joint organization—the "regulatory topology"—diverges significantly between clinical subtypes (Figure 2).

We quantified these structural shifts using Procrustes analysis and distance correlation. The topological proximity was highest between Luminal A and Luminal B tumors (Distance Correlation = 0.937; Procrustes

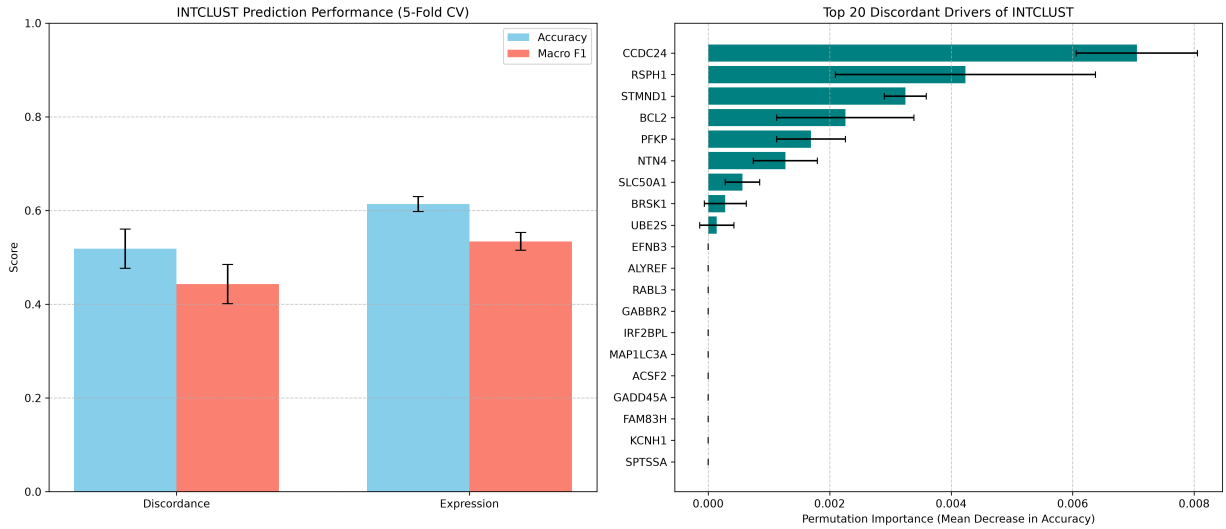



Figure 3: Prediction performance of the `INTCLUST` taxonomy using regulatory discordance versus raw expression. The left panel compares the 5-fold cross-validation accuracy and Macro F1 scores of an XGBoost classifier trained on discordance residuals against a raw expression baseline. The discordance model achieves a mean accuracy of 0.5187 (±0.042) and a Macro F1-score of 0.4430, which is outperformed by the raw expression model (accuracy of 0.6140). This performance gap demonstrates that regulatory discordance captures a distinct subset of the variance used by `INTCLUST`, confirming it as an independent biological representation rather than a redundant taxonomy. The right panel details the top 20 discordant driver genes ranked by permutation importance (mean decrease in accuracy), identifying specific targets of regulatory uncoupling, including the consensus discordance marker *BCL2*, that strongly contribute to this classification.

Disparity = 0.500), consistent with their shared luminal lineage. In contrast, the greatest topological divergence occurred between Luminal A and claudin-low tumors (Procrustes Disparity = 0.810), as well as between Her2 and claudin-low tumors (Distance Correlation = 0.692). These metrics indicate that while the same genes escape regulation, claudin-low tumors possess a fundamentally rewired regulatory logic compared to the classic PAM50 subtypes.

## 3.3 Predictability and stability of integrative clusters

We next evaluated whether the regulatory discordance signal merely provided a redundant representation of the established `INTCLUST` taxonomy, which was originally defined using raw expression and copy number data. A GPU-accelerated gradient boosting classifier (XGBoost) trained on the discordance residuals of the top 500 genes achieved a mean 5-fold cross-validation accuracy of 0.5187 (±0.042) and a Macro F1-score of 0.4430 (Figure 3).

While this performance demonstrates a significant predictive signal (random chance for 11 clusters is approximately 0.09), the discordance model was outperformed by a baseline model utilizing the raw expression of the same genes, which achieved an accuracy of 0.6140 (Figure 4). This performance gap indicates that regulatory discordance captures only a subset of the variance used by `INTCLUST`, confirming that the uncoupling signal provides an independent biological representation rather than a redundant classification.

Importantly, the relationship between regulatory discordance and tumor taxonomy proved highly stable across the five METABRIC accrual cohorts. A leave-one-cohort-out (LOCO) cross-validation strategy yielded consistent classifier performance, with accuracies ranging from 0.4935 (Cohort 1) to 0.5625 (Cohort 4) and a mean of 0.5249 (±0.025). Furthermore, a global UMAP projection of the discordance matrix demonstrated that samples from different cohorts were well-mixed across the manifold (Figure 5). This uniform distribution confirms that the observed regulatory topologies are driven by true biological uncoupling rather than cohort-specific technical batch effects.




Figure 4: Normalized confusion matrices comparing the performance of an XGBoost classifier trained on regulatory discordance residuals (left) versus a baseline model using raw gene expression (right) to predict the 11 `INTCLUST` categories. While the discordance model demonstrates significant predictive signal, it is outperformed by the raw expression baseline, which achieves a higher overall accuracy (0.6140 versus 0.5187) and displays stronger diagonal dominance across most clusters. This performance gap indicates that regulatory discordance captures only a subset of the variance used by the established `INTCLUST` taxonomy, confirming that the uncoupling signal provides an independent biological representation rather than a redundant classification.

### 3.4 Independent prediction of competing risks

The primary clinical utility of isolating localized regulatory discordance lies in its potential to delineate specific causes of death. In the METABRIC cohort, the probability of disease-specific death rises rapidly following diagnosis, whereas mortality from other causes accumulates steadily over time, largely driven by age (Figure 6).

To determine if the functional modules of uncoupling capture an orthogonal axis of localized molecular lethality, we constructed cause-specific Cox proportional hazards models adjusted for `AGE_AT_DIAGNOSIS` and `INTCLUST` (Figure 7).

The survival analysis revealed that specific discordance modules are independent predictors of disease-specific mortality, but not general mortality. Module M1 (luminal loss) was significantly associated with an increased risk of death from disease (Hazard Ratio [HR] = 1.042, 95% CI: 1.017–1.069, $p = 0.001$), yet showed no association with death from other causes (HR = 1.003, $p = 0.805$). Similarly, module M3 (proliferation) emerged as a strong predictor of disease-specific death (HR = 1.101, 95% CI: 1.060–1.144, $p < 0.0001$) without predicting other causes of death (HR = 1.001, $p = 0.952$). Conversely, module M5 exhibited a protective association specifically for disease-specific mortality (HR = 0.896, $p = 0.025$).

The overall concordance index for the disease-specific model was 0.6854, compared to 0.7722 for the model predicting death from other causes (which is dominated by `AGE_AT_DIAGNOSIS`). These results confirm that multi-layer regulatory discordance captures a distinct lethality signal that is orthogonal to the existing `INTCLUST` taxonomy and is specifically informative for cancer-related outcomes.

## 4 Conclusions

In this study, we established a multi-layer analytical framework to isolate localized transcriptional uncoupling from the broader noise of genomic instability and epigenetic alterations in breast cancer. By modeling gene expression residuals against discrete copy number states and promoter methylation, we identified a core

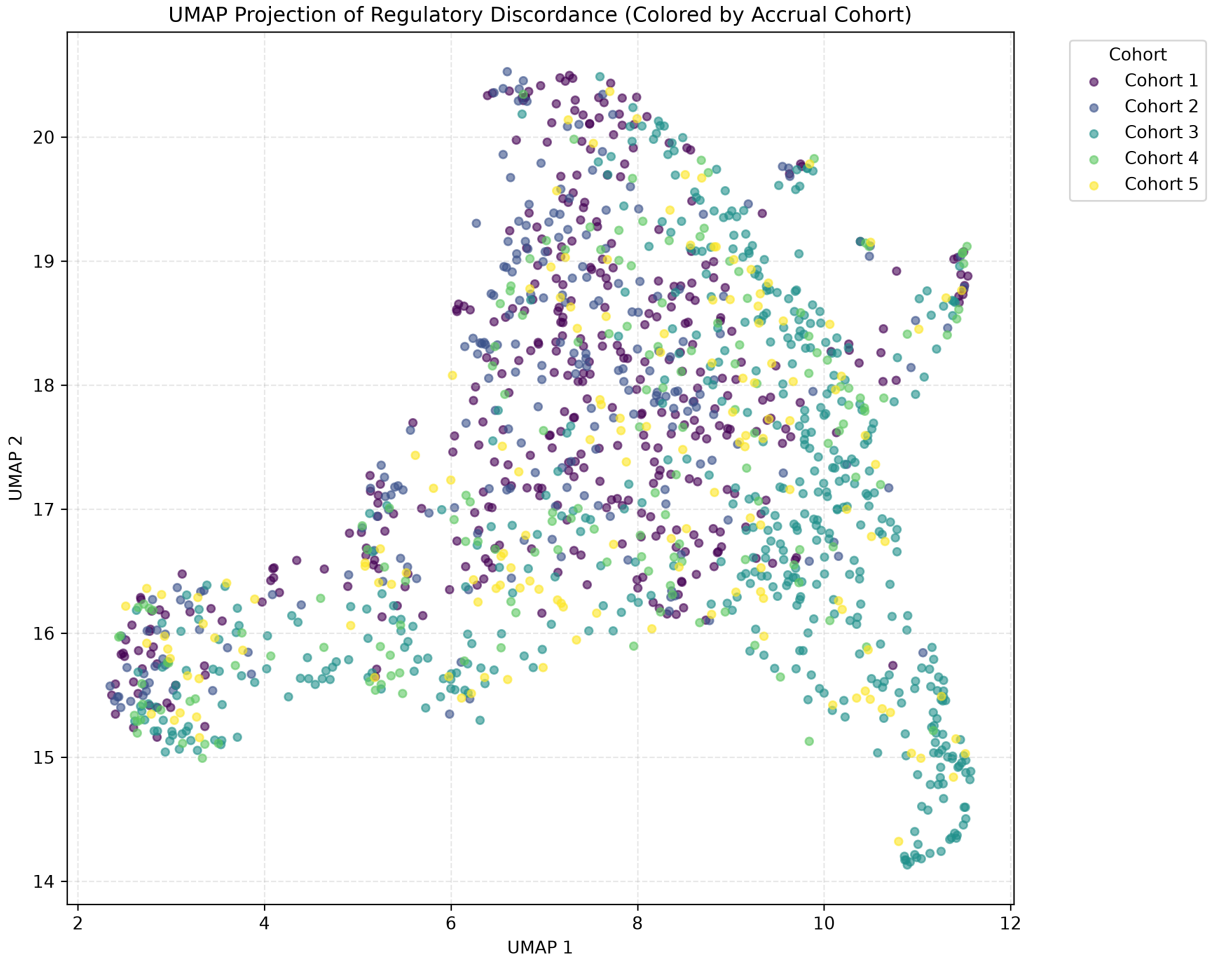



Figure 5: UMAP projection of the global regulatory discordance matrix colored by METABRIC accrual cohorts (1–5). The scatter plot demonstrates that samples from different cohorts are well-mixed across the lower-dimensional manifold. This uniform distribution confirms that the regulatory uncoupling signal captures true biological variation rather than being driven by cohort-specific technical batch effects, highlighting the structural stability of the discordance profiles across the dataset.

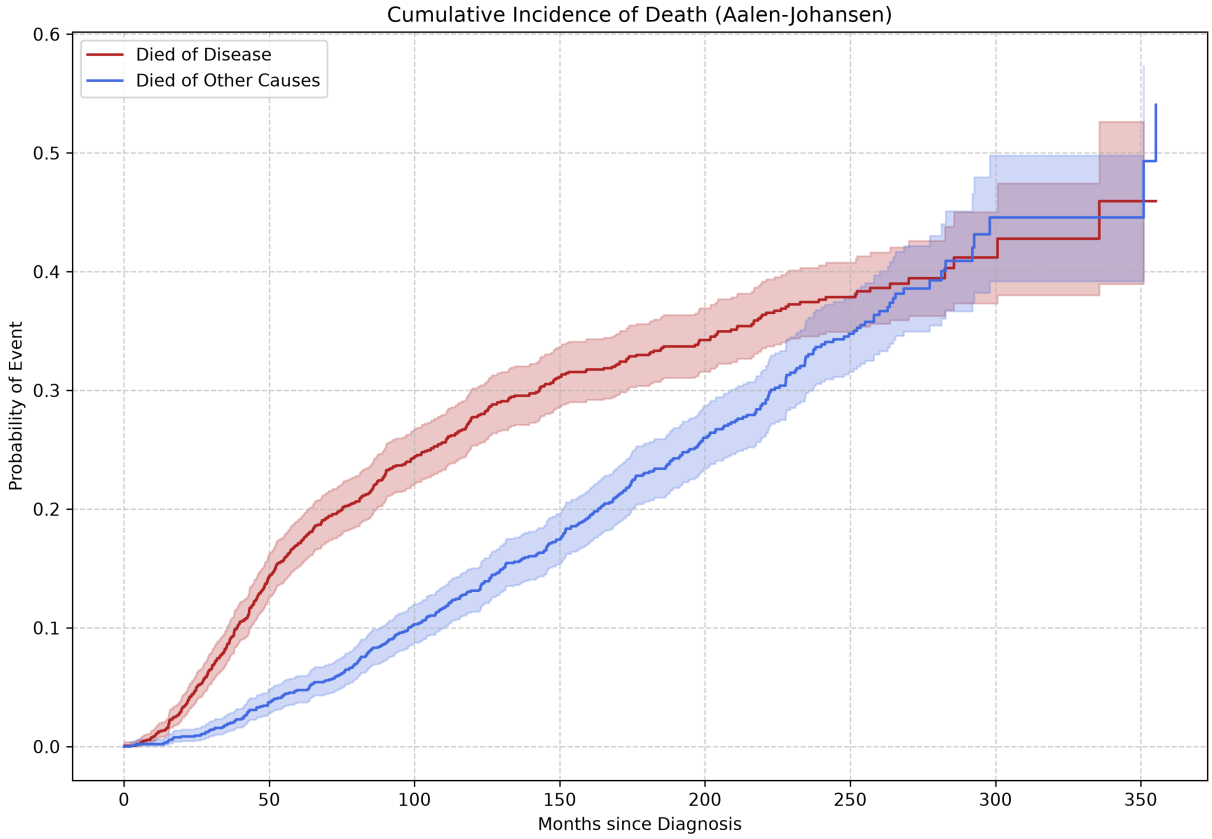



Figure 6: Cumulative incidence of death in the METABRIC cohort estimated via the Aalen-Johansen method, illustrating the competing risks of disease-specific mortality (red) and death from other causes (blue). The plot reveals distinct temporal dynamics: the probability of disease-specific death rises rapidly in the period immediately following diagnosis, whereas mortality from other causes accumulates steadily over time. This competing risks framework is essential for isolating cancer-specific lethality from general mortality, the latter of which is largely driven by `AGE_AT_DIAGNOSIS`. By separating these outcomes, cause-specific hazard models confirm that multi-layer regulatory discordance modules (such as the M1 luminal loss and M3 proliferation modules) specifically predict disease-specific mortality, providing an independent prognostic signal that is orthogonal to the established `INTCLUST` taxonomy.

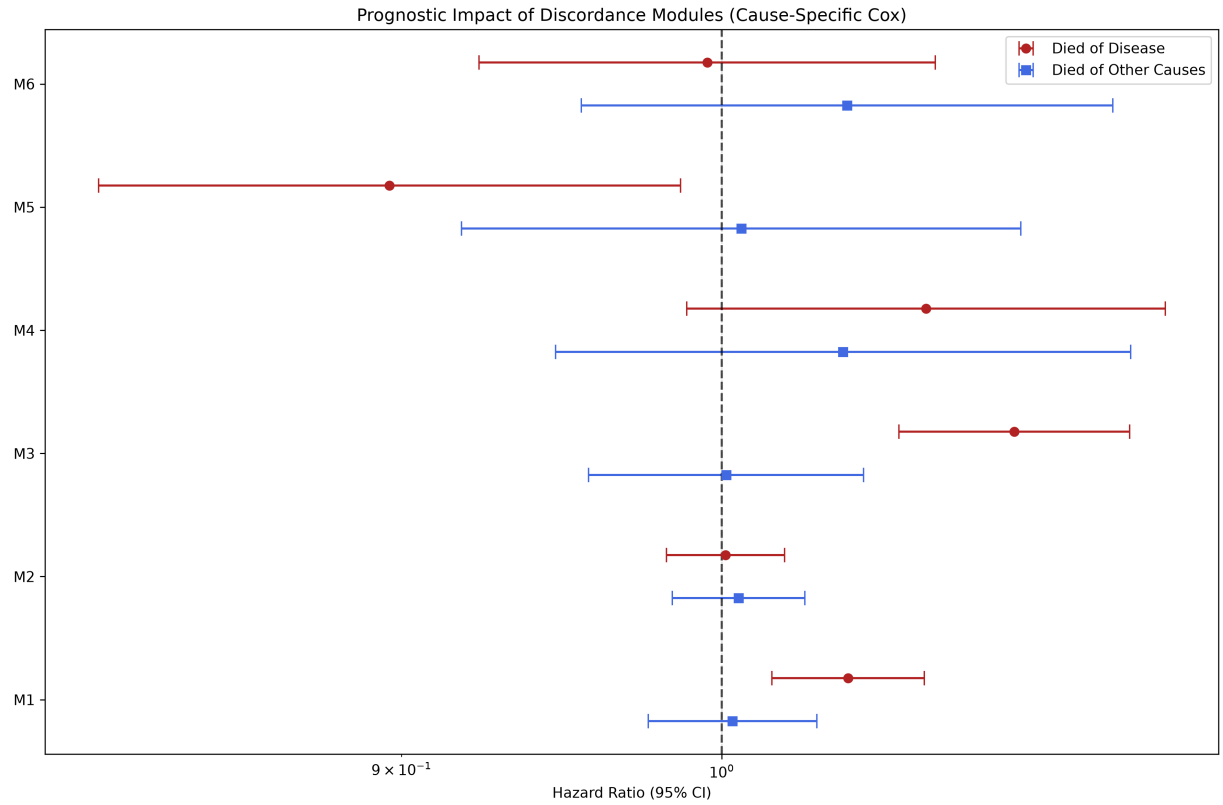



Figure 7: Forest plot detailing the prognostic impact of six regulatory discordance modules (M1–M6) using cause-specific Cox proportional hazards models adjusted for `AGE_AT_DIAGNOSIS` and `INTCLUST`. The plot displays hazard ratios and 95% confidence intervals for two competing risks: death from disease (red circles) and death from other causes (blue squares). The results demonstrate that specific modules are independent predictors of disease-specific mortality but not general mortality. Notably, module M1 (luminal loss) and module M3 (proliferation) are significantly associated with an increased risk of disease-specific death, whereas module M5 exhibits a protective association. None of the modules significantly predict death from other causes, indicating that multi-layer regulatory discordance captures a distinct lethality signal that is orthogonal to existing tumor taxonomies and specifically informative for cancer-related outcomes.

consensus of 271 genes that consistently escape regulatory control across all PAM50 subtypes. Crucially, while the individual targets of this regulatory discordance are highly conserved, their joint organization into lower-dimensional manifold topologies reveals profound structural divergence. This topological shift is most pronounced in claudin-low tumors, which exhibit a fundamentally rewired regulatory logic compared to classic luminal and basal phenotypes.

A central objective of this work was to determine whether localized regulatory discordance offers novel biological insights or merely provides a redundant representation of established molecular taxonomies. Our predictive modeling demonstrated that while discordance profiles can partially reconstruct the integrative clustering (`INTCLUST`) framework, they capture a distinct subset of variance that is fundamentally orthogonal to raw expression data. Furthermore, the structural stability of these discordance manifolds across five distinct METABRIC accrual cohorts confirms that this uncoupling signal is a robust biological phenomenon rather than an artifact of technical batch effects.

The clinical significance of this orthogonal signal is underscored by its ability to independently predict competing survival outcomes. By employing cause-specific hazard models adjusted for age and `INTCLUST`, we demonstrated that functional co-discordance modules are highly specific predictors of disease-specific mortality. Specifically, modules capturing the loss of luminal identity (M1) and dosage-independent proliferation (M3) significantly increase the hazard of cancer-related death without associating with other-cause mortality. Conversely, other modules, such as M5, capture protective regulatory adaptations that similarly apply only to disease-specific outcomes.

Ultimately, these findings highlight the critical role of localized regulatory failure in tumor evolution. By mapping the manifold topologies of multi-layer regulatory discordance, we provide a reproducible and independent axis for stratifying molecular lethality in breast cancer. This approach not only refines our understanding of transcriptional uncoupling but also offers a robust prognostic framework for isolating cancer-specific risks in highly heterogeneous patient populations.