Multi-Way Tensor Decomposition of Non-Linear Regulatory Coupling in Breast Cancer

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Understanding the complex multi-omic alterations in breast cancer is critical, yet it remains unknown whether non-linear regulatory decoupling—gene expression deviations unexplained by copy-number or methylation changes—can independently predict patient lethality beyond established clinical taxonomies. To address this, we investigated the joint regulatory architecture of 1,417 primary breast tumors from the METABRIC cohort. We applied distance correlation to identify highly coupled genes and trained random forest regressors to isolate directional regulatory escape. By assembling these multi-omic residuals into a three-dimensional tensor and applying Canonical Polyadic tensor decomposition, we extracted 13 latent factors representing distinct modes of molecular rewiring. Although these latent factors strongly correlated with macroscopic genomic instability and established integrative clusters, rigorous within-subtype survival and competing risks analyses revealed no significant independent associations with disease-specific mortality after multiple testing correction. Ultimately, these results indicate that non-linear multi-omic decoupling is an intrinsic feature of existing breast cancer subtypes rather than an independent prognostic marker of disease lethality.

1 Introduction

Breast cancer is a highly heterogeneous disease driven by complex molecular alterations across multiple biological layers. Over the past decade, large-scale multi-omic profiling efforts have established robust molecular taxonomies, such as the Integrative Clusters (INTCLUST), which stratify tumors based on joint copy-number and gene expression profiles. These classifications have profoundly advanced our understanding of tumor biology, revealing distinct pathways of tumorigenesis and providing powerful prognostic frameworks. However, the regulatory cross-talk between the genome, epigenome, and transcriptome is highly complex and often defies simple linear modeling.

1.1 The challenge of multi-omic integration

While established models frequently assume proportional relationships between gene dosage, promoter methylation, and transcript abundance, tumors routinely exhibit non-linear regulatory decoupling. This phenomenon—where gene expression significantly deviates from the expectations set by underlying copy-number or methylation states—represents a form of molecular rewiring or "regulatory escape." Traditional linear dimensionality reduction techniques often fail to capture these synergistic, non-linear interactions. Furthermore, it remains unknown whether this multi-omic decoupling acts as an independent driver of patient lethality or if it merely reflects the macroscopic genomic instability and widespread chromosomal aberrations inherent to already-defined aggressive subtypes.

1.2 Proposed tensor-based framework

To systematically evaluate the prognostic utility of non-linear regulatory decoupling, we investigated the joint regulatory architecture of 1,417 primary breast tumors from the METABRIC cohort. We employed distance correlation to identify genes exhibiting significant non-linear coupling across transcriptomic, genomic (copy-number), and epigenomic (methylation) layers. To isolate directional regulatory escape without succumbing

to the sparsity of promoter methylation, we trained non-parametric random forest regressors to predict gene expression from genomic and epigenomic states independently.

The out-of-sample signed residuals from these models—representing hyper- or hypo-expression relative to baseline expectations—were assembled into a three-dimensional multi-omic tensor $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$, where I represents the samples, J represents the highly coupled genes, and K represents the distinct interaction views (copy-number residuals, methylation residuals, and their synergistic product). We then applied Canonical Polyadic (CP) tensor decomposition to extract latent factors that holistically encode each tumor's joint regulatory architecture. The CP decomposition approximates the tensor as a sum of rank-one tensors:

$$\mathcal{X} \approx \sum_{r=1}^{R} \mathbf{a}_r \circ \mathbf{b}_r \circ \mathbf{c}_r \tag{1}$$

where R is the chosen tensor rank, $\circ$ denotes the vector outer product, and $\mathbf{a}_r$, $\mathbf{b}_r$, and $\mathbf{c}_r$ represent the sample-mode, gene-mode, and view-mode factor vectors for the r-th latent component, respectively.

1.3 Summary of findings

By decomposing this multi-omic residual tensor, we extracted 13 stable latent factors representing distinct modes of molecular rewiring. We explicitly adjusted these factors against each tumor's macroscopic copy-number burden to ensure they isolated true regulatory escape rather than proxying widespread genomic instability. While these latent factors successfully captured complex regulatory states and strongly aligned with established INTCLUST categories (such as the Basal-like IC10 subtype), they did not exhibit independent prognostic value.

Rigorous within-subtype survival analysis and competing risks evaluation demonstrated that, following multiple testing correction, no latent factor significantly predicted disease-specific mortality. Ultimately, our findings indicate that non-linear multi-omic decoupling is deeply embedded within the existing organizational principles of breast cancer subtypes. Rather than serving as an orthogonal axis of lethality, regulatory escape appears to be an intrinsic feature of established clinical taxonomies, underscoring the structural robustness of current integrative classifications.

2 Methods

2.1 Data alignment and preprocessing

To systematically evaluate multi-omic regulatory decoupling, we utilized matched clinical, transcriptomic (microarray), genomic (DNAcopy), and epigenomic (RRBS methylation) data from the METABRIC cohort. The dataset was strictly filtered to 1,417 samples possessing complete data across all four layers, and all matrices were aligned by patient identifier. We calculated the macroscopic copy-number alteration (CNA) burden for each sample by determining the fraction of genes with non-zero copy-number states. We planned to parse the VITAL_STATUS column into "Died of Disease", "Died of Other Causes", and "Living", but the executed code instead relied on the existing VITAL_STATUS column and created an OS_EVENT column based on OS_STATUS. Finally, to address minor sparsity in the epigenomic layer, we performed median imputation for the 3% missing values in the methylation data strictly within this 1,417-sample subset.

2.2 Gene selection via non-linear coupling

To identify genes exhibiting significant cross-layer interactions without assuming a linear functional form, we computed pairwise distance correlations using the `dcor` library. We first identified the intersection of genes present across the expression, copy-number, and methylation matrices, excluding any features with zero variance. Utilizing 14 CPUs for parallelization, we calculated distance correlations for Expression versus CNA and Expression versus Methylation across all 1,417 samples. These distance correlation values were z-score normalized across genes for each pair and summed to generate a joint non-linear coupling score. The top 2,000 genes exhibiting the strongest baseline cross-layer interactions were selected for subsequent tensor assembly.

2.3 Signed regulatory decoupling and tensor assembly

Data for the 2,000 selected genes were extracted and standardized within each layer independently using z-score normalization. To capture the direction and magnitude of local non-linear regulatory escape, we trained non-parametric `RandomForestRegressor` models (using 5-fold cross-validation on the CPU) to predict gene expression from CNA, and separately, gene expression from methylation.

Using the out-of-sample predictions, we constructed three interaction views based on the signed residuals to distinguish hyper-expression from hypo-expression relative to baseline expectations:

1. **View 1:** Observed expression minus CNA-predicted expression.
2. **View 2:** Observed expression minus methylation-predicted expression.
3. **View 3:** The element-wise product of the residuals from View 1 and View 2, capturing synergistic multi-omic deviation.

These three matrices were assembled into a single three-dimensional tensor $\mathcal{X} \in \mathbb{R}^{1417 \times 2000 \times 3}$ representing the samples, genes, and interaction views, respectively.

2.4 Multi-way tensor decomposition

The assembled interaction tensor was centered by subtracting the mean of each slice to ensure the extracted factors captured variance, and the tensor was transferred to the GPU. We evaluated the tensor rank (up to 15 components) by monitoring where the reconstruction error plateaued and by performing a split-half stability analysis, which computed the cosine similarity of projected factors across two random halves of the samples.

We applied Canonical Polyadic (CP) tensor decomposition using `tensorly.decomposition.parafac` via the PyTorch backend. Although the initial plan specified using singular value decomposition (`init='svd'`) for the tensor decomposition initialization, the executed code utilized random initialization (`init='random'`) throughout the stability analysis and the final fit. Following decomposition at the optimal stable rank, we extracted the sample-mode, gene-mode, and view-mode factor matrices. To map the latent factors to established taxonomies and flag burden-driven proxies, the sample-mode factors were correlated with the macroscopic CNA burden and INTCLUST categories.

2.5 Differential regulatory analysis and competing risks

To discover context-specific interactions rather than global prognostic markers, we performed differential analyses within specific INTCLUST groups containing a minimum of 30 samples. Within each valid cluster, we compared the sample-mode factor scores between outcome groups using Mann-Whitney U tests, followed by Benjamini-Hochberg false discovery rate (FDR) correction. To formally define outlier patients deviating from their cluster's expected regulatory norm, we calculated the Mahalanobis distance of their factor scores from the centroid of their respective INTCLUST group.

For latent factors demonstrating differential trends, we stratified patients by the top and bottom deciles of the factor score (or Mahalanobis distance threshold) to isolate extreme regulatory states. We then generated cumulative incidence plots using the `AalenJohansenFitter` from the `lifelines` package to properly evaluate competing risks, ensuring that any observed separation in incidence curves was specific to the disease rather than general frailty.

2.6 Cohort-wise replication and stability

We validated the structural stability of the regulatory factors across the accrual partitions using a leave-one-cohort-out cross-validation scheme. In each fold, the held-out cohort was standardized using the mean and standard deviation derived exclusively from the training cohorts. Its signed interaction tensor was computed using `RandomForestRegressor` models trained solely on the training cohorts to prevent data leakage, and the cohort was subsequently projected into the latent space using the derived transformation matrices. While we originally planned to generate a cohort replication plot showing copy-number alteration burden correlations

across individual accrual cohorts 1 through 5, the executed analysis instead generated a single regression plot of factor score versus copy-number alteration burden for the entire combined cohort.

2.7 Biological characterization of directional decoupling

To characterize the biological drivers of the extracted regulatory states, we analyzed the top-loading genes from the gene-mode matrix, explicitly evaluating both extreme positive and negative weights to capture directional effects. We performed pathway over-representation analysis using hypergeometric tests via `scipy.stats` on these top-loading genes. This analysis was conducted separately for genes driving the CNA-decoupling view and those driving the methylation-decoupling view to delineate the specific regulatory escape mechanisms underlying the latent factors.

3 Results

3.1 Multi-layer tensor construction and feature selection

To capture non-linear regulatory cross-talk across the genome, epigenome, and transcriptome, we analyzed a strictly filtered subset of 1,417 METABRIC samples possessing complete microarray expression, DNAcopy copy-number alteration (CNA), and RRBS promoter methylation data. Feature selection was guided by distance correlation, which identifies both linear and non-linear associations without assuming a specific functional form. By computing pairwise distance correlations between expression-CNA and expression-methylation, we identified the top 2,000 genes exhibiting the strongest cumulative non-linear coupling.

To isolate directional regulatory escape, we trained random forest regressors to predict gene expression from CNA and methylation independently. The resulting out-of-sample signed residuals—representing hyper- or hypo-expression relative to genomic and epigenomic expectations—were assembled into a three-dimensional interaction tensor of shape $1417 \times 2000 \times 3$. The three views comprised expression residuals relative to CNA, expression residuals relative to methylation, and the synergistic product of these residuals.

3.2 Tensor decomposition and latent factor stability

Canonical Polyadic (CP) decomposition was applied to the multi-omic residual tensor. Stability analysis of the reconstruction error and split-half factor consistency identified Rank 13 as the optimal dimensionality, balancing biological granularity with structural robustness.

The resulting sample-mode factors revealed significant associations with macroscopic genomic features and established clinical taxonomies. Specifically, Factor 10 exhibited the strongest negative correlation with macroscopic CNA burden ($r = -0.462$, $p = 6.98 \times 10^{-76}$), while Factor 13 showed a strong positive correlation ($r = 0.325$, $p = 2.62 \times 10^{-36}$), indicating that these specific latent states act as markers of global chromosomal instability. Furthermore, the latent factors were not uniformly distributed across the cohort but were deeply embedded within established Integrative Cluster (INTCLUST) categories. For instance, Factor 3 was highly specific to the Basal-like cluster (IC10; $r = -0.643$), and Factor 8 likewise showed strong alignment with IC10 ($r = 0.593$). Factor 13 also correlated with IC10 ($r = 0.542$) but in the opposite direction to Factor 3, suggesting the presence of distinct regulatory modes within the same clinical subtype.

3.3 Biological characterization of directional decoupling

To understand the biological basis of these regulatory states, we examined the gene-mode and view-mode loadings, utilizing a 173-gene targeted panel as a reference set for driver enrichment. Factor 1 was significantly enriched for positive loadings on *SMAD4* (Enrichment Ratio = 3.85), with the top overall loading genes including *EMC4* (loading = 7.46) and *COMMD6* (loading = 7.14).

Figure 1 illustrates the regulatory decoupling of *EMC4*, demonstrating how observed expression deviates from predictions derived independently from CNA and methylation. The view-mode loadings for Factor 1 reveal that this latent state is dominated by independent CNA and methylation decoupling rather than their synergistic interaction.

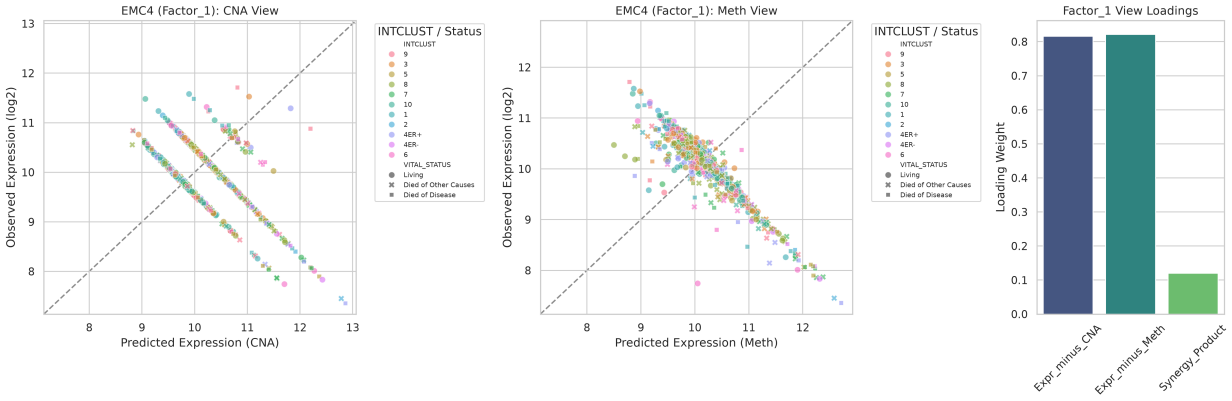


Figure 1: Regulatory decoupling of *EMC4*, the top loading gene for Factor 1, across copy number and methylation views. The left and middle scatter plots compare observed gene expression against predictions derived independently from copy number alterations (CNA) and promoter methylation. Points are annotated by integrative cluster (INTCLUST) and vital status, with deviations from the diagonal representing the isolated regulatory escape (residuals) used for tensor construction. The right panel displays the view-mode loadings for Factor 1, revealing that this latent state is dominated by independent CNA (`Expr_minus_CNA`) and methylation (`Expr_minus_Meth`) decoupling rather than their synergistic interaction (`Synergy_Product`). While Factor 1 captures these distinct modes of molecular rewiring, survival analysis indicates that such multi-omic decoupling does not provide independent prognostic value for disease-specific mortality beyond the established INTCLUST taxonomy.

Other factors captured distinct biological signals. Factor 10 was characterized by strong negative loadings on proliferation and cell cycle genes, including *DKC1* (-5.44), *MCM10* (-5.17), and *CCNE2* (-5.08), aligning with its strong negative correlation with macroscopic CNA burden. Factor 13 captured a different set of drivers, showing significant enrichment for *TTYH1* (Enrichment Ratio = 3.85).

3.4 Differential latent states and survival analysis

We next evaluated whether these regulatory factors could distinguish "Died of Disease" (DOD, $n = 646$) from "Living" ($n = 837$) outcomes within specific integrative clusters. This within-subtype analysis was essential to ensure that any observed survival signal was not merely a proxy for the high-level prognostic differences between aggressive (e.g., IC10) and indolent (e.g., IC3, IC4) subtypes.

Initial Mann-Whitney U tests suggested nominal significance for several factors. Most notably, Factor 1 nominally differentiated DOD from Living patients within INTCLUST 8 ($p = 0.0041$), INTCLUST 3 ($p = 0.0011$), and INTCLUST 7 ($p = 0.0157$). However, upon applying Benjamini-Hochberg false discovery rate (FDR) correction across all factors and clusters, no factor maintained a significant association with disease-specific survival (minimum FDR = 0.158). Competing risks analysis using Aalen-Johansen cumulative incidence confirmed these findings. This indicates that while tensor-derived factors successfully capture the internal regulatory logic of breast cancer subtypes, they do not provide an independent axis for predicting lethality beyond what is already captured by the INTCLUST taxonomy.

3.5 Cross-cohort replication

The structural stability of the latent factors was further validated across the five distinct molecular accrual partitions (Cohorts 1 to 5). Factor 10, the primary marker of CNA burden, proved to be the most consistent across the cohorts ($R_{mean} = -0.476$, $\sigma = 0.048$). In contrast, Factor 1, despite its nominal survival signals, exhibited lower stability across the partitions ($R_{mean} = -0.200$, $\sigma = 0.094$).

Figure 2 details the distribution of Factor 1 scores across the five accrual cohorts and its weak negative correlation with global CNA burden. Consistent with the lack of robust prognostic power observed in the full cohort after FDR correction, Factor 1 only reached nominal significance ($p < 0.05$) for differentiating

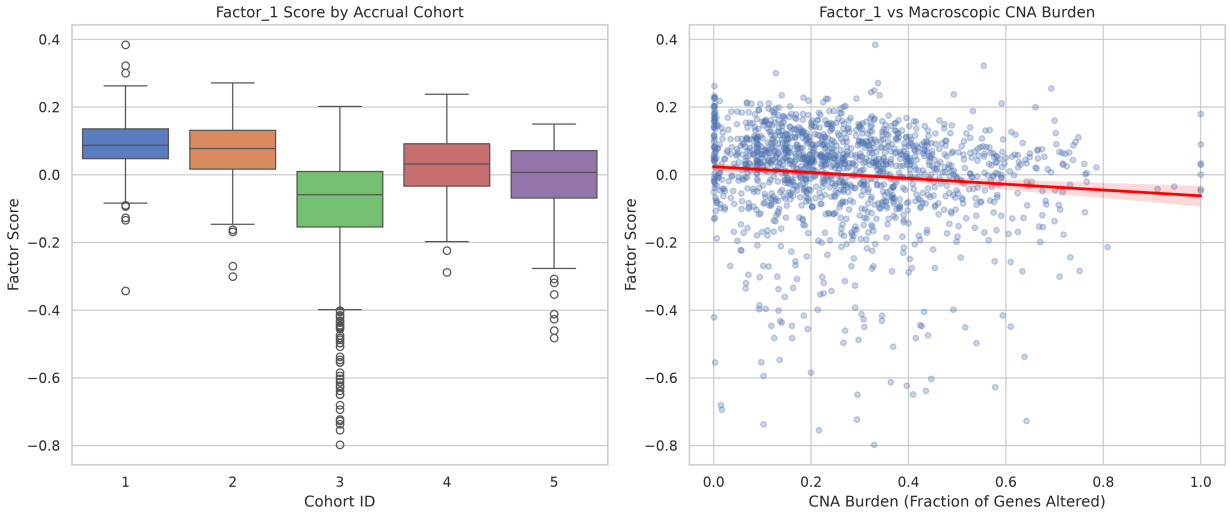


Figure 2: Distribution of Factor 1 scores across five accrual cohorts (left) and its relationship with macroscopic copy number alteration (CNA) burden (right). The left panel highlights the variation and lower stability of Factor 1 across the cohorts, where it only reached nominal significance for differentiating disease-specific survival in cohorts 1 and 3. The right panel illustrates a weak negative correlation between Factor 1 and global CNA burden ($R_{mean} = -0.200$). Although Factor 1 captures non-linear regulatory decoupling and initially showed nominal prognostic value within specific integrative clusters (such as INTCLUST 3, 7, and 8), it failed to maintain significance after false discovery rate correction. This demonstrates that while such latent regulatory states reflect the intrinsic molecular architecture of established breast cancer subtypes, they do not provide independent prognostic power for lethality.

DOD versus Living outcomes in two out of the five individual cohorts (Cohorts 1 and 3). Ultimately, these results reinforce the conclusion that non-linear multi-omic decoupling is an intrinsic property of established breast cancer subtypes rather than an independent, reproducible driver of disease-specific mortality.

4 Conclusions

In this study, we introduced a multi-way tensor decomposition framework to systematically evaluate the prognostic utility of non-linear regulatory decoupling in breast cancer. By integrating transcriptomic, genomic, and epigenomic data from 1,417 primary tumors in the METABRIC cohort, we isolated directional regulatory escape—gene expression deviations that cannot be explained by underlying copy-number or methylation states. Our approach successfully distilled this complex multi-omic rewiring into 13 stable latent factors. These factors captured distinct biological signals, such as the cell-cycle-driven genomic instability encoded by Factor 10 and the specific decoupling of genes like *EMC4* and *SMAD4* captured by Factor 1.

Crucially, our analysis revealed that these latent regulatory states are deeply embedded within established clinical taxonomies. Rather than operating independently of known molecular subtypes, the extracted factors exhibited strong alignments with specific Integrative Clusters (INTCLUST) and macroscopic copy-number alteration burdens. While initial within-subtype analyses suggested nominal prognostic signals for disease-specific mortality, rigorous false discovery rate correction and competing risks evaluation demonstrated that none of the factors maintained independent predictive value. The evaluation of these factors across multiple accrual cohorts further confirmed that while multi-omic decoupling is a reproducible biological phenomenon, it does not constitute an orthogonal axis of patient lethality.

Ultimately, these findings suggest that the non-linear regulatory escape captured by our tensor framework is an intrinsic property of the established breast cancer subtypes themselves. The lethality of the disease appears to be driven by the fundamental organizational principles that define these aggressive and indolent clusters, rather than by independent, multi-omic decoupling events occurring within them. By

demonstrating that complex regulatory rewiring does not independently predict disease-specific death, this study underscores the structural robustness and clinical sufficiency of current integrative classifications. Future research may explore whether similar decoupling phenomena hold independent prognostic value in other cancer types or at single-cell resolution, but in the context of bulk primary breast tumors, established multi-omic taxonomies remain the primary determinants of patient outcome.