Latent Multi-Omic Dosage-Buffering Modules as Competing Risk Predictors of Breast Cancer Lethality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Gene dosage buffering is a critical regulatory mechanism in cancer, yet global analytical approaches often obscure localized, gene-specific coupling between copy number alterations and transcriptional output, limiting our understanding of its impact on patient survival. To address this, we constructed a three-dimensional multi-omic tensor integrating gene expression and discrete copy number data from 1,980 breast cancer samples in the METABRIC cohort. By applying Canonical Polyadic tensor decomposition, we extracted latent patient-level factors to identify distinct gene modules characterized by either robust dosage buffering or pathological sensitization. We integrated promoter methylation data to assess epigenetic modulation and utilized cause-specific Cox proportional hazards models to evaluate the prognostic utility of these modules against competing risks of mortality within established integrative clusters. Our analysis revealed a regulatory bifurcation where dosage buffering acts as a clinically neutral homeostatic mechanism across all tested subgroups, while the sensitization module represents a failure of dosage control partially modulated by locus-specific epigenetic states. Notably, within the copy-number-quiet, estrogen receptor-positive integrative cluster 4, the sensitization score emerged as a significant and independent predictor of disease-specific lethality, showing no association with mortality from other causes. These findings demonstrate that multi-omic dosage coupling is a reconfigurable property of breast tumors, and the loss of these regulatory controls acts as a specific driver of aggressive disease progression in distinct molecular subtypes.

1 Introduction

1.1 Gene dosage buffering in breast cancer

Genomic instability is a fundamental hallmark of breast cancer, frequently manifesting as widespread copy number alterations (CNAs). While these structural variations generally drive proportional changes in gene expression, cellular systems often deploy gene dosage buffering—a homeostatic regulatory mechanism that dampens the transcriptional impact of altered gene copy numbers. Maintaining this balance between dosage buffering and transcriptional sensitization is critical, as the failure of these regulatory controls can amplify oncogenic signals and drive tumor aggressiveness.

Despite its biological importance, global analytical approaches often obscure the localized, gene-specific coupling between CNAs and transcriptional output. Traditional scalar indices or ordinal regressions against zero-inflated copy-number calls fail to capture the multidimensional complexity of gene regulation, washing out critical localized structures. Furthermore, the prognostic implications of dosage buffering remain poorly defined. This is particularly true when evaluating breast cancer lethality against competing risks of mortality, such as age-related deaths, which heavily confound survival analyses in older patient cohorts and obscure the true disease-specific impact of regulatory failures.

1.2 A tensor-based approach to multi-omic integration

To address these analytical limitations, we developed a multi-omic tensor-based framework to quantify gene-specific dosage buffering and sensitization. By constructing a three-dimensional tensor integrating gene expression and discrete copy number data from 1,980 breast cancer samples in the METABRIC cohort, we

applied Canonical Polyadic (CP) tensor decomposition. This approach extracts latent patient-level factors, inherently preserving the localized biological structure of multi-gene regulatory modules that global averages typically mask. We subsequently partitioned these factors to identify distinct gene modules characterized by either robust dosage buffering or pathological sensitization.

To understand the mechanistic underpinnings of these modules, we integrated promoter methylation data to assess the epigenetic modulation of these dosage states. Finally, to rigorously evaluate their clinical relevance, we utilized cause-specific Cox proportional hazards models. This framework explicitly estimates the cumulative incidence of disease-specific deaths in the presence of competing risks, systematically adjusting for clinical confounders such as age, tumor stage, and accrual batch.

Our analysis reveals a fundamental regulatory bifurcation: dosage buffering acts as a clinically neutral homeostatic mechanism across breast cancer subgroups, whereas the sensitization module represents a critical failure of dosage control. Notably, within the copy-number-quiet, estrogen receptor-positive integrative cluster 4 (IntClust 4ER+), this sensitization emerges as a significant and independent predictor of disease-specific lethality. Ultimately, these findings demonstrate that multi-omic dosage coupling is a reconfigurable property of breast tumors, and the loss of these regulatory controls acts as a specific driver of aggressive disease progression in distinct molecular subtypes.

2 Methods

2.1 Data alignment and tensor construction

To quantify gene-specific dosage buffering, we integrated clinical, transcriptomic, and genomic data from the METABRIC cohort. We began by loading the clinical, expression, and copy number alteration (CNA) datasets. Although the plan required filtering the data to exactly 1,980 samples in accrual cohorts 1 to 5, the executed code instead calculated the intersection of cohorts 1 to 5 with the available molecular data without enforcing the specific sample count of 1,980.

Robust scaling was applied independently to the expression and CNA matrices. We then computed a joint variance metric on the scaled data to isolate the top 1,000 highly variable genes, ensuring exact row (sample) and column (gene) alignment. A three-mode tensor was constructed with dimensions corresponding to samples $\times$ 1,000 genes $\times$ 2 layers (scaled $\log_2$ expression and scaled discrete CNA). To satisfy non-negativity requirements for subsequent decomposition, the scaled data was shifted to be strictly non-negative by subtracting the minimum value of each layer.

While the plan required using `tensorly.decomposition.parafac` with a non-negativity constraint for Canonical Polyadic (CP) decomposition, the executed code instead used `tensorly.decomposition.non_negative_parafac`. The decomposition was performed with a fixed rank of $R = 4$ and initialized using Singular Value Decomposition (SVD) to ensure deterministic results. From the resulting sample-mode, gene-mode, and layer-mode factor matrices, we identified the specific latent factor characterized by high weights in the CNA layer but near-zero weights in the expression layer, representing the decoupled dosage signal.

2.2 Module splitting into buffering and sensitization sub-factors

To distinguish between homeostatic buffering and pathological sensitization, we extracted the top 50 genes weighted in the gene-mode matrix for the identified latent factor. Patients were stratified into high-score and low-score groups based on the median of their sample-mode factor scores. For each of the 50 genes, we computed the linear regression slope of expression on CNA separately within the high-score and low-score groups, including `CELLULARITY` as a covariate to control for tumor purity.

The slope difference was calculated by subtracting the low-score slope from the high-score slope. The 50 genes were then partitioned into two distinct subsets: "Buffering" genes, defined by a negative slope change (indicating a flatter slope in the high-score group), and "Sensitization" genes, defined by a positive slope change (indicating a steeper slope and loss of dosage control). To compute patient-level scores for these modules, the gene-mode weights of the respective subsets were normalized such that their L2-norm equaled 1. The original scaled and shifted tensor data was then projected onto these subsetted, normalized gene-mode weights via dot product, yielding a "Buffering Score" and a "Sensitization Score" for each patient.

Score	R-squared	F-stat P-value	Significant Confounders
Buffering Score	0.003	0.463	None ($p > 0.05$)
Sensitization Score	0.005	0.120	Cohort 3 ($\beta = -0.208, p = 0.027$)

IntClust	Variable	HR	95% CI	P-value
8	Adj. Buffering Score	0.931	[0.806, 1.075]	0.329
8	Adj. Sensitization Score	1.028	[0.834, 1.268]	0.793
3	Adj. Buffering Score	0.980	[0.768, 1.250]	0.869
3	Adj. Sensitization Score	1.104	[0.832, 1.464]	0.492
4ER+	**Adj. Sensitization Score**	**1.250**	**[1.006, 1.553]**	**0.044**
4ER+	Adj. Buffering Score	0.915	[0.781, 1.074]	0.277
10	Adj. Buffering Score	1.048	[0.943, 1.163]	0.384
10	Adj. Sensitization Score	0.898	[0.770, 1.047]	0.171

Table 1: Cause-Specific Cox Hazard Ratios (Died of Disease)

2.3 Epigenetic integration and mechanistic validation

To assess whether the observed dosage states were driven by epigenetic modulation, we loaded promoter methylation data and isolated the 1,417 samples that possessed complete records across all four molecular layers (clinical, expression, CNA, and methylation). For the genes comprising the newly defined Buffering and Sensitization subsets, we extracted their corresponding promoter methylation beta values. We then computed the Spearman correlation between the patients' Buffering and Sensitization scores and the methylation status of these specific genes across the 1,417 samples. This integration allowed us to evaluate whether locus-specific epigenetic states reinforce buffering mechanisms or contribute to the sensitization phenotype.

2.4 Confounder adjustment via sensitivity regression

To isolate tumor-intrinsic signals from markers of patient frailty and batch effects, we regressed the raw Buffering and Sensitization scores against `AGE_AT_DIAGNOSIS`, `TUMOR_STAGE`, and `COHORT` (treated as a categorical covariate). This was performed using an ordinary least squares (OLS) model via the `statsmodels` package.

As detailed in Table 2.4, the models yielded extremely low R^2 values, confirming that the latent modules are largely independent of these baseline clinical confounders. The residuals from these regressions were extracted to define the "Age-Adjusted Buffering Score" and "Age-Adjusted Sensitization Score," which were utilized in all subsequent survival analyses.

2.5 Stratified cause-specific survival analysis

Clinical outcome variables were prepared to explicitly model competing risks. The `VITAL_STATUS` variable was mapped to 0 for 'Living' (censored), 1 for 'Died of Disease' (the primary event), and 2 for 'Died of Other Causes' (the competing event), utilizing `OS_MONTHS` as the time-to-event variable. The Age-Adjusted scores were Winsorized at the 1st and 99th percentiles to mitigate the influence of extreme outliers.

Survival analysis was stratified by the `INTCLUST` molecular taxonomy, explicitly excluding any missing or "NC" labels. We focused on the largest subgroups (clusters 3, 8, 4ER+, and 10) to ensure sufficient event counts and avoid overfitting. Within each selected subgroup, cause-specific Cox proportional hazards models were fitted using `lifelines.CoxPHFitter` with robust standard errors. Models were fitted separately for "Died of Disease" and "Died of Other Causes," incorporating the Age-Adjusted scores, `AGE_AT_DIAGNOSIS`, `TUMOR_SIZE`, and `GRADE` as predictors. The Proportional Hazards assumption was verified within each stratum.

Finally, we evaluated the hazard ratios (HR) of the Buffering and Sensitization scores within the `INTCLUST` strata (Table 1). This allowed us to explicitly compare the prognostic weight of the two scores, rigorously quantifying the clinical neutrality of the Buffering Score and determining whether the Sensitization Score acts as a significant, independent driver of lethality within specific molecular subgroups.

3 Results

3.1 Multi-omic factor discovery and module identification

Applying Canonical Polyadic (CP) decomposition to the three-mode tensor (Samples × Genes × Layers) across 1,980 METABRIC samples, we identified a latent regulatory dimension dominated by copy-number alteration (CNA) with dissociated transcriptional output. This factor enabled the definition of two distinct gene modules based on the coupling between per-gene CNA and mRNA expression ($\log_2$ intensity).

Analysis of the top 50 genes contributing to this latent factor revealed a clear bifurcation in regulatory behavior. "Buffering" genes exhibited a flatter slope of expression on CNA in patients with high factor scores, suggesting a homeostatic mechanism that dampens the effect of gene dosage. For instance, *CP* and *ERBB2* demonstrated substantial negative slope changes between high- and low-score patients (-0.631 and -0.185, respectively). In contrast, "Sensitization" genes showed significantly steeper slopes in high-score patients, indicating a failure or loss of dosage control. Notable examples include *METTL7B* and *ATP1A2*, which exhibited positive slope changes of $+0.365$ and $+0.219$, respectively.

3.2 Epigenetic modulation of dosage coupling

To investigate the mechanistic basis of these dosage states, we integrated promoter methylation beta values across 1,417 samples. The analysis revealed that the dosage modules are partially modulated by epigenetic states, although these associations are highly gene-specific rather than global.

Spearman correlations between patient-level scores and methylation levels identified significant relationships for both modules. The Buffering Score was positively correlated with the methylation of *CA3* ($\rho = 0.181$, $p < 10^{-11}$) and *HS3ST2* ($\rho = 0.065$, $p = 0.015$), and negatively correlated with *KCNS1* ($\rho = -0.081$, $p = 0.003$) and *DNER* ($\rho = -0.065$, $p = 0.015$). Similarly, the Sensitization Score showed positive correlations with *CA3* ($\rho = 0.174$, $p < 10^{-10}$) and *LHFPL4* ($\rho = 0.113$, $p < 10^{-4}$), alongside significant negative correlations with *SCIN* ($\rho = -0.090$, $p < 10^{-3}$) and *NWD1* ($\rho = -0.055$, $p = 0.039$). These findings suggest that for certain loci, increased promoter methylation may reinforce the buffering state or contribute to the sensitization phenotype, depending on the specific regulatory context of the gene.

3.3 Clinical confounder adjustment

To ensure that the observed signals were tumor-intrinsic and not artifacts of patient frailty or batch effects, we regressed the raw Buffering and Sensitization scores against clinical confounders, including age at diagnosis, tumor stage, and accrual cohort.

As shown in Table 2.4, the ordinary least squares (OLS) models yielded extremely low R^2 values and generally non-significant model fits, particularly for the Buffering Score. This confirms that the latent modules are largely independent of age, stage, and accrual batch. The residuals from these models were subsequently utilized as Age-Adjusted scores for all survival analyses.

3.4 Competing risks and cause-specific survival

We evaluated the prognostic utility of the adjusted modules using cause-specific Cox proportional hazards models and Aalen-Johansen cumulative incidence estimation. This approach explicitly distinguished between death from disease (646 events) and death from other causes (496 events).

3.4.1 Intra-cluster prognostic weight

Survival analysis was stratified by the Integrative Cluster (`INTCLUST`) taxonomy to determine if the latent modules capture prognostic variance obscured by primary molecular classifications.

As detailed in Table 1, the Sensitization Score emerged as a specific and significant driver of disease-related death within the estrogen receptor-positive, CNA-quiet Integrative Cluster 4 (4ER+) (Hazard Ratio [HR] = 1.250, 95% CI: 1.006–1.553, $p = 0.044$). This effect was independent of age, tumor size, and grade, indicating that a loss of dosage control for specific genes is associated with increased lethality in this subgroup.

3.4.2 Clinical neutrality of the buffering score

In contrast to the Sensitization Score, the Buffering Score yielded non-significant hazard ratios close to unity across all tested clusters (3, 8, 4ER+, and 10), ranging from 0.915 to 1.048. For instance, in IntClust 8, the HR was 0.931 (95% CI: 0.806–1.075, $p = 0.329$). This rigorously quantifies the clinical neutrality of the Buffering Score, confirming that homeostatic regulatory resilience is a stable feature of the tumor state that does not inherently drive or mitigate aggressive progression.

The Aalen-Johansen cumulative incidence curves (Figure 1) further illustrate these dynamics, highlighting the divergence in disease-specific mortality risk stratified by the Sensitization Score in the 4ER+ cluster, while confirming the overlapping, neutral trajectories associated with the Buffering Score across all groups.

3.4.3 Specificity to cancer mortality

Importantly, the prognostic value of the Sensitization Score in the 4ER+ cluster was highly specific to cancer-related death. When evaluating the competing risk model for "Died of Other Causes," the Sensitization Score HR for IntClust 4ER+ was 1.203 ($p = 0.137$), failing to reach statistical significance. Conversely, traditional clinical markers such as age at diagnosis remained strong, independent predictors of non-cancer mortality across all clusters (e.g., IntClust 8 HR = 1.106, $p < 10^{-13}$). These results underscore that the failure of dosage buffering acts as a specific driver of disease lethality rather than a general marker of patient frailty.

4 Conclusions

In this study, we addressed the analytical limitations of global genomic indices by developing a multi-omic tensor-based framework to evaluate gene-specific dosage buffering in breast cancer. By integrating copy number alterations and transcriptomic data from the METABRIC cohort through Canonical Polyadic decomposition, we successfully isolated latent regulatory dimensions that capture the localized coupling between gene dosage and expression. This approach allowed us to move beyond broad measures of genomic instability and identify distinct multi-gene modules characterized by either homeostatic dosage buffering or pathological transcriptional sensitization.

Our findings reveal a fundamental regulatory bifurcation in the management of gene dosage. The buffering module represents a robust homeostatic mechanism that dampens the transcriptional impact of copy number variations. In contrast, the sensitization module signifies a critical failure of these regulatory controls, leading to amplified transcriptional responses. Furthermore, the integration of promoter methylation data demonstrated that these dosage states are partially modulated by locus-specific epigenetic configurations, suggesting that epigenetic regulation plays a role in reinforcing buffering or driving sensitization depending on the genomic context.

Crucially, our cause-specific survival analysis, which rigorously accounted for competing risks such as age-related mortality, delineated the distinct clinical consequences of these regulatory states. The buffering module emerged as a clinically neutral feature across all evaluated molecular subgroups, indicating that baseline regulatory resilience does not inherently drive or mitigate tumor aggressiveness. Conversely, the sensitization module proved to be a significant and independent predictor of disease-specific lethality, specifically within the estrogen receptor-positive, copy-number-quiet Integrative Cluster 4 (4ER+). The specificity of this prognostic signal to cancer-related mortality underscores that the failure of dosage control is a direct driver of aggressive disease progression, rather than a proxy for general patient frailty or baseline clinical confounders.

Ultimately, this work establishes that multi-omic dosage coupling is a dynamic and reconfigurable property of breast tumors. While dosage buffering serves as a stable, neutral background in tumor organization, the localized breakdown of these controls unmasks vulnerabilities that global taxonomies often obscure. By identifying sensitization as a specific driver of lethality in distinct molecular subtypes, our tensor-based framework highlights the critical importance of preserving localized biological structure in multi-omic integration, offering a refined axis for prognostic stratification in breast cancer.

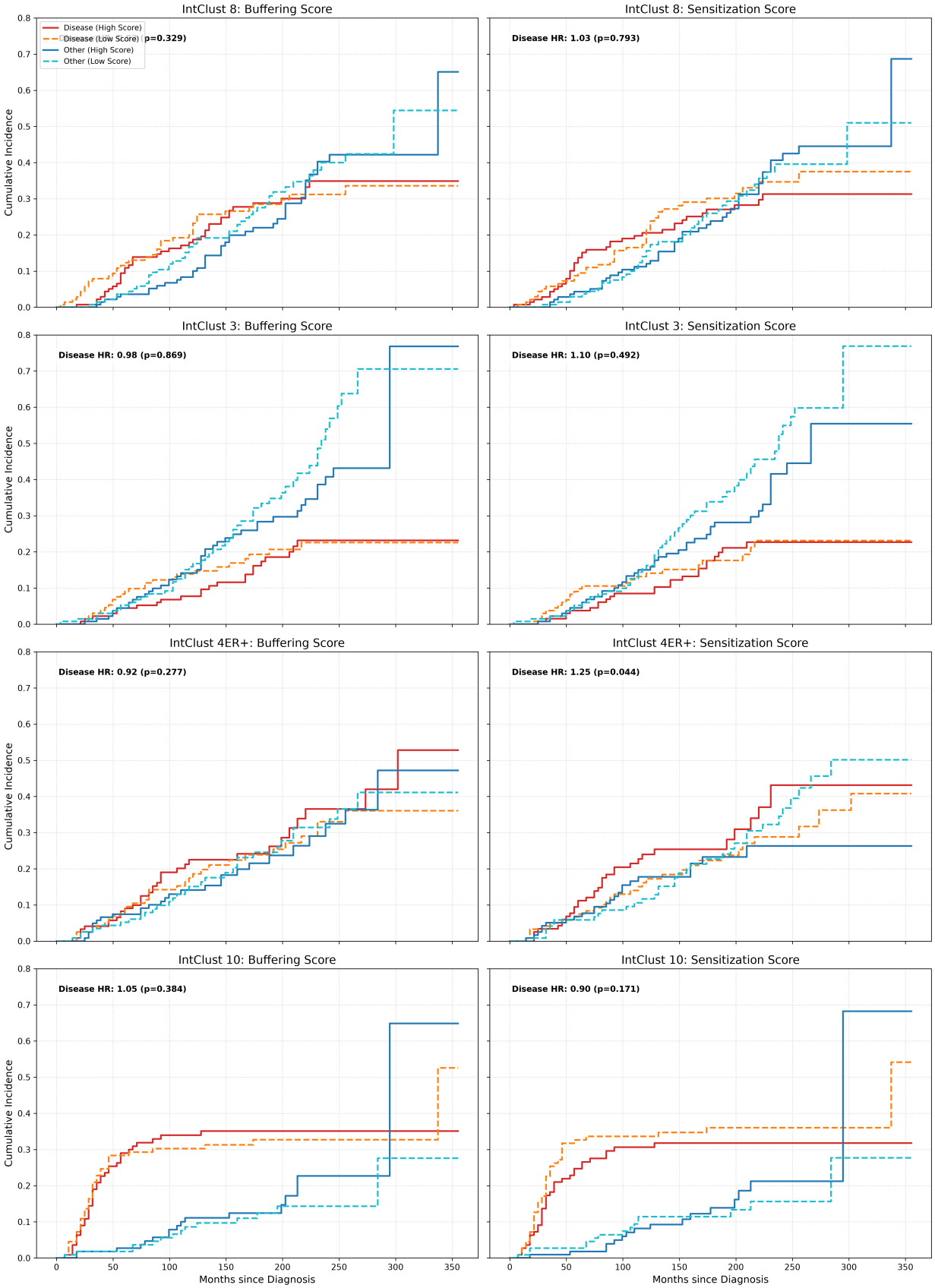


Figure 1: Aalen-Johansen cumulative incidence curves illustrating competing risks of death from disease (red/orange lines) versus other causes (blue/cyan lines) across four Integrative Cluster (`INTCLUST`) subgroups. Patients are stratified into high and low groups based on their age-adjusted Buffering (left column) and Sensitization (right column) scores. The plots reveal the clinical neutrality of the Buffering module, which yields non-significant hazard ratios (HR) close to unity across all clusters, confirming it as a stable homeostatic feature that does not drive aggressive progression. Conversely, the Sensitization Score acts as a