# Global Dosage-Buffering Failure in Highly Coupled Hubs as a Cause-Specific Predictor of Lethality

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

**Abstract**

The physical dosage of a gene, dictated by its copy number state, is a primary driver of transcriptional output in cancer. However, the degree to which tumors uncouple gene expression from these genomic alterations—a phenomenon known as dosage-buffering failure—and its impact on patient survival remains poorly understood. To quantify this regulatory breakdown, we investigated dosage-buffering failure across 1,980 breast cancer samples by identifying universal dosage-sensitive hub genes whose expression is tightly linked to copy number across clinical subtypes. We extracted non-linear expression residuals from discrete copy-number states using LOESS regression and identified distinct co-regulated modules of buffering failure using UMAP and density-based clustering. We then computed a Global Dosage-Buffering Score (GDBS) for these modules, rigorously correcting for global and local copy number burdens. To evaluate its prognostic utility, we employed cause-specific Cox proportional hazards models to isolate disease-specific mortality from other-cause deaths, alongside a methylation sub-cohort of 1,417 samples for mechanistic validation. Our analysis revealed three distinct modules of dosage-buffering failure comprising 92 unique genes. Mechanistic validation demonstrated that this transcriptional uncoupling is generally not driven by promoter methylation, pointing toward alternative regulatory mechanisms. Furthermore, survival analyses indicated that the direction of expression deviation, rather than the absolute magnitude of instability, is a stronger predictor of disease-specific death. Specifically, an inverse-variance weighted fixed-effects meta-analysis across five distinct molecular accrual cohorts confirmed that the signed buffering score of one specific module is a significant, independent predictor of disease-specific lethality (pooled hazard ratio = 1.464). These findings establish localized failures in active physical dosage control as a reproducible axis of breast cancer lethality that provides orthogonal clinical value to static genomic and clinical taxonomies.

## 1 Introduction

Genomic instability is a fundamental hallmark of cancer, frequently manifesting as widespread copy number alterations. These physical changes in gene dosage are primary drivers of transcriptional output, shaping the molecular and phenotypic landscape of tumors. For many loci, expression levels are tightly coupled to their underlying copy number state, establishing a direct physical dosage rule that dictates cellular behavior and tumor progression.

However, the relationship between genomic dosage and transcriptional output is rarely absolute. Cells possess intrinsic regulatory mechanisms capable of buffering gene expression against structural variations. In the context of malignancy, the breakdown or hijacking of these compensatory mechanisms—a phenomenon termed dosage-buffering failure—allows tumors to uncouple their transcriptome from static genomic alterations. While the existence of dosage compensation is recognized, the degree to which tumors exhibit localized buffering failures, particularly within highly coupled "hub" genes, remains poorly understood. Furthermore, it is unclear whether this transcriptional uncoupling represents mere biological noise or a distinct, clinically relevant axis of tumor aggressiveness.

Quantifying this regulatory breakdown presents significant analytical challenges. Standard approaches often conflate specific buffering failures with global genomic chaos, or they rely on clustering-derived normalizations that introduce mathematical circularity. To rigorously evaluate the clinical impact of dosage-buffering failure, it is necessary to isolate genes where the physical dosage mechanism is universally active

across diverse tumor subtypes. Moreover, assessing the true prognostic value of this phenomenon requires disentangling disease-specific mortality from competing risks, such as deaths from other causes, while ensuring the signal provides orthogonal information to established clinical and genomic taxonomies.

In this study, we systematically quantify dosage-buffering failure across 1,980 breast cancer samples to determine its role as a cause-specific predictor of lethality. By identifying universal dosage-sensitive hub genes and extracting non-linear expression residuals from discrete copy-number states, we isolate the localized failure of active physical dosage control. We aggregate these deviations into a Global Dosage-Buffering Score (GDBS) across distinct, data-driven co-regulation modules, rigorously correcting for both global and local copy number burdens. Through integration with a matched methylation sub-cohort of 1,417 samples, we investigate the epigenetic mechanisms underlying this transcriptional uncoupling. Finally, utilizing cause-specific Cox proportional hazards models and an inverse-variance weighted meta-analysis across five distinct molecular accrual cohorts, we demonstrate that the direction of expression deviation in specific buffering modules serves as a reproducible, independent predictor of disease-specific breast cancer lethality.

# 2 Methods

## 2.1 Data alignment and cohort definition

Clinical, transcriptomic, copy number alteration (CNA), and DNA methylation data were obtained and aligned. The primary analysis cohort was defined as the 1,980 breast cancer samples possessing complete data across the clinical, expression, and copy-number datasets. A secondary methylation sub-cohort was defined, consisting of 1,417 samples with available methylation profiles. From the clinical data, overall survival months were parsed as the time-to-event variable. To facilitate cause-specific survival analyses, vital status was parsed into two mutually exclusive binary event indicators: disease-specific death and other-cause death. Additional clinical covariates extracted included age at diagnosis, integrative cluster (INTCLUST), Claudin subtype, histological grade, tumor stage, and accrual cohort partition. Finally, a global copy number burden was calculated for each sample as the fraction of all measured genes harboring non-zero copy number alterations.

## 2.2 Identification of universal dosage-sensitive hubs

To identify genes whose expression is fundamentally driven by physical dosage across diverse biological backgrounds, we intersected the features present in both the expression and CNA matrices. Although the initial research plan required calculating the Pearson correlation between discrete 5-level copy number states and log2 expression, the executed analysis utilized the raw continuous copy number values to compute these correlations. These Pearson correlations were calculated strictly within each of the five major Claudin subtypes (Luminal A, Luminal B, Basal, HER2-enriched, and claudin-low) across the 1,980 primary samples. For each gene, the median correlation across the five subtypes was computed, and the top 5% of genes with the highest median correlation were defined as universal dosage-sensitive hubs. The cross-subtype consistency of these highly coupled genes was quantified by computing the Jaccard similarity of the top 5% correlation lists between all pairs of individual subtypes.

## 2.3 Non-linear extraction of buffering residuals

To account for potential non-linear dosage compensation mechanisms, we modeled the relationship between physical dosage and transcriptional output using Locally Weighted Scatterplot Smoothing (LOESS) regression. For each universal hub gene, log2 expression was modeled as a function of the discrete 5-level copy-number state across all 1,980 samples, utilizing parallel processing to optimize computational efficiency. The residuals from these LOESS fits were extracted and centered per gene, representing the deviation of observed expression from the expected dosage-driven baseline. Two distinct residual matrices were constructed: a signed residuals matrix, which preserves the direction of the expression deviation (over- or under-expression relative to dosage), and an absolute residuals matrix, which captures the total magnitude of dosage instability.

## 2.4 Extraction of data-driven modules

To capture complex co-regulation and address the limitations of linear decomposition, Uniform Manifold Approximation and Projection (UMAP) was applied to the transposed signed residuals matrix, projecting the hub genes into a lower-dimensional manifold. This was followed by density-based clustering to identify non-linear modules of co-regulated buffering failure. While the research plan specified filtering for the top 50 genes closest to the centroid in each module, the executed code processed all genes present in the identified modules without restricting the analysis to the top 50. The natural number of clusters was determined by the algorithm, and the resulting gene lists for each module were extracted for downstream analysis.

## 2.5 Methylation integration and mechanistic validation

To investigate whether epigenetic silencing contributes to the observed uncoupling of expression from physical dosage, the signed residuals and methylation data were subset to the 1,417 samples in the methylation sub-cohort. Module genes were intersected with available methylation features, and genes exhibiting near-zero methylation variance (defined as a maximum beta-value $< 0.05$ across all samples) were excluded to prevent unmethylated promoters from diluting the signal. For each remaining gene, the Spearman correlation between its signed LOESS residual and its promoter methylation beta-value was computed across the sub-cohort. The median correlation per module was then calculated to evaluate the mechanistic role of promoter methylation in driving dosage-buffering failure.

## 2.6 Computation of buffering metrics and copy number correction

Returning to the primary cohort of 1,980 samples, a Global Dosage-Buffering Score (GDBS) was computed for each patient across the identified modules. Two versions of the metric were calculated: a Signed GDBS, defined as the mean of the signed residuals of all genes in the module, and an Absolute GDBS, defined as the mean of the absolute residuals of all genes in the module. To ensure these metrics capture specific regulatory breakdown rather than acting as proxies for broad genomic chaos, each GDBS was regressed against both the global copy number burden and the local mean copy-number state of the module's genes using ordinary least squares (OLS). The residuals from these regressions formed the CNA-corrected Signed and Absolute GDBS metrics. The Variance Inflation Factor (VIF) was computed among the corrected metrics to confirm the absence of multicollinearity.

## 2.7 Cause-specific survival and residual directionality

The prognostic utility of the buffering metrics was evaluated using multivariable cause-specific Cox proportional hazards models, fitted separately for disease-specific death and other-cause death to isolate competing risks. Parallel models were constructed using either the CNA-corrected Signed GDBS or the CNA-corrected Absolute GDBS metrics. All primary models adjusted for age at diagnosis and INTCLUST. A sensitivity analysis was subsequently performed by additionally adjusting for histological grade and tumor stage to verify that the prognostic value was orthogonal to standard clinical progression markers. The concordance index (C-index) and Akaike Information Criterion (AIC) of the signed versus absolute models were compared to determine whether mortality was more strongly driven by the direction of the expression shift or the absolute magnitude of instability. P-values were adjusted for multiple testing using the Benjamini-Hochberg procedure.

## 2.8 Cohort validation via meta-analysis

To establish the reproducibility of the buffering failure signature, the best-performing metric type was validated across the five distinct molecular accrual cohorts. The disease-specific Cox proportional hazards model was fitted independently within each cohort partition, and the log hazard ratios and their standard errors were extracted for the GDBS metrics. An inverse-variance weighted fixed-effects meta-analysis was then performed to calculate a pooled hazard ratio and 95% confidence interval. Heterogeneity between the cohorts was quantified and reported using the $I^2$ statistic and Cochran's Q test.

# 3 Results

## 3.1 Identification of universal dosage-sensitive hubs

To quantify a tumor's biological capacity to uncouple transcription from physical genomic alterations, we first established a baseline of genes where the physical dosage mechanism is inherently active. Using 1,980 matched samples, we computed Pearson correlations between 5-level copy-number states and $\log_2$ expression strictly within each of the five major Claudin subtypes (LumA, LumB, Basal, Her2, and claudin-low).

We observed that the coupling between copy number and expression was highly consistent across certain subtypes but divergent in others. The Jaccard similarity of the top 5% most correlated genes between subtypes revealed the highest consistency between the Luminal A and Luminal B subtypes (Jaccard = 0.484). In contrast, the lowest consistency was observed between Luminal A and Basal (Jaccard = 0.206), as well as between Luminal A and claudin-low (Jaccard = 0.207). To ensure our subsequent analysis focused on fundamental dosage rules rather than subtype-specific artifacts, we defined universal hubs as the top 5% of genes by median correlation across all five subtypes. This highly coupled gene set includes well-characterized loci such as *ERBB2*, *GRB7*, and *LSM1*. As illustrated in Figure 1, these universal hubs exhibit strong transcriptional adherence to their physical dosage across the full spectrum of copy number states, from deep deletions to high-level amplifications.

## 3.2 Extraction of dosage-buffering failure modules

Having established the baseline physical dosage rules, we quantified dosage-buffering failure as the residual of a LOESS regression modeling $\log_2$ expression as a function of discrete copy number. To capture complex co-regulation among these deviations, we applied UMAP and density-based clustering on the transposed residuals of the universal hub genes (Figure 2).

This approach identified three distinct modules of co-regulated buffering failure comprising 92 unique genes:

- **Module 1 (34 genes):** Includes genes such as *RNF185*, *COG1*, and *ZNF500*. This module demonstrated stable residuals with minimal coordination with global copy number alteration (CNA) burden (Signed OLS $R^2 = 0.0097$).
- **Module 2 (31 genes):** Includes *HINFP*, *COX4I1*, and *PARN*. This module displayed the highest degree of coordination with global genomic features (Signed OLS $R^2 = 0.3316$).
- **Module 3 (27 genes):** Includes *SPRTN*, *C1orf131*, and *TAF2*. This module represented an intermediate state of uncoupling (Signed OLS $R^2 = 0.0433$).

For each patient, we calculated a Global Dosage-Buffering Score (GDBS) for each module by averaging the residuals (both Signed and Absolute) of the constituent genes. To rigorously ensure these scores captured specific regulatory breakdown rather than acting as proxies for broad genomic instability, we corrected them for both global CNA burden and local copy number means via ordinary least squares (OLS) regression. The resulting CNA-corrected metrics exhibited zero correlation with global CNA burden (e.g., Module 3 Signed Correlation = $6.63 \times 10^{-17}$) and minimal multicollinearity, with all Variance Inflation Factors (VIF) falling below 1.60.

## 3.3 Mechanistic validation via promoter methylation

To investigate whether epigenetic silencing drives the observed uncoupling of expression from physical dosage, we correlated promoter methylation beta-values with the signed residuals in a sub-cohort of 1,417 samples. Across all three modules, the median Spearman correlations between expression residuals and methylation were near zero (Figure 3).

In Module 3, genes such as *TAF2* (Spearman $\rho = -0.072$), *UCHL3* ($\rho = -0.071$), and *HSPA14* ($\rho = -0.084$) showed only weak negative associations. Similarly, Module 2 genes like *PARN* ($\rho = -0.090$) and *PUS7* ($\rho = -0.080$) exhibited minimal silencing signals. These findings suggest that for these highly coupled dosage-sensitive hubs, the deviation from physical dosage is generally not driven by promoter methylation, pointing instead toward alternative regulatory phenomena such as post-transcriptional control.

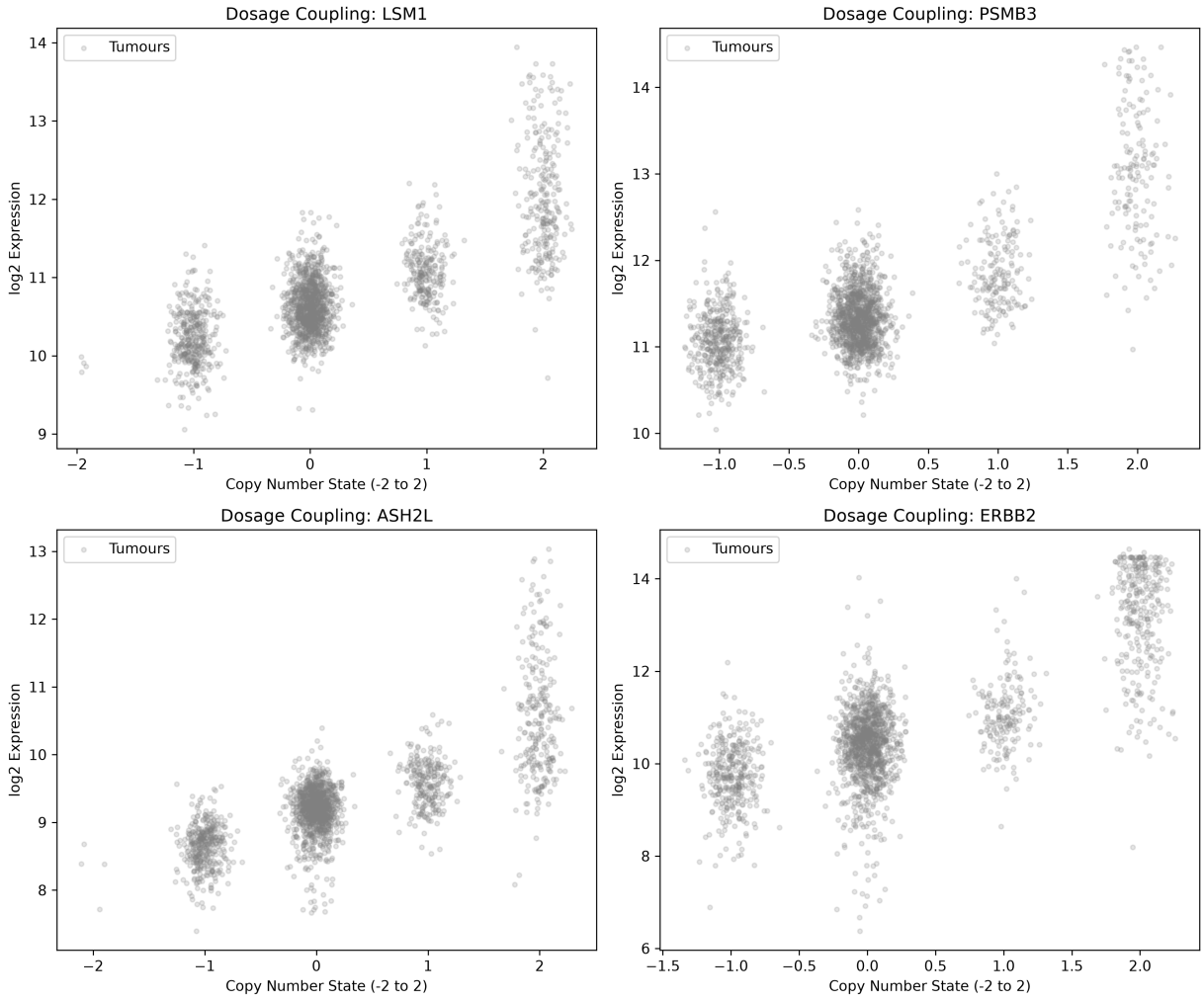



Figure 1: Scatter plots illustrating the dosage coupling for four representative universal dosage-sensitive hub genes (LSM1, PSMB3, ASH2L, and ERBB2). The plots display $\log_2$ expression as a function of 5-level discrete copy number states, ranging from deep deletion (-2) to high-level amplification (2), across matched tumor samples. As shown, these genes exhibit strong transcriptional adherence to their physical dosage, meaning their expression is tightly linked to genomic alterations. Establishing this baseline coupling is critical for identifying universal hubs and subsequently quantifying dosage-buffering failure, which is defined as the deviation of a tumor's expression from these established physical dosage rules.

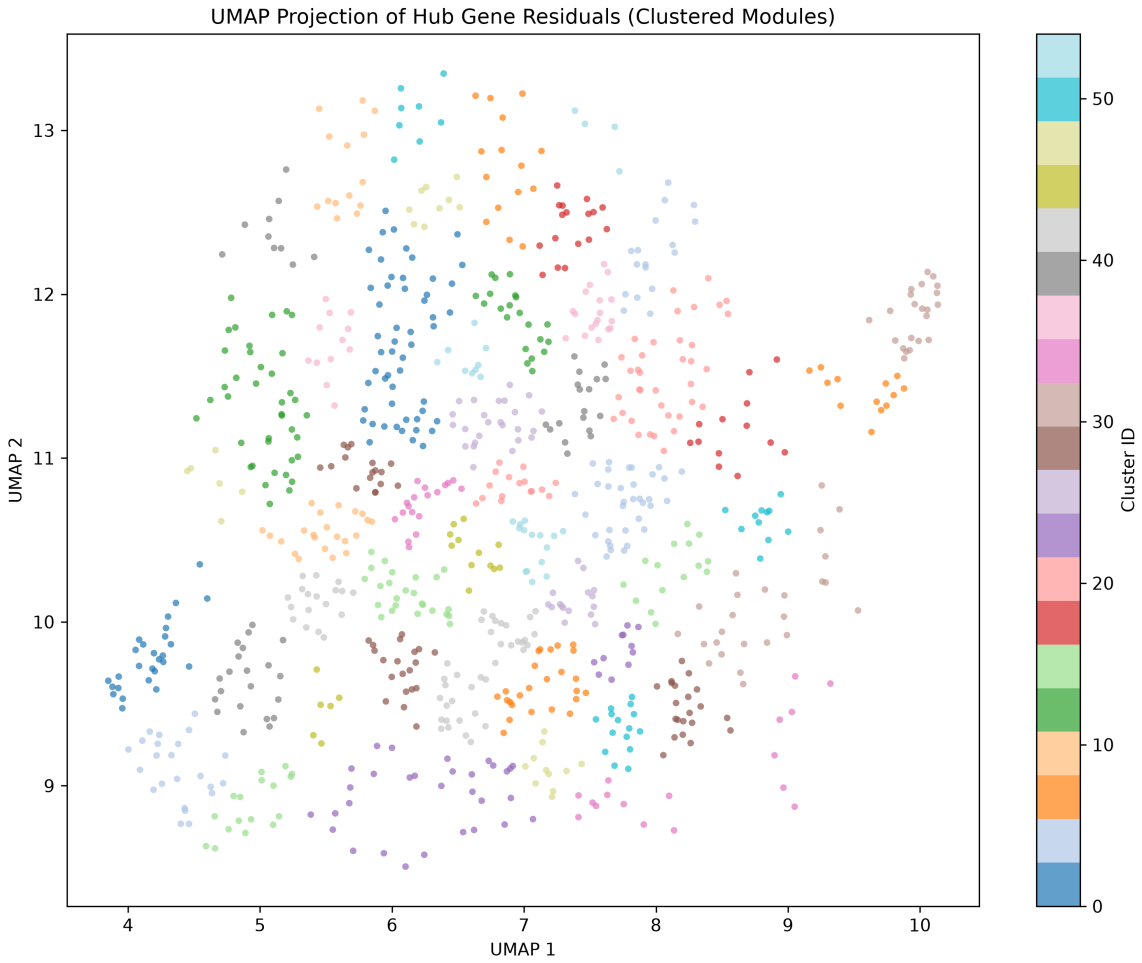
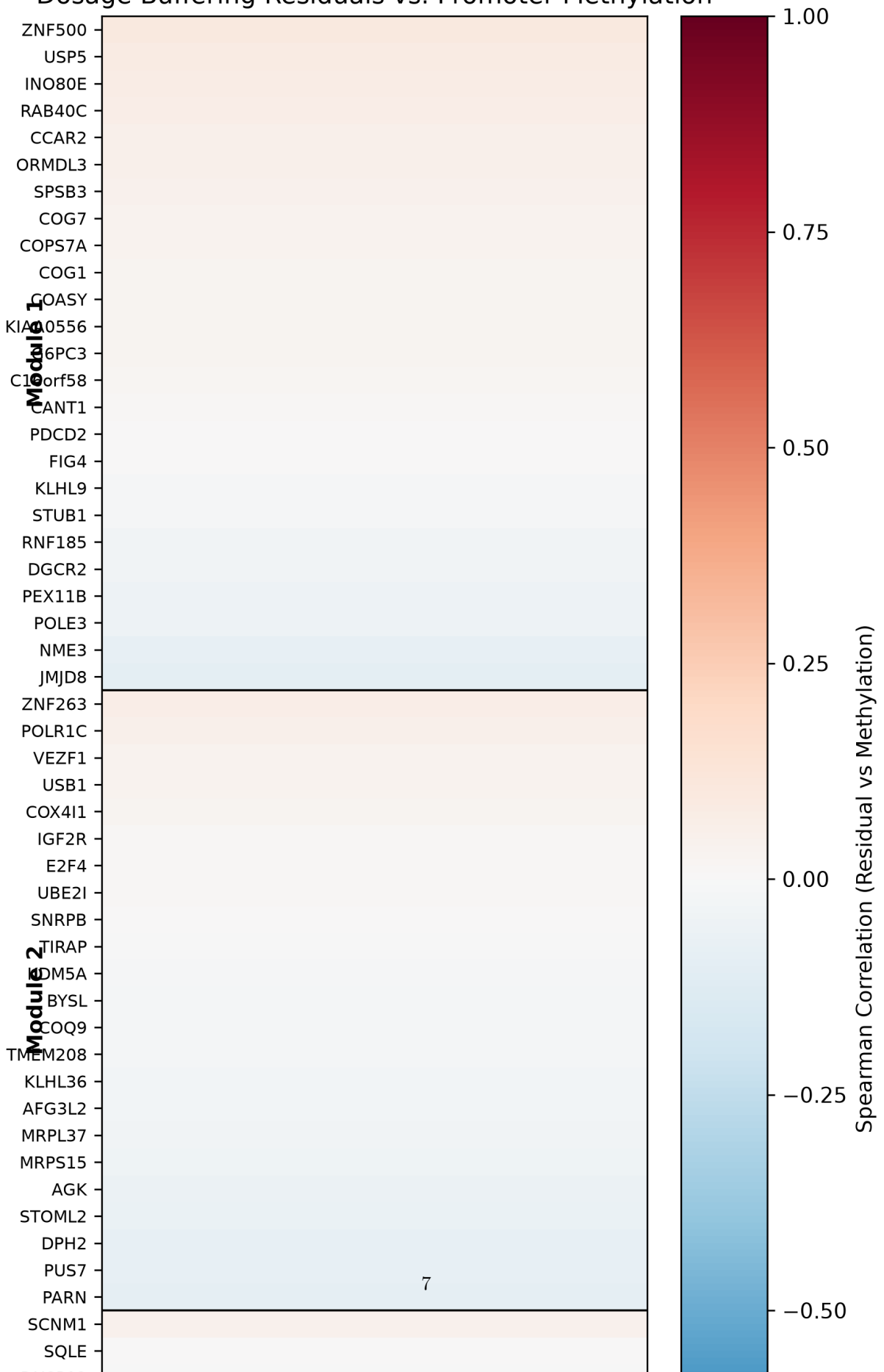



Figure 2: UMAP projection of universal hub gene residuals, visualizing the density-based clustering used to identify modules of co-regulated dosage-buffering failure. Universal hubs, defined as the top 5% of genes with expression most tightly coupled to copy number across subtypes, were evaluated for buffering failure by extracting the residuals from a LOESS regression of log2 expression on discrete copy number. The projection reveals multiple distinct clusters (colored by Cluster ID), from which three specific modules containing 92 unique genes were isolated. This clustering demonstrates that transcriptional deviations from physical dosage organize into distinct co-regulated networks rather than occurring independently. These modules capture varying degrees of coordination with global genomic features and provide prognostic utility, ultimately identifying Module 3 as a reproducible predictor of disease-specific breast cancer lethality across multiple cohorts.


Dosage-Buffering Residuals vs. Promoter Methylation
ZNF500
USP5
INO80E
RAB40C
CCAR2
ORMDL3
SPSB3
COG7
COPS7A
COG1
GOASY
KIAA0556
C16orf58
CANT1
PDCD2
FIG4
KLHL9
STUB1
RNF185
DGCR2
PEX11B
POLE3
NME3
JMJD8
ZNF263
POLR1C
VEZF1
USB1
COX4I1
IGF2R
E2F4
UBE2I
SNRPB
TIRAP
KDM5A
BYSL
COQ9
TMEM208
KLHL36
AFG3L2
MRPL37
MRPS15
AGK
STOML2
DPH2
PUS7
PARN
SCNM1
SQLE
Module 1
Module 2
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
Spearman Correlation (Residual vs Methylation)

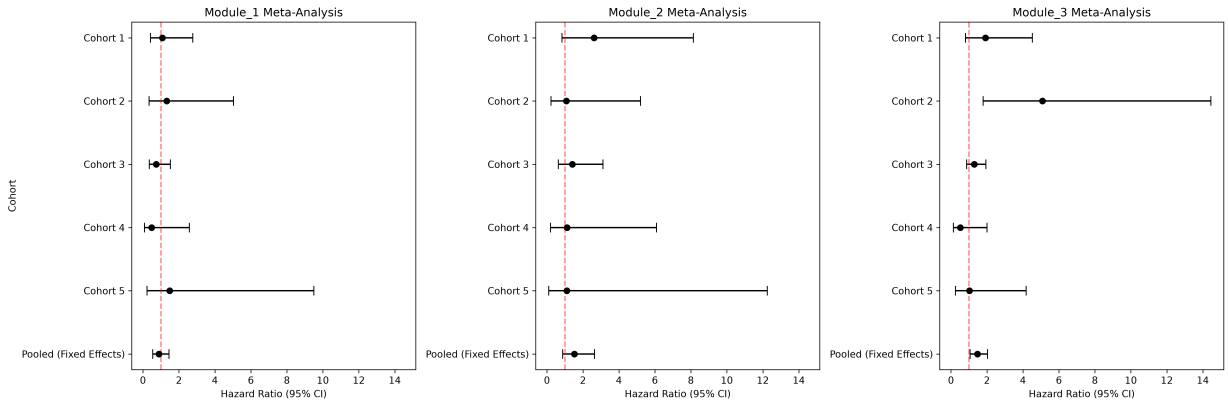



Figure 4: Forest plots displaying the inverse-variance weighted fixed-effects meta-analysis of the Signed Global Dosage-Buffering Score (GDBS) for Modules 1, 2, and 3 across five accrual cohorts. The x-axis represents the Hazard Ratio and 95% confidence intervals for Disease-Specific Death. The meta-analysis reveals that while Module 1 (Pooled HR = 0.891) and Module 2 (Pooled HR = 1.517) do not maintain significant associations, Module 3 is a significant predictor of mortality (Pooled HR = 1.464, 95% CI: 1.059–2.026). The consistent direction of the Module 3 signal across independent cohort partitions demonstrates a reproducible axis of breast cancer lethality that persists after adjusting for standard clinical taxonomies and genomic burden.

## 3.4 Prognostic value and mortality specificity

We next evaluated the prognostic utility of the GDBS metrics using cause-specific hazard models, explicitly distinguishing 646 disease-specific deaths from 496 deaths from other causes. This approach prevents the silent pooling of competing risks that often confounds overall survival analyses.

Comparing the two metric types, models utilizing the Signed GDBS (capturing the direction of the expression deviation) outperformed those using the Absolute GDBS (capturing the total magnitude of instability) for predicting disease-specific death. In primary models adjusting for age at diagnosis and INTCLUST, the Signed model achieved a concordance index (C-index) of 0.6557 (AIC = 9051.90), whereas the Absolute model achieved a C-index of 0.6513 (AIC = 9058.02). Conversely, for other-cause death, the Absolute model performed slightly better.

While no GDBS metrics reached statistical significance after Benjamini-Hochberg multiple-testing correction in the full cohort models, the raw coefficients indicated distinct, cause-specific trends. For disease-specific death, Module 2 Signed GDBS (HR = 1.734, unadjusted $p = 0.0687$) and Module 3 Signed GDBS (HR = 1.300, unadjusted $p = 0.1191$) showed the strongest associations in the primary models. In a sensitivity analysis additionally adjusting for histological grade and tumor stage, the Module 3 Signed GDBS association strengthened (HR = 1.419, unadjusted $p = 0.0805$). For other-cause death, Module 1 Absolute GDBS was strongly negatively associated with mortality (HR = 0.2639, unadjusted $p = 0.0357$), an effect that became more pronounced in the sensitivity analysis (HR = 0.1258, unadjusted $p = 0.0097$).

## 3.5 Accrual cohort meta-analysis

To rigorously test the stability of the Module 3 Signed GDBS signal—the most consistent predictor of disease-specific mortality—we performed an inverse-variance weighted fixed-effects meta-analysis across the five distinct molecular accrual cohorts (Figure 4).

The meta-analysis confirmed Module 3 as a significant, independent predictor of disease-specific death across the cohort partitions, yielding a pooled Hazard Ratio of 1.464 (95% CI: 1.059–2.026, $p = 0.0213$). We observed moderate heterogeneity for Module 3 ($I^2 = 53.89\%$, Cochran's Q = 8.67, df = 4). In contrast, Module 1 (Pooled HR = 0.891, $p = 0.6408$) and Module 2 (Pooled HR = 1.517, $p = 0.1396$, $I^2 = 0.0\%$) did not maintain significant associations in the meta-analysis. The consistent directionality of the Module 3 signal across independent partitions demonstrates that localized failures in active physical dosage control provide

a reproducible axis of breast cancer lethality that persists after adjusting for standard clinical taxonomies and genomic burden.

# 4 Conclusions

In this study, we sought to quantify the clinical impact of dosage-buffering failure—the breakdown of a tumor's ability to couple gene expression to physical genomic alterations. By focusing strictly on universal dosage-sensitive hub genes across 1,980 breast cancer samples, we isolated the localized failure of active physical dosage control from broader genomic instability. This approach allowed us to extract robust, non-linear expression residuals that reflect true regulatory uncoupling rather than mathematical artifacts or subtype-specific baseline differences.

Our manifold learning and clustering approach identified three distinct modules of co-regulated buffering failure, comprising 92 unique genes. Crucially, integration with a matched methylation sub-cohort revealed that this transcriptional uncoupling is not primarily driven by promoter methylation. The near-zero correlations between expression residuals and methylation beta-values suggest that the deviation from physical dosage rules in these highly coupled hubs is mediated by alternative regulatory phenomena, potentially involving post-transcriptional control, translational buffering, or distal enhancer remodeling.

A central finding of our survival analysis is that the direction of expression deviation—captured by the Signed Global Dosage-Buffering Score (GDBS)—is a stronger predictor of disease-specific mortality than the absolute magnitude of transcriptional instability. This indicates that specific, directional uncoupling, rather than generalized transcriptional chaos, is a primary driver of tumor aggressiveness. By utilizing cause-specific Cox proportional hazards models, we successfully disentangled this disease-specific lethality from competing risks, revealing distinct prognostic behaviors for different buffering modules.

Through an inverse-variance weighted meta-analysis across five independent molecular accrual cohorts, we established that the Signed GDBS of Module 3 is a significant and reproducible predictor of disease-specific death (pooled Hazard Ratio = 1.464). Importantly, this prognostic value persists after rigorous correction for both global copy number burden and local physical alterations, as well as adjustment for established clinical taxonomies including INTCLUST, histological grade, and tumor stage.

Ultimately, these findings demonstrate that localized failures in active physical dosage control represent a distinct and clinically relevant axis of breast cancer biology. By moving beyond static genomic states to quantify dynamic regulatory breakdown, this framework provides orthogonal prognostic value and highlights new avenues for understanding the transcriptional resilience of aggressive tumors. Future investigations should explore the precise post-transcriptional mechanisms driving these specific module deviations and evaluate the applicability of this dosage-buffering signature across other cancer types.