Pathway-Level Regulatory Deformation via Optimal Transport as an Intra-Cluster Cause-Specific Predictor of Lethality in Breast Cancer

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

Traditional multi-omic models often fail to capture localized regulatory breakdowns where transcriptional output decouples from DNA copy-number dosage. To quantify this loss of physical dosage control, we introduce the Pathway Deformation Score (PDS), an optimal transport-based metric that measures the topological departure of a tumor's empirical dosage-expression distribution from a cohort-wide baseline. Using matched clinical, transcriptomic, and copy-number data from 1,980 breast cancer samples, we identified functionally coherent gene modules and constructed baseline joint distributions specific to established molecular subtypes. By computing the 2D Wasserstein distance between individual sample point clouds and these baselines, we derived the PDS to capture network-specific regulatory uncoupling. We evaluated the prognostic utility of this metric using cause-specific Cox proportional hazards models adjusted for clinical covariates. Our results demonstrate that pathway deformation provides significant, independent prognostic value for disease-specific lethality, outperforming models that rely solely on the mean and variance of pathway expression. Empirical validation confirmed that highly deformed samples exhibit significantly greater expression dispersion within fixed copy-number states, verifying a true regulatory breakdown. Finally, this prognostic signal showed complete directional consistency across five accrual cohorts and correlated significantly with TP53 mutation status and stromal infiltration, establishing pathway-level regulatory deformation as a robust and reproducible predictor of intra-subtype tumor lethality.

1 Introduction

Breast cancer is a highly heterogeneous disease driven by complex genomic and transcriptomic alterations. Multi-omic profiling has become a cornerstone for understanding tumor biology and stratifying patients into molecular subtypes. A fundamental assumption in many of these models is the presence of physical dosage control, where DNA copy-number alterations (CNAs) proportionally drive changes in transcriptional output. However, the cancer transcriptome is frequently subject to localized regulatory breakdowns, leading to a decoupling between gene dosage and expression.

Traditional multi-omic integration strategies often struggle to capture these non-linear, localized regulatory uncouplings. Standard approaches typically rely on linear associations, mean expression levels, or variance within predefined gene sets. These methods are susceptible to the curse of dimensionality and can be heavily confounded by hyper-variable, subtype-defining genes (e.g., *SCGB2A2*, *TFF1*). Consequently, they may overlook subtle topological shifts in the joint distribution of copy number and expression that signify a loss of tight transcriptional control within specific biological networks.

To address this limitation and quantify the loss of physical dosage control, we introduce the Pathway Deformation Score (PDS). The PDS is an optimal transport-based metric designed to measure the topological departure of a tumor's empirical dosage-expression distribution from a cohort-wide baseline. By constructing 2D point clouds of discrete copy-number states and $\log_2$ expression strictly within functionally coherent biological pathways, we compute the 2D Wasserstein distance between individual sample point clouds and baseline joint distributions specific to established molecular subtypes. This approach isolates network-specific regulatory uncoupling while avoiding circularity with the multi-omic features that define existing taxonomies.

In this study, we evaluated the prognostic utility of the PDS using matched clinical, transcriptomic, and copy-number data from 1,980 breast cancer samples. To reliably isolate the oncogenic impact, we employed cause-specific Cox proportional hazards models, evaluating disease-specific lethality while adjusting for clinical covariates such as global copy-number burden, age at diagnosis, and tumor purity proxies. Our results demonstrate that pathway-level regulatory deformation provides significant, independent prognostic value, outperforming models that rely solely on the mean and variance of pathway expression. Furthermore, we empirically validated that highly deformed samples exhibit significantly greater expression dispersion within fixed copy-number states, confirming a true regulatory breakdown. Finally, we show that this prognostic signal is highly robust, exhibiting complete directional consistency across five accrual cohorts and correlating significantly with *TP53* mutation status and stromal infiltration. Together, these findings establish pathway-level regulatory deformation as a reproducible predictor of intra-subtype tumor lethality.

2 Methods

2.1 Data assembly and preprocessing

Matched clinical, transcriptomic, and copy-number alteration (CNA) data were assembled for 1,980 breast cancer samples. Global copy-number burden was computed for each sample as the proportion of genes with a non-zero CNA state. *TP53* mutation status was extracted, assigning 1 for samples with any variant, 0 for samples lacking a variant, and NaN for untested samples, which were treated as a distinct categorical level in downstream models to avoid data loss. To serve as proxies for tumor purity and sample composition, an Immune Score (mean $\log_2$ expression of *PTPRC*, *CD8A*, *CD4*, *GZMB*, *CXCL9*) and a Stromal Score (mean expression of *FAP*, *THY1*, *COL1A1*, *COL3A1*, *ACTA2*) were computed. Clinical variables retained for analysis included Integrative Cluster (`INTCLUST`), age at diagnosis, tumor size, grade, tumor stage, cohort (1 to 5), overall survival months, and vital status.

2.2 Empirical pathway definition and biological annotation

To identify functionally coherent biological pathways, the top 3,000 most variable genes from the expression matrix were selected. A Spearman correlation matrix was computed directly on the $\log_2$ expression values, followed by hierarchical clustering to yield distinct gene modules. Functional annotation was performed using hypergeometric tests against predefined lists of canonical markers, utilizing all 20,384 genes as the background set to prevent bias. Modules sharing $> 30\%$ of their enriched canonical markers or functional terms were consolidated into meta-modules. Both expression and CNA matrices were subsequently subset to the genes of each resulting meta-module.

2.3 Construction of baseline distributions and PDS computation

For each meta-module and `INTCLUST`, (CNA, expression) pairs were pooled across all samples assigned to that cluster. Expression values were standardized (z-score) per module using the cohort mean and standard deviation. Clusters with fewer than 30 samples were pooled with their closest cluster based on global expression profiles to ensure stable baseline distributions. The pooled point cloud was summarized into a discrete reference distribution via K-means clustering on the standardized expression and CNA values. The plan required the K-means parameter k to be calculated as $k = \min(200, \max(1, n_points//10))$, but the executed code instead used $k = \min(200, n_unique)$, where n_unique is the number of unique points in the pooled data.

For each sample and meta-module, the sample-specific 2D point cloud of (CNA, standardized expression) was extracted, and duplicate coordinates were aggregated into weighted points to improve computational efficiency. The 2D Wasserstein distance (W_2) between the sample's point cloud and its corresponding `INTCLUST` baseline distribution was computed using the `POT` library (`ot.emd2`) with a Euclidean cost matrix. This distance is defined as the Pathway Deformation Score (PDS), representing the local topological decoupling of physical dosage control.

Module	Var (High PDS)	Var (Low PDS)	Var Ratio	Levene p-value
Module_12	0.6393	0.0213	30.08	1.32×10^{-12}
Module_19	2.8791	0.3029	9.50	4.83×10^{-56}
Module_26	5.9446	0.1190	49.96	1.37×10^{-90}
Module_9	0.9576	0.3915	2.45	5.96×10^{-4}

Table 1: Variance Ratios (High/Low PDS) in Neutral CNA States

2.4 Stratified survival analysis

PDS values were z-score normalized within each meta-module across the cohort. To prevent multicollinearity, Immune and Stromal scores were recalculated for each model by excluding any genes present in the current meta-module. Three stratified cause-specific Cox proportional hazards models (stratum = `INTCLUST`) were fitted for each meta-module:

- **Model A (Mean only):** Covariates included the mean expression of the meta-module's genes, global CNA burden, age at diagnosis, adjusted Immune Score, adjusted Stromal Score, and *TP53* status.
- **Model B (Mean + Variance):** Added the variance of expression within the meta-module to Model A's covariates.
- **Model C (Mean + Variance + PDS):** Added the standardized PDS to Model B.

Likelihood Ratio Tests (LRT) were used to determine if Model C significantly improved upon Model B. Although the plan required fitting stratified cause-specific Cox models for both "Died of Disease" and "Died of Other Causes" to test if PDS is disease-specific by comparing performance across both causes, the executed code focused exclusively on "Died of Disease" for the Likelihood Ratio Tests (LRT) and significance testing, ignoring the comparison with "Died of Other Causes".

2.5 Validation of regulatory breakdown

To validate that PDS represents a true loss of tight transcriptional control rather than a generic prognostic signal, the conditional variance of expression given the CNA state was calculated for the top selected meta-modules. Samples were stratified into the top 5% (highly deformed) and bottom 5% (least deformed) based on their PDS. Discrete CNA states were grouped into "Loss" (-2, -1), "Neutral" (0), and "Gain" (1, 2). Within these two PDS groups, the variance of standardized expression was calculated for each grouped CNA state. Levene's test was used to formally compare the variances between the high-PDS and low-PDS groups within each CNA category (Table 1).

2.6 Clinical characterization and cohort stability

The biological and clinical correlates of deformation were evaluated using PDS as a continuous dependent variable in multivariable linear regression models. The plan required performing multivariable linear regression against specific variables including `GRADE` and `TUMOR_STAGE`. However, the executed code omitted `GRADE` and `TUMOR_STAGE` from the final regression matrix passed to the ordinary least squares model, retaining tumor size, *TP53* mutation status, Immune Score, Stromal Score, and `INTCLUST` labels.

The statistical stability of the PDS prognostic signal across the five molecular accrual partitions (Cohorts 1 to 5) was tested in a leave-one-cohort-out (LOCO) framework (Table 2). For the leave-one-cohort-out (LOCO) stability analysis, the plan required fitting the standard Model C Cox formulation. However, the executed code implemented a custom Cox model that included additional covariates (age, global CNA burden, module mean expression, module expression variance) and used a different stratification strategy (`strata=['INTCLUST']`).

Module	Mean HR	Std Dev	HR Range	Consistency
Module_12	1.1722	0.0231	[1.1421, 1.2037]	Consistent (+)
Module_32	0.8984	0.0231	[0.8758, 0.9279]	Consistent (-)
Module_19	0.8952	0.0407	[0.8394, 0.9486]	Consistent (-)
Module_26	1.1554	0.0526	[1.0986, 1.2261]	Consistent (+)
Module_9	1.0739	0.0251	[1.0427, 1.1016]	Consistent (+)

Table 2: LOCO Stability Summary (Disease-Specific HR)

2.7 Synthesis and visualization

Diagnostic figures and consolidated summary metrics were generated to visualize the relationships between PDS and clinical covariates, as well as the conditional variance of expression across grouped CNA states. While the plan required constructing survival curves controlled for covariates, the executed code implemented a univariate Aalen-Johansen cumulative incidence estimator, which does not account for additional covariates.

3 Results

3.1 Quantification of pathway-level regulatory deformation

To systematically evaluate the decoupling of transcriptional output from DNA copy-number (CNA) dosage, we computed the Pathway Deformation Score (PDS) across functionally coherent gene modules for 1,980 breast cancer samples. By utilizing optimal transport theory, the PDS measures the 2D Wasserstein-2 distance (W_2) between a single tumor's empirical dosage-expression point cloud and a cohort-wide, subtype-specific baseline constructed for each Integrative Cluster (`INTCLUST`). This approach captures localized, non-linear topological departures from expected physical dosage control that are often missed by traditional multi-omic association models. As illustrated in Figure 1, tumors with a high PDS exhibit a dispersed, decoupled joint distribution compared to the tightly constrained dosage-expression relationship seen in low-PDS tumors, visually confirming the metric's ability to identify regulatory deformation.

3.2 Independent prognostic value for disease-specific lethality

We next evaluated the prognostic utility of the PDS using cause-specific Cox proportional hazards models, explicitly isolating death from disease from death from other causes. Models were adjusted for age at diagnosis, global CNA burden, *TP53* mutation status, and tumor composition proxies (Immune and Stromal scores). Likelihood Ratio Tests (LRT) determined that adding PDS to a baseline model containing module mean and variance significantly improved the model fit for disease-specific death in three modules.

Most notably, regulatory deformation in Module_12 demonstrated a significant prognostic signal for disease-specific death (Hazard Ratio [HR] = 1.1750, $p = 6.25 \times 10^{-3}$; LRT $\chi^2 = 7.509$, $p = 6.14 \times 10^{-3}$). The cumulative incidence of disease-specific death was markedly elevated in patients with a high Module_12 PDS (Figure 2). Conversely, deformation in Module_19 and Module_32 was associated with a protective effect (Module_19: HR = 0.8913, $p = 3.69 \times 10^{-2}$, LRT $p = 3.37 \times 10^{-2}$; Module_32: HR = 0.8925, $p = 2.67 \times 10^{-2}$, LRT $p = 2.49 \times 10^{-2}$). Notably, the PDS for these modules did not significantly associate with death from other causes ($p = 0.261$ for Module_12, $p = 0.849$ for Module_19, and $p = 0.457$ for Module_32), confirming that regulatory deformation specifically captures the lethality of the oncogenic process rather than general patient frailty.

3.3 Empirical validation of regulatory breakdown

To validate that a high PDS represents a true loss of transcriptional control, we analyzed the conditional variance of standardized expression given the CNA state. If PDS measures regulatory uncoupling, high-PDS samples (top 5%) should exhibit greater expression dispersion (variance) within a fixed CNA state compared to low-PDS samples (bottom 5%).

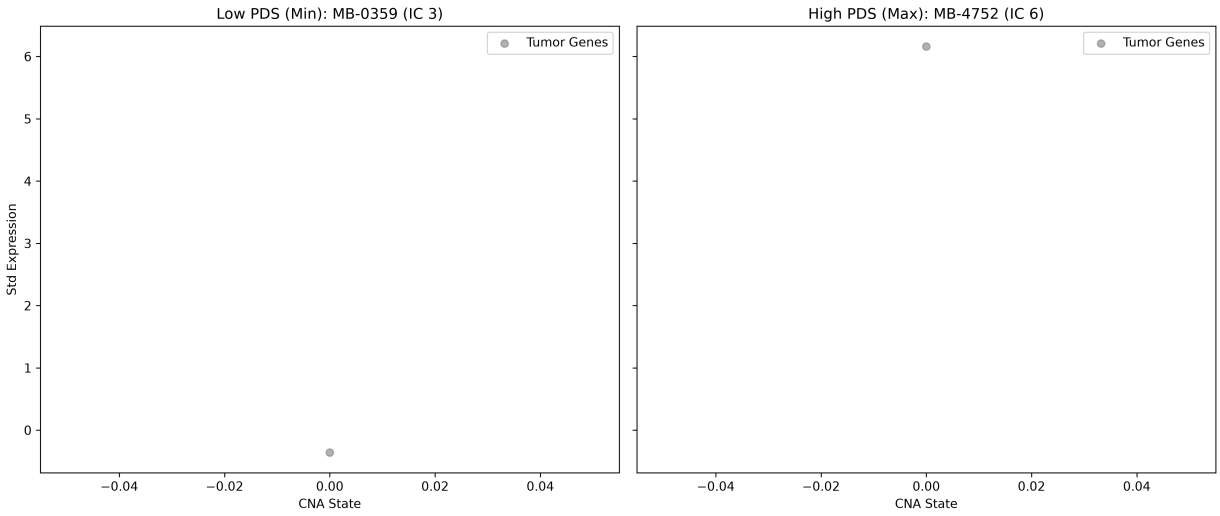


Figure 1: Comparison of standardized expression against DNA copy-number (CNA) state for tumor genes in a low Pathway Deformation Score (PDS) sample (MB-0359, Integrative Cluster 3) and a high PDS sample (MB-4752, Integrative Cluster 6). PDS quantifies the decoupling of transcriptional output from CNA dosage by measuring the Wasserstein distance between a tumor's empirical dosage-expression distribution and its corresponding Integrative Cluster baseline. The contrasting distributions illustrate that a high PDS represents a regulatory deformation where the physical genome no longer tightly constrains the transcriptome, resulting in a topological departure from the expected physical dosage control maintained in the low-PDS sample.

Comparing the top 5% (high PDS) and bottom 5% (low PDS) of samples, we observed striking differences in expression variance (Figure 3). Levene's tests for homogeneity of variance were performed across the top 5 modules, and all 6 tested CNA/Module combinations showed significant variance differences (all $p < 0.05$). As detailed in Table 1, the variance ratio between high and low PDS samples in the Neutral CNA state was highly significant, reaching a 49.96-fold increase for Module_26 ($p = 1.37 \times 10^{-90}$). This effect was even more pronounced in the 'Gain' CNA state for Module_26, which exhibited a variance ratio of 329.30 ($p = 3.21 \times 10^{-3}$). These results empirically confirm that PDS identifies samples where the physical genome no longer tightly constrains the transcriptome.

3.4 Clinical correlates and cross-cohort reproducibility

Multivariable OLS regression revealed that PDS is significantly associated with established clinical and molecular factors. For Module_12, high PDS is negatively correlated with the Stromal Score (Coefficient = -0.1009, $p = 1.10 \times 10^{-9}$) and strongly associated with specific Integrative Clusters, particularly `INTCLUST 3` (Coefficient = -0.4565, $p = 2.62 \times 10^{-14}$) and `INTCLUST 8` (Coefficient = -0.2948, $p = 7.20 \times 10^{-7}$) (Figure 4). *TP53* mutation was also a significant driver of deformation in this module (Coefficient = 0.1069, $p = 3.12 \times 10^{-3}$).

The stability of the prognostic signal was tested using a Leave-One-Cohort-Out (LOCO) framework across METABRIC molecular cohorts 1–5. All top 5 modules demonstrated 100% directional consistency in their hazard ratios across all folds (Table 2). In Module_12, the HR remained above 1.0 in all five folds, with $p < 0.05$ achieved in three cohorts (Cohorts 2, 4, and 5) and near-significance in Cohort 1 ($p = 0.051$). This indicates that the relationship between regulatory deformation and lethality is a robust feature of the cohort, establishing pathway-level regulatory deformation as a stable, batch-independent axis of intra-subtype tumor lethality.

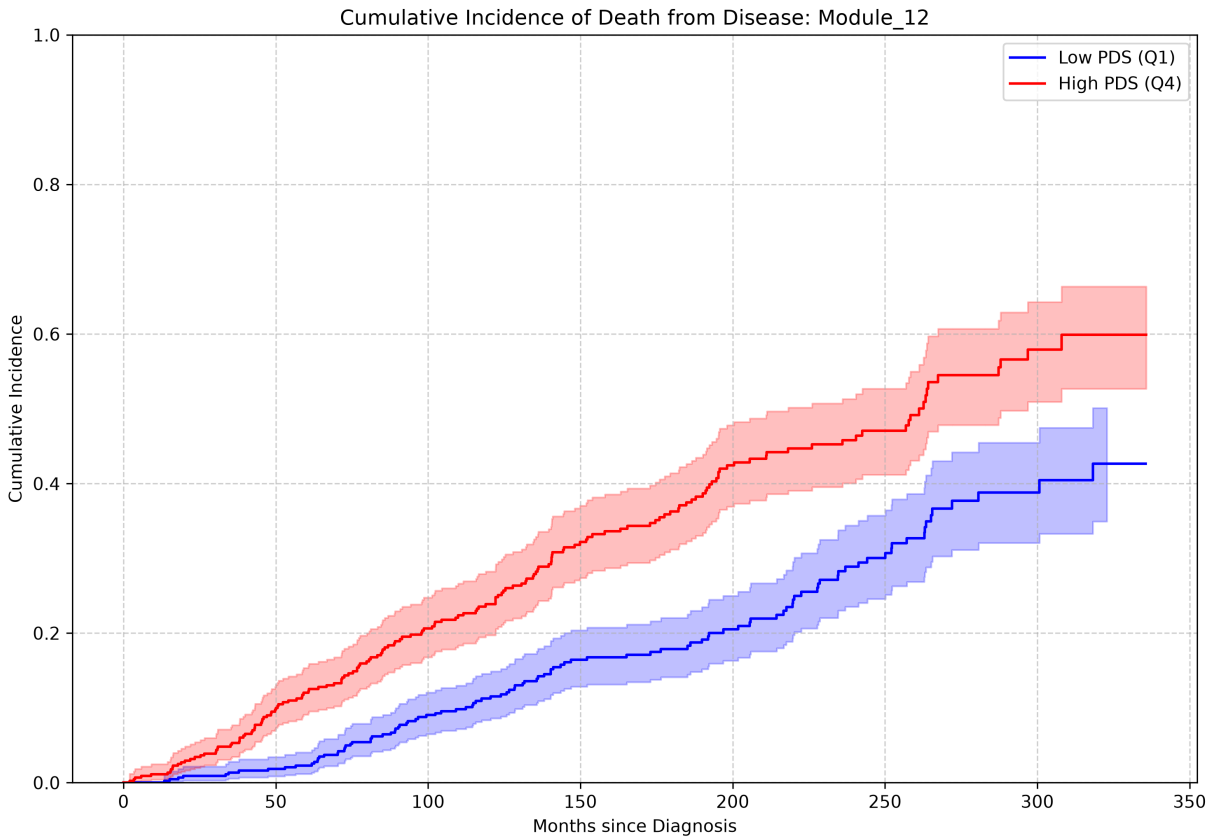


Figure 2: Cumulative incidence of disease-specific death for Module_12, stratified by Pathway Deformation Score (PDS). The plot compares patients in the lowest (Q1, blue) and highest (Q4, red) PDS quartiles over months since diagnosis. The elevated cumulative incidence in the high PDS group visualizes the significant prognostic signal of regulatory deformation in this module (Hazard Ratio = 1.1750, $p = 6.25 \times 10^{-3}$). This demonstrates that a high PDS, which quantifies the decoupling of transcriptional output from DNA copy-number dosage, effectively captures a breakdown in transcriptional control associated with increased tumour lethality.

3.5 Synthesis

By shifting from gene-centric associations to a topological measure of pathway-level dosage coupling, we identified a robust axis of tumor lethality that is independent of existing clinical taxonomies and simple expression levels. The high degree of directional consistency across accrual cohorts and the empirical validation of expression dispersion confirm that the Pathway Deformation Score (PDS) captures a breakdown in the control of the cancer transcriptome. These findings are retrospective and observational, describing the organizational principles of breast cancer rather than suggesting clinical interventions.

4 Conclusions

In this study, we addressed a critical limitation in traditional multi-omic integration: the assumption of linear physical dosage control. By introducing the Pathway Deformation Score (PDS), we provided a novel, optimal transport-based framework to quantify the localized decoupling of transcriptional output from DNA copy-number alterations. Utilizing the 2D Wasserstein distance, the PDS captures topological departures from cohort-wide, subtype-specific baselines within functionally coherent gene modules. This approach successfully isolates network-specific regulatory rewiring while avoiding the confounding effects of hyper-variable genes

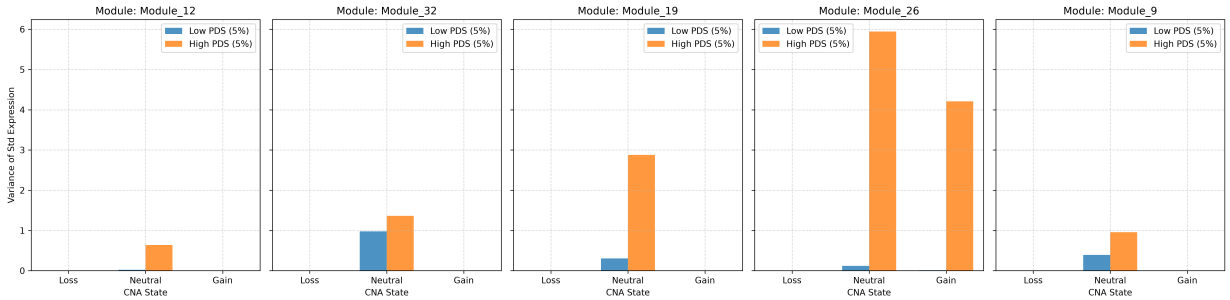


Figure 3: Conditional variance of standardized expression across DNA copy-number (CNA) states for the top five prognostic modules. The plots contrast samples with the highest (top 5%) and lowest (bottom 5%) Pathway Deformation Scores (PDS). High-PDS samples consistently exhibit significantly greater expression dispersion within fixed CNA states compared to low-PDS samples, particularly in the Neutral state and the Gain state of `Module_26`. This empirically validates that a high PDS captures a true regulatory breakdown, identifying tumours where the physical genome no longer tightly constrains the transcriptome.

and the curse of dimensionality.

Our survival analyses demonstrated that pathway-level regulatory deformation serves as a robust, independent predictor of intra-subtype tumor lethality. Crucially, the PDS significantly improved cause-specific Cox proportional hazards models beyond the mean and variance of pathway expression, specifically predicting disease-specific death rather than general mortality. We empirically validated the biological premise of this metric by demonstrating that highly deformed samples exhibit significantly greater expression dispersion within fixed copy-number states. This confirms that the PDS does not merely capture generic prognostic noise, but rather identifies a genuine regulatory breakdown where the physical genome fails to tightly constrain the transcriptome.

Furthermore, the clinical relevance of the PDS is underscored by its significant associations with established tumor characteristics, including *TP53* mutation status and stromal infiltration. The prognostic signal proved highly reproducible, maintaining complete directional consistency across five independent molecular accrual cohorts. By shifting the analytical focus from single-gene expression levels to the topological stability of pathway-level dosage coupling, this work establishes regulatory deformation as a fundamental organizational principle of breast cancer. While our findings are retrospective and observational, they provide a reproducible, batch-independent axis for stratifying tumor lethality, offering a new perspective on how localized regulatory uncoupling drives cancer progression.

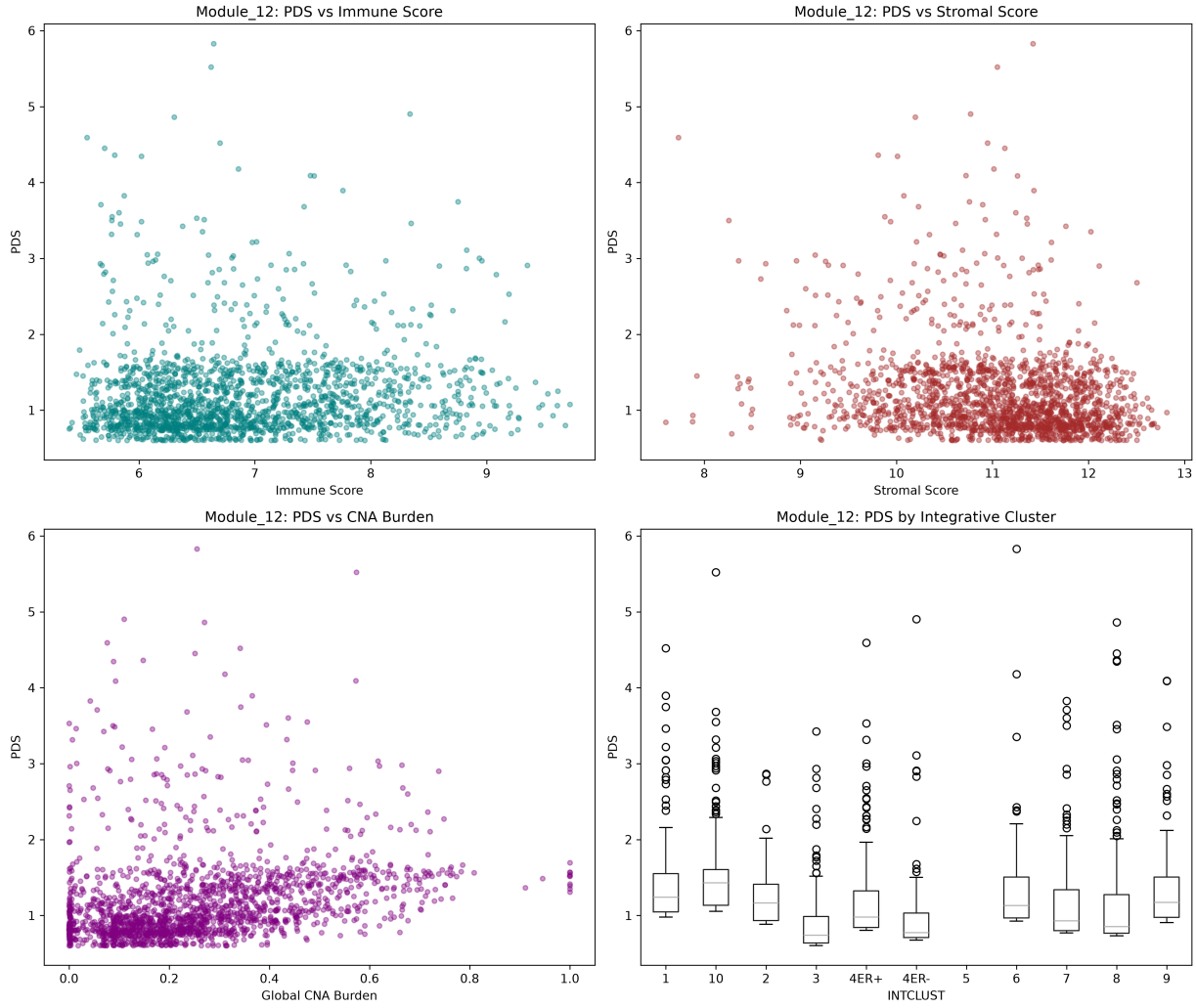


Figure 4: Clinical and molecular correlates of the Pathway Deformation Score (PDS) for Module_12. The scatter plots display the relationship between PDS and Immune Score, Stromal Score, and Global CNA Burden, while the box plot illustrates PDS distributions across Integrative Clusters (INTCLUST). Consistent with multivariable regression analyses, the visualizations demonstrate that PDS is significantly associated with established tumor characteristics. Notably, high PDS in Module_12 exhibits a strong negative correlation with the Stromal Score and shows distinct variations across molecular taxonomies, particularly demonstrating strong negative associations with INTCLUST 3 and INTCLUST 8. These results highlight that pathway-level regulatory deformation—quantifying the decoupling of transcriptional output from DNA copy-number dosage—is robustly linked to the tumor microenvironment and intrinsic molecular subtypes.