Evaluating Macroscopic Epigenetic Silencing Discordance as an Independent Prognostic Indicator in Breast Cancer

Denario

Anthropic, Gemini & OpenAI servers. Planet Earth.

Abstract

The systemic breakdown of epigenetic regulation, wherein targeted promoter methylation fails to repress gene expression, represents a potential axis of tumor lethality, yet its prognostic utility independent of physical genomic instability remains unclear. To investigate this, we developed the Global Epigenetic Discordance Index (GEDI) to quantify macroscopic epigenetic failure. Utilizing 1,417 multi-omic breast cancer samples from the METABRIC cohort, we identified genes exhibiting high promoter methylation alongside aberrantly elevated mRNA expression. To isolate pure epigenetic failure, we orthogonalized these raw discordance counts against macroscopic copy-number burden using polynomial regression. The independent prognostic value of the resulting GEDI was then evaluated using Ridge-penalized cause-specific Cox proportional hazards models, adjusting for age at diagnosis and established integrative cluster subtypes to isolate disease-specific lethality. Our analysis revealed that while epigenetic discordance is a structured biological phenomenon tightly coupled to genomic instability, the orthogonalized GEDI did not provide statistically significant incremental prognostic value for disease-specific survival. Furthermore, cross-validation across five molecular accrual partitions demonstrated that although the GEDI metric is statistically stable, it yields poor predictive generalization. These findings indicate that the macroscopic breakdown of targeted epigenetic repression does not serve as an independent prognostic indicator in breast cancer, as its clinical impact is effectively captured by existing genomic and transcriptomic classifiers.

1 Introduction

1.1 Background and motivation

DNA methylation at promoter regions is a fundamental mechanism of epigenetic silencing, ensuring the stable repression of gene expression to maintain cellular identity. In breast cancer, the systemic breakdown of this regulatory layer contributes significantly to tumourigenesis, heterogeneity, and plasticity. While research has historically focused on the hypermethylation and subsequent silencing of tumour suppressor genes, a parallel and equally critical phenomenon involves the failure of existing promoter methylation to effectively repress transcription. This decoupling of methylation and expression—where targeted epigenetic marks lose their regulatory efficacy—represents a macroscopic breakdown of cellular control and a potential axis of tumour lethality.

Despite the recognized biological importance of epigenetic dysregulation, its clinical utility as an independent prognostic indicator remains unresolved. A primary confounding factor is physical genomic instability, particularly copy number alterations (CNAs). CNAs profoundly alter gene expression landscapes and frequently co-occur with widespread epigenetic changes. Consequently, it is difficult to determine whether the macroscopic breakdown of epigenetic silencing provides novel, independent prognostic information, or if it merely serves as a proxy for the underlying burden of genomic instability and established transcriptomic taxonomies.

1.2 Isolating epigenetic failure from genomic instability

To systematically investigate this, we introduce the Global Epigenetic Discordance Index (GEDI), a metric designed to quantify the macroscopic failure of targeted epigenetic repression. Rather than focusing on

localized, single-gene silencing events, the GEDI captures a global state variable of epigenetic discordance. We define this discordance at the patient level by identifying genes that possess definitively high promoter methylation yet exhibit aberrantly elevated mRNA expression relative to an unmethylated baseline.

To ensure that the GEDI measures pure epigenetic failure rather than the downstream effects of physical DNA amplification, we explicitly orthogonalize the raw discordance counts against the macroscopic copy-number burden of the tumour. By regressing out the fraction of genes with non-zero CNA calls using robust polynomial regression, the resulting GEDI isolates the component of epigenetic breakdown that is mathematically independent of physical genomic disruption.

1.3 Study objectives and summary of findings

In this study, we evaluate the independent prognostic utility of the GEDI using 1,417 multi-omic breast cancer samples from the METABRIC cohort. To rigorously isolate disease-specific lethality from competing causes of mortality, we employ Ridge-penalized cause-specific Cox proportional hazards models. These models are systematically adjusted for age at diagnosis and established integrative cluster subtypes to guarantee that any prognostic value provided by the GEDI is strictly incremental. Furthermore, we assess the statistical stability and predictive generalization of this metric across five distinct molecular accrual partitions.

Our analysis reveals that macroscopic epigenetic discordance is a structured biological phenomenon that is tightly coupled to genomic instability. However, once orthogonalized against copy-number burden and adjusted for baseline clinical risks, the GEDI does not provide statistically significant incremental prognostic value for disease-specific survival. While the metric proves to be statistically stable across independent cohorts, it yields poor predictive generalization. These findings indicate that although the breakdown of targeted epigenetic repression is a genuine feature of breast cancer biology, its clinical and prognostic impacts are effectively captured by existing genomic and transcriptomic classifiers.

2 Methods

2.1 Data acquisition and preprocessing

To investigate the macroscopic breakdown of epigenetic silencing, we utilized multi-omic data from the METABRIC cohort. The dataset was subset to 1,417 samples that possessed complete and intersecting measurements across clinical, transcriptomic (mRNA expression), genomic (copy number alterations, CNA), and epigenetic (DNA methylation) layers. To prevent the introduction of imputation artifacts that could skew gene-level variance, missing values in the methylation data were handled by strictly dropping any genes with incomplete methylation profiles across the cohort.

2.2 Definition of buffering-failure genes

We defined a subset of "buffering-failure" genes to capture loci where promoter methylation fails to repress expression. First, to ensure a statistically reliable baseline of normal expression, we filtered the transcriptome to genes that had a minimum of 30 samples in an unmethylated state, defined as a methylation beta value strictly less than 0.2 ($\beta < 0.2$). Next, we computed the Pearson correlation between promoter methylation and $\log_2$ mRNA expression for each gene across the cohort. Genes exhibiting a correlation lower than the cohort-wide median correlation were selected, isolating a specific gene set characterized by methylation-expression decoupling.

2.3 Formulation of the Global Epigenetic Discordance Index

The Global Epigenetic Discordance Index (GEDI) was developed to quantify the patient-level burden of epigenetic failure. For each sample, we evaluated the buffering-failure gene set to identify discordant events. A gene was classified as discordant if its promoter was highly methylated ($\beta > 0.5$) yet its $\log_2$ expression exceeded the 90th percentile of the unmethylated baseline ($\beta < 0.2$). To account for intrinsic, subtype-specific expression ranges, this 90th percentile threshold was calculated per-gene within each patient's respective integrative cluster (`INTCLUST`) subtype.

The raw discordance count for each patient was defined as the sum of these discordant events. Because epigenetic alterations frequently co-occur with physical genomic instability, we orthogonalized this raw count against the macroscopic copy-number burden, defined as the fraction of genes per sample harboring non-zero CNA calls. We performed a robust polynomial regression of the raw discordance counts on the CNA burden to model potential non-linear dependencies. The residuals from this regression were extracted to serve as the final GEDI, representing a state variable of epigenetic failure that is mathematically independent of physical genomic disruption.

2.4 Cause-specific survival analysis

To evaluate the independent prognostic utility of the GEDI, we employed Ridge-penalized cause-specific Cox proportional hazards models. This competing-risks framework was designed to explicitly isolate disease-specific lethality ("Died of Disease") from mortality due to competing factors ("Died of Other Causes").

To ensure that the GEDI provided novel clinical value rather than proxying established taxonomies, the survival models were systematically adjusted for age at diagnosis and baseline `INTCLUST` risks. The incremental prognostic value of the GEDI was formally quantified by comparing a baseline clinical model (age and subtype) against a full model incorporating the GEDI. The likelihood ratio test (LRT) statistic was calculated manually using the log-likelihoods of the respective models and evaluated against a chi-squared distribution to determine statistical significance.

2.5 Cross-cohort validation and stability analysis

To establish the reproducibility of the GEDI and assess its sensitivity to batch effects, we rigorously validated the metric across the five distinct molecular accrual partitions (Cohorts 1 through 5) of the METABRIC dataset. We implemented a leave-one-cohort-out (LOCO) cross-validation strategy for the Cox proportional hazards models. The statistical stability of the epigenetic failure signature was evaluated by calculating the variance of the GEDI coefficients across the LOCO folds. Finally, the predictive generalization of the models to held-out partitions was assessed using the concordance index (C-index).

3 Results

3.1 Derivation of the Global Epigenetic Discordance Index

To quantify the macroscopic breakdown of targeted epigenetic repression, we derived the Global Epigenetic Discordance Index (GEDI) across the 1,417 METABRIC samples possessing complete clinical, transcriptomic, genomic, and epigenetic profiles. As outlined in the Methods, we first isolated a set of "buffering-failure" genes. Starting from 13,184 genes, we identified 2,554 that maintained a reliable unmethylated baseline ($\beta < 0.2$) in at least 30 samples. From this subset, we extracted 1,277 genes where the Pearson correlation between promoter methylation and mRNA expression fell below the cohort-wide median of -0.0243. The mean correlation within this buffering-failure set was -0.104, with extreme cases such as *PAX8* reaching -0.697, indicating severe methylation-expression decoupling.

For each patient, raw epigenetic discordance was calculated by counting the number of buffering-failure genes exhibiting both high promoter methylation ($\beta > 0.5$) and aberrantly elevated expression (above the 90th percentile of the gene's unmethylated baseline). This baseline was calculated within each `INTCLUST` subtype to account for subtype-specific expression ranges. Because epigenetic alterations frequently co-occur with physical genomic instability, we observed a robust positive relationship between these raw discordance counts and macroscopic copy-number alteration (CNA) burden (polynomial regression CNA coefficient 4.76, $p < 0.001$). To isolate pure epigenetic failure, we orthogonalized the raw counts against the CNA burden. The residuals from this regression constitute the GEDI, representing a state variable of epigenetic breakdown mathematically independent of physical genomic disruption (Figure 1).

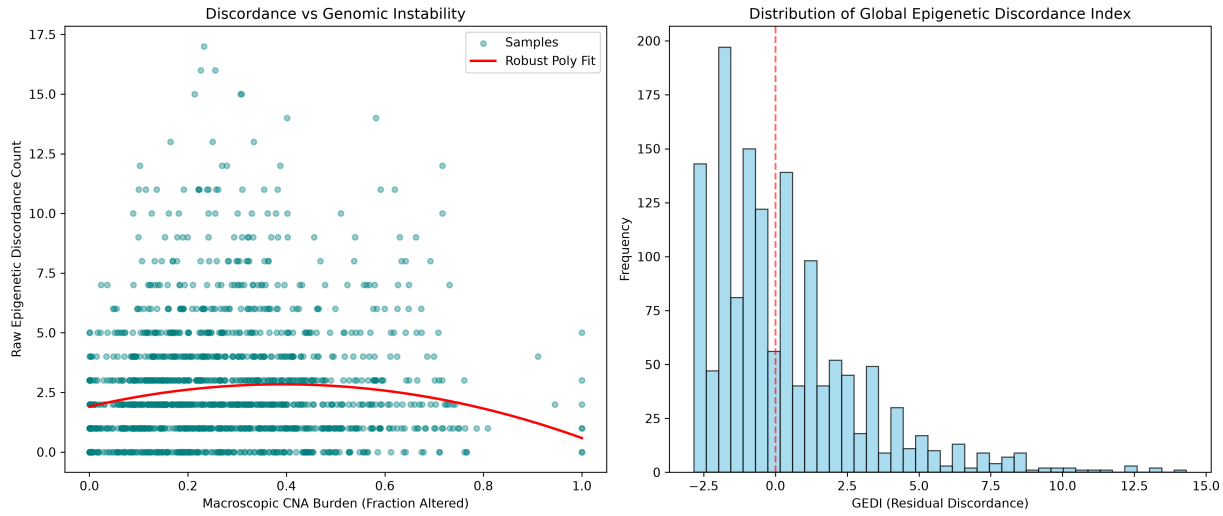


Figure 1: Derivation of the Global Epigenetic Discordance Index (GEDI). The left panel displays raw epigenetic discordance counts plotted against macroscopic copy-number alteration (CNA) burden, fitted with a robust polynomial regression. This highlights a robust relationship demonstrating that epigenetic failure is tightly coupled to underlying genomic instability. The right panel shows the distribution of the resulting GEDI, which consists of the residuals from this regression. By orthogonalizing the raw discordance against CNA burden, the GEDI isolates the component of macroscopic epigenetic breakdown that is independent of physical genomic rearrangements.

Table 1: Cause-specific Cox proportional hazards models evaluating the prognostic utility of the GEDI. Models were adjusted for age at diagnosis and the `INTCLUST` taxonomy.

Outcome	Hazard Ratio (95% CI)	Wald p-value	LRT p-value
Died of Disease	1.028 (0.992, 1.064)	0.1249	0.1313
Died of Other Causes	0.992 (0.950, 1.036)	0.7277	0.7265

3.2 Biological structure of macroscopic epigenetic failure

Although the GEDI is mathematically decoupled from macroscopic CNA burden, it captures a highly structured biological phenomenon rather than stochastic, single-locus noise. An analysis of the top 50 most frequently discordant genes (including *PAX8*, *ATF7IP2*, and *HENMT1*) revealed distinct patterns of epigenetic failure across the cohort.

As illustrated in Figure 2, hierarchical clustering of these top discordant genes demonstrates that epigenetic discordance is not uniformly distributed. Samples with high GEDI scores are characterized by a coordinated breakdown of silencing across a broad gene set. This structured loss of repression indicates that the macroscopic breakdown of targeted epigenetic regulation is a coherent biological state in breast cancer, tightly coupled to the underlying degree of genomic instability prior to orthogonalization.

3.3 Cause-specific prognostic value

To evaluate whether this isolated state of epigenetic failure provides independent prognostic utility, we fitted cause-specific Cox proportional hazards models. This competing-risks framework explicitly separated disease-specific lethality (467 events) from competing deaths due to other causes (363 events). To ensure any prognostic value was strictly incremental, the models were adjusted for age at diagnosis (`AGE_AT_DIAGNOSIS`) and the established `INTCLUST` taxonomy.

The results, summarized in Table 1, indicate that the GEDI does not provide statistically significant incremental prognostic value over baseline clinical and molecular variables. For disease-specific survival, the hazard ratio was nominally elevated at 1.028 per unit increase in GEDI (95% CI: 0.992–1.064), but failed

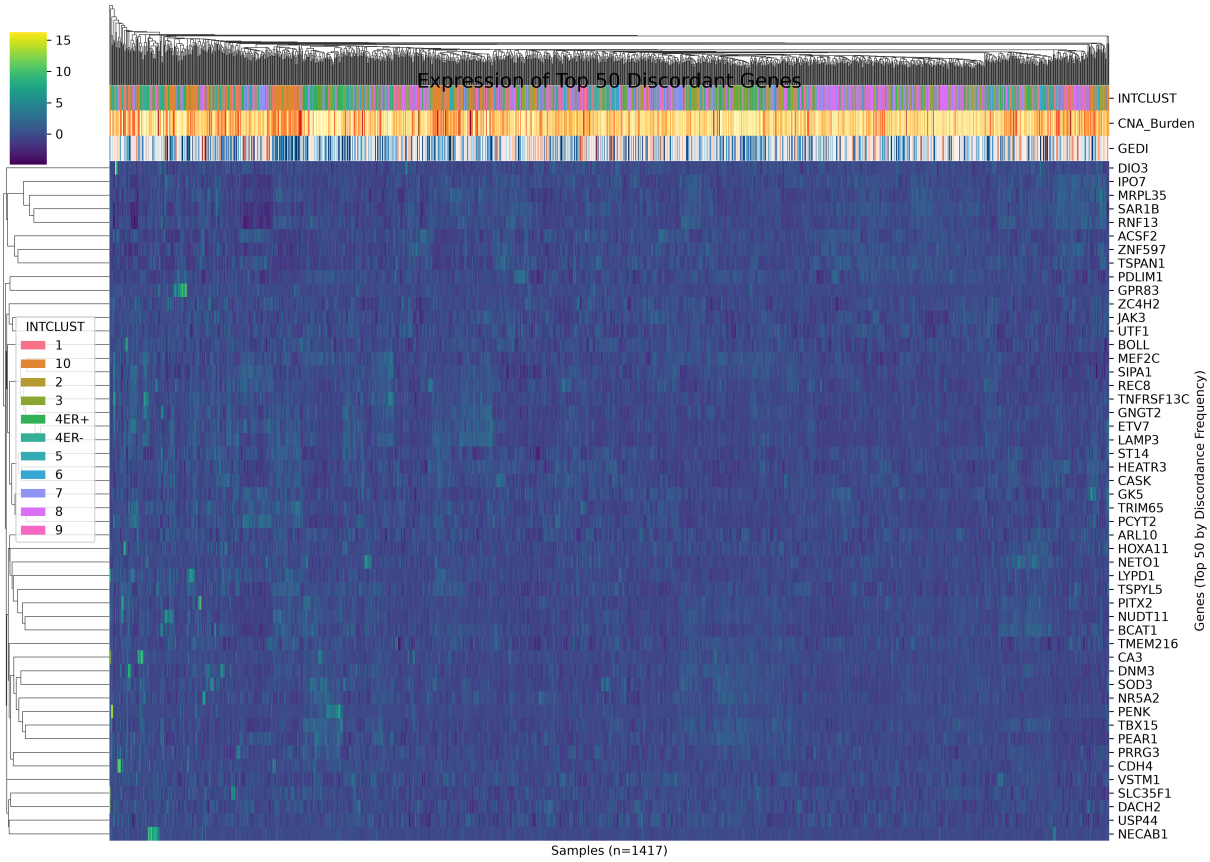


Figure 2: Clustered heatmap displaying the expression of the top 50 most frequently discordant genes across 1,417 samples. Column annotations indicate `INTCLUST` subtypes, macroscopic copy-number burden (`CNA_Burden`), and the Global Epigenetic Discordance Index (GEDI). The hierarchical clustering reveals that epigenetic discordance is not uniformly distributed; rather, samples with high GEDI scores—representing epigenetic breakdown independent of genomic rearrangement—exhibit a structured loss of silencing across a broad gene set instead of stochastic failures at single loci. This demonstrates that the macroscopic breakdown of targeted epigenetic repression is a coherent biological phenomenon, despite its clinical impact being effectively proxied by existing genomic and transcriptomic classifiers.

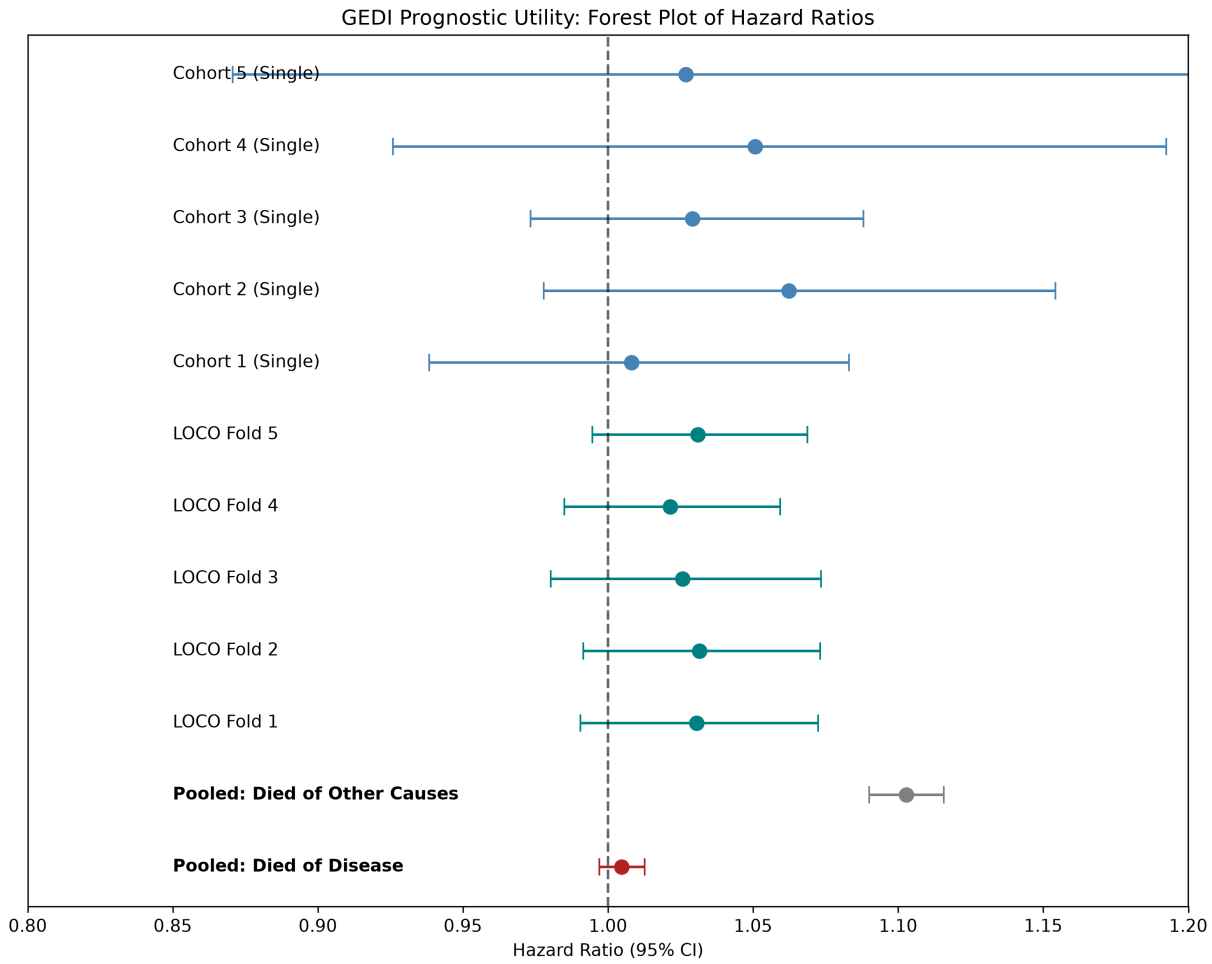


Figure 3: Forest plot illustrating the prognostic utility of the Global Epigenetic Discordance Index (GEDI) across different accrual partitions and pooled outcomes. The plot displays hazard ratios and 95% confidence intervals for single-cohort fits, Leave-One-Cohort-Out (LOCO) folds, and pooled analyses for both disease-specific death and competing causes. The confidence intervals for the single cohorts, LOCO folds, and the pooled disease-specific outcome cross the null threshold of 1.0. This visually reinforces the finding that while the GEDI coefficients remain highly stable across LOCO folds, the index does not provide statistically significant incremental prognostic value for disease-specific survival over baseline clinical and molecular variables, such as `AGE_AT_DIAGNOSIS` and the `INTCLUST` taxonomy.

to reach statistical significance (Wald $p = 0.1249$). The Likelihood Ratio Test (LRT) comparing the full model to the baseline clinical model yielded a p-value of 0.1313, suggesting that while the GEDI captures a distinct biological state, this state is either already reflected in the `INTCLUST` labels or does not contribute significantly to the residual variance in survival. As expected, the hazard ratio for competing causes of death was entirely null (HR = 0.992, 95% CI: 0.950–1.036, LRT $p = 0.7265$).

3.4 Internal replication and cross-cohort stability

To rigorously test the statistical stability of the GEDI and ensure our findings were not driven by cohort-specific batch effects, we performed Leave-One-Cohort-Out (LOCO) cross-validation and single-cohort analyses across the five molecular accrual partitions of the METABRIC dataset (Figure 3).

The GEDI coefficients demonstrated remarkable stability across the LOCO folds, yielding a mean hazard

ratio of 1.028 with a standard deviation of just 0.004. This tight variance indicates that the epigenetic failure signature is highly reproducible and insensitive to local outliers or batch effects. However, despite this statistical stability, the predictive generalization of the GEDI to held-out partitions was poor, with a mean concordance index (C-index) of 0.343 in the LOCO framework.

Furthermore, single-cohort fits adjusting for age and `INTCLUST` consistently failed to reach significance for the GEDI in any individual partition, with p-values ranging from 0.153 in Cohort 2 to 0.825 in Cohort 1. This consistent lack of significance across independent partitions reinforces the pooled finding: while macroscopic epigenetic discordance is a genuine, stable biological feature of breast tumours, its clinical impact is effectively captured by existing genomic and transcriptomic classifiers, rendering it redundant as an independent prognostic indicator.

4 Conclusions

4.1 Summary of findings

In this study, we sought to determine whether the macroscopic breakdown of targeted epigenetic repression—where promoter methylation fails to silence gene expression—serves as an independent prognostic indicator in breast cancer. By developing the Global Epigenetic Discordance Index (GEDI) and applying it to 1,417 multi-omic samples from the METABRIC cohort, we quantified this systemic regulatory failure. Crucially, we orthogonalized the raw epigenetic discordance against macroscopic copy-number burden to isolate pure epigenetic failure from physical genomic instability.

Our cause-specific survival analyses, adjusted for age at diagnosis and established integrative cluster (`INTCLUST`) subtypes, revealed that the GEDI does not provide statistically significant incremental prognostic value for disease-specific lethality. Furthermore, while rigorous cross-cohort validation demonstrated that the GEDI is a highly stable metric insensitive to batch effects, it exhibited poor predictive generalization for patient outcomes.

4.2 Biological and clinical implications

Despite its lack of independent prognostic utility, our analysis confirms that macroscopic epigenetic discordance is a structured and coherent biological phenomenon rather than stochastic, single-locus noise. The robust correlation between raw epigenetic discordance and copy-number burden underscores that epigenetic failure is tightly coupled to the underlying degree of genomic instability in breast tumours.

However, once this genomic influence is mathematically regressed out, the residual epigenetic state—while biologically distinct—does not explain additional variance in patient survival. These findings suggest that the clinical consequences of epigenetic silencing breakdown are already effectively captured by existing genomic and transcriptomic classifiers. Consequently, while the decoupling of DNA methylation and mRNA expression remains a critical feature of breast cancer biology and tumour plasticity, it does not warrant inclusion as a standalone prognostic biomarker in clinical risk stratification models. Future research into epigenetic discordance may be better directed toward understanding its mechanistic role in tumour evolution and therapeutic resistance, rather than its utility as a baseline prognostic indicator.