Entropy and Stemness Index–Based Quantification of Epigenetic Dysregulation for Machine Learning–Driven Stratification of Representative Cancer Types

Djansel Bukovec[1][a], Goran Kungulovski[2][b], Sonja Gievska[1][c] and Monika Simjanoska Misheva[1][d]
[1]*Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University, Skopje, North Macedonia*
[2]*Fingerprint Diagnostics LLC, Skopje, North Macedonia*

Keywords: DNA Methylation Entropy, Epigenetic Heterogeneity, Cancer Prognosis, Stemness Index, Hypoxia, Machine Learning.

Abstract: Epigenetic dysregulation, quantified by DNA methylation entropy and the methylation-derived stemness index (mDNAsi), has been proposed as a marker of tumor aggressiveness; however, its prognostic relevance in established cancers remains unclear. It was examined whether epigenetic instability represents a hallmark of malignant transformation and whether its variation is associated with tumor outcomes. A three-state Markov model of unmethylated/intermediate/methylated (U–I–M) transitions was fitted to raw β values from 275 paired tumor–normal samples across four representative TCGA cohorts, and the overall adjacent switching rate (s_{any}) was compared between tissues. Consistently higher local switching was observed in tumors than in matched normal tissues, and a strong correlation between genome-wide Shannon entropy and s_{any} was detected. In 2,661 primary tumors, z-scored entropy and mDNAsi were evaluated in multivariable models of primary lymph-node positivity, fraction genome altered (FGA), and overall survival (OS), with adjustment for age and cancer type. After adjustment, independence association of entropy or mDNAsi with nodal involvement was identified (entropy OR per 1-SD ≈ 1.14, $p = 0.466$), and survival effects were found to be limited and cancer-type dependent. In linear models of FGA, entropy contributed little after covariate adjustment, whereas higher mDNAsi remained positively associated with increased copy-number burden. Mantel tests further indicated that epigenetic/genomic distance structures aligned with hypoxia but were largely uncoupled from nodal-status distances. These findings suggest that methylation entropy is unlikely to serve as a universal marker of aggressiveness in established tumors and may instead reflect evolutionary history and microenvironmental stress.

1 INTRODUCTION

Cell identity emerges from gene-regulatory networks (GRNs), whose behavior depends on specific combinations of genetic and epigenetic features. Genetic and epigenetic states jointly modulate network activity and effective connectivity by altering regulatory-element function and the ability of regulators to access and bind DNA. Among epigenetic mechanisms, DNA methylation is the most common covalent DNA modification, involving the addition of a methyl group to cytosines in CpG dinucleotides (Vaidya et al., 2023; Lee et al., 2023). The resulting CpG-methylation patterns form part of an "epigenetic code" that is mitotically heritable but not encoded in the primary DNA sequence (Vaidya et al., 2023). By establishing tissue-specific DNA-methylation patterns, cells with the same genetic background can adopt distinct phenotypes. Consequently, DNA methylation plays a key role in cellular differentiation, repression of transposable elements, transcriptional regulation, and developmental programming (Sheffield et al., 2017), and these patterns are written and maintained by DNA methyltransferases, DNMTs. When this epigenetic machinery becomes dysregulated, the balance between DNA methylation and demethylation is disrupted, creating

[a] https://orcid.org/0009-0008-1522-0678
[b] https://orcid.org/0000-0001-6673-1795
[c] https://orcid.org/0000-0002-3411-2399
[d] https://orcid.org/0000-0002-5028-3841

Bukovec, D., Kungulovski, G., Gievska, S. and Misheva, M. S.
Entropy and Stemness Index–Based Quantification of Epigenetic Dysregulation for Machine Learning–Driven Stratification of Representative Cancer Types.
DOI: 10.5220/0014418400004070

In *Proceedings of the 19th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2026) - Volume 2: BIODEVICES, BIOIMAGING, BIOINFORMATICS*, pages 527-538
ISBN: 978-989-758-802-0; ISSN: 2184-4305

a permissive landscape for cancer development. The disruption commonly manifests itself as promoter hypermethylation of tumor-suppressor genes, leading to their silencing and uncontrolled proliferation, together with global hypomethylation, which destabilizes the genome, contributes to chromosomal instability, and can activate oncogenes (Vaidya et al., 2023; Lee et al., 2023).

As tumors evolve, epigenetic dysregulation is closely related to changes in cellular identity. Cancer progression is often accompanied by a gradual loss of differentiated identity and the acquisition of progenitor or stem-cell–like features. Therefore, stemness indices derived from epigenetic data, such as methylation-based stemness (mDNAsi), have been proposed to quantify oncogenic dedifferentiation, capture intratumoral heterogeneity, and highlight potential therapeutic vulnerabilities (Vaidya et al., 2023; Chan et al., 2025). According to this view, undifferentiated primary tumors are more prone to metastasis and form therapy-resistant metastases, which are typically associated with a poor prognosis (Monyak et al., 2025). Although stemness indices summarize how "stem-like" a tumor is, they mainly reflect dedifferentiation rather than dysregulation of methylation patterns in the cell mixture in a bulk sample. Thus, stemness and methylation heterogeneity represent distinct and potentially decoupled axes of tumor biology, motivating the use of multiple complementary metrics when characterizing tumor aggressiveness.

Entropy provides a complementary view of epigenetic dysregulation by summarizing how methylation states are distributed throughout the cell population in bulk methylation profiles (Vaidya et al., 2023; Chan et al., 2025). In any given genomic region, methylation can range from fully methylated to fully unmethylated (β from 0 to 1) with intermediate β-states indicating cell-to-cell variability (Vaidya et al., 2023). DNA methylation heterogeneity at a specific locus is defined as variation in methylation patterns at that locus within a cell population and can arise from genetic and epigenetic factors (Lee et al., 2023; Chan et al., 2025). By tracking variation in methylation patterns, methylation heterogeneity captures the "fingerprints" of epigenetic influences throughout development and disease (Sheffield et al., 2017; Ma et al., 2024). Intratumoral heterogeneity, therefore, reflects the combined effects of genetic and epigenetic instability, and can be summarized at the sample level by epigenetic randomness measures, such as Shannon entropy (Chan et al., 2025), which quantify the degree of unpredictability in methylation patterns within a cell population. Shannon entropy typically ranges from 0 to 1, with values closer to 1 indicating higher randomness or variability. This quantitative approach may allow us to compare the epigenetic instability between different samples and conditions.

An additional local descriptor of epigenetic dysregulation can be obtained by summarizing local methylation instability using a three-state Markov model with statesunmethylated (U), intermediate/partially methylated (I), and methylated (M), which estimates transition probabilities between adjacent CpGs and thereby captures fine-scale state switching along the genome. These epigenetic features may related to clinical hallmarks of aggressiveness, as clinical behavior is shaped by genomic instability, primary lymph-node involvement, and microenvironmental stressors such as hypoxia (Bakhoum and Landau, 2017; Teng et al., 2020; Hosea et al., 2024). Jointly considering entropy, mDNAsi, local switching, and genomic instability therefore enables integrated comparisons across tumors and facilitates interpretation in clinically relevant contexts (Vaidya et al., 2023; Monyak et al., 2025).

In this study, an entropy–stemness framework is developed to quantify epigenetic dysregulation in cancer, treating genome-wide methylation entropy and mDNAsi as core covariates in multivariable clinical–genomic models with local switching as a complementary descriptor. The first objective was to determine whether epigenetic instability formalized as genome-wide methylation entropy and local U–I-M transition rates, distinguishes tumors from their patient-matched normal tissues. A second objective was to characterize how the variation in entropy and mDNAsi between tumors is related to nodal invasion, chromosomal instability, survival, and hypoxia, thus defining the contexts in which these epigenetic readouts act as markers of tumor aggressiveness.

The rest of the paper is organized as follows. Section 2 details the data sources, preprocessing, feature construction (U/I/M transition metrics, entropy, mDNAsi), and modeling procedures. Section 3 reports tumor–normal analyses and associations with lymph-node status, chromosomal instability, survival, and hypoxia. Section 4 provides biological and clinical interpretationsons. Conclusions recap the main contributions and future directions.

2 METHODS

2.1 Study Cohort and Data Preprocessing

DNA methylation and clinical data were obtained from The Cancer Genome Atlas (TCGA) (National Cancer Institute and National Human Genome Research Institute, 2006). We analyzed pan-cancer cohorts with a focus on lung adenocarcinoma (LUAD), breast carcinoma (BRCA), thyroid carcinoma (THCA), and kidney renal clear-cell carcinoma (KIRC) as representative tumor types. Primary tumor samples with complete information on age at diagnosis, lymph-node status (node-positive vs. node-negative), overall survival time and status (OS), fraction of genome altered (FGA), and Ragnum hypoxia scores were retained (Ragnum et al., 2015). The four-cancer tumor cohort comprised 2,661 primary tumors (BRCA = 1,084; LUAD = 566; THCA = 499; KIRC = 512), with the effective sample size varying by endpoint. For analyses requiring matched tumor–normal DNA-methylation β-value matrices, the subset with paired samples was used (BRCA = 90 pairs; KIRC = 158; LUAD = 23; THCA = 56; total = 275 pairs). All continuous variables were z-standardized before modeling.

2.2 Markov Model of Methylation Dynamics

For each tumor–normal sample, β-values were categorized into three methylation states: unmethylated (U, $\beta < 0.20$), intermediate (I, $0.20 \leq \beta \leq 0.60$), and methylated (M, $\beta > 0.60$). These thresholds were selected based on established literature, as the 0.20 and 0.60 boundaries are widely recognized for their biological relevance and consistency across samples (Novakovic et al., 2011; Shao et al., 2017). A first-order Markov model estimated transition probabilities between consecutive CpGs, yielding a 3×3 transition matrix $(P_{XY},\ X,Y \in \{U,I,M\})$. Summary instability metrics were calculated, including any state change (s_{any}), defined as the sum of off-diagonal transition probabilities. Differences in switching metrics and transition probabilities between tumor and normal samples were tested within each cancer type using a two-sided Mann-Whitney U.

2.3 Methylation Entropy and Stemness Index

For each CpG probe, the β-value (the proportion of methylated DNA intensity over total intensity) was treated as a per-sample estimate of the probability that the CpG is methylated. To avoid undefined logarithms at $\beta = 0$ and $\beta = 1$, we clipped β to the interval $[10^{-6}, 1-10^{-6}]$. The Shannon entropy for a single CpG i in sample j (H_{ij}) was then computed acomputed from the methylation β-value (β_{ij}) as:

$$H_{ij} = -\left[\beta_{ij}\log_2(\beta_{ij}) + (1-\beta_{ij})\log_2(1-\beta_{ij})\right].$$

Sample-level global methylation entropy H_j was defined as the mean H_{ij} across all autosomal CpG probes with non-missing β in that sample j. mDNAsi scores were obtained from the Pan-Cancer Atlas publication (Malta et al., 2018). The published mDNAsi values were matched to the TCGA barcodes. To determine whether this global entropy measure aligns with local methylation instability, the association between sample-level entropy and the three-state Markov switching metric (s_{any}) was assessed using Spearman's rank correlation.

2.4 Statistical Analyses

2.4.1 Lymph-Node Involvement

A multivariable logistic regression model was fitted to predict primary lymph-node positivity (node-positive vs node-negative). Predictors were z-scored entropy and the DNA methylation–based stemness index (mDNAsi); age and cancer type were included as categorical covariates. Effect sizes are reported as odds ratios (ORs) per 1-standard deviation (SD) increase with 95% confidence intervals (CIs). Model discrimination was assessed using the area under the receiver operating characteristic curve (AUC-ROC), and calibration was evaluated with the Hosmer–Lemeshow test.

2.4.2 Chromosomal Instability

The association between epigenetic instability and chromosomal instability was examined by regressing fraction genome altered (FGA) on z-scored entropy, mDNAsi, and age, with cancer-type fixed effects, using ordinary least-squares regression (OLS) with robust standard errors. Coefficients are expressed per 1-SD increase.

2.4.3 Overall Survival

Cox proportional-hazards models were fitted separately for each cancer type, with overall survival

(OS) time and status as the outcome and z-scored entropy, mDNAsi, and age as covariates. Hazard ratios (HR) per 1-SD increase, 95% confidence intervals, and concordance indices (C-index) are reported. Proportional-hazards assumptions were evaluated with Schoenfeld residuals.

2.4.4 Mantel Tests

Euclidean distance matrices were constructed from z-standardized entropy, FGA, and hypoxia scores. The Hamming distance matrix was built from binary node status. Pairwise Spearman Mantel tests were performed to assess correlations between epigenetic/genomic and clinical distance structures in the pan-cancer cohort and within each cancer type. Across the five variables (entropy, mDNAsi, FGA, hypoxia, node status) we evaluated eight pairwise Mantel tests per analysis and controlled multiple testing using the Benjamini–Hochberg false discovery rate (FDR). Unless stated otherwise, p-values for Mantel tests are FDR-adjusted. We report the Mantel correlation coefficient (r), the permutation p-value, and the FDR-adjusted p-value.

3 RESULTS

3.1 Tumors Exhibit Increased Methylation Switching

To characterize dynamic, state-switching features of DNA methylation, a three-state Markov model was applied to paired normal and tumor β -value profiles, with states defined as U (unmethylated), I (intermediate or partially methylated), and M (methylated). Instability was quantified for each sample using complementary switching metrics, including s_{any}, which represents the probability of transition to any other state (U↔I, I↔M, U↔M). Tumor-normal differences were summarized as $\Delta s = s_{\text{tumor}} - s_{\text{normal}}$ for each cancer type. Among the four cancer types analyzed (BRCA, KIRC, LUAD, and THCA), tumors showed higher switching rates than matched normal tissues, reflecting more frequent local-state changes and reduced methylation stability. Stratification by cancer type revealed a significant tumor–normal increase in s_{any} for each cancer, with the most pronounced difference in KIRC and significant effects in BRCA, LUAD, and THCA (Figure 1,Table S1). Statistical significance for tumor–normal comparisons was determined using a two-sided Mann–Whitney U test. Collectively, these findings indicate that local methylation state switching is consistently elevated in tumors across these cancer types.

To assess whether the local switching metric (s_{any}) reflected broader methylation heterogeneity, its relationship with genome-wide entropy was examined. Entropy showed a strong positive correlation with three-state Markov switching (Spearman $\rho = 0.634$, Pearson $r = 0.622$; (Figure 1), suggesting that increased global methylation randomness is accompanied by increased local U–I–M state switching. Given its sample-level interpretability and suitability for global trend analyses, entropy was prioritized in subsequent multivariable models, with Markov switching used as a complementary local measure.

3.2 Epigenetic Dysregulation and Early Metastatic Spread

Among 327 primary tumors in the matched-pairs subset with available entropy and mDNAsi values (BRCA = 90; KIRC = 158; LUAD = 23; THCA = 56), 275 had both primary lymph-node status and age recorded and were therefore eligible for multivariable logistic regression. As summarized in Table 1 and visualized in Figure 3, completeness was high for THCA (55/56) and KIRC (154/158), moderate for BRCA (66/90), and absent for LUAD (0/23), such that LUAD did not contribute node-labeled cases to this analysis.

In unadjusted comparisons pooling all cancers, node-positive (N+) tumors showed higher distributions of methylation entropy and mDNAsi than node-negative (N0) tumors (Figure 2), consistent with greater epigenetic heterogeneity and stemness-like profiles in the N+ group at the descriptive level. However, these differences were not retained as independent predictors after accounting for cancer type. A multivariable logistic regression model was fitted to predict primary lymph-node positivity (N+ vs N0) using z-scored entropy, mDNAsi, and age as continuous covariates and cancer type as a categorical covariate; because LUAD had no node-status labels, the fitted model effectively included BRCA, KIRC, and THCA only. After adjustment, neither methylation entropy (OR = 1.14 per SD, 95% CI 0.80–1.64, $p = 0.466$) nor mDNAsi (OR = 0.93 per SD, 95% CI 0.68–1.26, $p = 0.629$) was significantly associated with node-positive status, and age was also not significant (OR = 0.99, 95% CI 0.68–1.44, $p = 0.940$) (Table 1). In contrast, cancer type showed strong associations with nodal involvement: relative to BRCA (reference), the odds of nodal positivity were markedly lower in KIRC (OR ≈ 0.0095, 95% CI 0.0020–0.045, $p < 0.001$) and THCA (OR ≈ 0.087, 95% CI 0.017–0.456, $p = 0.004$), indicating that nodal status in this

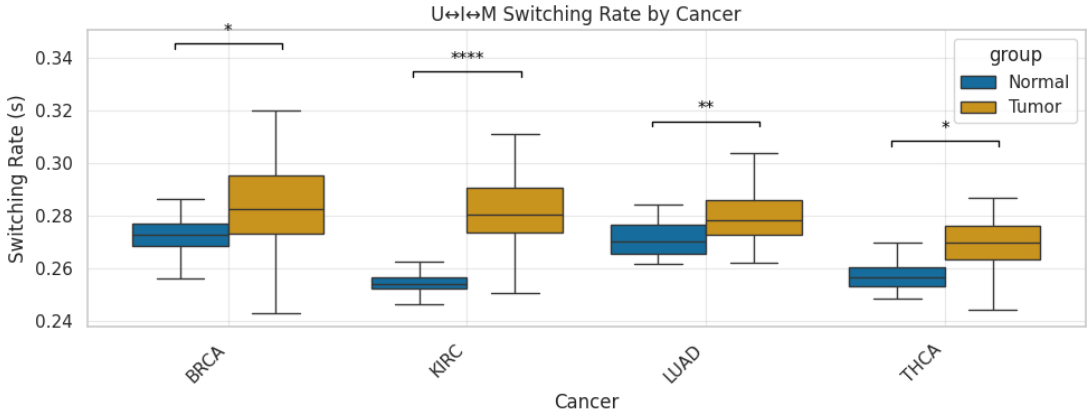

Figure 1: U↔I↔M switching rate by cancer. Boxplots show the distribution of the 3-state Markov switching rate s_{any} (any change among unmethylated U, intermediate I, and methylated M states) for normal (blue) and tumor (orange) samples across BRCA, KIRC, LUAD, and THCA. Boxes indicate the median and interquartile range; whiskers extend to 1.5×IQR. Asterisks denote the significance of tumor–normal differences within each cancer ($^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, $^{****}p < 10^{-4}$; Mann–Whitney U test).

Table 1: Logistic regression for primary lymph-node positivity (N+ vs. N0).

Predictor	OR	95% CI	p-value
Entropy (z)	1.14	0.80–1.64	0.466
mDNAsi (z)	0.93	0.68–1.26	0.629
Age (z)	0.99	0.68–1.44	0.940
KIRC vs. BRCA (ref.)	9.50×10^{-3}	2.00×10^{-3}–0.045	< 0.001
THCA vs. BRCA (ref.)	0.087	0.017–0.457	0.004

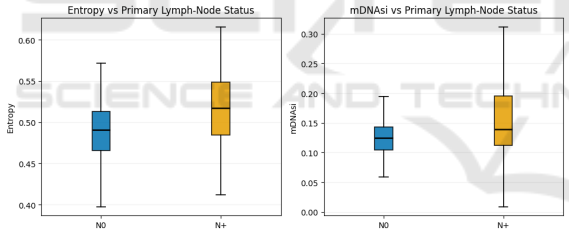

Figure 2: Entropy and mDNAsi by primary lymph-node status. Boxplots compare methylation entropy (left) and mDNAsi (right) between node-negative (N0) and node-positive (N+) primary tumors. Boxes indicate the interquartile range with the median shown as a horizontal line; whiskers extend to 1.5×IQR. Colors denote nodal status (N0, blue; N+, orange).

cohort is dominated by cancer-type-specific baseline differences rather than a universal, feature-driven effect of entropy or mDNAsi. Overall model fit was moderate (Cox–Snell pseudo-$R^2 \approx 0.39$), and discrimination was good (5-fold CV AUC $= 0.83 \pm 0.11$; however, given the absence of LUAD labels and the strong cancer-type coefficients, this discrimination is expected to reflect primarily between-cancer separation, with minimal incremental contribution from entropy or mDNAsi beyond cancer type.

3.3 Association with Chromosomal Instability

Next, it was evaluated whether epigenetic instability was associated with chromosomal instability, quantified as the fraction of genome altered (FGA). After stratification by cancer type (Figure 3), the association between Shannon entropy and FGA was found to be heterogeneous: a weak positive slope was observed in BRCA, whereas KIRC and LUAD exhibited flat to mildly negative slopes, and THCA was characterized by uniformly low FGA values with little apparent dependence on entropy. By contrast, mDNAsi was more consistently positively associated with FGA within BRCA, KIRC, and LUAD, while THCA again remained near the floor of the FGA distribution with minimal variation. In accordance with these stratified patterns, the multivariable ordinary least-squares model adjusted for age and cancer type indicated that any independent contribution of entropy to FGA was small and not robust after adjustment, whereas higher mDNAsi was retained as a modest positive correlate of increased FGA (Figure 3). Collectively, these results suggest that copy-number burden is more closely aligned with stemness-related dedifferentiation than with global methylation disorder, and that the pooled relationship is strongly shaped by cancer-

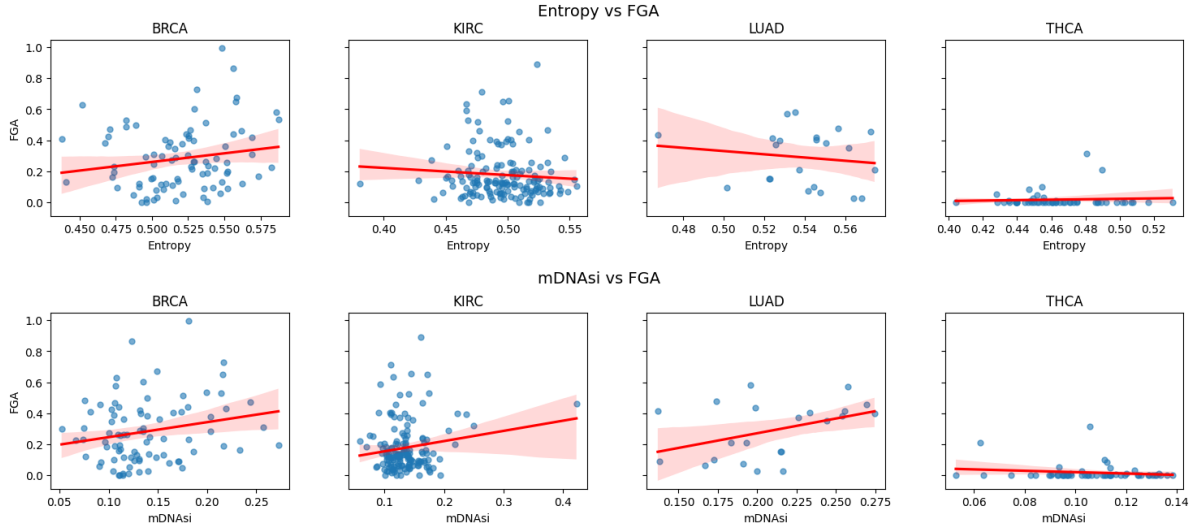

Figure 3: Cancer-stratified associations of epigenetic instability and stemness with chromosomal instability. Scatter plots of FGA versus Shannon entropy (top row) and FGA versus mDNAsi (bottom row) are shown for BRCA, KIRC, LUAD, and THCA (left to right). The red line denotes the within-cancer ordinary least-squares fit and the shaded band indicates the 95% confidence interval. THCA is characterized by uniformly low FGA values, whereas more variable FGA distributions are observed in BRCA, KIRC, and LUAD.

specific structure, including the low-FGA distribution observed in THCA.

3.4 Impact on Overall Survival

To examine the association between global methylation dysregulation and clinical outcome, we performed Kaplan–Meier (KM) overall-survival (OS) analyses in four TCGA cohorts by dichotomizing patients into low vs high groups using a within-cohort median split of Shannon entropy and mDNAsi (Figure 5). Consistent with the KM curves, the entropy-based stratification did not yield statistically significant OS differences in any cohort (BRCA log-rank $p = 0.411$, KIRC $p = 0.261$, LUAD $p = 0.607$, THCA $p = 0.348$). For mDNAsi, separation was also generally weak, with only a borderline trend in BRCA (log-rank $p = 0.055$), while KIRC ($p = 0.181$), LUAD ($p = 0.216$), and THCA ($p = 0.845$) showed no evidence of group-wise OS differences (Figure 5).

We then fitted cancer-specific Cox proportional-hazards models with OS time and status as the outcome and within-cancer z-scored entropy, mDNAsi, and age as covariates. The resulting hazard-ratio patterns were modest in magnitude and clearly cancer-dependent (Figure 4). For entropy, LUAD showed the largest positive association (HR > 1 per SD), BRCA showed a smaller positive association, KIRC was approximately neutral (HR ≈ 1), and THCA showed an inverse association (HR < 1). For mDNAsi, BRCA showed a positive association (HR > 1), whereas LUAD showed a negative association (HR < 1), and KIRC/THCA were closer to null with wide uncertainty (Figure 4). Overall, these results indicate that neither entropy nor mDNAsi behaves as a consistent pan-cancer prognostic marker in this four-cohort set; instead, their associations with OS are weak-to-moderate and heterogeneous across tumor types, in agreement with the largely non-significant KM median-split comparisons.

Finally, to probe whether any single cohort could drive apparent nonlinear/pan-level shapes, we performed leave-one-cancer-out sensitivity analyses (Figure 4). The entropy curve retained a similar overall shape across exclusions, suggesting that the pan-level nonlinear pattern is not dominated by a single cancer. In contrast, the mDNAsi curve changed more noticeably when BRCA was excluded, indicating greater sensitivity of the pan-level mDNAsi pattern to cohort composition (Figure 4). These sensitivity results reinforce that any pooled interpretation should be treated cautiously and that cohort-specific biology can strongly influence apparent prognostic shapes.

3.5 Correlation with the Tumor Microenvironment

To relate global genomic/epigenomic disorder to clinical and microenvironmental burden, we compared pairwise distance matrices derived from tumor en-

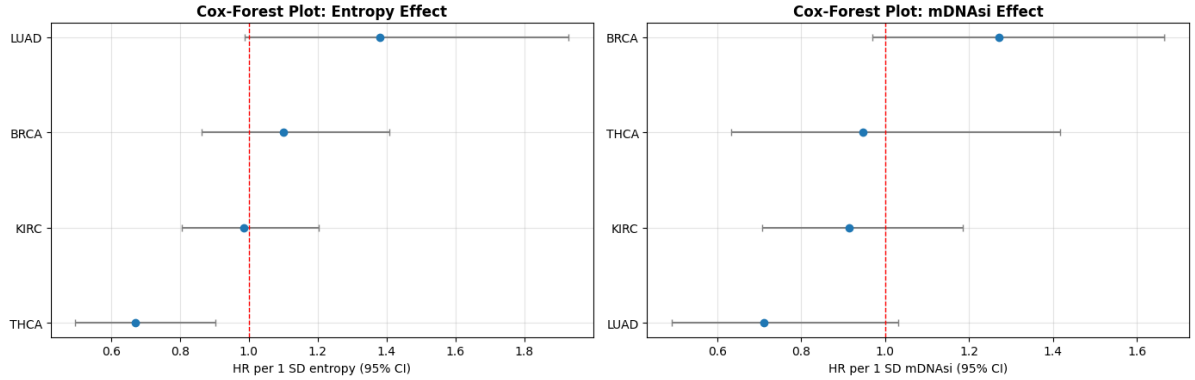

Figure 4: Cox forest plots for overall survival across four TCGA cohorts. Cancer-specific Cox proportional-hazards models were fit with overall survival (OS) as the outcome and within-cancer z-scored entropy, mDNAsi, and age as covariates. Points indicate hazard ratios (HRs) per $+1$ SD of the indicated marker; horizontal bars denote 95% confidence intervals. The red dashed vertical line marks HR$=1$ (no association).

tropy, mDNAsi, and fraction of genome altered (FGA) against distance structures from Ragnum hypoxia scores and lymph-node status in the pooled four-cancer cohort ($N = 273$) using Mantel tests (Table S4; Figure 6). Entropy distance showed a modest but significant alignment with hypoxia distance (Mantel $r = 0.156$, permutation $p = 0.001$), whereas FGA distance was more strongly concordant with hypoxia ($r = 0.310$, $p = 0.001$). mDNAsi distances also tracked hypoxia significantly, but with a smaller effect size than FGA (Table S4; Figure 6). In contrast, lymph-node status was largely decoupled from these global measures: node-distance was essentially unrelated to entropy (Mantel $r = 0.002$, $p = 0.589$) and only trivially associated with FGA ($r = 0.022$, $p = 0.002$), consistent with a statistically detectable but biologically very small effect at this sample size. The Mantel link diagram summarizes this pattern, showing strong links to hypoxia and weak or non-significant links to node status (Figure 6). A complementary Pearson correlation heat map of the underlying variables recapitulates the same structure: entropy and mDNAsi correlate moderately with hypoxia (Pearson $r = 0.42$ and $r = 0.34$, respectively), FGA correlates more strongly with hypoxia ($r = 0.53$), while node positivity shows near-zero linear correlations with entropy, mDNAsi, FGA, and hypoxia (all $|r| \leq 0.09$; Figure 6). Together, these analyses suggest that global genomic and epigenomic disorder covary with hypoxic tumor microenvironments, but are largely uncoupled from early nodal spread in this pan-four-cancer setting. Mechanistically, chronic hypoxia may impair the fidelity of maintenance methylation for example, via reduced *DNMT1* and *UHRF1* activity at hemimethylated CpGs thereby promoting passive demethylation, increased U-I-M switching, and elevated methylation entropy in hypoxic tumor regions (Skowronski et al., 2010; Wang et al., 2021).

4 DISCUSSION

Our entropy–stemness framework links DNA methylation entropy (mDNAsi), chromosomal instability, and hypoxia to clinical endpoints across four representative TCGA cancers. Using a three-state Markov model of unmethylated/intermediate/methylated (U/I/M) dynamics, tumors exhibited consistently higher local switching (s_{any}) than matched normal tissues across BRCA, KIRC, LUAD, and THCA, with the most pronounced differences in KIRC. Importantly, entropy tracked these local dynamics: genome-wide entropy correlated strongly with s_{any}, supporting entropy as a global summary of methylation "switching noise." Collectively, these results indicate that malignant transformation is generally accompanied by reduced methylation stability, consistent with epigenetic instability as a hallmark of oncogenesis.Notably, LUAD still showed a significant tumor–normal switching increase despite its small matched-pairs sample size, suggesting that limited power may primarily constrain downstream clinical associations rather than the presence of transformation-linked instability per sample.

In the multivariable logistic regression model for primary lymph-node involvement, neither methylation entropy nor mDNAsi showed a significant independent effect after adjustment for age and cancer type.Because LUAD contributed no node-labeled cases, the adjusted model effectively compared BRCA, KIRC, and THCA. Although node-

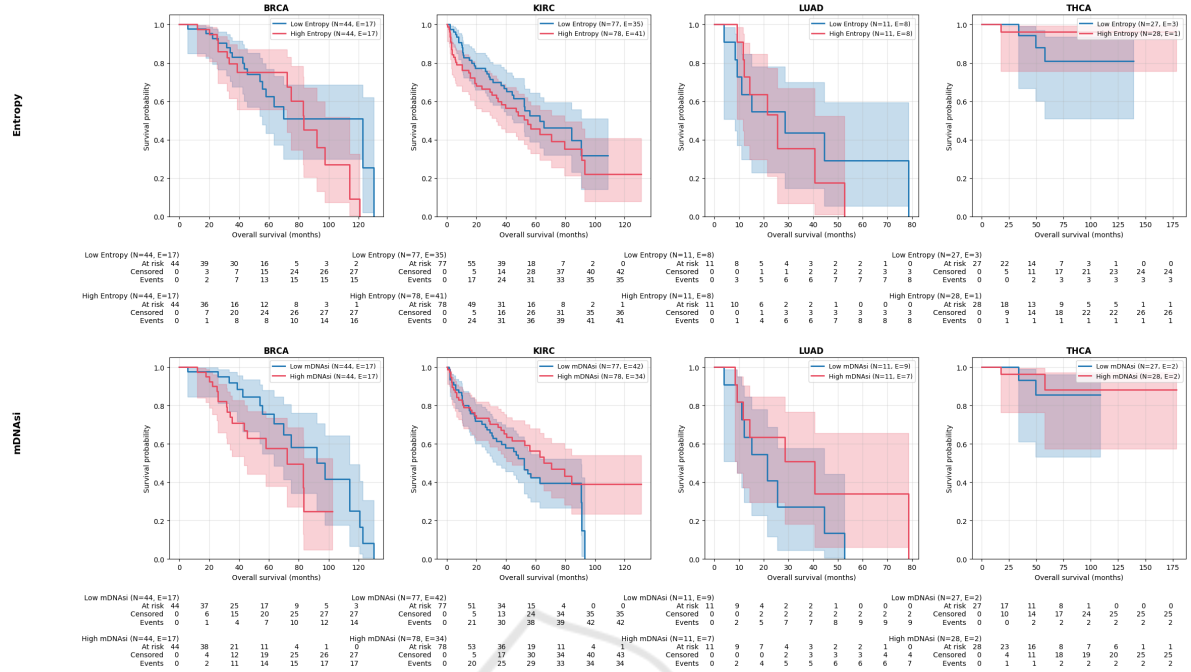

Figure 5: Kaplan–Meier overall survival curves stratified by entropy and stemness across four TCGA cohorts. For each cancer type (BRCA, KIRC, LUAD, THCA), patients were dichotomized into low vs high groups using a median split of the corresponding marker (top row: Shannon entropy; bottom row: mDNAsi). Shaded regions indicate 95% confidence intervals, and numbers at risk are shown below each panel. Log-rank p-values are reported within panels.

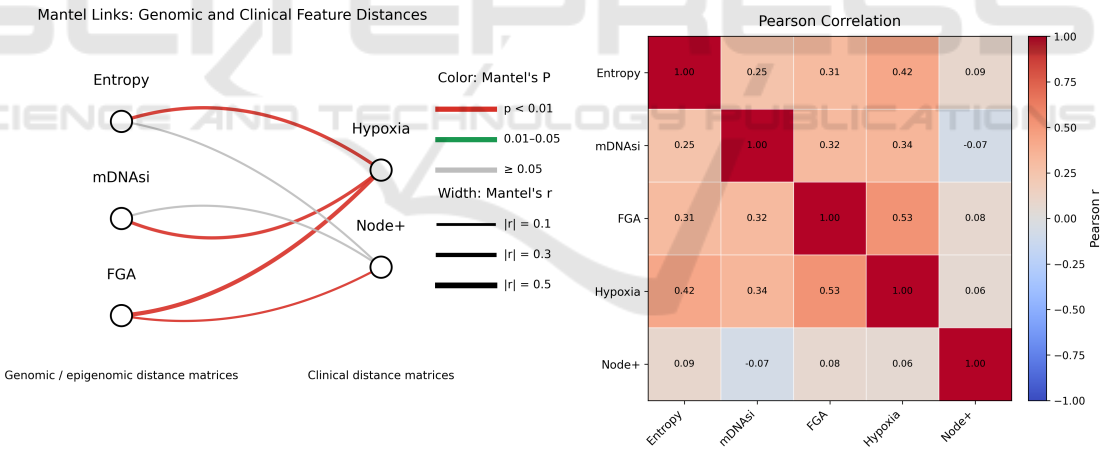

Figure 6: Mantel links and Pearson correlations between genomic/epigenomic disorder and clinical features **.Left:** Mantel test network comparing pairwise distance matrices derived from tumor entropy, mDNAsi and fraction of genome altered (FGA) against distances from Ragnum hypoxia scores and lymph-node status (Node+). Edge colour encodes permutation significance (red: $p < 0.01$; green: $0.01 \leq p < 0.05$; grey: $p \geq 0.05$), and edge width is proportional to $|r|$ (Mantel correlation). **Right:** Pearson correlation matrix of the same variables; numbers denote Pearson's r and the colour scale indicates correlation strength.

positive tumors showed higher entropy and mDNAsi distributions descriptively, neither retained an independent association with nodal positivity after adjustment; instead, cancer type dominated nodal status. Model discrimination was good, consistent with performance largely driven by between-cancer separation rather than incremental signal from global epigenetic disorder. A plausible biological explanation is that nodal spread at diagnosis is shaped primarily by tissue-of-origin programs, anatomic drainage constraints, and subtype-specific invasion dynamics, which may not be encoded by a single bulk, genome-

wide heterogeneity summary.

The relationship between entropy and chromosomal instability (FGA) was modest and cancer-type dependent. Stratified analyses showed a weak positive entropy–FGA trend in BRCA, flat to mildly negative trends in KIRC and LUAD, and uniformly low FGA in THCA with minimal dependence on entropy. In contrast, mDNAsi showed more consistently positive within-cancer associations with FGA in BRCA, KIRC, and LUAD, while THCA remained near the floor. In agreement with these patterns, multivariable OLS suggested that any independent contribution of entropy to FGA was small and not robust after adjustment for age and cancer type, whereas higher mDNAsi persisted as a modest positive correlate of increased FGA. One mechanistic interpretation is that copy-number burden may align more closely with dedifferentiation-related proliferative stress and genome-maintenance failure than with the cumulative "epigenetic drift" component summarized by entropy (Malta et al., 2018; Baba et al., 2024).

Entropy also showed modest and heterogeneous associations with overall survival. Median-split Kaplan–Meier analyses did not show statistically significant OS differences for entropy in any cohort, and mDNAsi stratification was similarly weak overall. Cox models nevertheless suggested cancer-dependent continuous effects on entropy, reinforcing the notion that the prognostic meaning of global entropy is context-specific rather than universal. Given cohort-specific event rates and the sensitivity of cohort-level hazard ratio estimates to sparse events in some strata, these survival findings should be treated as exploratory.

Global disorder measures were more closely aligned with the hypoxic tumor microenvironment than with nodal status. Mantel tests showed that entropy distance aligned modestly with hypoxia distance, whereas FGA showed a stronger association; node distance was essentially unrelated to entropy and only trivially associated with FGA. These results support a model in which entropy reports cumulative stress exposure and diversification history, while FGA reflects concurrent genomic fragmentation in hypoxic tumors, rather than either metric serving as a direct readout of metastatic capacity. More broadly, several non-mutually-exclusive factors could explain weak and heterogeneous clinical associations: entropy is a genome-wide aggregate that can dilute locus and program-specific signals relevant to invasion; bulk methylation entropy can be confounded by variable tumor purity and immune/stromal admixture; and dominant, cancer-type-specific biology, including lineage programs, exposure-driven mutagenesis, and hypoxia signaling, can overwhelm marginal effects of global epigenetic disorder.

5 CONCLUSIONS

This study shows that global DNA methylation entropy and U–I–M switching rates robustly distinguish tumors from matched normal tissues across multiple cancer types, supporting epigenetic instability as a core feature of malignant transformation. However, within established tumors, entropy provided only weak and cancer-type–dependent signal for lymph-node involvement and overall survival after covariate adjustment. mDNAsi captured a complementary axis of oncogenic dedifferentiation and, together with FGA and hypoxia, supported a multidimensional landscape of tumor disorder in which entropy is best interpreted as a readout of evolutionary history and cumulative stress exposure rather than a standalone prognostic marker. These findings support using entropy as one component in integrated models that combine epigenetic, genetic, and microenvironmental features.Extending this framework to single-cell and longitudinal data will be essential for translating entropy–stemness metrics into tools that can track tumor trajectories, refine risk stratification, and design interventions that exploit vulnerabilities created by epigenetic dysregulation.

6 FUTURE WORK

In future work, cohort size and diversity will be expanded by combining bulk and single-cell data, immune deconvolution will be integrated, and longitudinal samples collected at recurrence or metastasis will be analyzed. A broader feature set will be incorporated and cancer-specific models will be developed to reflect distinct tumor origins and evolutionary trajectories, while entropy and stemness will be evaluated against additional endpoints, including immunotherapy response and richer, tumor-type–specific clinical covariates. Findings will be validated in independent external datasets, and the resulting structured signals will be used to train and benchmark LLM-assisted pipelines for future clinical reporting and decision support.

ACKNOWLEDGEMENTS

The authors used the language model ChatGPT (OpenAI, 2025) to assist with paraphrasing and re-ordering

some sentences in the manuscript. All scientific content, analyses, and final wording were reviewed and approved by the authors, who take full responsibility for any remaining errors.

FUNDING

Funded by the European Union under Horizon Europe (project ChatMED Grant Agreement ID: 101159214).

REFERENCES

Baba, S. A., Zakeri, A., and Desgrosellier, J. S. (2024). Chromosomal instability as an architect of the cancer stemness landscape. *Frontiers in Cell and Developmental Biology*, 12:1450614.

Bakhoum, S. F. and Landau, D. A. (2017). Chromosomal instability as a driver of tumor heterogeneity and evolution. *Cold Spring Harbor perspectives in medicine*, 7(6):a029611.

Chan, J., Rubbi, L., and Pellegrini, M. (2025). Dna methylation entropy is a biomarker for aging. *Aging (Albany NY)*, 17(3):685.

Hosea, R., Hillary, S., Naqvi, S., Wu, S., and Kasim, V. (2024). The two sides of chromosomal instability: drivers and brakes in cancer. *Signal transduction and targeted therapy*, 9(1):75.

Lee, H., Lin, P.-Y., and Chen, P.-Y. (2023). There's more to it: Uncovering genomewide dna methylation heterogeneity. *Epigenomics*, 15(13):687–691.

Ma, S., Pan, X., Gan, J., Guo, X., He, J., Hu, H., Wang, Y., Ning, S., and Zhi, H. (2024). Dna methylation heterogeneity attributable to a complex tumor immune microenvironment prompts prognostic risk in glioma. *Epigenetics*, 19(1):2318506.

Malta, T. M., Sokolov, A., Gentles, A. J., Burzykowski, T., Poisson, L., Weinstein, J. N., Kamińska, B., Huelsken, J., Omberg, L., Gevaert, O., et al. (2018). Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell*, 173(2):338–354.

Monyak, D. L., Holloway, S. T., Gumbert, G. J., Grimm, L. J., Hwang, E. S., Marks, J. R., Shibata, D., and Ryser, M. D. (2025). Mapping the temporal landscape of breast cancer using epigenetic entropy. *Communications Biology*, 8(1):1477.

National Cancer Institute and National Human Genome Research Institute (2006). The cancer genome atlas (tcga) program. Accessed 2025-11-11.

Novakovic, B., Yuen, R. K., Gordon, L., Penaherrera, M. S., Sharkey, A., Moffett, A., Craig, J. M., Robinson, W. P., and Saffery, R. (2011). Evidence for widespread changes in promoter methylation profile in human placenta in response to increasing gestational age and environmental/stochastic factors. *BMC genomics*, 12(1):529.

OpenAI (2025). Chatgpt (gpt-5 thinking). Large language model.

Ragnum, H., Vlatkovic, L., Lie, A., Axcrona, K., Julin, C., Frikstad, K., Hole, K., Seierstad, T., and Lyng, H. (2015). The tumour hypoxia marker pimonidazole reflects a transcriptional programme associated with aggressive prostate cancer. *British journal of cancer*, 112(2):382–390.

Shao, C., Dai, W., Li, H., Tang, W., Jia, S., Wu, X., and Luo, Y. (2017). The relationship between rassf1a gene promoter methylation and the susceptibility and prognosis of melanoma: A meta-analysis and bioinformatics. *PloS one*, 12(2):e0171676.

Sheffield, N. C., Pierron, G., Klughammer, J., Datlinger, P., Schönegger, A., Schuster, M., Hadler, J., Surdez, D., Guillemot, D., Lapouble, E., et al. (2017). Dna methylation heterogeneity defines a disease spectrum in ewing sarcoma. *Nature medicine*, 23(3):386–395.

Skowronski, K., Dubey, S., Rodenhiser, D. I., and Coomber, B. (2010). Ischemia dysregulates dna methyltransferases and p16ink4a methylation in human colorectal cancer cells. *Epigenetics*, 5(6):547–556.

Teng, H., Xue, M., Liang, J., Wang, X., Wang, L., Wei, W., Li, C., Zhang, Z., Li, Q., Ran, X., et al. (2020). Inter- and intratumor dna methylation heterogeneity associated with lymph node metastasis and prognosis of esophageal squamous cell carcinoma. *Theranostics*, 10(7):3035.

Vaidya, H., Jeong, H. S., Keith, K., Maegawa, S., Calendo, G., Madzo, J., Jelinek, J., and Issa, J.-P. J. (2023). Dna methylation entropy as a measure of stem cell replication and aging. *Genome biology*, 24(1):27.

Wang, Y., Zhang, Y., Huang, Y., Chen, C., Zhang, X., Xing, Y., Gu, Y., Zhang, M., Cai, L., Xu, S., et al. (2021). Intratumor heterogeneity of breast cancer detected by epialleles shows association with hypoxic microenvironment. *Theranostics*, 11(9):4403.

APPENDIX

Supplementary Tables

Table S1: Three-state Markov model: switching rate s_{any} (U/I/M) in normal vs. tumor.

Cancer	N_{normal}	N_{tumor}	Median s_{any} (normal)	Median s_{any} (tumor)	p-value
BRCA	90	90	0.2728	0.2824	0.0107
KIRC	158	158	0.2542	0.2802	1.55×10^{-13}
LUAD	23	23	0.2703	0.2784	0.00603
THCA	56	56	0.2566	0.2698	0.0391

Table S2: Per-cancer logistic regression for node positivity (N+ vs. N0) using z-standardized entropy, mDNAsi, and age.

Cancer	N	N_{N+}	Predictor	OR	95% CI	p-value
BRCA	66	64	Entropy (z)	3.15	0.134–73.8	0.477
BRCA	66	64	mDNAsi (z)	1.06	0.0611–18.3	0.969
BRCA	66	64	Age (z)	3.90×10^{-3}	1.78×10^{-6}–8.55	0.158
KIRC	154	34	Entropy (z)	1.08	0.73–1.60	0.700
KIRC	154	34	mDNAsi (z)	0.905	0.59–1.39	0.649
KIRC	154	34	Age (z)	1.20	0.815–1.76	0.359
THCA	55	39	Entropy (z)	1.08	0.598–1.96	0.795
THCA	55	39	mDNAsi (z)	0.873	0.481–1.59	0.656
THCA	55	39	Age (z)	0.981	0.548–1.76	0.950

Table S3: Cox models: hazard ratio for entropy per cancer type.

Cancer	N	HR (Entropy, z)	95% CI (lower)	95% CI (upper)	p-value	C-index
LUAD	22	1.60	0.98	2.61	0.06	0.63
BRCA	88	1.11	0.85	1.45	0.44	0.66
KIRC	155	0.98	0.76	1.27	0.88	0.59
THCA	55	0.57	0.37	0.87	0.009	0.93

Table S4: Mantel correlations between genomic/epigenomic disorder and clinical burden.

Comparison	Mantel r	Permutation p	N
Entropy vs Hypoxia	0.1779659136	0.001	315
mDNAsi vs Hypoxia	0.1630901434	0.001	315
FGA vs Hypoxia	0.2871189631	0.001	315
Entropy vs Node+	0.002270698151	0.585	272
mDNAsi vs Node+	0.005814954402	0.077	272
FGA vs Node+	0.0240629010	0.002	272

Supplementary Figures

Figure 1: Entropy correlates with three-state Markov switching. Scatter plot of genome-wide methylation entropy versus the three-state Markov switching rate s_{any} (any adjacent state change among unmethylated U, intermediate I, and methylated M states), shown for normal and tumor samples. A fitted linear trend is shown in black. Spearman and Pearson correlation coefficients are reported in the inset.

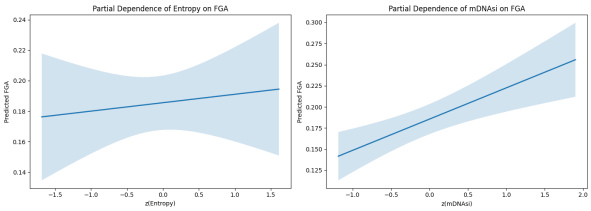

Figure 2: Adjusted partial dependence of entropy and stemness on chromosomal instability. Partial dependence plots from the multivariable linear model of FGA, adjusted for age and cancer type, are shown for Shannon entropy (left) and mDNAsi (right). Curves represent the mean predicted FGA across the four-cancer dataset as the corresponding z-standardized predictor is varied, with shaded bands indicating 95% confidence intervals.

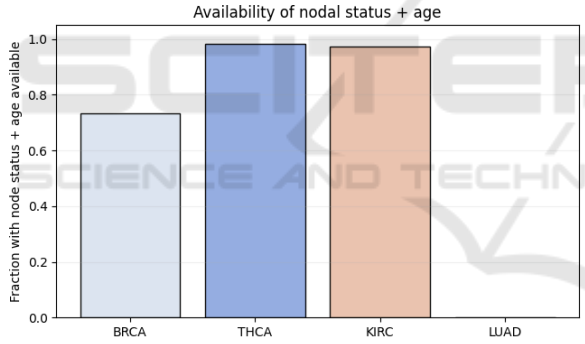

Figure 3: Availability of nodal status and age in the matched-pairs subset. Bars show, for each cancer type, the fraction of primary tumors with both primary lymph-node status (N0/N+) and age at diagnosis recorded among tumors with available entropy and mDNAsi in the matched tumor–normal (paired) cohort. LUAD contributes no node-labeled cases, whereas completeness is high for THCA and KIRC and moderate for BRCA.

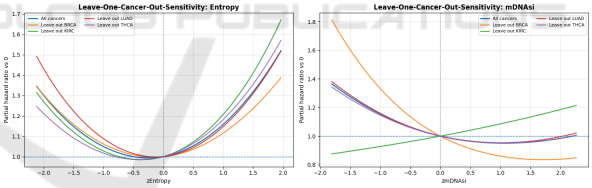

Figure 4: Leave-one-cancer-out sensitivity of the pan-4 stratified spline Cox model. Partial hazard-ratio curves from the pan-4 cancer Cox model with natural cubic splines are shown for entropy (left; zEntropy) and stemness (right; zmDNAsi), each plotted relative to the reference value 0. The blue curve uses all cancers; additional curves show refits after excluding one cancer cohort at a time, highlighting whether the overall shape is dominated by a single cohort. Dashed lines indicate HR= 1 and the reference point at 0.