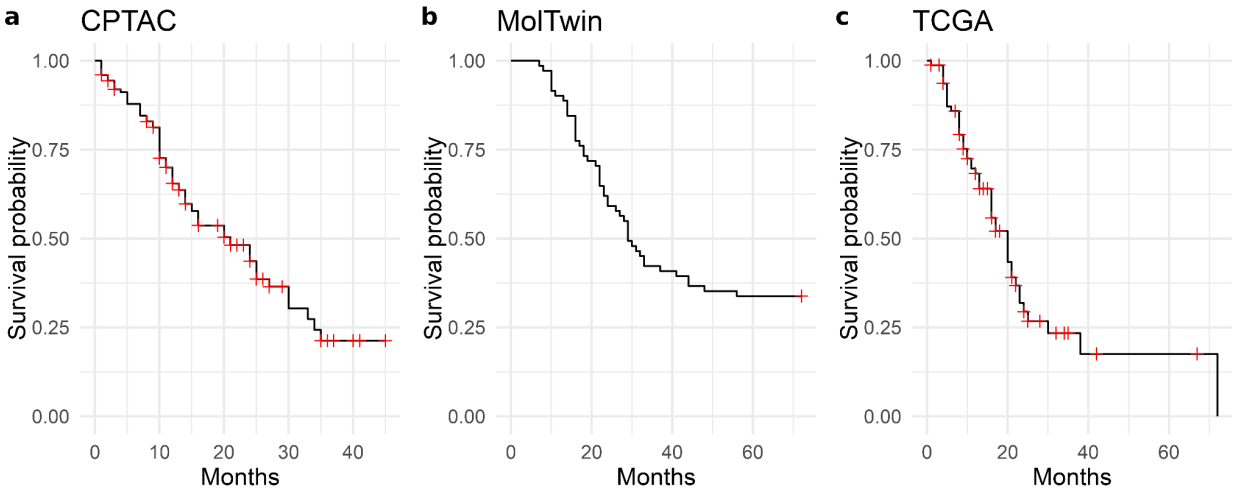
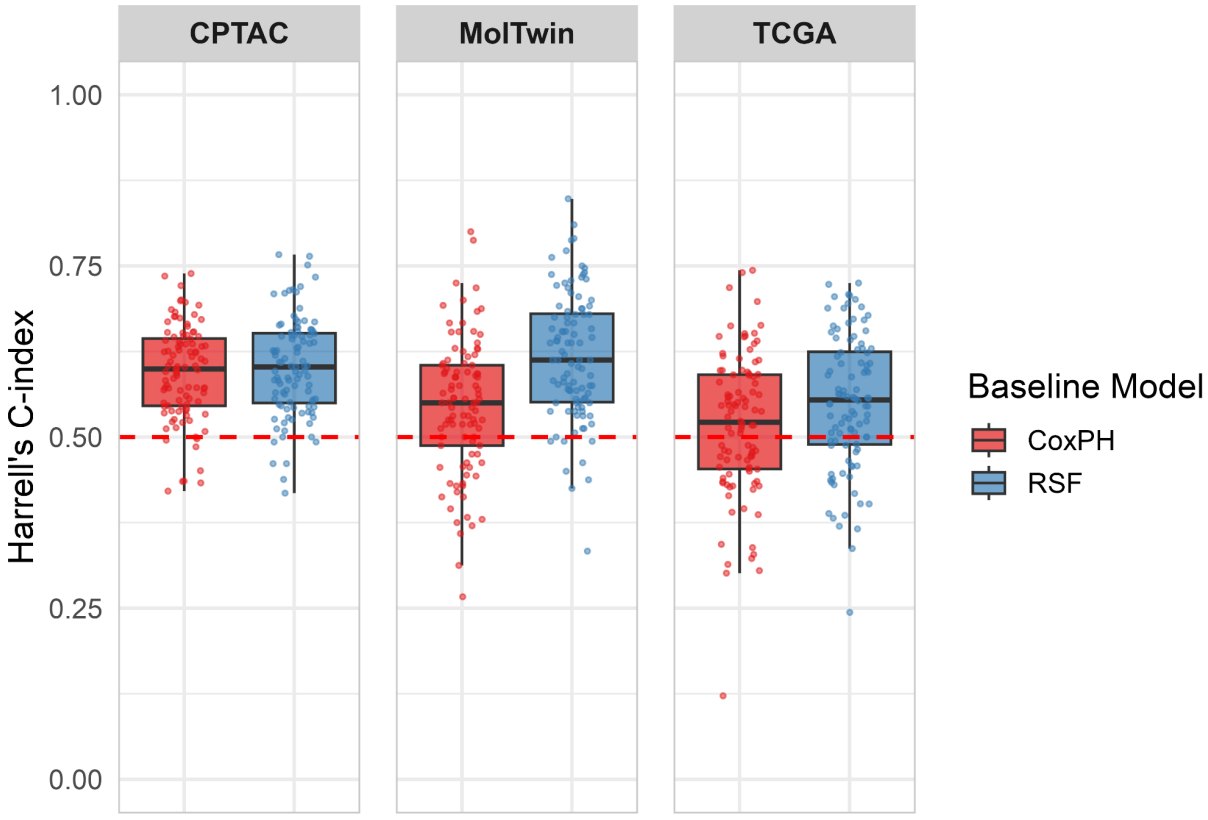
Supplementary Material

Supplementary Figure 1



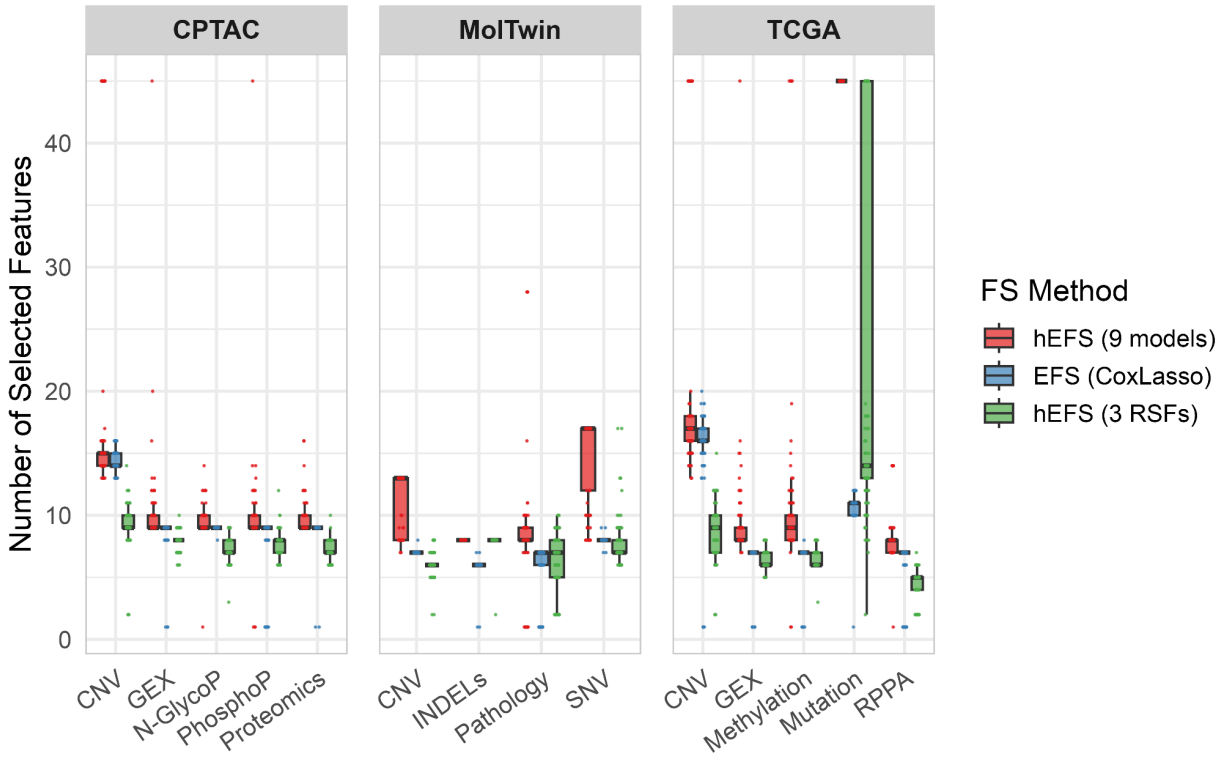
Supplementary Figure 1. Kaplan–Meier survival curves for the three PDAC cohorts used in this study. (**a**) CPTAC-PDAC [1] cohort (**b**) MolTwin [2] cohort; (**c**) TCGA-PDAC [3] cohort. Each plot displays the estimated survival probability over time (in months) for patients with pancreatic ductal adenocarcinoma (PDAC). Red ticks indicate censored observations. The MolTwin cohort exhibits administrative censoring at the study's maximum follow-up time of 72 months, whereas censoring in the TCGA and CPTAC cohorts is distributed throughout the follow-up period.

Supplementary Figure 2



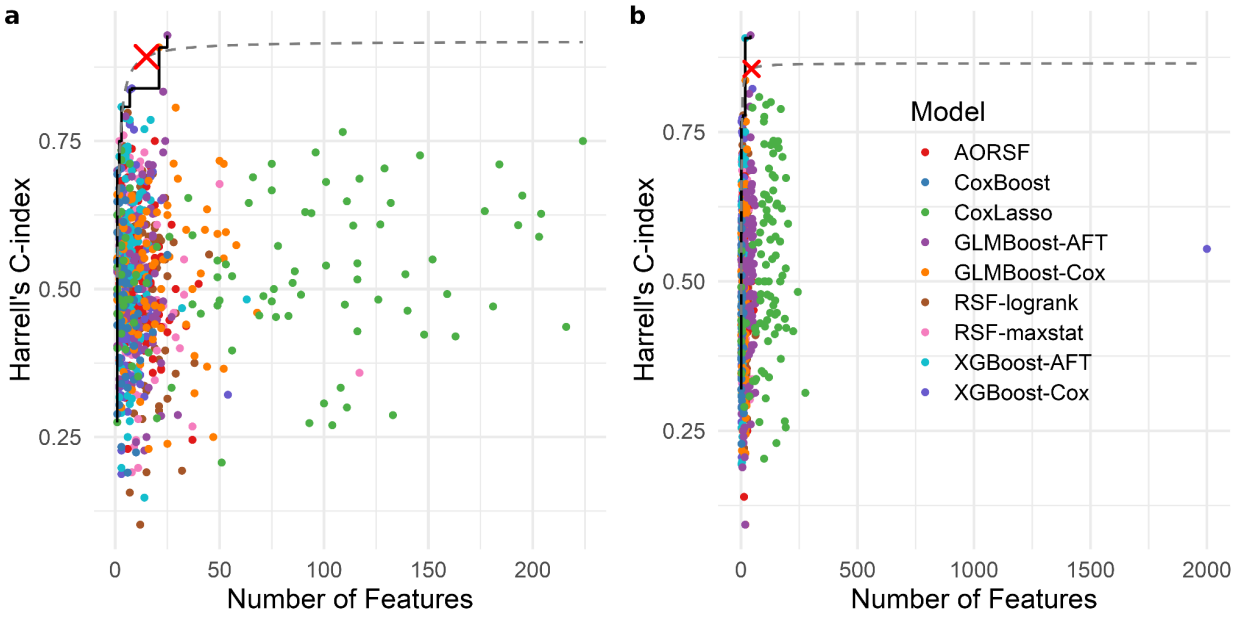
Supplementary Figure 2. Comparison of baseline survival models trained on clinical features only. Harrell's C-index distribution for Cox proportional hazards (CoxPH) and random survival forest (RSF) models trained exclusively on clinical variables, across 100 Monte Carlo cross-validation (MC-CV) iterations. Results are shown separately for the CPTAC [1], MolTwin [2] and TCGA [3] PDAC cohorts. RSF consistently outperforms CoxPH in two out of three datasets, motivating its use as the reference integration model in our benchmark. The red dashed line marks the random discriminatory performance with a C-index of 0.5.

Supplementary Figure 3



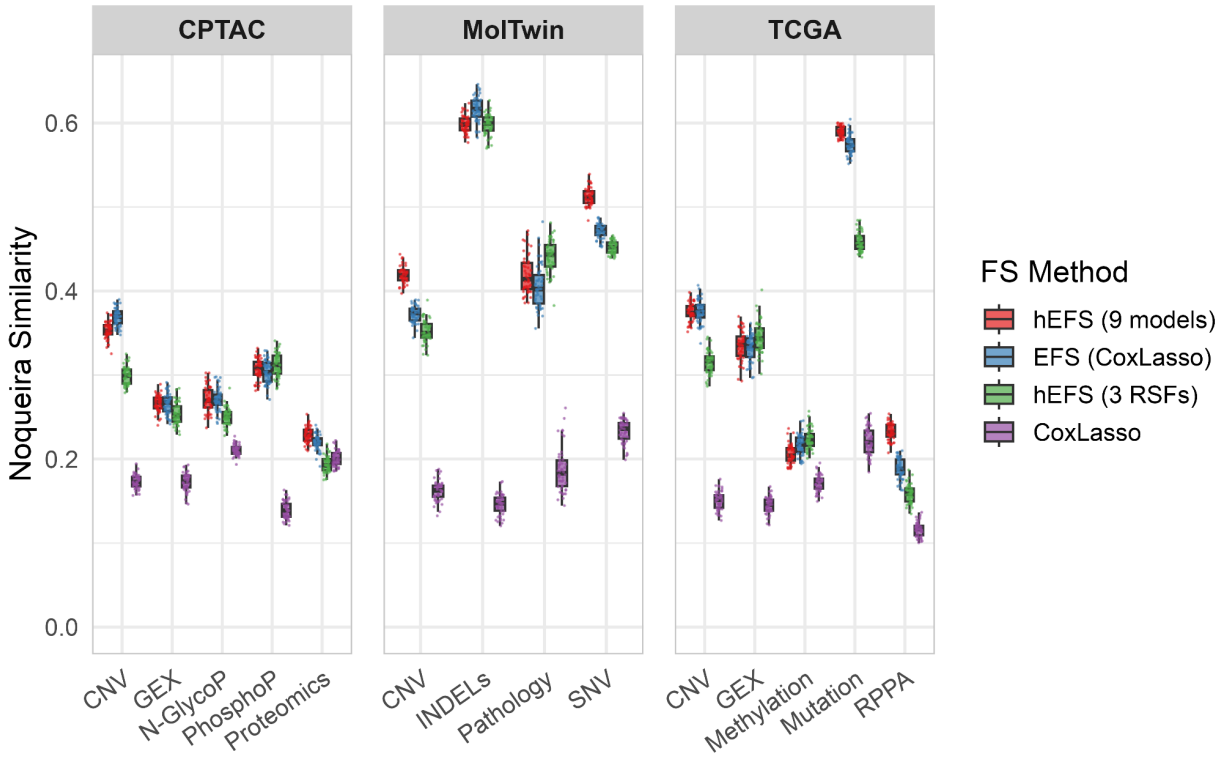
Supplementary Figure 3. Per-omic feature sparsity across three PDAC multi-omics datasets, excluding the CoxLasso baseline. This figure replicates the analysis from **Fig. 3a** but excludes the CoxLasso baseline to better visualize the lower feature counts of the hybrid ensemble feature selection (hEFS) methods. Across 100 Monte Carlo cross-validation iterations, the hEFS (3 RSF) yields the most sparse selections, followed by the EFS (CoxLasso), and then the full hEFS with all nine models—regardless of the original omic dimensionality. In most cases, fewer than 15 features on average are selected per omic layer.

Supplementary Figure 4



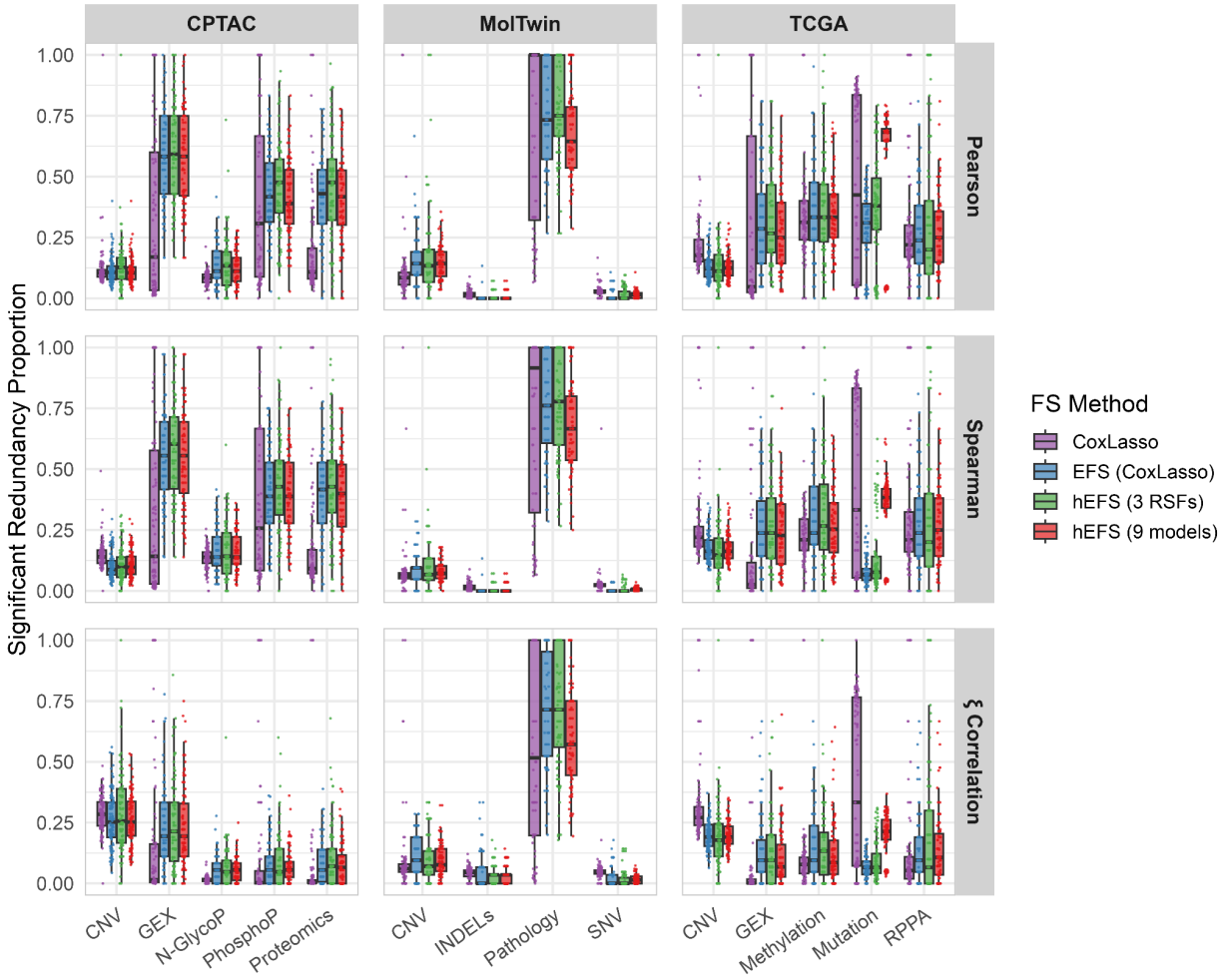
Supplementary Figure 4. Biomarker panel size variability driven by the shape of the estimated Pareto front. Pareto fronts (empirical: stepwise, black; estimated via inverse-number-of-features weighting: dashed, grey; **Methods**) and selected knee points (red crosses) from two outer resampling iterations using mutation data from the TCGA cohort [3]. Each case corresponds to 65 patients and the 2000 most variable features. These illustrate how the number of selected features in the hEFS (9 models) variant can vary depending on the structure of the Pareto front. **(a)** The Pareto front yields a compact solution with a knee point at 15 features. Points not on the empirical front—such as those from CoxLasso—are excluded from the estimated front, so the knee point is calculated solely using the remaining Pareto-optimal solutions (grey dashed line). **(b)** The inclusion of a model–data pair with substantially more features (e.g., XGBoost-Cox in one inner resampling) 'stretches' the estimated Pareto front, shifting the knee point to a higher value (45 features). This highlights how such outliers can influence the Pareto front geometry and increase the selected biomarker panel size.

Supplementary Figure 5



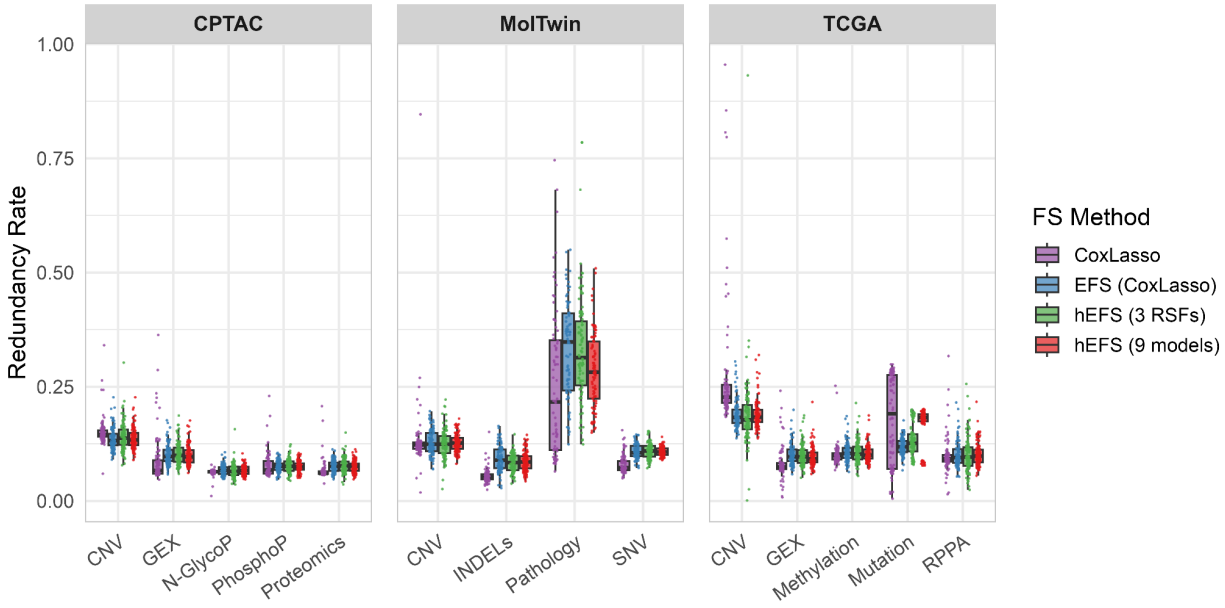
Supplementary Figure 5. Robustness of stability estimates across Monte Carlo cross-validation (MC-CV) iterations. For each dataset, omic type, and feature selection method combination, we randomly sampled 50 MC-CV runs (and their corresponding selected feature subsets) from the original 100 runs, and repeated this resampling procedure 50 times. Nogueira similarity scores were computed for each resampled set, and boxplots summarize the distribution of these scores across replicates. The consistently low variability (interquartile ranges typically <0.05) confirms the robustness of the stability patterns observed in **Fig. 4a**.

Supplementary Figure 6



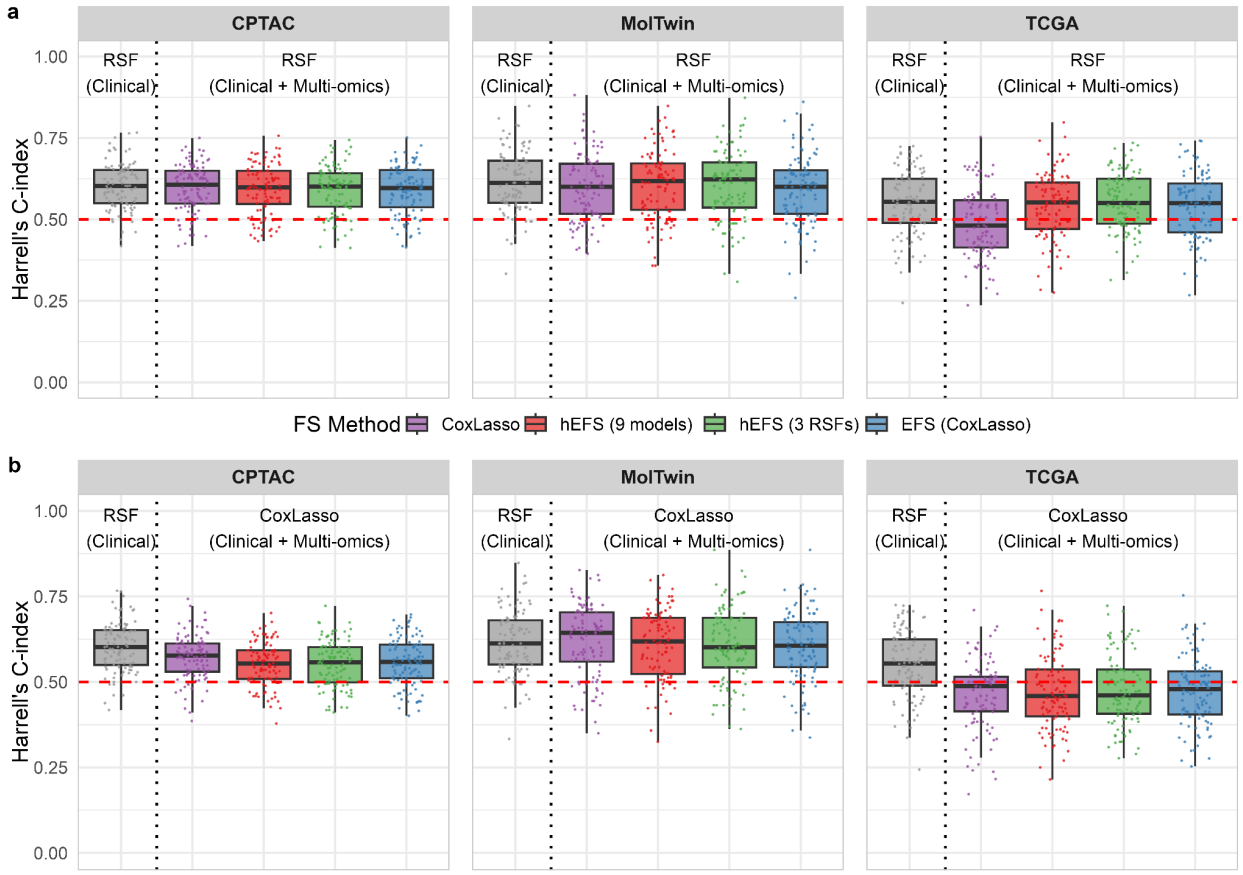
Supplementary Figure 6. Significant Redundancy Proportion across Omic Types and Correlation Metrics. Each panel shows the proportion of feature pairs identified as significantly redundant (FDR-adjusted $p < 0.05$) for each feature selection (FS) method, stratified by omic type (x-axis) and PDAC cohort (columns). Redundancy was assessed using three different correlation measures (rows): Pearson, Spearman, and ξ correlation [4]. Trends across correlation metrics were consistent, with FS methods showing broadly similar relative patterns within each omic type. Notably, ξ correlation (bottom row; same as in **Fig. 4b**) yielded consistently lower redundancy proportions across omics and datasets, reflecting its sensitivity to more general forms of statistical dependence rather than strict linear or monotonic associations. Boxplots summarize variability across 100 Monte Carlo cross-validation iterations.

Supplementary Figure 7



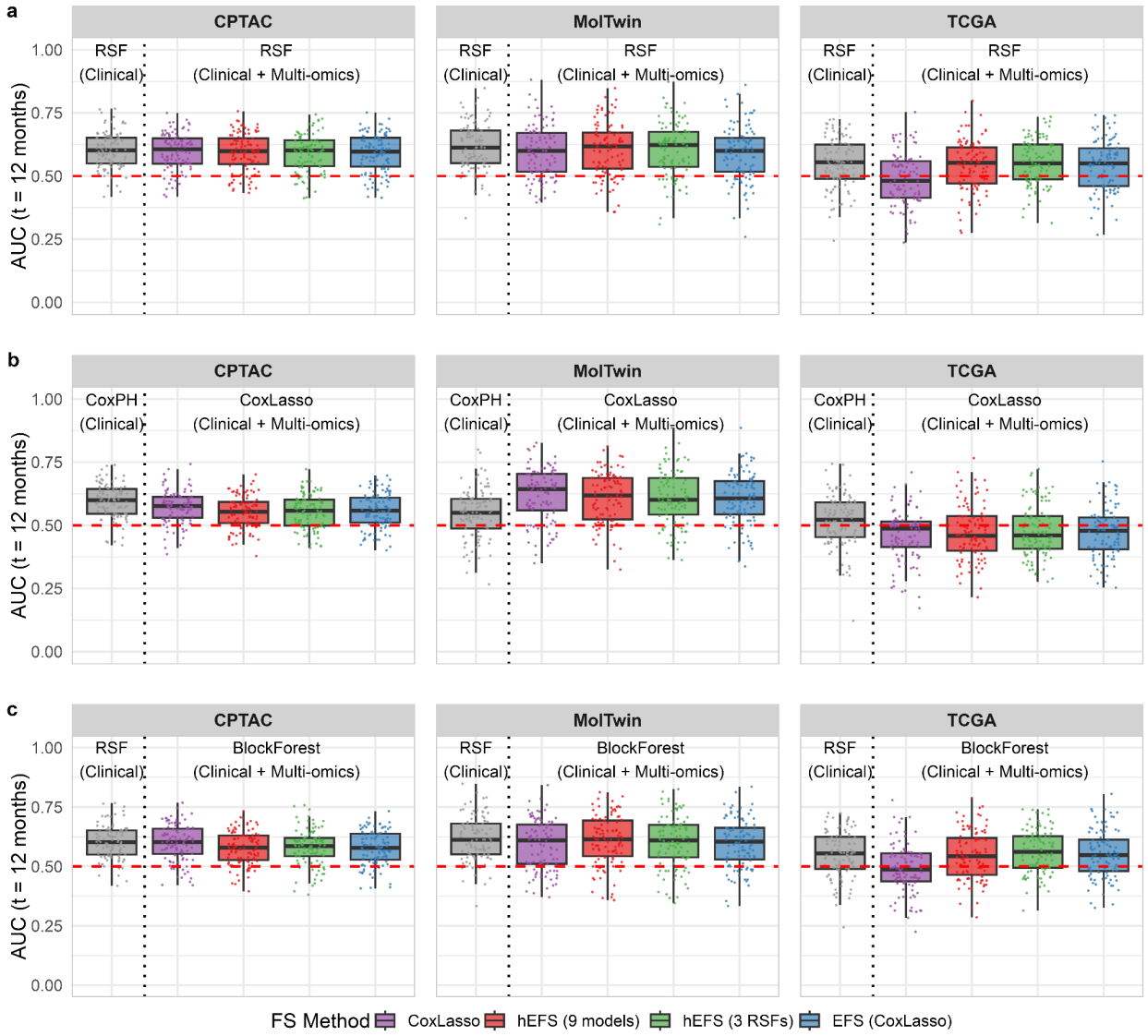
Supplementary Figure 7. Redundancy rate across omic types based on mean absolute ξ correlation. Boxplots show the distribution of mean absolute ξ correlation scores across 100 Monte Carlo cross-validation iterations, stratified by feature selection (FS) method, omic type, and PDAC cohort. Lower values indicate lower average redundancy among selected feature pairs. Across most omics and cohorts, redundancy rates remained low (median ≤ 0.2) and comparable between FS methods.

Supplementary Figure 8



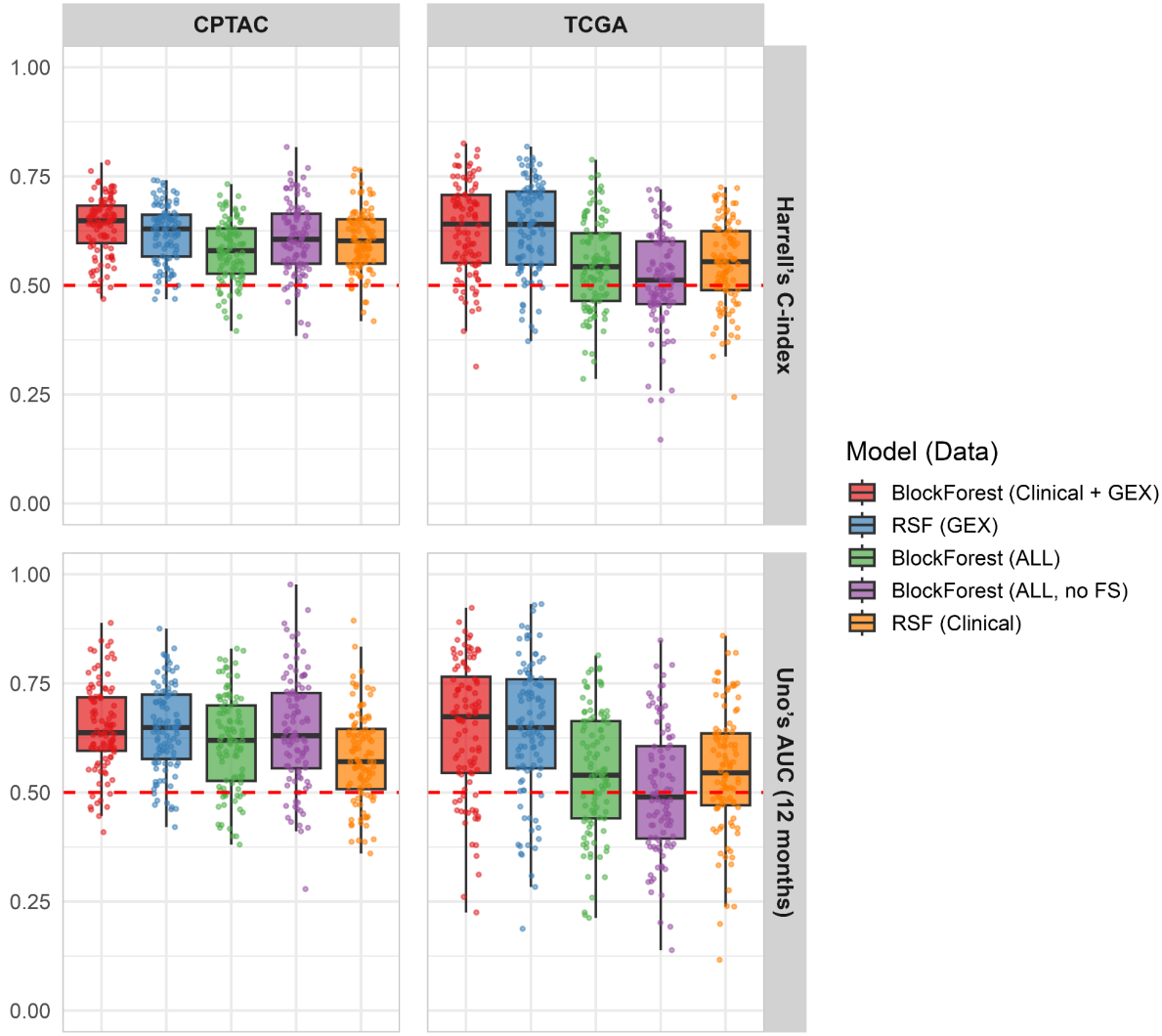
Supplementary Figure 8. Predictive performance (C-index) across feature selection methods and PDAC cohorts using group-naive integration models. Discriminatory performance of (**a**) Random Survival Forests (RSF) and (**b**) CoxLasso models, that integrate clinical data with late-fused, feature-selected multi-omics matrices, benchmarked against a baseline RSF trained only on clinical features (grey). The red dashed line indicates random discriminatory ability (C-index = 0.5). Higher C-index, better discriminatory performance. Boxplots summarize variability across 100 Monte Carlo cross-validation iterations.

Supplementary Figure 9



Supplementary Figure 9. Time-dependent predictive performance (AUC) across feature selection methods and PDAC cohorts using three integration models. Discriminatory performance is shown for two group-naïve models—Random Survival Forests (RSF; **a**) and CoxLasso (**b**)—and one group-aware model, BlockForest (**c**). All approaches use clinical data combined with late-fused, feature-selected multi-omics matrices, and are benchmarked against a baseline RSF trained on clinical features only (grey). The red dashed line denotes random discriminatory performance (AUC = 0.5). Uno's AUC was evaluated at 12 months. Boxplots summarize performance variability across 100 Monte Carlo cross-validation iterations; higher AUC indicates better discrimination.

Supplementary Figure 10



Supplementary Figure 10. **Comparison of discriminatory performance between single-omic (gene expression), clinical-only, clinical + gene expression (GEX), and multi-omic models across the CPTAC [1] and TCGA [3] PDAC cohorts**. For each cohort, performance is shown for both Harrell's C-index and Uno's time-dependent AUC evaluated at 12 months. 'ALL' denotes models using clinical data combined with all available multi-omics data for the respective cohort. Feature selection with hEFS (all 9 models) was applied to the GEX data and used in RSF (GEX; blue), BlockForest (Clinical + GEX; red), and BlockForest (ALL; green). For comparison, BlockForest (ALL, no FS) used the full set of omics plus clinical data without feature selection. Boxplots summarize variability across 100 Monte Carlo cross-validation iterations. The MolTwin cohort is not shown, as GEX data were not used due to sample size constraints.

Supplementary Table 1

Supplementary Table 1. Average execution times (in seconds) with standard deviations for feature selection methods across PDAC cohorts (ordered by decreasing sample size), averaged over omic data types and Monte Carlo CV iterations.

PDAC Dataset	CoxLasso	EFS (CoxLasso)	hEFS (3 RSFs)	hEFS (9 models)
CPTAC [1]	2.40 ± 1.50	16.99 ± 2.62	115 ± 32	836 ± 182
TCGA [3]	2.24 ± 1.02	13.79 ± 1.98	77 ± 16	634 ± 99
MolTwin [2]	1.72 ± 0.67	11.76 ± 3.00	50 ± 16	473 ± 86

References

[1] Cao, L., Huang, C., Cui Zhou, D., Hu, Y., Lih, T. M., Savage, S. R., Krug, K., Clark, D. J., Schnaubelt, M., Chen, L., da Veiga Leprevost, F., Eguez, R. V., Yang, W., Pan, J., Wen, B., Dou, Y., Jiang, W., Liao, Y., Shi, Z., … Zhao, G. (2021). Proteogenomic characterization of pancreatic ductal adenocarcinoma. *Cell*, *184*(19), 5031-5052.e26. https://doi.org/10.1016/J.CELL.2021.08.023

[2] Osipov, A., Nikolic, O., Gertych, A., Parker, S., Hendifar, A., Singh, P., Filippova, D., Dagliyan, G., Ferrone, C. R., Zheng, L., Moore, J. H., Tourtellotte, W., Van Eyk, J. E., & Theodorescu, D. (2024). The Molecular Twin artificial-intelligence platform integrates multi-omic data to predict outcomes for pancreatic adenocarcinoma patients. *Nature Cancer 2024*, 1–16. https://doi.org/10.1038/s43018-023-00697-7

[3] Wissel, D., Rowson, D., & Boeva, V. (2023). Systematic comparison of multi-omics survival models reveals a widespread lack of noise resistance. *Cell Reports Methods*, *3*(4), 100461. https://doi.org/10.1016/j.crmeth.2023.100461

[4] Chatterjee, S. (2021). A New Coefficient of Correlation. *Journal of the American Statistical Association*, *116*(536), 2009–2022. https://doi.org/10.1080/01621459.2020.1758115