Epigenetic profile drives accurate survival prediction in breast cancer via a multi-omics machine learning model

Guillaume Mestrallet[1], Alexandre Pierga[2], Paul Fremont[3]

[1] Division of Hematology and Oncology, Hess Center for Science & Medicine, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

[2] Université de Paris, Imagine Institute, Laboratory of Neurogenetics and Neuroinflammation, Institut National de la Santé et de la Recherche Médicale, Unité mixte de recherche 1163, Paris, France.

[3] Department of Biology, University of Maryland, College Park, MD, USA.

Abstract

Accurate overall survival (OS) prediction is key for personalized treatment in breast cancer, but mutation burden alone is insufficient. To improve prognostic accuracy, we integrated genomic, transcriptomic, proteomic, epigenetic, and clinical features from 802 breast cancer patients to develop BANDOL (Breast cancer Analysis with Neoplastic Data and Omics Learning), a Random Survival Forest model. BANDOL correctly predicted survival ranking in 75% of patient pairs and outperformed mutation-based models (time-dependent AUC: 0.9–1 vs. 0.4–0.9). Immune activation signatures correlated with a longer OS after therapy, while a shorter OS was linked to TREM2 myeloid cells, B cells, and leptin signaling. Model robustness was confirmed in two independent The Cancer Genome Atlas (TCGA) cohorts (uterine, glioma cancers). Epigenetic features were the strongest OS predictors for BANDOL. Current therapies may be combined with strategies to target the methylations of ME3, PPARG, OLIG3 and SLC25A22 genes. This study demonstrates that multi-omics integration via machine learning enhances survival prediction and reveals actionable biomarkers.

Key words

Epigenetic, protein, RNA, mutation, cancer, overall survival, machine learning, multi-omics

Introduction

The accumulation of somatic mutations and the resulting increase in tumor mutational burden (TMB) have emerged as important biological features with potential clinical implications across multiple cancer types. High TMB is often associated with an increased likelihood of generating neoantigens, which may enhance tumor immunogenicity and facilitate immune-mediated tumor clearance. Consequently, a growing body of evidence suggests that elevated TMB correlates with improved responses to immune checkpoint blockade therapies and, in some cancers, with better overall survival (OS) [1–4]. OS is the length of time from either the date of diagnosis or the start of treatment for a disease, such as cancer, that patients diagnosed with the disease are still alive. Building on this observation, several studies have utilized mutational data in combination with machine learning techniques to predict patient outcomes and therapeutic responses. These approaches have demonstrated that predictive modeling using mutational features can yield clinically relevant insights and may inform personalized treatment strategies [5–7]. However, the prognostic significance of TMB is not universal. For instance, in breast cancer and other malignancies, high TMB does not reliably predict improved survival outcomes [8,9]. This inconsistency underscores the limitations of relying solely on TMB or mutation count as prognostic biomarkers, particularly in cancers with low immunogenicity or distinct tumor microenvironments.

To address this limitation, more comprehensive models that integrate multi-dimensional tumor features are needed. A growing consensus in the field supports the incorporation of transcriptomic, proteomic, and epigenetic data, such as RNA and protein expression levels and DNA methylation patterns, alongside mutational and clinical features to improve prognostic accuracy [10,11]. These data layers provide complementary insights into tumor biology, capturing regulatory, functional, and structural alterations that are not reflected in mutational profiles alone. While previous studies have identified specific epigenetic or transcriptomic signatures associated with prognosis, these efforts have generally focused on isolated data types and have not employed integrative, machine learning–based approaches that could exploit the full spectrum of molecular data [12]. For example in this study the authors identified an epigenetic signature associated with OS in breast cancer, and used it to design a monogram stratifying low and high-risk groups, but this approach is not as precise as machine learning approaches and does not incorporate all individual features. As such, the predictive potential of combining multi-omic features with advanced computational models remains partially underexplored, particularly in breast cancer.

In this study, we aim to fill this gap by developing a robust machine learning framework for survival prediction in breast cancer patients. We focus on breast cancer not only because of its clinical relevance and heterogeneity but also due to the absence of a clear association between TMB and patient survival, making it an ideal setting in which to evaluate the added value of integrative modeling (**Figure 1**). To this end, we leveraged data from The Cancer Genome Atlas

(TCGA), which provides a rich compendium of molecular and clinical data, including tumor RNA and protein expression levels, DNA methylation profiles, mutational information, and detailed patient annotations such as OS following treatment [13,14]. We trained Random Survival Forest (RSF) models, a machine learning method known for its robustness in handling high-dimensional and censored survival data, and assessed their ability to predict OS using this integrative dataset [7,15]. To evaluate how well our approach generalizes, we further tested our models on additional TCGA datasets from independent patient cohorts. Our findings demonstrate that integrative, multi-omic modeling substantially improves survival prediction in breast cancer patients compared to models based solely on TMB or clinical features. This approach not only advances the methodological toolkit for precision oncology but also provides new biological insights into the determinants of survival in breast cancer.

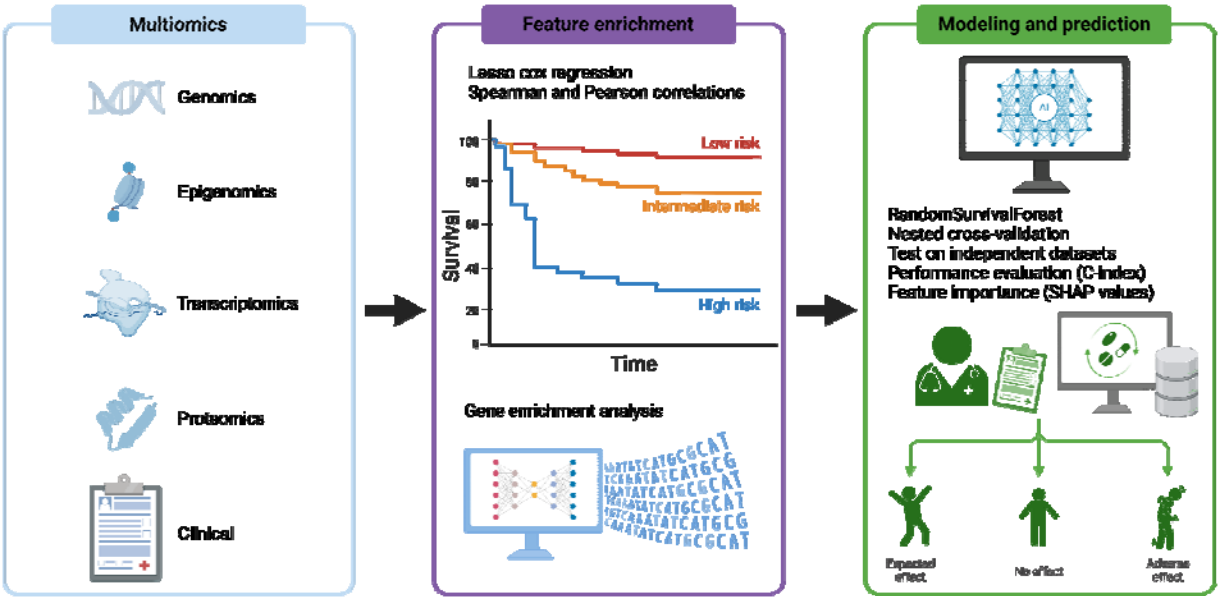


Figure 1) Proposed workflow to predict breast cancer patient survival using multi-omics. We analyzed multi-omic and clinical data from 802 female breast cancer patients in The Cancer Genome Atlas (TCGA), integrating genomic, transcriptomic, proteomic, epigenetic, and clinical features. Using Random Survival Forest models, we predicted overall survival based on features identified through correlation analyses, Lasso-Cox regression, and hazard ratio assessments. Model performance was evaluated via nested cross-validation and measured using the Concordance Index. Time-dependent receiver operating characteristic (ROC) curves and cumulative dynamic areas under the curve (AUC) were computed at multiple time points. External validation was performed using two independent TCGA cohorts (uterine (n=227), glioma (n=189) cancers). To interpret the model and identify key prognostic drivers, we computed Shapley values for each feature.

Results

A limited number of mutations and clinical features correlate with OS in breast cancer.

We hypothesized that the accumulation of mutations and a high TMB would not be associated with a better overall survival (OS) in breast cancer. When plotting Pearson and Spearman correlations between mutational features and OS in months, non synonymous TMB significantly negatively correlates with a longer OS, and no correlation was observed after applying the Bonferroni correction (**Figure 2 A**). Moreover, very few mutations (5 among 24496 mutations) are significantly linked to a difference in OS in months in breast cancer, and no correlation was observed after applying the Bonferroni correction (**Figure 2 B**). Of note, no mutations were associated with a shorter survival, and most of the ones significantly potentially associated with a longer survival are fusion mutations (**Figure 2 B**). Asian genetic ancestry correlates with a shorter OS after applying the Bonferroni correction, while the use of doxorubicin (blocking topoisomerase 2), paclitaxel (stopping mitosis by interfering with microtubules) or cyclophosphamide (DNA replication inhibitor) significantly correlates with a longer survival in this cohort (**Figure 2 A**). When calculating the hazard ratios and performing cox regressions, no mutations were significantly associated with a higher risk of dying in this cohort. Overall, a limited number of clinical features and no mutational features significantly correlate with OS in breast cancer.

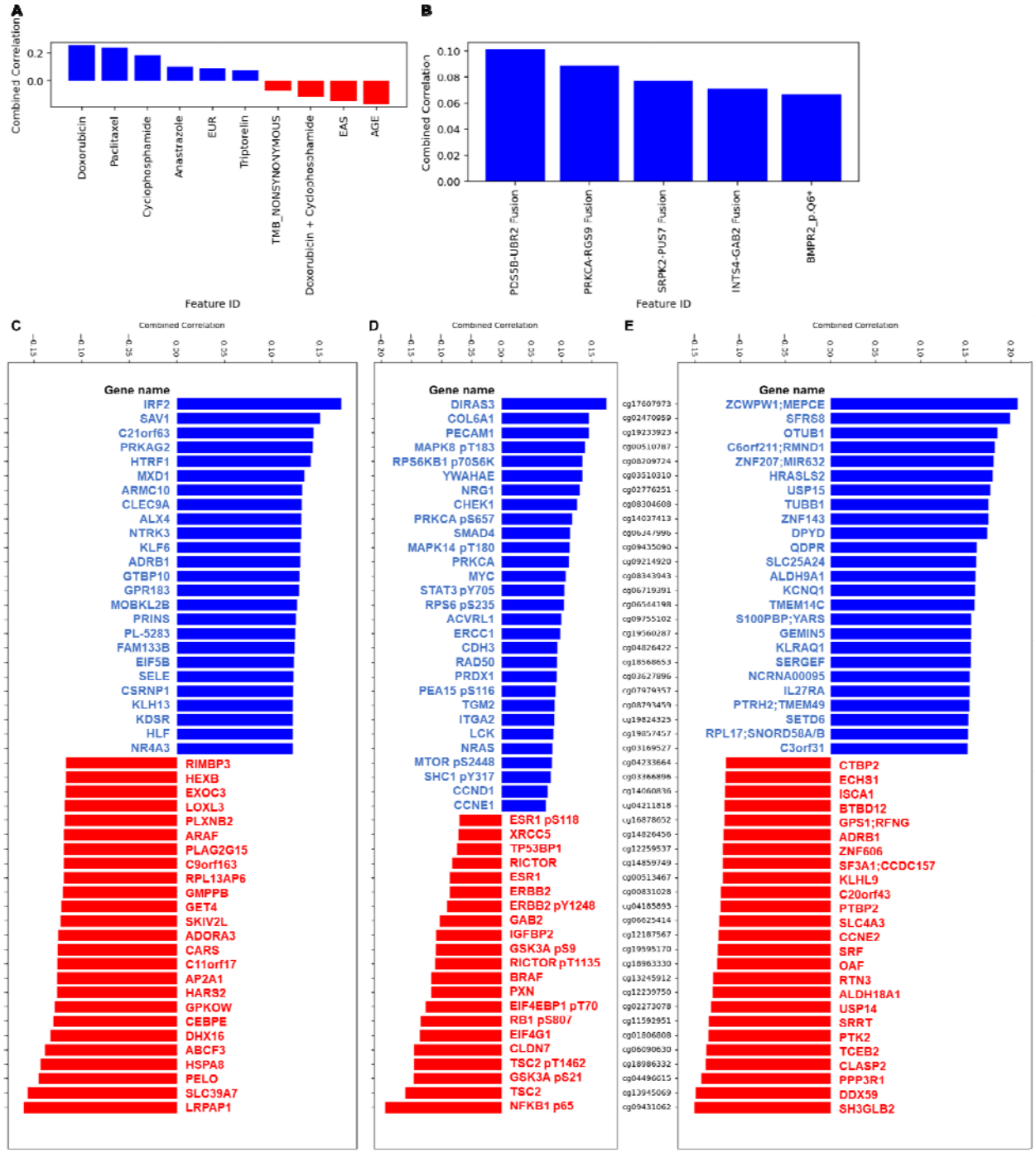


Figure 2) Breast cancer patient multi-omics and clinical features associated with a difference in overall survival.

A. Breast cancer patient clinical features significantly associated with a longer or shorter overall survival (for both Pearson and Spearman correlations). Features include clinical variables, tumor mutational burden (TMB), age, genetic ancestry (EUR = europe, EAS = asia), final overall survival status (living, deceased), drug used (doxorubicin, paclitaxel, cyclophosphamide, triptorelin). The only variables plotted are the ones associated with a difference in survival using

both Pearson and Spearman correlations. **B**. Breast cancer patient mutational features significantly associated with a difference in overall survival (for both Pearson and Spearman correlations). N=802 breast cancer patients. The only variables plotted are the ones associated with a difference in survival using both Pearson and Spearman correlations. **C**. Breast cancer patient mRNA transcripts (RSEM, batch normalized from Illumina HiSeq_RNASeqV2) with expression significantly associated with a longer or shorter overall survival for both Pearson and Spearman correlations. N=802 breast cancer patients. Highest and lowest 25 correlations (average between Spearman and Pearson) were plotted. The only variables plotted are the ones associated with a difference in survival using both Pearson and Spearman correlations. **D**. Breast cancer patient protein expression, measured by reverse-phase protein array (RPPA), significantly associated with a longer or shorter overall survival for both Pearson and Spearman correlations. N=802 breast cancer female patients. Highest and lowest correlations (average between Spearman and Pearson) were plotted. The only variables plotted are the ones associated with a difference in survival using both Pearson and Spearman correlations. **E**. Breast cancer patient methylations between-platform (hm27 and hm450) normalization values significantly associated with a longer or shorter overall survival for both Pearson and Spearman correlations. N=802 breast cancer patients. Highest and lowest 25 correlations (average between Spearman and Pearson) were plotted. The only variables plotted are the ones associated with a difference in survival using both Pearson and Spearman correlations. Gene enrichment and cell type signature analysis with these top/bottom genes was performed using metascape in (**Supplementary Figure 1/2**).

An integrated genomic, transcriptomic, epigenomic, and proteomic signature predicts OS in breast cancer.

We hypothesized that other multi-omic features may be associated with differences in OS and allow a better prediction in breast cancer. Importantly, we found that mRNA expression (**Figure 2 C**), protein levels (**Figure 2 D**) and DNA methylation levels (**Figure 2 E**) were significantly associated with differences in OS in breast cancer. Moreover, many of these features have a coefficient of correlation with OS between 0.1 and 0.2, while the mutation ones are all below 0.1, suggesting that mRNA, protein and DNA methylation levels are stronger predictors of OS than mutations.

In more details, we found that the mRNA expression level of IRF2, SAV1, C21orf63, PRKAG2, HTRF1, MXD1, ARMC10, CLEC9A, ALX4, NTRK3, KLF6, ADRB1, GTBP10, GPR183, MOBKL2B, PRINS, PL-5283, FAM133B, EIF5B, SELE, CSRNP1, KLH13, KDSR, HLF and NR4A3 genes significantly correlates with a longer OS (**Figure 2 C**). By applying gene set enrichment and cell type signature analysis with metascape [16], we found that this mRNA signature is close to the cytokine production (GO:0001819) and platelet-derived growth factor receptor signaling (GO:0048008) pathways, and suggests a tumor infiltration by myeloid cells

(**Supplementary Figure 2)**. On the other hand, the mRNA expression level of RIMBP3, HEXB, EXOC3, LOXL3, PLXNB2, ARAF, PLAG2G15, C9orf163, RPL13AP6, GMPPB, GET4, SKIV2L, ADORA3, CARS, C11orf17, AP2A1, HARS2, GPKOW, CEBPE, DHX16, ABCF3, HSPA8, PELO, SLC39A7 and LRPAP1 genes significantly correlates with a shorter OS (**Figure 2 C**). This signature is close to mRNA metabolomic processes (GO:0016071) and translation regulation (GO:0006412) signatures (**Supplementary Figure 1)**. Overall, longer OS significantly correlated with the upregulation of myeloid, cytokine and platelet immune mRNA signatures, while shorter OS significantly correlated with the upregulation of mRNA metabolomic and translation regulation signatures.

Next, using the reverse-phase protein array (RPPA) assay, we found that the protein expression levels of DIRAS3, COL6A1, PECAM1, MAPK8, RPS6KB1, YWAHAE, NRG1, CHEK1, PRKCA, SMAD4, MAPK14, MYC, STAT3, RPS6, ACVRL1, ERCC1, CDH3, RAD50, PRDX1, PEA15, TGM2, ITGA2, LCK, NRAS, MTOR, SHC1, CCND1 and CCNE1 significantly correlate with a longer OS (**Figure 2 D**). By applying gene set enrichment and cell type signature analysis with metascape, we found that this proteomic signature is close to the IL-2, and type 1 IFN pathways and suggests a tumor infiltration by naive T cells (**Supplementary Figure 2)**. On the other hand, the protein expression levels of ESR1, XRCC5, TP53BP1, RICTOR, ESR1, ERBB2, ERBB2, GAB2, IGFBP2, GSK3A, RICTOR, BRAF, PXN, EIF4EBP1, RB1, EIF4G1, CLDN7, TSC2, GSK3A, TSC2 and NFKB1 significantly correlate with a shorter OS (**Figure 2 D**). This signature is close to the leptin and multiple cancer immunosuppressive signatures (**Supplementary Figure 1)**. Overall, longer OS significantly correlated with the upregulation of IL-2 and type 1IFN pathway proteomic signatures, while shorter OS significantly correlated with the upregulation of leptin and multiple cancer immunosuppressive proteomic signatures.

Finally, we found that the DNA methylation level of ZCWPW1, MEPCE, SFRS8, OTUB1, C6orf211, RMND1, ZNF207, MIR632, HRASLS2, USP15, TUBB1, ZNF143, DPYD, QDPR, SLC25A24, ALDH9A1, KCNQ1, TMEM14C, S100PBP, YARS, GEMIN5, KLRAQ1, SERGEF, NCRNA00095, IL27RA, PTRH2, TMEM49, SETD6, RPL17, SNORD58A, SNORD58B and C3orf31 genes significantly correlates with a longer OS (**Figure 1 E**). On the other hand, the DNA methylation level of CTBP2, ECHS1, ISCA1, BTBD12, GPS1, RFNG, ADRB1, ZNF606, SF3A1, CCDC157, KLHL9, C20orf43, PTBP2, SLC4A3, CCNE2, SRF, OAF, RTN3, ALDH18A1, USP14, SRRT, PTK2, TCEB2, CLASP2, PPP3R1, DDX59 and SH3GLB2 genes significantly correlates with a shorter OS (**Figure 2 E**). Thus, differences in OS are significantly associated with multiple DNA methylation signatures.

Treatment response is linked to innate and adaptive immune activation, while resistance correlates with immunosuppressive leptin signatures.

To better characterize the pathways associated with a difference in survival, we performed a fully integrated gene enrichment and cell type signature analysis with these genomic, transcriptomic and proteomic signatures combined using metascape (**Figure 3 A**). From an immunological perspective, we observed that a longer OS was associated with the presence of naive T cells and myeloid cells (monocytes and dendritic cells), IL-2 and IFN type 1 signaling pathways, while a shorter OS was associated with TREM2 immunosuppressive myeloid cell [17,18] and B cell signatures, and the leptin pathway known to be key in immune regulation [19] (**Figure 3 A**).

To investigate it further, we looked at the direct protein expression using mass spectrometry and not RPPA as we did before, and their level of phosphorylation in the tumors of the 56 patients that were profiled in the cohort (**Figure 3 B/C**). We confirmed that longer survival was associated with antigen presentation to T cells (HLA-DPB1 expression), T cell infiltration (CD3E expression) and NK cell infiltration (HLA-E and CD2 expression) (**Figure 3 B/C**). Moreover, MAVS phosphorylation and downstream Interferon Stimulated Genes (ISGS) pathway (MAVS, TRAF3IP3, IFIT3, IFIT5, PARP14, FADD, CAMK2D, RFX5, GBP4, PSMB8 and BST2) activation significantly correlate with survival in breast cancer. The MAVS pathway, essential for the antiviral immune response, functions as a central hub that connects virus recognition to innate immunity [20]. TRAF3IP3 is a protein that interacts with TRAF3 and is involved in downstream MAVS signaling, mediating the recruitment of TRAF3 to MAVS for effective antiviral responses [21]. IFIT5 plays a significant role in recognizing viral RNA and participating in downstream signaling that enhances antiviral mechanisms [22]. Finally, PARP14 regulates immune signaling pathways, including those mediated by MAVS, influencing antiviral responses [23,24]. Overall, response to therapy significantly correlated with the activation of both innate and adaptive immunity, while resistance significantly correlated with TREM2+ myeloid, B cell and immunosuppressive leptin signatures.

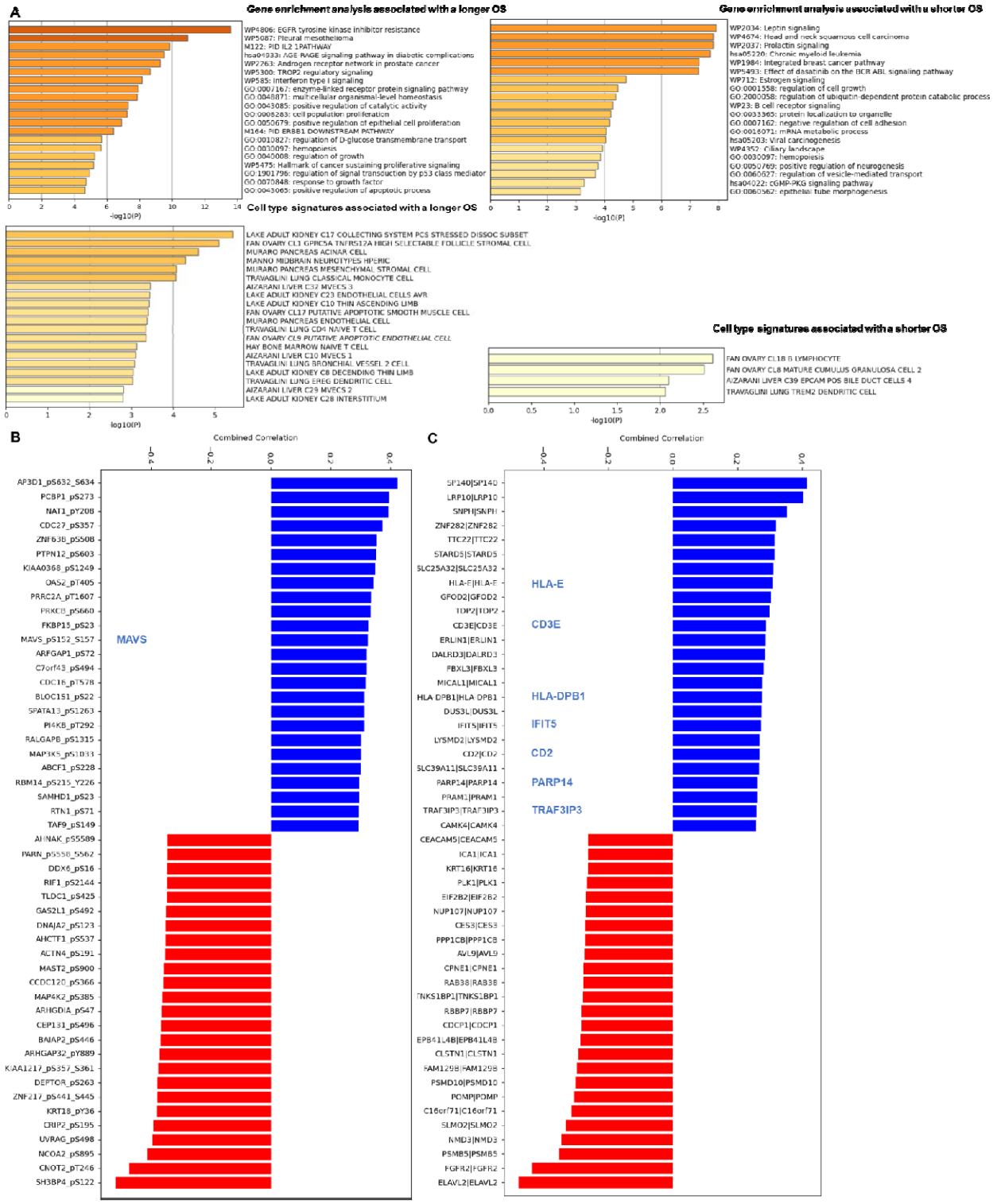


Figure 3) Treatment response is linked to innate and adaptive immune activation, while resistance correlates with immunosuppressive leptin signatures.

A. Gene enrichment and cell type signature analysis significantly associated with a difference in survival. Gene enrichment and cell type signature analysis with these top/bottom genes (from the RNA dataset of Figure 1 C, from the methylation dataset of Figure 1 D and from the protein

dataset of Figure 1 E) was performed using metascape [16]. **B**. Breast cancer patient protein expression by mass spectometry and their phosphorylation levels significantly associated with a difference in survival. **B/C**. Breast cancer patient protein expression (**B**) and their level of phosphorylation (**C**), directly measured on 56 patients, significantly associated with a longer or shorter overall survival for both Pearson and Spearman correlations. N=56 female breast cancer patients. Highest and lowest correlations (average between Spearman and Pearson) were plotted. The only variables plotted are the ones associated with a difference in survival using both Pearson and Spearman correlations.

Epigenetic profile drives accurate survival prediction in breast cancer via a multi-omics machine learning model.

We hypothesized that we may leverage machine learning algorithms trained on the multi-omics features that we identified in **Figure 2** as associated with a higher or lower risk to better predict cancer patient outcome. First, we investigated more deeply which multi-omics breast cancer patient features are significantly associated with a difference in survival in this cohort of 802 patients using a Lasso cox regression. This test takes into account not only the OS time component but also the event/status of the patient at the end point (deceased/living). Lower risks were observed in presence of higher ESRRG, IRF2 and MAP2K6 RNA levels, and cg04515986 (FTHL17 gene), cg06531741 (RIN2 gene) and cg17016000 (HTR3B gene) methylation levels, and the use of Cyclophosphamide (**Figure 4 A**). On the contrary, increased risks were observed for older patients, and patients with higher PELO and MAFA RNA levels, phosphorylated NFKB1 protein levels and cg09645888 (ME3 gene), cg03889226 (OLIG3 gene), cg04632671 (PPARG gene) and cg10970251 (SLC25A22 gene) methylation levels (**Figure 4 A**).

We used a Random Survival Forest model trained on these features (**Table 1**) with nested cross-validation and parameter optimization to predict the OS status (0 for living, 1 for deceased) and OS in months. The outer loop consisted of a 5-fold cross-validation to estimate the generalization performance of the model, while the inner loop performed 5-fold cross-validation on the training data to select optimal hyperparameters via grid search. Model performance was evaluated using the concordance index (C-index) computed on held-out test folds in the outer loop. To obtain a final model for downstream application, we retrained the RSF model on the entire dataset using the hyperparameter set that achieved the best average cross-validation performance in the inner loop. We obtained a c-Index of 0.75 for the model trained on multi-omics data vs 0.53 for the one trained only on mutation data, meaning that the multi-omics model was more efficient to predict OS (**Figure 4 B**). It indicates that in 75% of cases, the model correctly predicts that a patient with a shorter predicted survival time indeed has a shorter observed survival time compared to another patient, and globally outperformed models trained on mutation profile.

As an example, we plotted the prediction of OS and the cumulative risk (estimating the probability of experiencing the event (e.g., death) by a certain time) based on ESRRG, IRF2 and MAP2K6 RNA levels, and cg04515986, cg06531741 and cg17016000 methylation levels, and the use of Cyclophosphamide for several patients (**Figure 4 C**). On each graph, patients 0, 1 and 2 have the lowest levels of the feature of interest, while patients 3, 4 and 5 have the highest levels. It shows that, as expected, patients with lower levels of these features have an increased predicted risk compared to patients with higher levels of these features.

Then, we tested this multi-omics model on 2 other independent datasets with different cancer types, and obtained a c-Index of 0.57 for 227 uterine cancer (UEC) patients and a c-Index of 0.62 for 189 lower grade glioma (LGG) patients (**Figure 4 D**). Thus, this model may be used in different cancer settings but may not be as accurate as it is for breast cancer patients.

We further investigated which features drive the prediction of survival in our model by calculating the Shapley values (**Figure 4 E**). We observed that the features with the highest importance to drive the model performance are the epigenetic ones. In more details cg06531741 (HTR3B), cg17016000 (RIN2), cg09645888 (ME3) and cg03889226 (OLIG3) methylations predict overall survival by our machine learning model (**Figure 4 E**). Altogether, training machine-learning models on clinical and multi-dimensional genomic features could better predict overall survival, and DNA methylations lead these predictions.

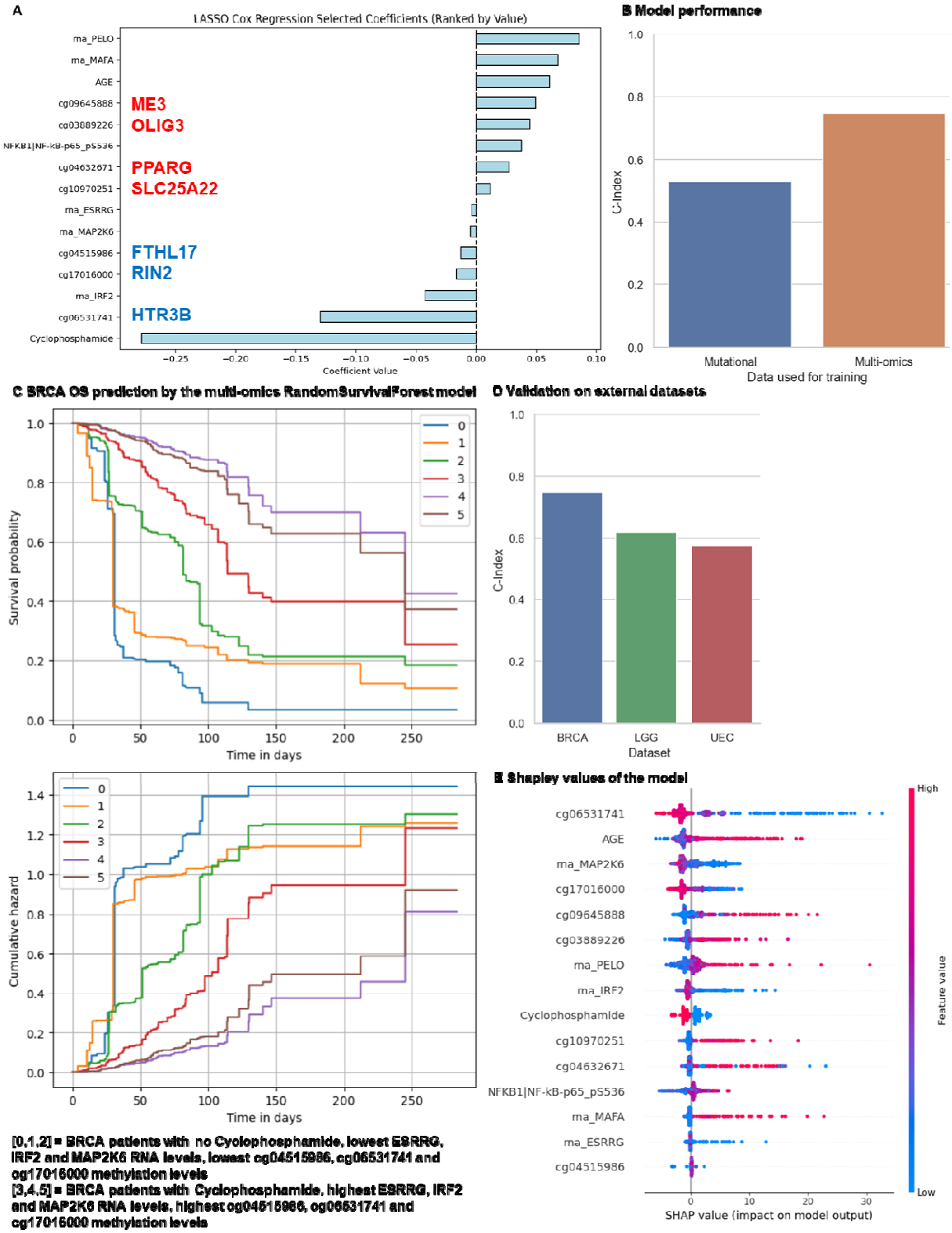


Figure 4) Epigenetic profile drives accurate survival prediction in breast cancer via a multi-omics machine learning model.

A. Cancer patient features associated with an increased or decreased hazard ratio (Lasso cox regression). N=802 breast cancer female patients. **B**. Prediction of overall survival of cancer patients based on their methylation, RNA, protein and clinical features identified in panel 3 A. A RandomSurvivalForest model was used with nested cross validation. Survival was predicted using the Kaplan-Meier estimator. Model performance (C-Index) according to the dataset used for training (multi-omics vs mutational data). **C**. Prediction of OS and the cumulative risk (estimating the probability of experiencing the event (e.g., death) by a certain time) based on ESRRG, IRF2 and MAP2K6 RNA levels, and cg04515986, cg06531741 and cg17016000 methylation levels, and the use of Cyclophosphamide. **D**. The model was tested on 3 other independent datasets (227 uterine cancer (UEC), 189 lower grade glioma (LGG) female patients). **E**. Features predicting response by the model (Shapley values).

To evaluate the prognostic performance of our models, we generated time-dependent receiver operating characteristic (ROC) curves and computed the cumulative dynamic area under the curve (AUC) at multiple time points (**Figure 5**). The time-dependent AUC reflects the model's ability to discriminate between individuals who experience an event (e.g., death) by a given time versus those who do not, accounting for censoring. An AUC close to 1 indicates strong predictive performance, while 0.5 corresponds to random chance. Additionally, we plotted the Kaplan–Meier survival curve based on observed outcomes in the test cohorts, alongside the average predicted survival curve derived from the model. This comparison provides a visual assessment of how closely the model's predicted survival aligns with actual patient outcomes over time. In accordance with the c-Index results, the multi-omics model outperformed the mutational model for test BRCA data, with time-dependent AUC between 0.9 and 1 (**Figure 5 A**) vs 0.4 and 0.9 (**Figure 5 B**). Overall, our multi-omics model outperformed the mutational model performances to predict OS.

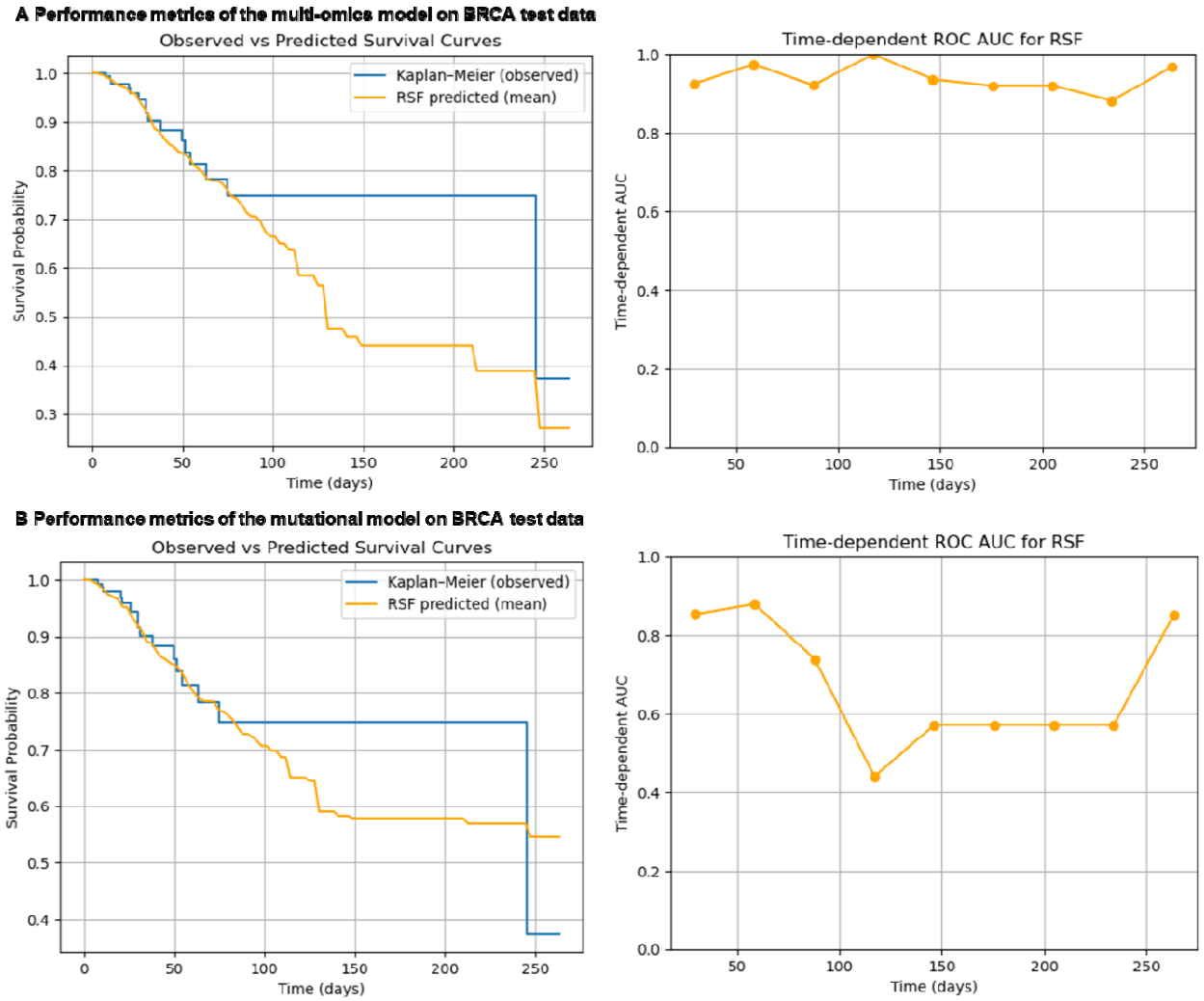


Figure 5) Performance metrics of the multi-omics and mutational models on the test data. **A**. Performance metrics of the multi-omics model on BRCA test data. **B**. Performance metrics of the mutational model on BRCA test data. Time-dependent receiver operating characteristic (ROC) curves and cumulative dynamic areas under the curve (AUC) were computed at multiple time points. Kaplan–Meier survival curves based on observed outcomes in the test cohorts were also plotted alongside the average predicted survival curves derived from the models.

Discussion

Accurate prediction of overall survival (OS) in breast cancer patients is crucial for personalized treatment strategies. However, mutation profiling and a high TMB do not predict a better OS in all cancers. To better predict breast cancer patient outcomes, we considered approaches integrating clinical features with multi-dimensional genomic (mutations), epigenetic (DNA methylations) and proteomic features from patient tumors.

We observed that lower risks were observed in presence of higher ESRRG, IRF2 and MAP2K6 RNA levels. In several cancers, Interferon regulatory factor 2 (IRF2) downregulation was associated with worst outcomes by limiting MHC class I antigen presentation and increased PD-L1 expression [25]. However, in esophageal cancers, IRF2 promotes resistance to IFN-II and immune escape by inhibiting transcription of the IFNγR [26]. Similarly, estrogen-related receptor gamma (ESRRG) expression was associated with better outcomes in several cancers such as gastric cancer (GC) [27], but not in esophageal squamous cell carcinoma (ESCC) [28]. This gene encodes a member of the estrogen receptor-related receptor (ESRR) family. Finally, mitogen-activated protein kinase kinase 6 (MAP2K6) expression is downregulated in non-responder colorectal cancer patients of oxaliplatin therapy [29]. MAP2K6 phosphorylates and activates p38 MAP kinase in response to inflammatory cytokines or environmental stress, and regulates stress induced cell cycle arrest, transcription activation and apoptosis. Lower risks were also observed in presence of higher cg04515986 (FTHL17 gene), cg06531741 (RIN2 gene) and cg17016000 (HTR3B gene) methylation levels. In breast cancer, the expression of 5-hydroxytryptamine receptor (HTR2A/2B) was positively and significantly correlated with the infiltration of immune cells such as CD8+ T cells and macrophages [30]. It was also shown that NAA20 promotes the growth and invasion capability of triple-negative breast cancer cells by recruiting Ras and Rab interactor 2 (RIN2) to inhibit Rab5A-mediated activation of EGFR signaling [31]. Moreover, ferritin heavy polypeptide-like 17 (FTHL17) gene is a cancer/testis antigen gene that is overexpressed in cancer cells and in testis [32]. Thus, epigenetic regulation of these genes may disrupt the underlying immune escape mechanisms, and response to therapy significantly correlated with the activation of both innate and adaptive immunity.

On the other hand, increased risks were observed for older patients, and patients with higher PELO and MAFA RNA levels, as well as for patients with elevated phosphorylation of NFKB1. Interestingly, PELO was recently identified as a promising drug target for MSI-H patients with deleterious TTC37 mutations or with biallelic 9p21.3 deletions involving FOCAD [33]. MAFA is a MAF transcription factor and acts as a transactivator of insulin and several genes involved in glucose-stimulated insulin secretion [34]. MAFA is up-regulated in 50% of human multiple myelomas and 60% of angioimmunoblastic T cell lymphomas [35]. In breast cancer, NFKB1, which encodes the p50 subunit of the NF-κB protein complex, is often dysregulated and phosphorylated, leading to the activation of the NF-κB signaling pathway [36]. This activation is frequently associated with tumor growth, metastasis, and treatment resistance. Higher risks were also observed in patients with higher cg09645888 (ME3 gene), cg03889226 (OLIG3 gene), cg04632671 (PPARG gene) and cg10970251 (SLC25A22 gene) methylation levels. It was shown that targeting Malic Enzyme 3 (ME3) may be a strategy to treat pancreatic ductal adenocarcinoma (PDAC), as overexpression of ME3 promotes pancreatic tumor proliferation, invasiveness and metastasis [37]. ME3 encodes a mitochondrial NADP(+)-dependent enzyme that catalyzes the oxidative decarboxylation of malate to pyruvate, a crucial step in energy production and biosynthesis. OLIG3 is a transcription factor that regulates early cerebellar development, cell

proliferation and differentiation [38], and other studies have established functional roles of OLIG1 and OLIG2 in brain cancer [39]. Peroxisome proliferator-activated receptor gamma (PPARG) is a ligand-dependent transcription factor expressed that regulates tumor development, progression, and metastasis in breast cancer [40]. It was shown that epigenetic derepression converts PPARG into a druggable target in triple-negative and endocrine-resistant breast cancers [41]. It was also shown that targeting of SLC25A22, a mitochondrial glutamate carrier, boosts the immunotherapeutic response in KRAS-mutant colorectal cancer by suppressing CXCL1 production and impairing MDSC recruitment [42]. Thus, targeting these molecules in high risk patients may improve their OS.

Altogether, response to therapy significantly correlated with the activation of both innate and adaptive immunity in the tumor, while resistance significantly correlated with TREM2+ myeloid, B cell and immunosuppressive leptin signatures. Leveraging on this multi-omics signature, we developed BANDOL, a RandomSurvivalForest model that estimates OS of 802 breast cancer patients following training on these clinical and multi-omics data associated with a difference in OS. In 75% of cases, BANDOL correctly predicts that a patient with a shorter predicted survival time indeed has a shorter observed survival time compared to another patient. This model was also effective in other cancer types, but with a lower predictive capacity probably due to specific characteristics for each cancer type. Thus, BANDOL can compute a robust 'survival probability' or 'risk' for each patient at a given time after it enters treatment. Overall, integrating not only genomic, but also transcriptomic, proteomic, and most importantly epigenetic profiles using machine learning is promising to better predict cancer patient OS. We showed that epigenetic features are the most important ones to drive the prediction by our model. Current therapies may be combined with strategies to target ME3, PPARG, OLIG3 and SLC25A22 gene methylations, as well as strategies to dephosphorylate NFKB1 and target MAFA and PELO.

References

1. Reck, M., Rodríguez–Abreu, D., Robinson, A.G., Hui, R., Csőszi, T., Fülöp, A., Gottfried, M., Peled, N., Tafreshi, A., Cuffe, S., et al. (2019). Updated Analysis of KEYNOTE-024: Pembrolizumab Versus Platinum-Based Chemotherapy for Advanced Non–Small-Cell Lung Cancer With PD-L1 Tumor Proportion Score of 50% or Greater. J. Clin. Oncol. *37*, 537–546. https://doi.org/10.1200/JCO.18.00149.

2. Silva, I.P. da, Ahmed, T., Reijers, I.L.M., Weppler, A.M., Warner, A.B., Patrinely, J.R., Serra-Bellver, P., Allayous, C., Mangana, J., Nguyen, K., et al. (2021). Ipilimumab alone or ipilimumab plus anti-PD-1 therapy in patients with metastatic melanoma resistant to anti-PD-(L)1 monotherapy: a multicentre, retrospective, cohort study. Lancet Oncol. *22*, 836–847. https://doi.org/10.1016/S1470-2045(21)00097-8.

3. Dantoing, E., Piton, N., Salaün, M., Thiberville, L., and Guisier, F. (2021). Anti-PD1/PD-L1 Immunotherapy for Non-Small Cell Lung Cancer with Actionable Oncogenic Driver Mutations. Int. J. Mol. Sci. *22*, 6288. https://doi.org/10.3390/ijms22126288.

4. Mestrallet, G., Brown, M., Bozkus, C.C., and Bhardwaj, N. (2023). Immune escape and resistance to immunotherapy in mismatch repair deficient tumors. Front. Immunol. *14*.

5. Jee, J., Fong, C., Pichotta, K., Tran, T.N., Luthra, A., Waters, M., Fu, C., Altoe, M., Liu, S.-Y., Maron, S.B., et al. (2024). Automated real-world data integration improves cancer outcome prediction. Nature, 1–9. https://doi.org/10.1038/s41586-024-08167-5.

6. Mestrallet, G. (2024). Leveraging Tumor Mutation Profiles to Forecast Immune Checkpoint Blockade Resistance in Melanoma, Lung, Head and Neck, Bladder and Renal Cancers. Onco *4*, 439–457. https://doi.org/10.3390/onco4040031.

7. Mestrallet, G. (2025). Fraction of Genome Altered, Age, Microsatellite Instability Score, Tumor Mutational Burden, Cancer Type, Metastasis Status, and Choice of Cancer Therapy Predict Overall Survival in Multiple Machine Learning Models. Onco *5*, 8. https://doi.org/10.3390/onco5010008.

8. McGrail, D.J., Pilié, P.G., Rashid, N.U., Voorwerk, L., Slagter, M., Kok, M., Jonasch, E., Khasraw, M., Heimberger, A.B., Lim, B., et al. (2021). High tumor mutation burden fails to predict immune checkpoint blockade response across all cancer types. Ann. Oncol. *32*, 661–672. https://doi.org/10.1016/j.annonc.2021.02.006.

9. Samstein, R.M., Lee, C.-H., Shoushtari, A.N., Hellmann, M.D., Shen, R., Janjigian, Y.Y., Barron, D.A., Zehir, A., Jordan, E.J., Omuro, A., et al. (2019). Tumor mutational load predicts survival after immunotherapy across multiple cancer types. Nat. Genet. *51*, 202–206. https://doi.org/10.1038/s41588-018-0312-8.

10. Tran, D., Nguyen, H., Pham, V.-D., Nguyen, P., Nguyen Luu, H., Minh Phan, L., Blair DeStefano, C., Jim Yeung, S.-C., and Nguyen, T. (2025). A comprehensive review of cancer survival prediction using multi-omics integration and clinical variables. Brief. Bioinform. *26*, bbaf150. https://doi.org/10.1093/bib/bbaf150.

11. Menyhárt, O., and Győrffy, B. (2021). Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. Comput. Struct. Biotechnol. J. *19*, 949–960. https://doi.org/10.1016/j.csbj.2021.01.009.

12. Bao, X., Anastasov, N., Wang, Y., and Rosemann, M. (2019). A novel epigenetic signature for overall survival prediction in patients with breast cancer. J. Transl. Med. *17*, 380. https://doi.org/10.1186/s12967-019-2126-6.

13. Ciriello, G., Gatza, M.L., Beck, A.H., Wilkerson, M.D., Rhie, S.K., Pastore, A., Zhang, H., McLellan, M., Yau, C., Kandoth, C., et al. (2015). Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. Cell *163*, 506–519. https://doi.org/10.1016/j.cell.2015.09.033.

14. Koboldt, D.C., Fulton, R.S., McLellan, M.D., Schmidt, H., Kalicki-Veizer, J., McMichael, J.F., Fulton, L.L., Dooling, D.J., Ding, L., Mardis, E.R., et al. (2012). Comprehensive molecular portraits of human breast tumours. Nature *490*, 61–70. https://doi.org/10.1038/nature11412.

15. Yoo, S.-K., Fitzgerald, C.W., Cho, B.A., Fitzgerald, B.G., Han, C., Koh, E.S., Pandey, A., Sfreddo, H., Crowley, F., Korostin, M.R., et al. (2025). Prediction of checkpoint inhibitor

immunotherapy efficacy for cancer using routine blood tests and clinical data. Nat. Med., 1–12. https://doi.org/10.1038/s41591-024-03398-5.

16. Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1523. https://doi.org/10.1038/s41467-019-09234-6.

17. Binnewies, M., Pollack, J.L., Rudolph, J., Dash, S., Abushawish, M., Lee, T., Jahchan, N.S., Canaday, P., Lu, E., Norng, M., et al. (2021). Targeting TREM2 on tumor-associated macrophages enhances immunotherapy. Cell Rep. *37*, 109844. https://doi.org/10.1016/j.celrep.2021.109844.

18. Mestrallet, G., Brown, M., Vaninov, N., Cho, N.W., Velazquez, L., Ananthanarayanan, A., Spitzer, M., Vabret, N., Cimen Bozkus, C., Samstein, R.M., et al. (2025). Coordinated macrophage and T cell interactions mediate response to checkpoint blockade in colorectal cancer. Preprint at Immunology, https://doi.org/10.1101/2025.02.12.637954 https://doi.org/10.1101/2025.02.12.637954.

19. Park, H.-K., and Ahima, R.S. (2014). Leptin signaling. F1000Prime Rep. *6*, 73. https://doi.org/10.12703/P6-73.

20. Ren, Z., Ding, T., Zuo, Z., Xu, Z., Deng, J., and Wei, Z. (2020). Regulation of MAVS Expression and Signaling Function in the Antiviral Innate Immune Response. Front. Immunol. *11*, 1030. https://doi.org/10.3389/fimmu.2020.01030.

21. Zhu, W., Li, J., Zhang, R., Cai, Y., Wang, C., Qi, S., Chen, S., Liang, X., Qi, N., and Hou, F. (2019). TRAF3IP3 mediates the recruitment of TRAF3 to MAVS for antiviral innate immunity. EMBO J. *38*, e102075. https://doi.org/10.15252/embj.2019102075.

22. Zhang, B., Liu, X., Chen, W., and Chen, L. (2013). IFIT5 potentiates anti-viral response through enhancing innate immune signaling pathways. Acta Biochim. Biophys. Sin. *45*, 867–874. https://doi.org/10.1093/abbs/gmt088.

23. Du, Q., Miao, Y., He, W., and Zheng, H. (2023). ADP-Ribosylation in Antiviral Innate Immune Response. Pathogens *12*, 303. https://doi.org/10.3390/pathogens12020303.

24. Parthasarathy, S., and Fehr, A.R. (2022). PARP14: A key ADP-ribosylating protein in host–virus interactions? PLoS Pathog. *18*, e1010535. https://doi.org/10.1371/journal.ppat.1010535.

25. Kriegsman, B.A., Vangala, P., Chen, B.J., Meraner, P., Brass, A.L., Garber, M., and Rock, K.L. (2019). Frequent Loss of IRF2 in Cancers Leads to Immune Evasion through Decreased MHC Class I Antigen Presentation and Increased PD-L1 Expression. J. Immunol. Baltim. Md 1950 *203*, 1999–2010. https://doi.org/10.4049/jimmunol.1900475.

26. Lukhele, S., Rabbo, D.A., Guo, M., Shen, J., Elsaesser, H.J., Quevedo, R., Carew, M., Gadalla, R., Snell, L.M., Mahesh, L., et al. (2022). The transcription factor IRF2 drives interferon-mediated CD8+ T cell exhaustion to restrict anti-tumor immunity. Immunity *55*, 2369-2385.e10. https://doi.org/10.1016/j.immuni.2022.10.020.

27. Kang, M.-H., Choi, H., Oshima, M., Cheong, J.-H., Kim, S., Lee, J.H., Park, Y.S., Choi, H.-S., Kweon, M.-N., Pack, C.-G., et al. (2018). Estrogen-related receptor gamma functions as a tumor suppressor in gastric cancer. Nat. Commun. *9*, 1920. https://doi.org/10.1038/s41467-018-04244-2.

28. Wang, T., Zhu, Y., Chen, L., Zhang, W., Qi, H., Shi, X., Zhong, M., Chen, H., and Li, Q. (2023). ESRRG-PKM2 axis reprograms metabolism to suppress esophageal squamous carcinoma progression and enhance anti-PD-1 therapy efficacy. J. Transl. Med. *21*, 605. https://doi.org/10.1186/s12967-023-04347-5.

29. Rasmussen, M.H., Lyskjær, I., Jersie-Christensen, R.R., Tarpgaard, L.S., Primdal-Bengtson, B., Nielsen, M.M., Pedersen, J.S., Hansen, T.P., Hansen, F., Olsen, J.V., et al. (2016). miR-625-3p regulates oxaliplatin resistance by targeting MAP2K6-p38 signalling in human colorectal adenocarcinoma cells. Nat. Commun. *7*, 12436. https://doi.org/10.1038/ncomms12436.

30. Zhan, D., Wang, X., Zheng, Y., Wang, S., Yang, B., Pan, B., Wang, N., and Wang, Z. (2023). Integrative dissection of 5-hydroxytryptamine receptors-related signature in the prognosis and immune microenvironment of breast cancer. Front. Oncol. *13*, 1147189. https://doi.org/10.3389/fonc.2023.1147189.

31. Qiao, L., Dong, C., Jia, W., and Ma, B. (2023). NAA20 recruits Rin2 and promotes triple-negative breast cancer progression by regulating Rab5A-mediated activation of EGFR signaling. Cell. Signal. *112*, 110922. https://doi.org/10.1016/j.cellsig.2023.110922.

32. Aoki, N., Mochizuki, K., and Matsui, Y. (2017). DNA methylation of the Fthl17 5'-upstream region regulates differential Fthl17 expression in lung cancer cells and germline stem cells. PLoS ONE *12*, e0172219. https://doi.org/10.1371/journal.pone.0172219.

33. Borck, P.C., Boyle, I., Jankovic, K., Bick, N., Foster, K., Lau, A.C., Parker-Burns, L.I., Lubicki, D.A., Li, T., Borah, A.A., et al. (2025). SKI complex loss renders 9p21.3-deleted or MSI-H cancers dependent on PELO. Nature *638*, 1104–1111. https://doi.org/10.1038/s41586-024-08509-3.

34. Fottner, C., Sollfrank, S., Ghiasi, M., Adenaeuer, A., Musholt, T., Schad, A., Miederer, M., Schadmand-Fischer, S., Weber, M.M., Lackner, K.J., et al. (2022). Second MAFA Variant Causing a Phosphorylation Defect in the Transactivation Domain and Familial Insulinomatosis. Cancers *14*, 1798. https://doi.org/10.3390/cancers14071798.

35. Iacovazzo, D., Flanagan, S.E., Walker, E., Quezado, R., de Sousa Barros, F.A., Caswell, R., Johnson, M.B., Wakeling, M., Brändle, M., Guo, M., et al. (2018). MAFA missense mutation causes familial insulinomatosis and diabetes mellitus. Proc. Natl. Acad. Sci. *115*, 1027–1032. https://doi.org/10.1073/pnas.1712262115.

36. Pavitra, E., Kancharla, J., Gupta, V.K., Prasad, K., Sung, J.Y., Kim, J., Tej, M.B., Choi, R., Lee, J.-H., Han, Y.-K., et al. (2023). The role of NF-κB in breast cancer initiation, growth, metastasis, and resistance to chemotherapy. Biomed. Pharmacother. *163*, 114822. https://doi.org/10.1016/j.biopha.2023.114822.

37. Grell, T.A.J., Mason, M., Thompson, A.A., Gómez-Tamayo, J.C., Riley, D., Wagner, M.V., Steele, R., Ortiz-Meoz, R.F., Wadia, J., Shaffer, P.L., et al. (2022). Integrative structural

and functional analysis of human malic enzyme 3: A potential therapeutic target for pancreatic cancer. Heliyon *8*, e12392. https://doi.org/10.1016/j.heliyon.2022.e12392.

38. Lowenstein, E.D., Rusanova, A., Stelzer, J., Hernaiz-Llorens, M., Schroer, A.E., Epifanova, E., Bladt, F., Isik, E.G., Buchert, S., Jia, S., et al. (2021). Olig3 regulates early cerebellar development. eLife *10*, e64684. https://doi.org/10.7554/eLife.64684.

39. Szu, J.I., Tsigelny, I.F., Wojcinski, A., and Kesari, S. (2023). Biological functions of the Olig gene family in brain cancer and therapeutic targeting. Front. Neurosci. *17*, 1129434. https://doi.org/10.3389/fnins.2023.1129434.

40. Augimeri, G., Giordano, C., Gelsomino, L., Plastina, P., Barone, I., Catalano, S., Andò, S., and Bonofiglio, D. (2020). The Role of PPARγ Ligands in Breast Cancer: From Basic Research to Clinical Studies. Cancers *12*, 2623. https://doi.org/10.3390/cancers12092623.

41. Loo, S.Y., Syn, N.L., Koh, A.P.-F., Teng, J.C.-F., Deivasigamani, A., Tan, T.Z., Thike, A.A., Vali, S., Kapoor, S., Wang, X., et al. (2021). Epigenetic derepression converts PPARγ into a druggable target in triple-negative and endocrine-resistant breast cancers. Cell Death Discov. *7*, 1–15. https://doi.org/10.1038/s41420-021-00635-5.

42. Zhou, Q., Peng, Y., Ji, F., Chen, H., Kang, W., Chan, L.-S., Gou, H., Lin, Y., Huang, P., Chen, D., et al. (2023). Targeting of SLC25A22 boosts the immunotherapeutic response in KRAS-mutant colorectal cancer. Nat. Commun. *14*, 4677. https://doi.org/10.1038/s41467-023-39571-6.

43. Jardillier, R., Koca, D., Chatelain, F., and Guyon, L. (2022). Prognosis of lasso-like penalized Cox models with tumor profiling improves prediction over clinical data alone and benefits from bi-dimensional pre-screening. BMC Cancer *22*, 1045. https://doi.org/10.1186/s12885-022-10117-1.

44. Barraclough, H., Simms, L., and Govindan, R. (2011). Biostatistics Primer: What a Clinician Ought to Know: Hazard Ratios. J. Thorac. Oncol. *6*, 978–982. https://doi.org/10.1097/JTO.0b013e31821b10ab.

45. Cawley, G.C., and Talbot, N.L.C. (2010). On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. J Mach Learn Res *11*, 2079–2107.

46. Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions. J. R. Stat. Soc. Ser. B Methodol. *36*, 111–147.

Methods

Patient cohorts

We analyzed data from The Cancer Genome Atlas (TCGA) Breast Invasive Carcinoma (BRCA) cohort, a publicly available, large-scale dataset that provides multi-omic and clinical information for breast cancer patients. This cohort comprises over 1,000 primary breast cancer samples from female patients and is uniquely suited for integrative analyses aimed at uncovering molecular determinants of clinical outcomes.

The TCGA BRCA dataset includes several molecular data types:

- Genomic data: Whole-exome sequencing (WES) data were used to identify somatic single nucleotide variants (SNVs), small insertions and deletions (indels), and copy number alterations (CNAs), allowing for comprehensive mutational profiling of tumors.

- Transcriptomic data: RNA sequencing (RNA-seq) and microarray data were used to quantify gene expression levels across the transcriptome.

- Epigenomic data: Genome-wide DNA methylation profiles were obtained using Illumina Infinium HumanMethylation arrays, providing insight into regulatory changes associated with gene expression and silencing.

- Proteomic data: Reverse-phase protein array (RPPA) data measured relative protein abundance and post-translational modifications across signaling and metabolic pathways. Mass spectrometry was also performed on 56 patients.

- Clinical data: Curated clinical annotations include tumor stage, histologic grade, hormone receptor status (estrogen receptor [ER], progesterone receptor [PR]), HER2 status, treatment regimens, and patient outcomes, including overall survival (OS), disease-free survival (DFS), and response to therapy. OS is the length of time from either the date of diagnosis or the start of treatment for a disease, such as cancer, that patients diagnosed with the disease are still alive.

To test the generalizability of our model and evaluate its robustness across cancer types, we included two independent validation cohorts also obtained from TCGA (227 uterine cancer (UEC), 189 lower grade glioma (LGG) patients).

All patients included in this study were biologically female, and informed consent and ethical approvals were handled by the original TCGA consortium.

Statistical analyses

To identify tumor and clinical features associated with patient survival, we applied multiple statistical approaches. First, Pearson and Spearman correlation analyses were used to examine linear and rank-based associations, respectively, between individual features and overall survival time. This allowed us to assess both non-linear (Spearman) and linear (Pearson) relationships between continuous molecular measurements (e.g., gene expression, methylation levels, protein abundance) and survival duration. Bonferroni correction was also applied for multiple comparisons.

Next, we performed Lasso penalized Cox proportional hazards regression to identify the most predictive features while controlling for multicollinearity and reducing overfitting in high-dimensional datasets [43]. This is a statistical method that combines the Cox proportional hazards model with Lasso regularization, enabling variable selection and shrinkage in survival analysis. This approach is particularly useful when dealing with high-dimensional data, where the number of potential predictor variables exceeds the number of observations. Lasso regularization was used to shrink less informative feature coefficients toward zero, thereby enabling robust feature selection across multiple data modalities.

We also calculated hazard ratios (HRs) and corresponding 95% confidence intervals for individual features to estimate their relative impact on the risk of death. A HR measures how often a particular event happens in one group compared to how often it happens in another group, over time [44]. A hazard ratio of one means that there is no difference in survival between the two groups. A hazard ratio of greater than one or less than one means that survival was better in one of the groups. These analyses included a range of data types—clinical variables (e.g., age, tumor stage), mutational features (e.g., TMB, driver mutations), transcriptomic and proteomic levels, and epigenetic marks (e.g., DNA methylation levels). Statistical analyses were performed using Python libraries (scipy, statsmodels, lifelines and sksurv). The final curated dataset included N = 802 breast cancer patients with complete clinical and molecular data.

Machine-learning models for survival prediction

To develop predictive models of overall survival (OS) in breast cancer, we implemented machine learning workflows using features identified in the previous statistical analyses (**Table 1**). Our goal was to construct models that could capture complex, nonlinear relationships between multi-omic tumor features and patient outcomes.

We employed the Random Survival Forest (RSF) algorithm [15], an ensemble tree-based method well-suited for high-dimensional survival data with right-censoring. RSF extends the standard random forest algorithm to survival analysis by growing survival trees using log-rank splitting criteria. This method does not assume proportional hazards and is robust to multicollinearity and missing data.

The input features (**Table 1**) for the model included:

- Clinical variables: age at diagnosis, treatment information
- Mutational features: TMB, specific driver gene mutations
- Transcriptomic features: normalized RNA expression levels
- Proteomic features: RPPA-based protein levels
- Epigenetic features: normalized methylation beta-values for selected CpG sites

To assess the predictive performance of clinical and molecular features on overall survival, we implemented a nested cross-validation approach using the RSF model [45,46]. The outer loop consisted of a 5-fold cross-validation to estimate the generalization performance of the model, while the inner loop performed 5-fold cross-validation on the training data to select optimal hyperparameters via grid search. Model performance was evaluated using the concordance index (C-index) [15] computed on held-out test folds in the outer loop, a standard metric for survival prediction that evaluates the proportion of correctly ranked survival times. A C-index of 0.5 corresponds to random prediction, while a value of 1.0 indicates perfect prediction. To obtain a final model for downstream application, we retrained the RSF model on the entire dataset using the hyperparameter set that achieved the best average cross-validation performance in the inner loop. All analyses were performed using Python (scikit-learn and scikit-survival packages), with joblib for parallelization. Then we tested the final model on two external cohorts (UEC, LGG).

To interpret the model and identify key molecular and clinical drivers of survival, we computed Shapley Additive Explanation (SHAP) values [7,15], which quantify the contribution of each feature to the model's prediction for each individual. SHAP values offer a unified approach to feature attribution and facilitate biological interpretation by ranking features based on their average impact on model outputs.

To evaluate the prognostic performance of our Random Survival Forest (RSF) model, we generated time-dependent receiver operating characteristic (ROC) curves and computed the cumulative dynamic area under the curve (AUC) at multiple time points (at 10% of the time scale for each cohort, then 20%, 30%, 40%, 50%, 60%, 70%, 80% and 90%). The time-dependent AUC reflects the model’s ability to discriminate between individuals who experience an event (e.g., death) by a given time versus those who do not, accounting for censoring. An AUC close to 1 indicates strong predictive performance, while 0.5 corresponds to random chance. Additionally, we plotted the Kaplan–Meier survival curve based on observed outcomes in the test cohort, alongside the average predicted survival curve derived from the RSF model. This comparison provides a visual assessment of how closely the model’s predicted survival aligns with actual patient outcomes over time.

Acknowledgements

The authors thank Yan Souza and Diego Chowell for helpful discussions.

Data Availability

All data are openly available on the TCGA website.

https://portal.gdc.cancer.gov/analysis_page?app=

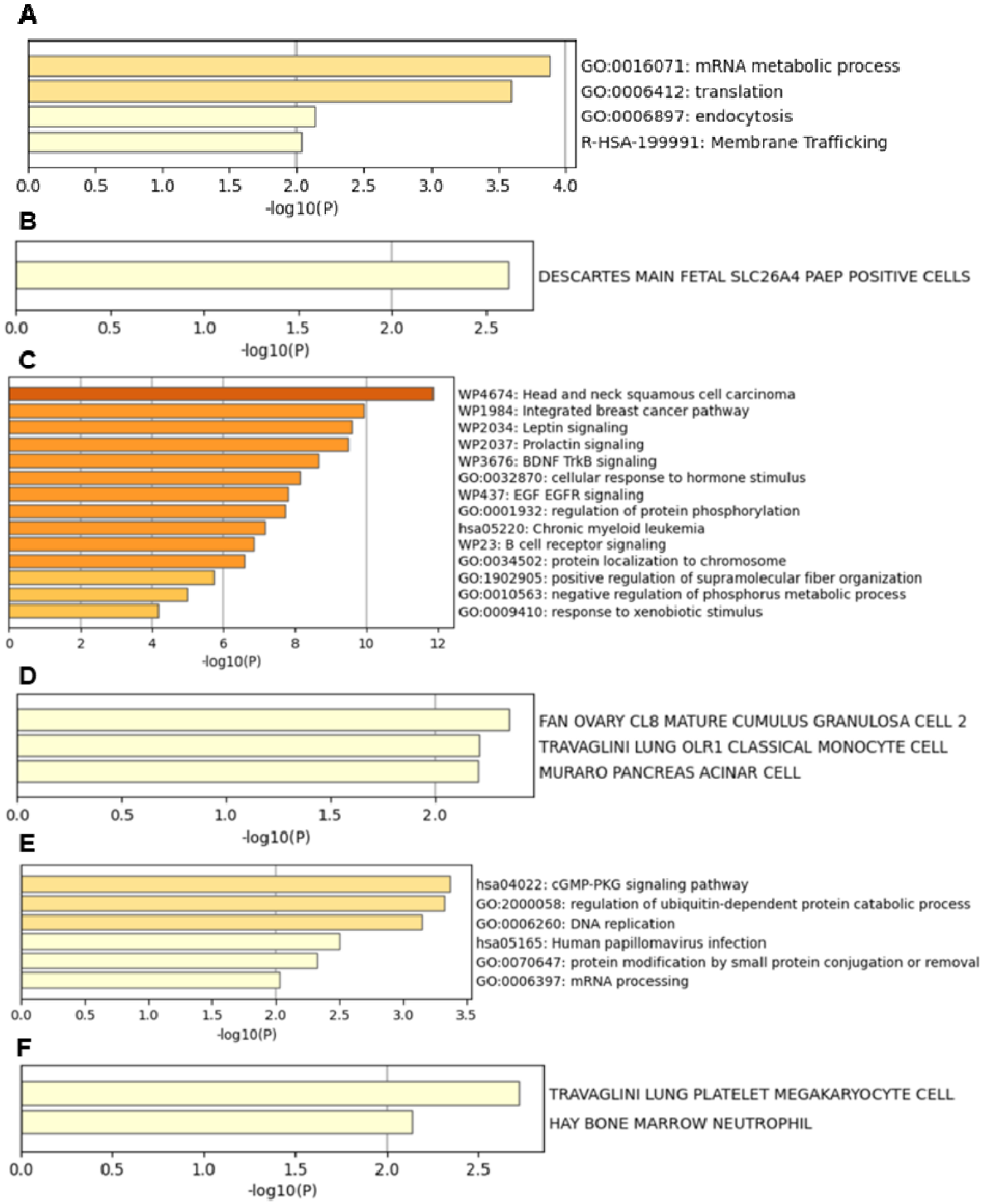
Code is available on Github.

https://github.com/gmestrallet/brcatcga

Supplementary data

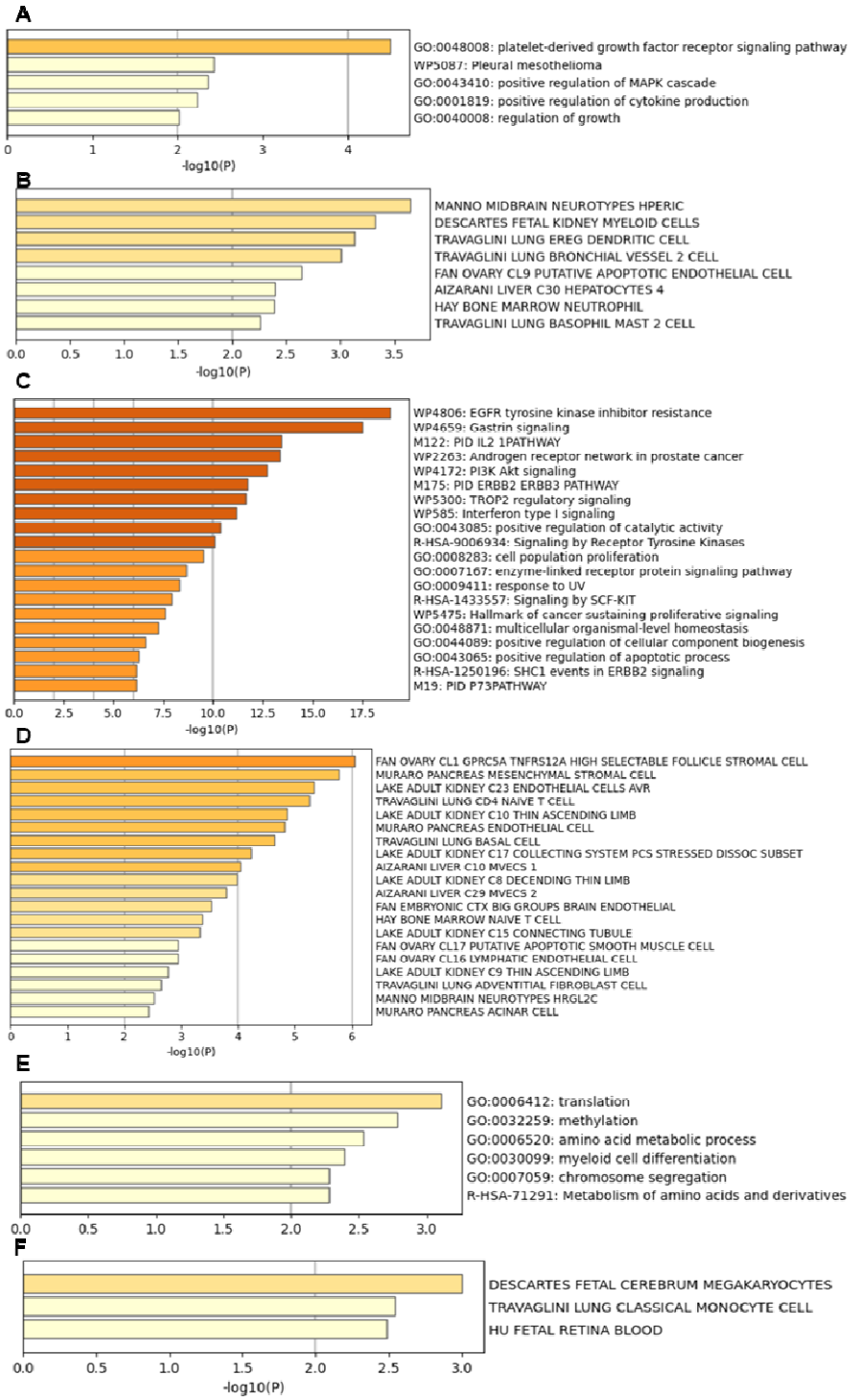
Table 1) Features selected for BANDOL training

Feature type							
Methylation	**SLC25A22**	**PPARG**	**OLIG3**	**FTHL17**	**ME3**	**RIN2**	**HTR3B**
mRNA	**PELO**	**MAFA**	**ESRRG**	**MAP2K6**	**IRF2**		
Clinical	**Age**	**Cyclophosphamide**					
Protein	**NFKB1**						



Supplementary Figure 1) Breast cancer patient multi-omics and clinical features associated with a shorter survival and gene enrichment analysis.

Gene enrichment (**A/C/E**) and cell type signature (**B/D/F**) analysis with these top/bottom genes (**A/B** from the RNA dataset of Figure 1 C, **C/D** from the methylation dataset of Figure 1 D and **E/F** from the protein dataset of Figure 1 E) was performed using metascape [16].



Supplementary Figure 2) Breast cancer patient multi-omics and clinical features associated with a longer survival and gene enrichment analysis.

Gene enrichment (**A/C/E**) and cell type signature (**B/D/F**) analysis with these top/bottom genes (**A/B** from the RNA dataset of Figure 1 C, **C/D** from the methylation dataset of Figure 1 D and **E/F** from the protein dataset of Figure 1 E) was performed using metascape [16].

Conflicts of Interest

The authors do not have any conflict of interest to declare.

Funding

Not applicable.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.