
| RESEARCH ARTICLE**Knowledge-Guided, Causally Constrained Multi-Omic Graph Learning for Patient-Specific Signaling-Driver Inference in Breast Cancer****Shaila Shahid***Department of Biochemistry, Bangladesh Medical University***Corresponding Author:** Shaila Shahid, **E-mail:** shaila07shahid@gmail.com

| ABSTRACT

Background: The transcriptomic subtypes of breast cancers can have distinct signaling states because RNA levels do not adequately reflect the levels of proteins and phosphoproteins. Although cohort-level prediction-focused classifiers are prevalent in the field of graph-based multi-omic classifiers, directionally constrained and patient-level network scores are necessary for clinically interpretable prioritization. Methods: Sample-level harmonization of matched RNA sequencing, reverse-phase protein array (RPPA), and clinical matrices was performed. The main analysis comprised of 742 primary tumours, 20,530 transcript measurements, and 131 RPPA analytes. This resulted in 103 nodes representing genes being mapped with RPPA analytes. A structural-equation model was constrained by a 132-edge acyclic signaling prior spanning model that included programs for 13 different pathways: receptor tyrosine kinase, PI3K-AKT-mTOR, MAPK, hormone-receptor, DNA-damage, apoptosis, Hippo, WNT, TGF-beta, adhesion, and cell-cycle. After the heteroskedasticity-robust testing and Benjamini-Hochberg correction, prior-sign agreement was performed, and 200 bootstrap resamples were used to keep the edges. The patient specific putative-driver scores were composed of role-oriented protein activity, structural residual activation, and total downstream influence. The classification of PAM50 was evaluated on 673 tumours; cross-validation was done five times on each tumour, with leakage control. The associations between overall-survival and variables were tested in 582 tumours and 70 deaths. Results: RNA abundance correlated significantly with 111 of 131 RPPA analytes at a false-discovery rate <0.05 . The median RNA-protein correlation for total proteins, cleaved proteins, and phosphoproteins was 0.370, 0.375, and 0.143, respectively, with both total and cleaved proteins showing significantly greater levels of correlation than phosphoproteins (Spearman ρ 0.370; 0.375; 0.143; $P=3.04 \times 10^{-5}$ and 1.12×10^{-4}). Of the 132 previous edges, 52 were both directionally concordant and FDR-significant and were also bootstrap-stable. Supported relationships included ESR1-GATA3 (standardized $\beta = 0.481$), AKT1-AKT1S1 ($\beta = 0.533$), MAP2K1-MAPK1 ($\beta = 0.497$), MET-RAF1 ($\beta = 0.444$), CHEK2-TP53 ($\beta = 0.448$), and MTOR-RPS6KB1 ($\beta = 0.297$). The early-fusion RNA-RPPA model achieved balanced accuracy 0.867 (95% CI 0.863-0.872) and macro-AUROC 0.975 (95% CI 0.973-0.976). Recall was 0.958 for Basal, 0.836 for HER2-enriched, 0.906 for Luminal A, and 0.765 for Luminal B disease. ERBB2 regained control at the patient level on HER2-enriched tumours, INPP4B lost control on Basal tumours and ESR1-associated programs were regained on luminal disease. There was no pathway level score that survived multiplicity correction that had an association with overall survival. Conclusions: Stable and directionally constrained signaling graph and heterogeneous patient-specific driver profiles were supported by matched transcriptomic and proteomic data. The findings illustrate the interpretive benefits of proteomic integration, as well as highlighting that cross-sectional TCGA data enable causally constrained inferences, but do not allow for definitive causal identification. External validation and genomic perturbation anchors remain necessary before clinical deployment.

| KEYWORDS

Breast cancer; TCGA; multi-omics; RPPA; graph learning; structural equation model; patient-specific network; PAM50; causal constraints.

| ARTICLE INFORMATION**ACCEPTED:** July 30, 2025 **PUBLISHED:** December 15, 2025 **DOI:** <https://doi.org/10.61424/ijmhr.v3i3.1107>

1. Introduction

Breast cancer is a highly heterogeneous disease with major differences in lineage, proliferation and clinical behaviour represented by intrinsic expression subtypes^{1,2}. Subsequent large-scale genomic analyses showed that there was also some heterogeneity within the individual PAM50 classes, with new copy-number and mutation

Copyright: © 2025 The Author(s). This open access article is distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC-BY-NC-SA) License (<https://creativecommons.org/licenses/by-nc-sa/4.0/>), published by Bluemark Publishers.

states, plus distinct immune and proteomic states ²⁻⁵. These results lead to analytical frameworks that do not minimize molecular variation within each tumour by assigning it a single subtype label, but maintain patient level molecular variation.

The abundance of transcripts is not an accurate measure of signaling activity. The measurement of total protein and total phosphoprotein can be decoupled due to RNA ²⁻⁵, protein degradation, post-translational modification, and subcellular localization, or feedback in the pathways. Reverse-phase protein arrays offer focused and quantitative measurements of oncogenic and tumour-suppressive signaling nodes, such as phosphorylation states which are more distal to the execution of the pathway than mRNA levels. The combination of RNA and RPPA is then a tractable environment to evaluate the translation of transcript-level programs into signaling phenotypes.

Latent-factor models, similarity-network fusion, sparse supervised integration, and graph neural networks are all methods used in multi-omic integration ⁶⁻¹³. The utility of graph-based integration in enhancing breast-cancer classification into subtypes and discovery of biomarkers is illustrated by GAIN-BRCA, MOGONET, moBRCA-net, and related approaches ⁶⁻¹⁰. Even when the predictive accuracy is high, however, it does not necessarily reflect causal structure or point to the molecular alteration that is most responsible for an individual tumour state. Most of the patient-specific driver methods used to address this interpretive objective currently focus on mutation, CN analysis, or large transcriptomic networks, but not matched signaling measurements ¹⁴.

In the present study, we tested whether a small, knowledge-driven graph that was constrained by a canonical direction of signaling could integrate matched TCGA-BRCA RNA and RPPA data, recover reproducible relations among the pathways, create patient-specific signaling-driver scores, and preserve predictive information for PAM50 subtype classification. Causally constrained means that causality and sign of the edge were determined by biological knowledge and estimated by the structural equations. It is not an assumption that a single cross sectional observational dataset would identify intervention effects. The study is of an empirical secondary analysis character type, explicitly falsified and leakage resistant, and claim boundaries are conservative.

2. Materials and methods

2.1 Study design and data provenance

Three matrices from the UCSC Xena TCGA-BRCA, provided for analysis, were used in this completed secondary analysis: TCGA.BRCA.sampleMap_HiSeqV2.gz, TCGA.BRCA.sampleMap_RPPA_RBN.gz, and TCGA.BRCA.sampleMap_BRCA_clinicalMatrix. The files were those that contained gene-level RNA-sequencing information, total-protein/phosphoprotein values from RPPA and sample level clinical annotations. The integrity of the files was confirmed through the use of SHA-256 hashes, as shown in Appendix A. In this study, no patient recruitment, biospecimen processing or laboratory assays were done.

Identifiers of the samples were harmonized precisely. The RNA data had 1,218 samples in it, the RPPA data had 747 samples in it, and the clinical matrix had 1,247 records. There was 100% RNA and clinical match for all 747 RPPA samples. Primary tumours were selected for the main analysis (742) with the remaining five metastatic samples kept as a sensitivity set and not used to train the models. Then, the normal-like and missing labels were excluded from the PAM50 classification and the remaining 673 tumours were classified as: Luminal A (330), Luminal B (162), Basal (120) and HER2-enriched (61).

Cohort assembly and analysis sets

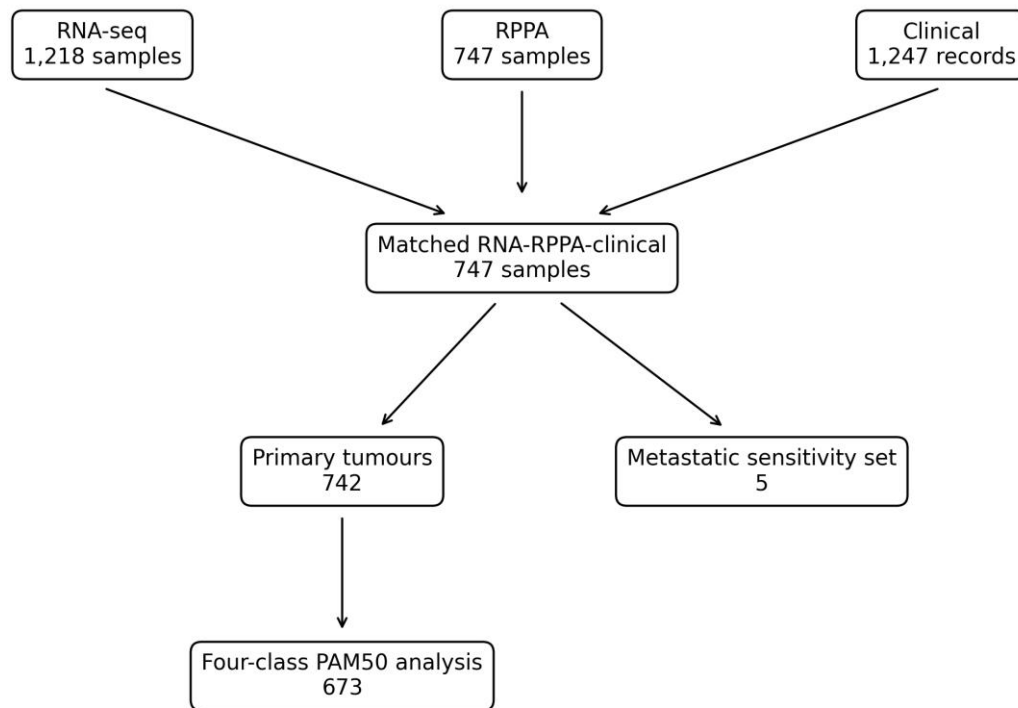


Figure 1. Cohort assembly. Exact sample identifiers linked RNA, RPPA, and clinical records. Primary tumours formed the main analysis set; metastatic samples were excluded from model fitting.

2.2 Molecular preprocessing and feature mapping

There were 20,530 different gene symbols in the RNA matrix with no missing values. In the matched primary-tumour cohort, there were 131 analytes in the RPPA matrix without missing values. One or more RNA genes were mapped to each RPPA analyte - composite antibodies to a canonical graph node, while preserving all mapped RNA genes for transcript aggregation. A mapping process resulted in 103 gene-level signaling nodes. Multiple antibodies were averaged together for each node, and within each training set, a normalization step was performed on the values of each analyte. The value of RNA for each node was averaged and normalized in the training set. This ordering avoided that test folds would get into pre-processing estimates.

Spearman correlation was used to evaluate the concordance level between RNA and protein at the analytes level. Serine, threonine, or tyrosine residue-labeled as a notation in the RPPA analyte label identified phosphoproteins. The Benjamini-Hochberg procedure was used to correct for multiple comparison of correlation P values¹⁵. Median correlations of assay classes were compared using the Mann-Whitney test.

2.3 Knowledge-guided signaling graph

A directed acyclic prior graph contained 132 edges across 87 measured nodes. Edges represented canonical receptor, kinase, transcriptional, apoptotic, adhesion, and cell-cycle relationships supported by pathway resources and cancer knowledge bases¹⁶⁻²¹. Each edge carried an expected activating or inhibitory sign. The graph was intentionally restricted to measured RPPA nodes to avoid imputing unobserved intermediates. The complete edge list and analyte-to-gene mapping are reported in Appendix A.

For target protein j in tumour i , the structural equation was

$$P_{ij} = \alpha_j + \gamma_j R_{ij} + \sum_k \beta_{kj} P_{ik} + \varepsilon_{ij}$$

Where R_{ij} was the matched transcript value and k indexed prespecified upstream protein parents. The coefficients were estimated using ordinary least squares with HC3 heteroskedasticity-robust standard errors. FDR adjusted p value < 0.05 with the same FDR sign as the previous sign was considered an edge. The 95% percentile intervals and sign stability were estimated based on 200 tumour level bootstraps. Prior-sign stability of 80% or more and a bootstrap interval not including zero were needed to have a stable edge.

The directionally constrained model is similar to a linear structural causal model but causal interpretation is only possible if the graph is specified correctly, there are no unmeasured confounding of each parent-target relation, consistency and positivity²². These assumptions have not been measurable with cross-sectional TCGA measurements. Accordingly, the estimated coefficients are reported as structurally conditioned signaling associations.

2.4 Patient-specific signaling-driver score

The stable edge coefficients were used to construct a weighted acyclic matrix B , and then the total downstream influence was obtained as $T=(I-B)^{-1}-I$ and normalized by node numbers. After removing the prediction of the structural equation, the target-protein residuals have been standardized for each tumour. The known cancer roles were used to orient activity, positive ones were given for oncogenic nodes, negative ones were given for tumour suppressing nodes, and absolute deviation were given for nodes that had both roles. The patient-specific score was $D_{ig} = w_g(0.6A_{ig} + 0.4U_{ig})$ where w was the normalized total downstream influence, A was the role-oriented protein activity and U was the role-oriented unexplained residual activity. Ranked scores were obtained within each tumour for curated oncogenic and tumour-suppressor candidates.

The top five driver frequency and standardized mean difference between one subtype and the remaining three subtypes were used to summarize subtype enrichment. Within the subtypes, multiple tests were controlled by Benjamini-Hochberg procedure. Convergence was quantified as pairwise Jaccard similarity between patient top five sets, both within and between subtypes.

2.5 Subtype prediction and leakage control

Five features sets were tested: mapped nodes of RNA, nodes of the RPPA at the gene level, early fusion of RNA and RPPA, patient-specific network-driver scores and network-augmented fusion. The five-fold cross-validation with repeated ten times was used to evaluate a multinomial logistic-regression model weighted by the class. Analyte standardisation, node aggregation, structural-equation fitting, residual scaling, calculation of downstream influences and fitting of driver scores were fitted in the training partition and applied to the held-out partition in every outer fold. The performance measures included: balanced accuracy, macro-F1, macro one-versus-rest AUROC, and overall accuracy. Repeat-level means were used to calculate confidence intervals.

The normalized confusion matrix and AUROCs for subtypes were made using a separate 5-fold out-of-fold prediction. For the negative control, 50 times the labels were permuted and all the classification procedures were repeated. Model reporting was based on relevant TRIPOD, PROBAST and MI-CLAIM principles^{23,24}.

2.6 Survival analysis

Overall-survival time and event status were taken from the harmonized clinical fields. The adjusted survival cohort comprised 582 four-class PAM50 tumours with positive follow up time and 70 deaths. Pathway scores are the average of standardized patient driver scores for genes grouped into each of the pathways. Age, pathologic stage, and markers for Basal, HER2-enriched, and Luminal B subtype were used in Cox proportional-hazards models with Luminal A as reference²⁵. The hazard ratios were the ones standard deviation increase in pathway score. Seventeen

pathway or global-burden hypotheses were corrected by the Benjamini-Hochberg method. Because of the exploratory nature of gene-level models, no statistical tests were performed.

There is no survival model used for the selection and tuning of the graph. This separation helped to avoid overfitting the network on outcomes. Proportionality and transportability remain limitations because the clinical matrix contains relatively few deaths and heterogeneous follow-up.

2.7 Falsification and sensitivity analyses

The edge falsification test was conducted by permuting edges of each protein node within tumours, while keeping the transcript values and repeating the estimation of the edge structure 100 times. Subtype labels were randomly shuffled 50 times for the classification falsification test. These analyses estimated the number of apparently supported edges and predicted classification accuracy if the molecular structure and/or the labels were destroyed. All random operations were seeded with 20260722.

2.8 Software and reproducibility

Analyses were executed in Python 3.13 using pandas 2.2.3, NumPy 2.3.5, SciPy 1.17.0, scikit-learn 1.8.0, statsmodels 0.14.6, NetworkX 3.6.1, and Matplotlib 3.10.8. The manuscript is accompanied by machine-readable tables, patient-level scores, figures, an analysis summary, and an executable workflow. File hashes and definitions of models are given in Appendix A and expanded empirical results are given in Appendix B.

3. Results

3.1 Cohort characteristics

A total of 747 samples were matched exactly across the three matrices of which 742 were primary tumours and 5 were metastatic. There were 673 primary tumours in the four-class PAM50 cohort. The median age of diagnosis was 58 years (48.8-67.0 years). The distribution of clinical subtypes and the number of outcomes available are provided in Table 1. After excluding missing or zero follow-up, 582 tumours and 70 deaths were included in the survival analysis.

Table 1. Characteristics of the four-class PAM50 analysis cohort.

Group	N	Age median	Age Q1	Age Q3	Stage I	Stage II	Stage III	Stage IV	ER+	HER2 +	OS available	Deaths
Overall	673	58.00 0	48.7 50	67.00 0	106	388	154	14	455	93	636	70
LumA	330	60.00 0	49.0 00	67.00 0	70	176	72	6	282	22	306	27
LumB	162	58.00 0	48.2 50	69.00 0	18	92	47	4	144	28	154	19
Basal	120	54.00 0	47.0 00	62.25 0	14	85	17	2	12	2	116	14
Her2	61	58.00 0	51.0 00	64.00 0	4	35	18	2	17	41	60	10

Counts for receptor status reflect available clinical annotations and therefore do not sum to group totals.

3.2 RNA-protein concordance is strong for lineage proteins and weaker for phosphorylation states

After FDR correction, a significant correlation was observed between RNA abundance and 111 of 131 RPPA analytes. 84 of 92 total or cleaved protein, and 27 of 39 phosphoproteins were significant. Median Spearman correlation was

observed to be 0.370 for the total and cleaved proteins, and 0.143 for the phosphoproteins (Mann-Whitney $P=3.04 \times 10^{-5}$; Figure 2).

The strongest concordance occurred for ESR1/ER-alpha ($\rho=0.889$), BCL2 ($\rho=0.841$), GATA3 ($\rho=0.785$), PGR/PR ($\rho=0.765$), INPP4B ($\rho=0.762$), IGFBP2 ($\rho=0.748$), CCNB1/cyclin B1 ($\rho=0.732$), ASNS ($\rho=0.704$), CHEK2/CHK2 ($\rho=0.675$), and KIT/c-KIT ($\rho=0.674$). In contrast, several phosphorylation-specific markers showed weak or inverse transcript coupling, including RPS6 pS240/244 ($\rho=-0.112$), MAP2K1 pS217/221 ($\rho=-0.046$), CDKN1B pT157 ($\rho=-0.045$), and RPS6KB1 pT389 ($\rho=-0.045$). These results point towards direct measurement of the signaling states and not transcript-only substitution.

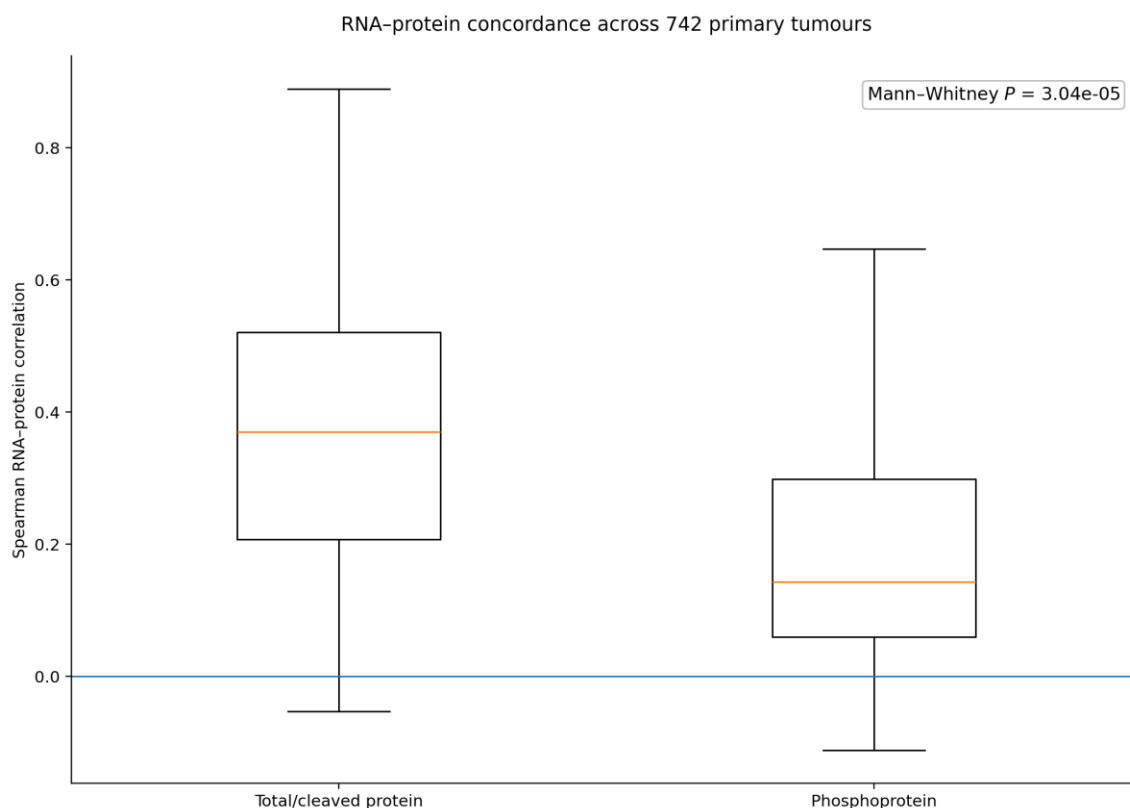


Figure 2. Matched RNA-protein concordance. Total or cleaved protein measurements showed higher median transcript correlation than phosphorylation-specific measurements.

3.3 Directionally constrained structural modeling retains stable pathway relations

For a structural model, 132 prior directed edges were tested. The number of FDR-significant and prior-sign concordant 53 edges was also found to meet the bootstrap interval and $\geq 80\%$ sign-stability criteria. The median R-squared for target-protein was 0.182 (interquartile range 0.083-0.335). The 100 protein-permutation controls had a median of 0 edges (range 0 to 1) while the observed data had 52 stable edges.

The strongest supported associations included AKT1-AKT1S1 ($\beta=0.533$, 95% bootstrap CI 0.474-0.592), MAP2K1-MAPK1 ($\beta=0.497$, 0.440-0.560), ESR1-GATA3 ($\beta=0.481$, 0.436-0.533), CHEK2-TP53 ($\beta=0.448$, 0.382-0.523), MET-RAF1 ($\beta=0.444$, 0.367-0.527), CHEK1-TP53 ($\beta=0.335$, 0.269-0.396), MTOR-RPS6KB1 ($\beta=0.297$, 0.252-0.355), SMAD4-FN1 ($\beta=0.263$, 0.181-0.356), ERBB2-ERBB3 ($\beta=0.243$, 0.180-0.317), and the expected inhibitory BCL2-CASP7 relation ($\beta=-0.303$, -0.370 to -0.237).

The pathways with the highest pathway retention rates were TGF-beta (5/5), DNA-damage signaling (6/8), Notch (2/3) and cell-cycle signaling (4/6). A few canonical priors were not kept since their observed conditional coefficient had the reverse sign. For instance, the association of CDH1 and CTNNB1 was positive after the conditioning,

suggesting that the association is likely due to common confounding from the epithelial-state, not the direct inhibitory relation proposed by the prior. These disagreements were removed, not included, in the patient score.

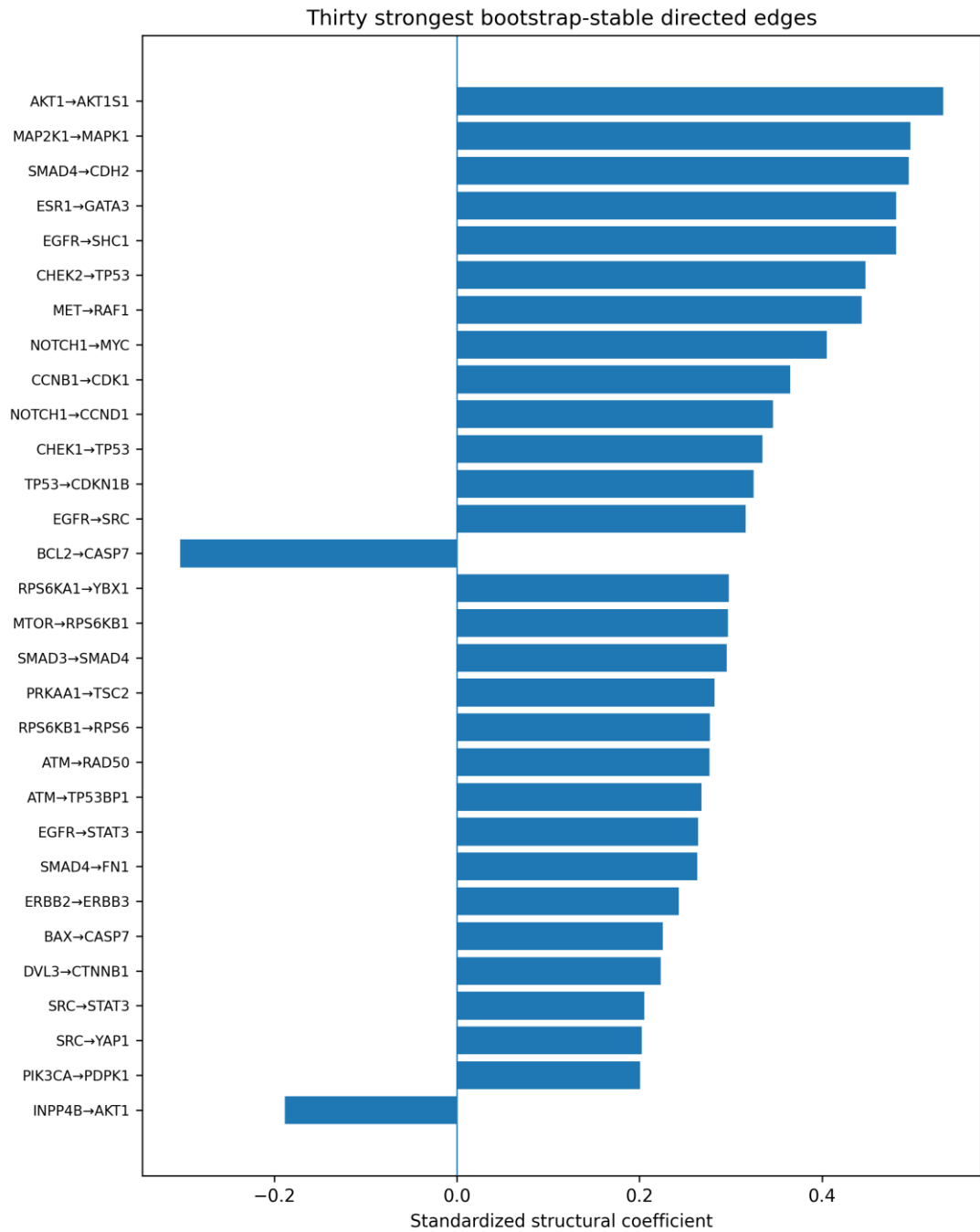


Figure 3. Strongest bootstrap-stable directed edges. Bars show standardized structural coefficients after conditioning on the target transcript and other prespecified parents.

3.4 Patient-specific scores recover subtype biology while preserving within-subtype heterogeneity

Expected subtype-enriched patterns were generated using patient-specific scores. The contrast for ERBB2 was the highest in HER2-enriched tumours (standardized mean difference [SMD] 2.313; $q=6.68\times10^{-14}$), and ERBB2 was one of the top five candidate drivers in 60.7% of HER2-enriched tumours. Basal tumours were characterized by the loss-oriented INPP4B score (SMD 1.451; $q=4.43\times10^{-36}$), CCNB1 (SMD 1.297; $q=6.36\times10^{-23}$), CCNE1 (SMD 1.267; $q=6.18\times10^{-24}$), NOTCH1 (SMD 0.947; $q=1.13\times10^{-16}$), and KIT (SMD 0.923; $q=5.84\times10^{-12}$).

Luminal B tumours showed enrichment for IGFBP2 (SMD 0.479; $q=5.0 \times 10^{-6}$), ESR1 (SMD 0.444; $q=1.0 \times 10^{-6}$), CDKN1B loss-oriented activity (SMD 0.416; $q=7.0 \times 10^{-5}$), STK11 loss-oriented activity (SMD 0.399; $q=6.0 \times 10^{-6}$), and GATA3 (SMD 0.374; $q=6.0 \times 10^{-6}$). The luminal A tumours had relatively higher MAPK9, STAT3, PGR, MAPK14, CHEK2 and BCL2 scores. The latter pattern indicates that a patient-specific influence score is not identical to differential expression: it incorporates residual activation and downstream network reach.

The mean pairwise Jaccard similarities among the within-subtype top five driver sets were 0.096; between subtypes, 0.074. The small absolute similarity only indicates a large degree of patient heterogeneity even in the presence of the enrichment of the patients by the subtype. This is reflective of the intended use of the score as a personal prioritisation system and not a substitute for PAM50 classification.

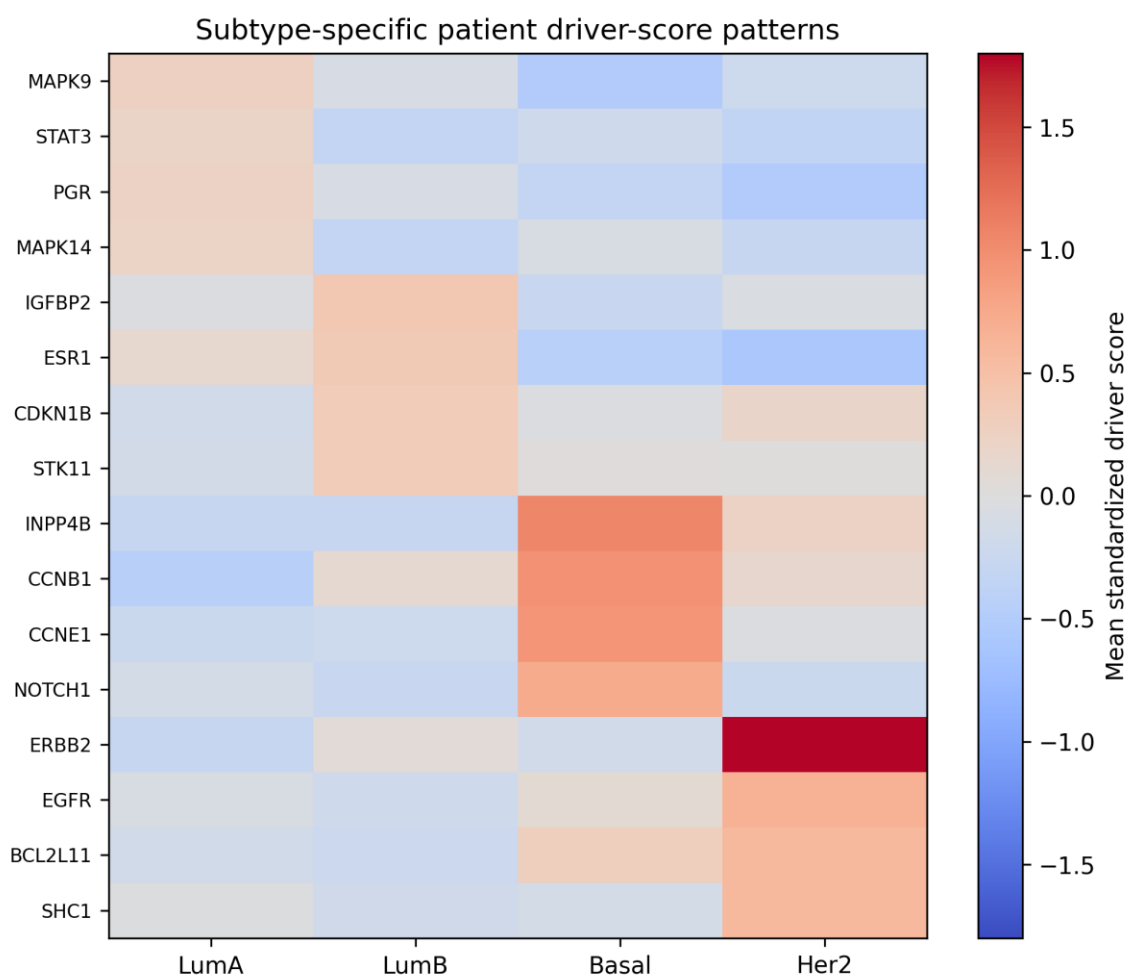


Figure 4. Subtype-specific driver-score patterns. Values are mean standardized patient-specific scores for the strongest positively enriched genes in each subtype. Tumour-suppressor scores are loss-oriented.

3.5 RNA-RPPA integration provides high subtype discrimination

In repeated five-fold cross-validation, early RNA-RPPA fusion achieved the highest balanced accuracy, 0.867 (95% CI 0.863-0.872), and macro-AUROC, 0.975 (95% CI 0.973-0.976). The balanced accuracy of RNA alone was 0.862 and macro-AUROC was 0.974; RPPA alone was 0.779 and 0.929, respectively. Balanced accuracy and macro-AUROC of the patient-specific network-driver matrix were 0.679 and 0.875 respectively. Early fusion with network scores did not lead to an improvement (balanced accuracy 0.861), showing that interpretability was the main interest of the driver representation, not additional subtype classification.

The out-of-fold early-fusion model correctly recalled 95.8%, 83.6%, 90.6% and 76.5% of Basal, HER2-enriched, Luminal A and Luminal B tumours, respectively. Subtype-specific AUROCs were 0.997, 0.990, 0.969, and 0.941, respectively. Misclassification was targeted in the Luminal A to B classification. The multiclass Brier score was 0.191 and log loss was 0.397. The mean balanced accuracy for the label-permutation control was 0.253 (95% empirical interval 0.215-0.280), while the observed five-fold split had an accuracy of 0.869.

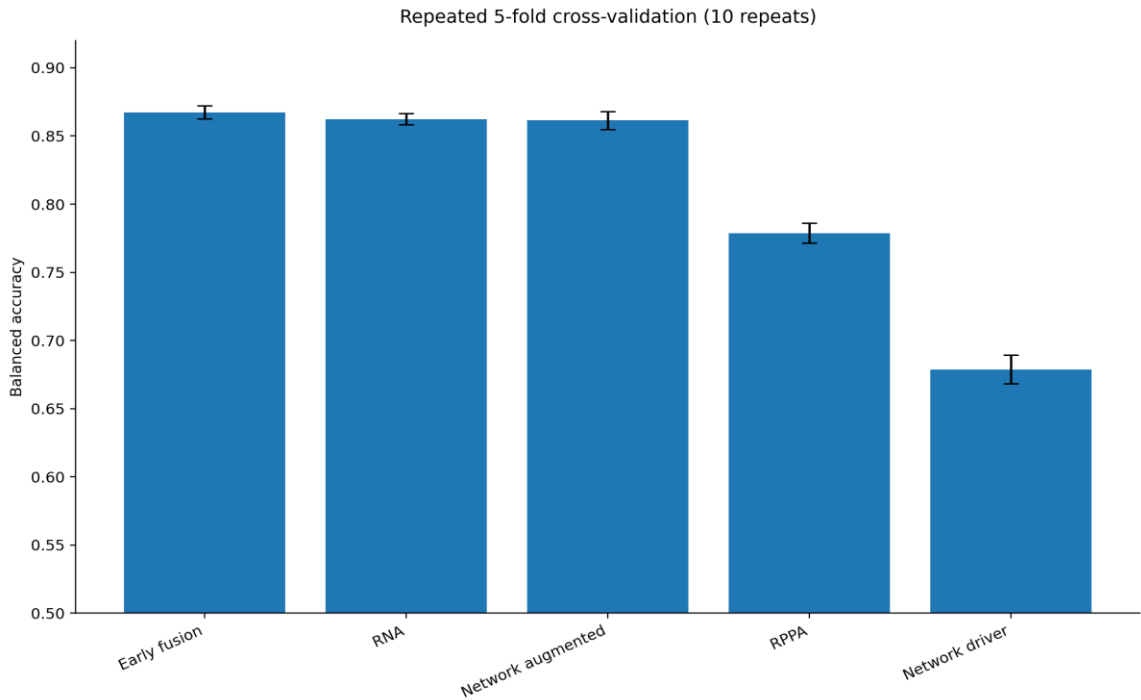


Figure 5. Repeated cross-validation performance. Bars show mean balanced accuracy across ten repeats; error bars show 95% confidence intervals from repeat-level means.

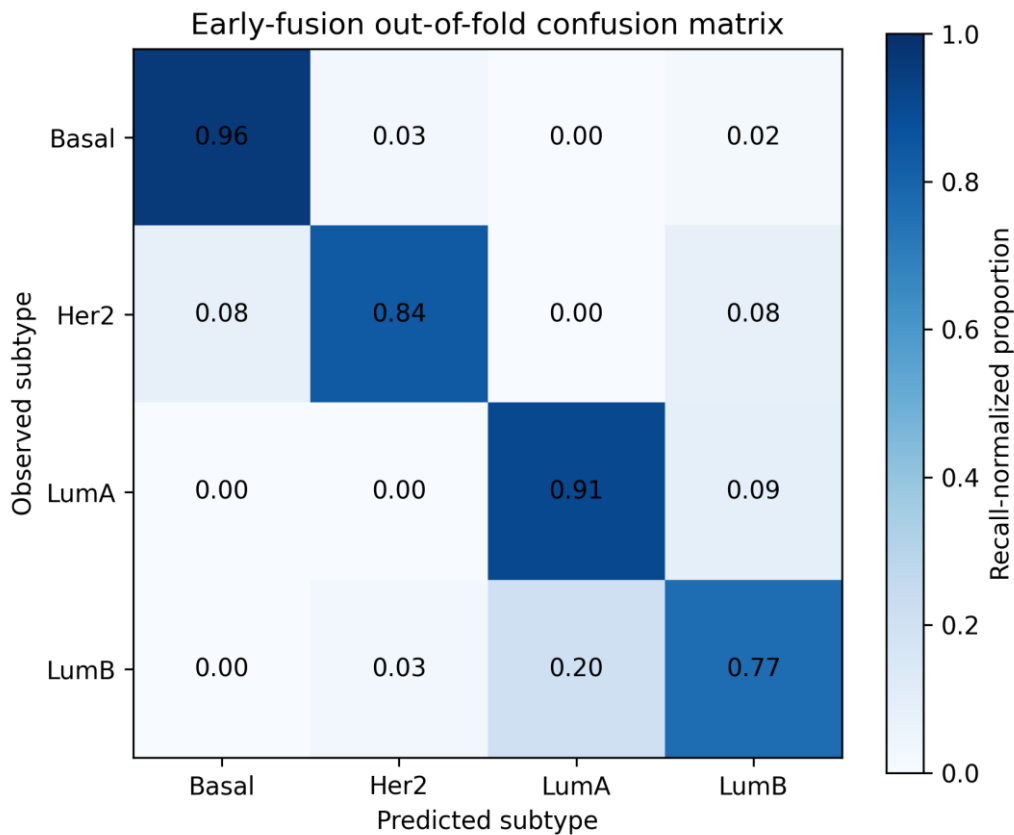


Figure 6. Recall-normalized out-of-fold confusion matrix for early RNA-RPPA fusion. Luminal A-Luminal B separation was the principal residual error mode.

3.6 Patient-specific pathway scores are not independently prognostic after multiplicity correction

This survival cohort was adjusted and included 582 tumours and 70 deaths. All pathway and global driver-burden scores failed FDR. The smallest nominal P value occurred for PI3K-AKT signaling (hazard ratio per standard deviation 0.789, 95% CI 0.609-1.022; $P=0.073$; $q=0.449$), followed by AMPK signaling (HR 0.800, 95% CI 0.624-1.027; $P=0.080$; $q=0.449$). There was also no result with FDR association at the gene level, with the smallest nominal association being SYK (HR 0.652, 95% CI 0.473-0.899; $P=0.009$; $q=0.668$).

The non-significant results for survival after multiplicity adjustment are significant. They suggest that the impact of network influence and subtype discrimination is not necessarily independent of prognosis in a small sample of events. The manuscript does not make any prognosis claim.

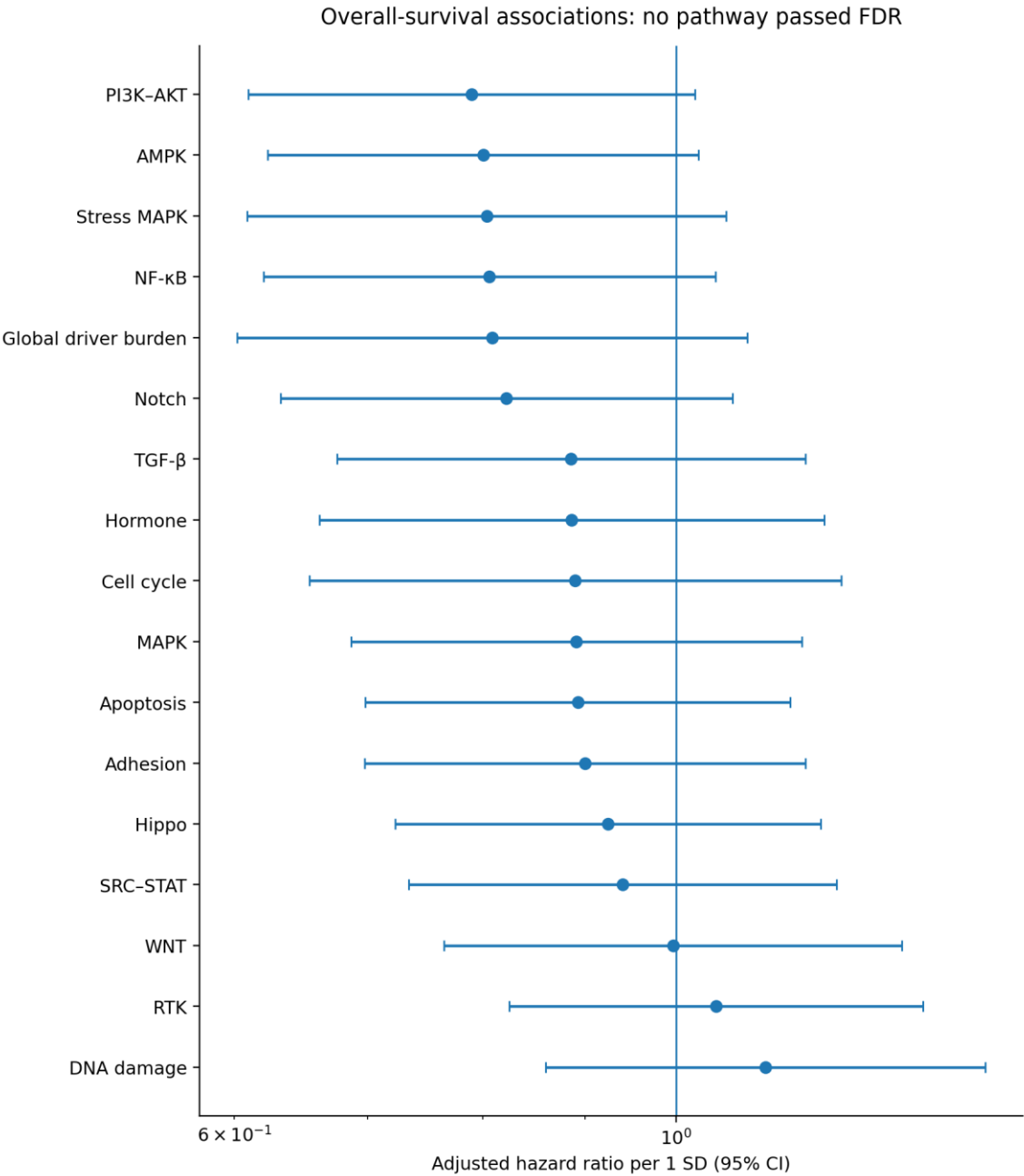


Figure 7. Adjusted overall-survival associations. Hazard ratios are per one-standard-deviation pathway score and are adjusted for age, stage, and PAM50 subtype. No pathway passed FDR correction.

4. Discussion

This analysis highlights the ability of a compact, knowledge-guided graph to embed matched measurements of the TCGA-BRCA transcript and signaling in stable relations on the cohort-level and patient-level prioritization scores. There are three key findings. First, the RNA-protein coupling was significant for the lineage- and abundance-driven proteins, and was significantly lower for phosphorylation states. Second, a conservative combination of prior direction, strong inference, FDR control and bootstrap stability was able to keep 52 from 132 candidate relations and reject biologically plausible relations that were not supported by the observational data. Third, patient-specific network scores retrieved major drivers associated with each major subtype but showed a high degree of heterogeneity in each PAM50 group.

The RNA-protein results are in line with the proteogenomic studies that revealed that post-transcriptional regulation and phosphorylation mould the functional consequences of genomic and transcriptomic alterations²⁻⁵. While there was strong transcript-protein correlation for ESR1, PGR, GATA3, BCL2, and ERBB2, several phospho-epitopes demonstrated weak correlation with the corresponding mRNA. This is the reason why RNA was nearly as good as early fusion for the classification of subtype but was not able to give the same signaling-state interpretation. The improvement in the early fusion, as measured by balanced accuracy, was small numerically since the PAM50 labels were themselves expression-based, and the RPPA layer provided another type of biological evidence, instead of a separate label definition.

The stable edge set recapitulated several canonical signaling pathways, such as MEK-ERK, AKT-PRAS40, mTOR-p70S6K, ERBB2-ERBB3 relations, checkpoint-TP53, and SMAD-fibronectin relations. The results agree with knowledge constrained estimation as compared with unconstrained network discovery^{16-18,22,26}. Meanwhile, some edges were not signed before, which is why knowledge should limit, but shouldn't dictate, the results. Lineage, feedback, measurement composition and unmeasured confounding can all be represented by cross sectional co-variation. For instance, a positive conditional association of CDH1 and CTNNB1 (the CDH1 and CTNNB1 are measured at the level of the cohort, but do not point to that known molecular mechanism under the equation fitted).

The patient-specific score was created to differentiate between activity and influence. A downstream node with a very unusual terminal marker was not automatically dominant, even if it had a short reach downstream; on the other hand, a node upstream (in the direction of the flow) could gain influence through stable edges. Adding residual activation further highlighted proteins for which the observed state was not predicted by their measured parents and transcript. For ERBB2, this design recovers the ERBB2 enriched disease, and loss-oriented INPP4B was recovered from the proliferative nodes in Basal disease. However, the scores are only putative signaling-driving priorities, and not proven therapeutic dependencies. Patient-specific network scores such as DriverNet, DawnRank, OncoIMPACT, LIONESS, and PARADIGM also show that the scores must be validated orthogonal to the network scores¹⁴.

The objective performance of the strong subtype is similar to graph-based multi-omic classifiers like GAIN-BRCA, MOGONET, moBRCA-net, and multi-view GNN approaches⁶⁻¹⁰, although not directly in data split or feature definition. The present study does not rely on a deep classification architecture, but rather focuses on a transparent structural graph and scores on the patient level. However, the network-driver representation was not as predictive as RNA or early fusion, which is expected given that it is a compression of measurements into the role-based deviations and downstream influence. It's not about the highest possible accuracy of labels, it's about mechanical interpretation.

There is a critical boundary imposed by the survival results. None of the pathways or gene scores passed multiplicity correction with age, stage and subtype adjusted. The power to predict prognosis is limited because of the limited number of events, the diversity of treatments, the quality of study endpoints and the use of baseline measures from the primary tumour. The study hence provides further basis for the inference of the subtypes and signaling states and not for clinical outcome prediction. As opposed to the selective omission of this negative finding, this one is kept.

4.1 Causal interpretation

Causal is a term that is sometimes used loosely in the study of graph learning. In this case, the direction and sign of candidate relations was imposed by known signaling biology and incorporated into structural equations, and is therefore not causal. The analysis did not make use of observational data to identify a causal graph, nor was it randomized or longitudinal interventions, germline instruments, somatic copy-number anchors, or perturbation-response data used. With strong assumptions on the structure, the coefficients can be interpreted as conditional direct effects and without those assumptions the coefficients are still directed associations. It is thus preferable to employ the term "causally constrained," instead of "causal discovery."

4.2 Strengths and limitations

Strengths are the ability to match samples exactly, to provide complete RNA and RPPA matrices, to preprocess the data using the training-fold, and to pre-specify pathway directions, to offer robust standard errors, to apply FDR correction, to provide 200 bootstrap resamples, and to explicitly include negative controls, as well as repeated cross validation and release of patient-based scores and machine-readable output. The study also removes the distinction between the prediction of subtypes and survival testing and provides null prognostic results.

Limitations are substantial. Only transcriptomic data and RPPA were available with no integration of mutation, copy-number, methylation or treatment data. Analytes covered in RPPA were limited to 131 and gene-level scoring were based on combined total and phospho measurements. The graph has been manually curated and does not capture all the context-specific feedback. TCGA is retrospective and cross-sectional in nature. PAM50 labels are derived from expression - favoring RNA-based models. Cross-cohort validation was not done because there was no METABRIC or CPTAC matrix provided. There were not enough metastatic samples to be meaningful. Lastly, patient-specific scores were not associated with level of functional perturbation or drug-response outcome.

4.3 Implications and next steps

A high-impact extension requires independent validation rather than additional complexity alone. In addition to testing protein and phosphoprotein reproducibility 3-5, CPTAC-BRCA can test the transport of transcript-associated driver modules and clinical subgroup patterns, while METABRIC can test the transport of transcript-associated driver modules and clinical subgroup patterns. Genomic anchors for ERBB2, PIK3CA, PTEN, TP53, CCND1 and others can be provided by copy-number and mutation layers. Perturbation/Drug-sensitivity datasets can be used to determine if top-ranked patient drivers predict dependency. Unless these additions are made, the model cannot be used for assignment of treatment.

5. Conclusions

Matched TCGA-BRCA RNA and RPPA data were used to recover an orientationally consistent signaling graph, 52 of whose edges were found to be bootstrap stable, and patient-specific putative-driver profiles. Early RNA-RPPA fusions accurately classified PAM50 subtypes, and network scores revealed signaling heterogeneity not captured by subtype classifications. The study offers an empirical framework that is transparent and a full package of reproducibility. It is not asserting external validation, causal discovery, or prognosis or clinical readiness.

Data and code availability: The analysis used publicly available TCGA-BRCA matrices distributed by UCSC Xena. The exact input filenames and SHA-256 hashes are listed in Appendix A. The accompanying analysis package contains the Python workflow, executed result tables, patient-level driver scores, figures, and an analysis summary. No controlled-access sequencing files were used.

Key points

- Protein and phosphoprotein measurements added biologically interpretable information not captured by RNA alone.
- Fifty-two directed signaling edges passed FDR, prior-sign, and 200-bootstrap stability criteria.
- Early RNA-RPPA fusion classified four PAM50 subtypes with 0.867 balanced accuracy and 0.975 macro-AUROC.

Ethics statement: This study was a secondary analysis of de-identified, publicly available data. No new human participants were recruited and no new biospecimens were collected. Institutional requirements for secondary analysis of public de-identified data apply at the submitting institution.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

Acknowledgements: The authors acknowledge the patients and investigators of The Cancer Genome Atlas and the UCSC Xena team for public data harmonization.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers

Artificial Intelligence (AI) Use Disclosure: The authors declare that no artificial intelligence tools were used in the preparation of this manuscript.

References

- [1] Argelaguet R, Velten B, Arnol D, et al. (2018) Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. *Mol Syst Biol.* 2018;14(6):MSB178124. doi:10.15252/msb.20178124
- [2] Benjamini Y, Hochberg Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Ser B Stat Methodol.* 1995;57(1):289-300. doi:10.1111/j.2517-6161.1995.tb02031.x
- [3] Chakravarty D, Gao J, Phillips S, et al. OncoKB (2017) : A Precision Oncology Knowledge Base. *JCO Precis Oncol.* 2017;(1):1-16. doi:10.1200/PO.17.00011
- [4] Choi JM, Chae H. (2023) moBRCA-net: a breast cancer subtype classification framework based on multi-omics attention neural networks. *BMC Bioinformatics.* 2023;24(1):169. doi:10.1186/s12859-023-05273-5
- [5] Cox DR. (1972) Regression Models and Life-Tables. *J R Stat Soc Ser B Stat Methodol.* 1972;34(2):187-202. doi:10.1111/j.2517-6161.1972.tb00899.x
- [6] Curtis C, Shah SP, Chin SF, et al. (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature.* 2012;486(7403):346-352. doi:10.1038/nature10983
- [7] Gillespie M, Jassal B, Stephan R, et al. (2022) The reactome pathway knowledgebase 2022. *Nucleic Acids Res.* 2022;50(D1):D687-D692. doi:10.1093/nar/gkab1028
- [8] Huang HY, Lin YCD, Li J, et al. (2019) miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res.* Published online October 24, 2019;gkz896. doi:10.1093/nar/gkz896
- [9] Krug K, Jaehnig EJ, Satpathy S, et al. (2020) Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. *Cell.* 2020;183(5):1436-1456.e31. doi:10.1016/j.cell.2020.10.036
- [10] Kuijjer ML, Hsieh PH, Quackenbush J, Glass K, Ionescu R (2019) : single sample network inference in R. *BMC Cancer.* 2019;19(1):1003. doi:10.1186/s12885-019-6235-7
- [11] Norgeot B, Quer G, Beaulieu-Jones BK, et al. (2020) Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med.* 2020;26(9):1320-1324. doi:10.1038/s41591-020-1041-y
- [12] Parker JS, Mullins M, Cheang MCU, et al. (2009) Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *J Clin Oncol.* 2009;27(8):1160-1167. doi:10.1200/JCO.2008.18.1370
- [13] Patel JC, Shakyawar SK, Sethi S, Guda C. (2024) GAIN-BRCA: a graph-based AI-net framework for breast cancer subtype classification using multiomics data. Adebisi M, ed. *Bioinforma Adv.* 2024;5(1):vbaf116. doi:10.1093/bioadv/vbaf116
- [14] Pereira B, Chin SF, Rueda OM, et al. (2016) The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat Commun.* 2016;7(1):11479. doi:10.1038/ncomms11479
- [15] Peters J, Bühlmann P, Meinshausen N. (2016) Causal Inference by using Invariant Prediction: Identification and Confidence Intervals. *J R Stat Soc Ser B Stat Methodol.* 2016;78(5):947-1012. doi:10.1111/rssb.12167
- [16] Pratapa A, Jalihal AP, Law JN, Bharadwaj A, Murali TM. (2020) Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data. *Nat Methods.* 2020;17(2):147-154. doi:10.1038/s41592-019-0690-6
- [17] Ren Y, Gao Y, Du W, et al. (2024) Classifying breast cancer using multi-view graph neural network based on multi-omics data. *Front Genet.* 2024;15:1363896. doi:10.3389/fgene.2024.1363896
- [18] Rohart F, Gautier B, Singh A, Lê Cao KA. (2017) mixOmics: An R package for 'omics feature selection and multiple data integration. Schneidman D, ed. *PLOS Comput Biol.* 2017;13(11):e1005752. doi:10.1371/journal.pcbi.1005752
- [19] Singh A, Shannon CP, Gautier B, et al. (2019) DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays. Birol I, ed. *Bioinformatics.* 2019;35(17):3055-3062. doi:10.1093/bioinformatics/bty1054
- [20] Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. (2018) The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer.* 2018;18(11):696-705. doi:10.1038/s41568-018-0060-1
- [21] Szklarczyk D, Kirsch R, Koutrouli M, et al. (2023) The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic Acids Res.* 2023;51(D1):D638-D646. doi:10.1093/nar/gkac1000
- [22] The Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature.* 2012;490(7418):61-70. doi:10.1038/nature11412
- [23] Türe D, Valdeolivas A, Gul L, et al. (2021) Integrated intra- and intercellular signaling knowledge for multicellular omics analysis. *Mol Syst Biol.* 2021;17(3):MSB20209923. doi:10.15252/msb.20209923
- [24] Wang T, Shao W, Huang Z, et al. (2021) MOGONET integrates multi-omics data using graph convolutional networks allowing patient classification and biomarker identification. *Nat Commun.* 2021;12(1):3445. doi:10.1038/s41467-021-23774-w
- [25] Wolff RF, Moons KGM, Riley RD, et al. (2019) PROBAST: A Tool to Assess the Risk of Bias and Applicability of Prediction Model Studies. *Ann Intern Med.* 2019;170(1):51-58. doi:10.7326/M18-1376

[26] Yan H, Weng D, Li D, Gu Y, Ma W, Liu Q. (2024) Prior knowledge-guided multilevel graph neural network for tumor risk prediction and interpretation via multi-omics data integration. *Brief Bioinform.* 2024;25(3):bbae184. doi:10.1093/bib/bbae184

Appendix A. Complete technical specification

A1. Input files, provenance, and integrity

Modality	Filename	Parsed dimensions	SHA-256
RNA sequencing	TCGA.BRCA.sampleMap_HiSeqV2.gz	20,530 x 1,218	263bf67245cc4b9062676583c0ff0306f08471a26aafd5504037e1da22133746
Clinical	TCGA.BRCA.sampleMap_BRCA_clinicalMatrix	1,247 x 194	39eb3be0fb86e6a577bd2cc01502a7fa5a271e1e1cba294e9dc644ad99580d7f
RPPA	TCGA.BRCA.sampleMap_RPPA_RBN.gz	131 x 747	7823a7a375495f7e5f114d0c7ec8773e070cb9ac409785f98045d12ab71195e6

All identifiers were matched as complete TCGA sample barcodes. Primary tumour barcodes ending in sample-type code 01 formed the principal cohort. Five code-06 metastatic samples were excluded from training.

A2. RPPA analyte-to-gene mapping

RPPA analyte	Graph node	RNA gene(s)
1433EPSILON	YWHAЕ	YWHAЕ
4EBP1	EIF4EBP1	EIF4EBP1
4EBP1PS65	EIF4EBP1	EIF4EBP1
4EBP1PT37T46	EIF4EBP1	EIF4EBP1
53BP1	TP53BP1	TP53BP1
ACCP579	ACACA	ACACA
ACC1	ACACA	ACACA
AKT	AKT1	AKT1
AKTPS473	AKT1	AKT1
AKTPT308	AKT1	AKT1
AMPKALPHA	PRKAA1	PRKAA1
AMPKPT172	PRKAA1	PRKAA1
AR	AR	AR
ATM	ATM	ATM
BAK	BAK1	BAK1

BAX	BAX	BAX
BCL2	BCL2	BCL2
BCLXL	BCL2L1	BCL2L1
BECLIN	BECN1	BECN1
BETACATENIN	CTNNB1	CTNNB1
BID	BID	BID
BIM	BCL2L11	BCL2L11
CJUNPS73	JUN	JUN
CKIT	KIT	KIT
CMETPY1235	MET	MET
CMYC	MYC	MYC
CRAF	RAF1	RAF1
CRAFPS338	RAF1	RAF1
CASPASE7CLEAVEDD198	CASP7	CASP7
CAVEOLIN1	CAV1	CAV1
CD31	PECAM1	PECAM1
CD49B	ITGA2	ITGA2
CDK1	CDK1	CDK1
CHK1	CHEK1	CHEK1
CHK1PS345	CHEK1	CHEK1
CHK2	CHEK2	CHEK2
CHK2PT68	CHEK2	CHEK2
CIAP	BIRC2	BIRC2
CLAUDIN7	CLDN7	CLDN7
COLLAGENVI	COL6A1	COL6A1
CYCLINB1	CCNB1	CCNB1
CYCLIND1	CCND1	CCND1
CYCLINE1	CCNE1	CCNE1
DJ1	PARK7	PARK7
DVL3	DVL3	DVL3

ECADHERIN	CDH1	CDH1
EEF2	EEF2	EEF2
EEF2K	EEF2K	EEF2K
EGFRPY1068	EGFR	EGFR
EGFRPY1173	EGFR	EGFR
EIF4E	EIF4E	EIF4E
ERALPHA	ESR1	ESR1
ERALPHAPS118	ESR1	ESR1
ERK2	MAPK1	MAPK1
FIBRONECTIN	FN1	FN1
FOXO3A	FOXO3	FOXO3
GAB2	GAB2	GAB2
GATA3	GATA3	GATA3
GSK3ALPHABETA	GSK3A	GSK3A GSK3B
GSK3ALPHABETAPS21S9	GSK3A	GSK3A GSK3B
HER2	ERBB2	ERBB2
HER2PY1248	ERBB2	ERBB2
HER3	ERBB3	ERBB3
HER3PY1298	ERBB3	ERBB3
HSP70	HSPA1A	HSPA1A HSPA1B
IGFBP2	IGFBP2	IGFBP2
INPP4B	INPP4B	INPP4B
IRS1	IRS1	IRS1
JNK2	MAPK9	MAPK9
KU80	XRCC5	XRCC5
LKB1	STK11	STK11
LCK	LCK	LCK
MAPKPT202Y204	MAPK1	MAPK1 MAPK3
MEK1	MAP2K1	MAP2K1
MEK1PS217S221	MAP2K1	MAP2K1

MIG6	ERRFI1	ERRFI1
MRE11	MRE11A	MRE11A
MTOR	MTOR	MTOR
NCADHERIN	CDH2	CDH2
NFKBP65PS536	RELA	RELA
NF2	NF2	NF2
NOTCH1	NOTCH1	NOTCH1
PCADHERIN	CDH3	CDH3
P27	CDKN1B	CDKN1B
P27PT157	CDKN1B	CDKN1B
P38MAPK	MAPK14	MAPK14
P38PT180Y182	MAPK14	MAPK14
P53	TP53	TP53
P70S6K	RPS6KB1	RPS6KB1
P70S6KPT389	RPS6KB1	RPS6KB1
P90RSKPT359S363	RPS6KA1	RPS6KA1
PAXILLIN	PXN	PXN
PCNA	PCNA	PCNA
PDK1PS241	PDPK1	PDPK1
PEA15	PEA15	PEA15
PI3KP110ALPHA	PIK3CA	PIK3CA
PKCALPHA	PRKCA	PRKCA
PKCALPHAPS657	PRKCA	PRKCA
PKCDELTAPS664	PRKCD	PRKCD
PR	PGR	PGR
PRAS40PT246	AKT1S1	AKT1S1
PTEN	PTEN	PTEN
RAD50	RAD50	RAD50
RAD51	RAD51	RAD51
RBPS807S811	RB1	RB1

S6	RPS6	RPS6
S6PS235S236	RPS6	RPS6
S6PS240S244	RPS6	RPS6
SHCPY317	SHC1	SHC1
SMAD1	SMAD1	SMAD1
SMAD3	SMAD3	SMAD3
SMAD4	SMAD4	SMAD4
SRC	SRC	SRC
SRCPY416	SRC	SRC
SRCPY527	SRC	SRC
STAT3PY705	STAT3	STAT3
STAT5ALPHA	STAT5A	STAT5A
STATHMIN	STMN1	STMN1
SYK	SYK	SYK
TUBERIN	TSC2	TSC2
VEGFR2	KDR	KDR
XRCC1	XRCC1	XRCC1
YAP	YAP1	YAP1
YAPPS127	YAP1	YAP1
YB1	YBX1	YBX1
YB1PS102	YBX1	YBX1
JNKPT183Y185	MAPK8	MAPK8 MAPK9
PAI1	SERPINE1	SERPINE1
MTORPS2448	MTOR	MTOR
ASNS	ASNS	ASNS
EGFR	EGFR	EGFR

Composite or pan-protein antibodies were represented by a canonical graph node. When multiple RNA genes corresponded to an antibody, RNA abundance was averaged before standardization.

A3. Complete 132-edge signaling prior

Source	Target	Expected sign	Pathway
ERRFI1	EGFR	-1	RTK

EGFR	SHC1	1	RTK
EGFR	GAB2	1	RTK
EGFR	PIK3CA	1	RTK
EGFR	SRC	1	RTK
EGFR	STAT3	1	RTK
EGFR	RAF1	1	RTK
ERBB2	ERBB3	1	RTK
ERBB2	GAB2	1	RTK
ERBB2	PIK3CA	1	RTK
ERBB2	SRC	1	RTK
ERBB2	RAF1	1	RTK
ERBB3	GAB2	1	RTK
ERBB3	PIK3CA	1	RTK
MET	GAB2	1	RTK
MET	PIK3CA	1	RTK
MET	SRC	1	RTK
MET	RAF1	1	RTK
MET	STAT3	1	RTK
KIT	PIK3CA	1	RTK
KIT	SRC	1	RTK
KIT	RAF1	1	RTK
KDR	PIK3CA	1	RTK
KDR	SRC	1	RTK
KDR	RAF1	1	RTK
IRS1	PIK3CA	1	PI3K_AKT
IGFBP2	PIK3CA	1	PI3K_AKT
SYK	PIK3CA	1	PI3K_AKT
LCK	PIK3CA	1	PI3K_AKT
ESR1	PGR	1	Hormone
ESR1	GATA3	1	Hormone

ESR1	BCL2	1	Hormone
ESR1	CCND1	1	Hormone
ESR1	MYC	1	Hormone
AR	CCND1	1	Hormone
AR	MYC	1	Hormone
GATA3	CDH1	1	Hormone
PIK3CA	PDPK1	1	PI3K_AKT
PIK3CA	AKT1	1	PI3K_AKT
PDPK1	AKT1	1	PI3K_AKT
PDPK1	PRKCA	1	PI3K_AKT
PTEN	AKT1	-1	PI3K_AKT
INPP4B	AKT1	-1	PI3K_AKT
PARK7	AKT1	1	PI3K_AKT
AKT1	AKT1S1	1	PI3K_AKT
AKT1	MTOR	1	PI3K_AKT
AKT1	GSK3A	-1	PI3K_AKT
AKT1	FOXO3	-1	PI3K_AKT
AKT1	TSC2	-1	PI3K_AKT
AKT1	CDKN1B	-1	PI3K_AKT
STK11	PRKAA1	1	AMPK
PRKAA1	TSC2	1	AMPK
PRKAA1	MTOR	-1	AMPK
PRKAA1	ACACA	-1	AMPK
TSC2	MTOR	-1	PI3K_AKT
MTOR	EIF4EBP1	1	PI3K_AKT
MTOR	RPS6KB1	1	PI3K_AKT
RPS6KB1	RPS6	1	PI3K_AKT
RPS6KB1	EEF2K	1	PI3K_AKT
EIF4EBP1	EIF4E	-1	PI3K_AKT
SHC1	RAF1	1	MAPK

GAB2	RAF1	1	MAPK
SRC	RAF1	1	MAPK
RAF1	MAP2K1	1	MAPK
MAP2K1	MAPK1	1	MAPK
MAPK1	RPS6KA1	1	MAPK
MAPK1	JUN	1	MAPK
MAPK1	MYC	1	MAPK
MAPK1	CCND1	1	MAPK
MAPK9	JUN	1	Stress_MAPK
MAPK14	JUN	1	Stress_MAPK
MAPK14	STAT3	1	Stress_MAPK
RPS6KA1	YBX1	1	MAPK
SRC	STAT3	1	SRC_STAT
SRC	CTNNB1	1	SRC_STAT
SRC	PXN	1	Adhesion
SRC	YAP1	1	Hippo
CAV1	SRC	1	Adhesion
FN1	ITGA2	1	Adhesion
ITGA2	SRC	1	Adhesion
CDH2	PXN	1	Adhesion
STAT3	BCL2L1	1	SRC_STAT
STAT3	MYC	1	SRC_STAT
STAT3	CCND1	1	SRC_STAT
STAT3	SERPINE1	1	SRC_STAT
STAT5A	BCL2	1	SRC_STAT
STAT5A	CCND1	1	SRC_STAT
RELA	BCL2L1	1	NFkB
RELA	BIRC2	1	NFkB
RELA	SERPINE1	1	NFkB
CDH1	CTNNB1	-1	WNT

DVL3	CTNNB1	1	WNT
CTNNB1	MYC	1	WNT
CTNNB1	CCND1	1	WNT
NF2	YAP1	-1	Hippo
YAP1	MYC	1	Hippo
YAP1	CCND1	1	Hippo
YAP1	BCL2L1	1	Hippo
YAP1	SERPINE1	1	Hippo
SMAD1	SMAD4	1	TGFb
SMAD3	SMAD4	1	TGFb
SMAD4	FN1	1	TGFb
SMAD4	CDH2	1	TGFb
SMAD4	SERPINE1	1	TGFb
NOTCH1	MYC	1	Notch
NOTCH1	CCND1	1	Notch
NOTCH1	BCL2	1	Notch
ATM	CHEK2	1	DNA_damage
ATM	TP53BP1	1	DNA_damage
ATM	MRE11A	1	DNA_damage
ATM	RAD50	1	DNA_damage
CHEK2	TP53	1	DNA_damage
CHEK1	TP53	1	DNA_damage
TP53BP1	TP53	1	DNA_damage
TP53	BAX	1	Apoptosis
TP53	BAK1	1	Apoptosis
TP53	BCL2L11	1	Apoptosis
TP53	BID	1	Apoptosis
TP53	CDKN1B	1	DNA_damage
BCL2L11	BAX	1	Apoptosis
BCL2L11	BAK1	1	Apoptosis

BAX	CASP7	1	Apoptosis
BAK1	CASP7	1	Apoptosis
BID	CASP7	1	Apoptosis
BCL2	CASP7	-1	Apoptosis
BCL2L1	CASP7	-1	Apoptosis
MYC	CCNE1	1	Cell_cycle
MYC	PCNA	1	Cell_cycle
CCNB1	CDK1	1	Cell_cycle
CCND1	RB1	1	Cell_cycle
CCNE1	RB1	1	Cell_cycle
CDK1	RB1	1	Cell_cycle

Expected sign was +1 for activating relations and -1 for inhibitory relations. Only measured nodes were included. The graph was acyclic to permit total-effect propagation.

A4. Statistical and computational definitions

Correlation analysis: two-sided Spearman rank correlation for each matched RNA-analyte pair; Benjamini-Hochberg FDR across 131 analytes.

Structural equations: standardized target protein regressed on the target transcript and all prespecified parent proteins. HC3 robust standard errors were used for main inference.

Edge retention: $q < 0.05$, estimated coefficient sign equal to prior sign, bootstrap sign stability ≥ 0.80 , and 95% percentile bootstrap interval excluding zero.

Patient score: role-oriented protein activity and structural residual were combined at weights 0.6 and 0.4, then multiplied by normalized total downstream influence.

Subtype validation: repeated stratified 5-fold cross-validation with 10 repeats. Every molecular transformation was fitted inside the outer training fold.

Survival analysis: Cox regression adjusted for standardized age, standardized stage, and PAM50 indicators. Pathway tests were FDR corrected.

Random seed: 20260722 for all bootstrap, cross-validation, and permutation procedures.

Appendix B. Complete empirical result tables

B1. All 131 matched RNA-RPPA correlations

RPPA feature	Gene	Spearman rho	P	q	q<0.05	Assay type
ERALPHA	ESR1	0.889	6.57e-253	8.61e-251	1	total_or_cleaved

BCL2	BCL2	0.841	1.01e-199	6.61e-198	1	total_or_cleaved
GATA3	GATA3	0.785	2.58e-156	1.13e-154	1	total_or_cleaved
PR	PGR	0.765	1.53e-143	5.02e-142	1	total_or_cleaved
INPP4B	INPP4B	0.762	7.52e-142	1.97e-140	1	total_or_cleaved
IGFBP2	IGFBP2	0.748	4.61e-134	1.01e-132	1	total_or_cleaved
CYCLINB1	CCNB1	0.732	1.24e-125	2.33e-124	1	total_or_cleaved
ASNS	ASNS	0.704	4.23e-112	6.93e-111	1	total_or_cleaved
CHK2	CHEK2	0.675	7.14e-100	1.04e-98	1	total_or_cleaved
CKIT	KIT	0.674	1.70e-99	2.23e-98	1	total_or_cleaved
AR	AR	0.670	5.50e-98	6.55e-97	1	total_or_cleaved
CYCLINE1	CCNE1	0.669	2.11e-97	2.30e-96	1	total_or_cleaved
4EBP1	EIF4EBP1	0.653	3.18e-91	3.20e-90	1	total_or_cleaved
ERALPHAPS118	ESR1	0.647	3.54e-89	3.31e-88	1	phospho
HER2	ERBB2	0.634	9.73e-85	8.49e-84	1	total_or_cleaved
GAB2	GAB2	0.615	1.47e-78	1.21e-77	1	total_or_cleaved
P70S6K	RPS6KB1	0.608	3.91e-76	3.02e-75	1	total_or_cleaved
HER2PY1248	ERBB2	0.598	3.08e-73	2.24e-72	1	phospho
IRS1	IRS1	0.598	4.52e-73	3.12e-72	1	total_or_cleaved
LCK	LCK	0.582	1.97e-68	1.29e-67	1	total_or_cleaved
P27	CDKN1B	0.576	9.85e-67	6.14e-66	1	total_or_cleaved
JNK2	MAPK9	0.562	5.76e-63	3.43e-62	1	total_or_cleaved
SYK	SYK	0.559	2.50e-62	1.43e-61	1	total_or_cleaved
SRC	SRC	0.556	1.48e-61	8.06e-61	1	total_or_cleaved
YAPPS127	YAP1	0.550	5.30e-60	2.78e-59	1	phospho
ACCP579	ACACA	0.540	2.08e-57	1.05e-56	1	phospho
ECADHERIN	CDH1	0.537	1.51e-56	7.33e-56	1	total_or_cleaved
PKCALPHAPS657	PRKCA	0.518	4.59e-52	2.15e-51	1	phospho
ACC1	ACACA	0.516	1.23e-51	5.54e-51	1	total_or_cleaved
STAT5ALPHA	STAT5A	0.493	1.04e-46	4.54e-46	1	total_or_cleaved
EGFR	EGFR	0.491	3.24e-46	1.37e-45	1	total_or_cleaved

CAVEOLIN1	CAV1	0.475	4.63e-43	1.90e-42	1	total_or_cleaved
PKCALPHA	PRKCA	0.471	2.65e-42	1.05e-41	1	total_or_cleaved
FIBRONECTIN	FN1	0.471	2.83e-42	1.09e-41	1	total_or_cleaved
SMAD1	SMAD1	0.469	8.52e-42	3.19e-41	1	total_or_cleaved
53BP1	TP53BP1	0.463	1.02e-40	3.70e-40	1	total_or_cleaved
EEF2K	EEF2K	0.450	3.42e-38	1.21e-37	1	total_or_cleaved
BIM	BCL2L11	0.447	1.11e-37	3.84e-37	1	total_or_cleaved
DJ1	PARK7	0.432	4.42e-35	1.49e-34	1	total_or_cleaved
PTEN	PTEN	0.429	1.60e-34	5.24e-34	1	total_or_cleaved
CD49B	ITGA2	0.427	2.55e-34	8.16e-34	1	total_or_cleaved
CRAF	RAF1	0.426	5.38e-34	1.68e-33	1	total_or_cleaved
CLAUDIN7	CLDN7	0.421	2.78e-33	8.48e-33	1	total_or_cleaved
YAP	YAP1	0.418	8.42e-33	2.51e-32	1	total_or_cleaved
HER3	ERBB3	0.412	1.03e-31	3.00e-31	1	total_or_cleaved
PEA15	PEA15	0.409	2.56e-31	7.28e-31	1	total_or_cleaved
AKT	AKT1	0.407	6.35e-31	1.77e-30	1	total_or_cleaved
NOTCH1	NOTCH1	0.406	7.10e-31	1.94e-30	1	total_or_cleaved
PCNA	PCNA	0.387	7.24e-28	1.94e-27	1	total_or_cleaved
MEK1	MAP2K1	0.385	1.38e-27	3.61e-27	1	total_or_cleaved
4EBP1PS65	EIF4EBP1	0.383	2.48e-27	6.36e-27	1	phospho
DVL3	DVL3	0.381	4.27e-27	1.08e-26	1	total_or_cleaved
4EBP1PT37T46	EIF4EBP1	0.378	1.34e-26	3.30e-26	1	phospho
SMAD3	SMAD3	0.358	8.23e-24	2.00e-23	1	total_or_cleaved
CHK2PT68	CHEK2	0.356	1.27e-23	3.03e-23	1	phospho
CYCLIND1	CCND1	0.354	2.89e-23	6.76e-23	1	total_or_cleaved
PAI1	SERPINE1	0.346	2.76e-22	6.35e-22	1	total_or_cleaved
PCADHERIN	CDH3	0.345	3.87e-22	8.64e-22	1	total_or_cleaved
PAXILLIN	PXN	0.345	3.89e-22	8.64e-22	1	total_or_cleaved
ATM	ATM	0.344	4.97e-22	1.09e-21	1	total_or_cleaved
NF2	NF2	0.342	8.06e-22	1.73e-21	1	total_or_cleaved

EIF4E	EIF4E	0.342	9.67e-22	2.04e-21	1	total_or_cleaved
STAT3PY705	STAT3	0.341	1.01e-21	2.10e-21	1	phospho
CDK1	CDK1	0.338	2.75e-21	5.64e-21	1	total_or_cleaved
CMYC	MYC	0.318	5.91e-19	1.19e-18	1	total_or_cleaved
GSK3ALPHABETA	GSK3A	0.317	9.14e-19	1.81e-18	1	total_or_cleaved
FOXO3A	FOXO3	0.316	1.18e-18	2.30e-18	1	total_or_cleaved
PDK1PS241	PDPK1	0.305	1.97e-17	3.80e-17	1	phospho
AMPKALPHA	PRKAA1	0.304	2.22e-17	4.22e-17	1	total_or_cleaved
CHK1PS345	CHEK1	0.293	4.19e-16	7.84e-16	1	phospho
AMPKPT172	PRKAA1	0.285	2.41e-15	4.45e-15	1	phospho
P53	TP53	0.276	1.77e-14	3.21e-14	1	total_or_cleaved
RAD51	RAD51	0.276	2.14e-14	3.84e-14	1	total_or_cleaved
XRCC1	XRCC1	0.271	5.78e-14	1.02e-13	1	total_or_cleaved
P90RSKPT359S363	RPS6KA1	0.267	1.31e-13	2.29e-13	1	phospho
MIG6	ERRFI1	0.259	8.17e-13	1.41e-12	1	total_or_cleaved
STATHMIN	STMN1	0.256	1.34e-12	2.28e-12	1	total_or_cleaved
EGFRPY1068	EGFR	0.254	2.16e-12	3.63e-12	1	phospho
CJUNPS73	JUN	0.239	4.66e-11	7.72e-11	1	phospho
YB1PS102	YBX1	0.236	7.68e-11	1.26e-10	1	phospho
RAD50	RAD50	0.234	1.02e-10	1.65e-10	1	total_or_cleaved
LKB1	STK11	0.233	1.33e-10	2.12e-10	1	total_or_cleaved
COLLAGENVI	COL6A1	0.232	1.50e-10	2.37e-10	1	total_or_cleaved
EGFRPY1173	EGFR	0.230	2.23e-10	3.48e-10	1	phospho
ERK2	MAPK1	0.226	5.03e-10	7.75e-10	1	total_or_cleaved
CASPASE7CLEAVEDDD198	CASP7	0.215	3.22e-09	4.91e-09	1	total_or_cleaved
PKCDELTAPS664	PRKCD	0.196	7.88e-08	1.19e-07	1	phospho
PI3KP110ALPHA	PIK3CA	0.184	4.32e-07	6.43e-07	1	total_or_cleaved
BCLXL	BCL2L1	0.175	1.53e-06	2.26e-06	1	total_or_cleaved
1433EPSILON	YWHAE	0.168	4.23e-06	6.16e-06	1	total_or_cleaved
CIAP	BIRC2	0.160	1.16e-05	1.68e-05	1	total_or_cleaved

CD31	PECAM1	0.146	6.45e-05	9.18e-05	1	total_or_cleaved
TUBERIN	TSC2	0.145	7.50e-05	1.06e-04	1	total_or_cleaved
SHCPY317	SHC1	0.143	9.51e-05	1.33e-04	1	phospho
AKTPT308	AKT1	0.143	9.78e-05	1.35e-04	1	phospho
BAX	BAX	0.142	1.01e-04	1.37e-04	1	total_or_cleaved
AKTPS473	AKT1	0.140	1.37e-04	1.86e-04	1	phospho
SMAD4	SMAD4	0.136	1.96e-04	2.62e-04	1	total_or_cleaved
CRAFPS338	RAF1	0.120	0.001	0.001	1	phospho
P38MAPK	MAPK14	0.120	0.001	0.001	1	total_or_cleaved
MAPKPT202Y204	MAPK1	0.113	0.002	0.003	1	phospho
KU80	XRCC5	0.109	0.003	0.004	1	total_or_cleaved
RBPS807S811	RB1	0.103	0.005	0.006	1	phospho
CHK1	CHEK1	0.092	0.012	0.015	1	total_or_cleaved
BETACATENIN	CTNNB1	0.092	0.012	0.015	1	total_or_cleaved
S6	RPS6	0.089	0.015	0.019	1	total_or_cleaved
GSK3ALPHABETAPS21S9	GSK3A	0.088	0.016	0.020	1	phospho
CMETPY1235	MET	0.087	0.018	0.022	1	phospho
BID	BID	0.083	0.024	0.028	1	total_or_cleaved
MTOR	MTOR	0.078	0.034	0.041	1	total_or_cleaved
JNKPT183Y185	MAPK8	0.070	0.058	0.068	0	phospho
YB1	YBX1	0.069	0.060	0.070	0	total_or_cleaved
MTORPS2448	MTOR	0.068	0.062	0.072	0	phospho
NCADHERIN	CDH2	0.065	0.076	0.087	0	total_or_cleaved
SRCPY527	SRC	0.059	0.105	0.119	0	phospho
NFKBP65PS536	RELA	0.059	0.111	0.124	0	phospho
PRAS40PT246	AKT1S1	0.041	0.265	0.285	0	phospho
EEF2	EEF2	0.031	0.393	0.411	0	total_or_cleaved
P38PT180Y182	MAPK14	0.028	0.441	0.459	0	phospho
SRCPY416	SRC	0.019	0.600	0.618	0	phospho
BAK	BAK1	0.017	0.635	0.650	0	total_or_cleaved

HER3PY1298	ERBB3	0.004	0.911	0.918	0	phospho
BECLIN	BECN1	0.002	0.946	0.946	0	total_or_cleaved
MRE11	MRE11A	-0.008	0.827	0.840	0	total_or_cleaved
VEGFR2	KDR	-0.035	0.340	0.359	0	total_or_cleaved
S6PS235S236	RPS6	-0.039	0.291	0.310	0	phospho
P70S6KPT389	RPS6KB1	-0.045	0.225	0.244	0	phospho
P27PT157	CDKN1B	-0.045	0.217	0.237	0	phospho
MEK1PS217S221	MAP2K1	-0.046	0.209	0.230	0	phospho
HSP70	HSPA1A	-0.053	0.149	0.165	0	total_or_cleaved
S6PS240S244	RPS6	-0.112	0.002	0.003	1	phospho

B2. All 52 bootstrap-stable directed edges

Source	Target	Sign	Beta	q	Boot low	CI high	Boot high	CI low	Sign stability	Target R2
ESR1	GATA3	1	0.481	3.75e-70	0.436	0.533	1.000	0.779		
AKT1	AKT1S1	1	0.533	4.11e-53	0.474	0.592	1.000	0.287		
MAP2K1	MAPK1	1	0.497	4.51e-36	0.440	0.560	1.000	0.267		
MET	RAF1	1	0.444	7.00e-33	0.367	0.527	1.000	0.452		
CHEK2	TP53	1	0.448	7.86e-29	0.382	0.523	1.000	0.534		
SMAD4	CDH2	1	0.495	7.17e-26	0.414	0.592	1.000	0.255		
MTOR	RPS6KB1	1	0.297	1.92e-23	0.252	0.355	1.000	0.396		
DVL3	CTNNB1	1	0.223	1.22e-22	0.182	0.268	1.000	0.687		
CHEK1	TP53	1	0.335	3.35e-21	0.269	0.396	1.000	0.534		
NOTCH1	MYC	1	0.405	3.48e-18	0.311	0.493	1.000	0.321		
BCL2	CASP7	-1	-0.303	1.06e-16	-0.370	-0.237	1.000	0.190		
EGFR	SHC1	1	0.481	3.82e-16	0.367	0.592	1.000	0.248		
RPS6KA1	YBX1	1	0.298	1.74e-15	0.224	0.352	1.000	0.134		
CCNB1	CDK1	1	0.365	1.94e-15	0.280	0.441	1.000	0.148		
TP53	CDKN1B	1	0.325	1.23e-14	0.243	0.390	1.000	0.252		
PRKAA1	TSC2	1	0.282	5.25e-13	0.202	0.344	1.000	0.122		
NOTCH1	CCND1	1	0.346	1.06e-12	0.240	0.435	1.000	0.370		
RPS6KB1	RPS6	1	0.277	1.61e-12	0.206	0.354	1.000	0.077		

ATM	RAD50	1	0.277	6.06e-11	0.212	0.352	1.000	0.138
SMAD3	SMAD4	1	0.296	9.12e-11	0.212	0.380	1.000	0.142
ERBB2	ERBB3	1	0.243	8.89e-10	0.180	0.317	1.000	0.154
SMAD4	FN1	1	0.263	2.98e-09	0.181	0.356	1.000	0.264
EGFR	SRC	1	0.316	9.78e-09	0.236	0.442	1.000	0.357
ATM	TP53BP1	1	0.268	2.15e-08	0.179	0.362	1.000	0.259
SRC	YAP1	1	0.202	7.00e-08	0.128	0.263	1.000	0.355
GATA3	CDH1	1	0.153	1.01e-06	0.105	0.206	1.000	0.486
SRC	STAT3	1	0.205	1.06e-06	0.123	0.279	1.000	0.297
BAX	CASP7	1	0.226	1.18e-06	0.137	0.322	1.000	0.190
INPP4B	AKT1	-1	-0.189	1.61e-06	-0.255	-0.128	1.000	0.159
MAPK14	STAT3	1	0.178	1.67e-06	0.107	0.249	1.000	0.297
ESR1	PGR	1	0.135	1.76e-05	0.074	0.192	1.000	0.552
EGFR	STAT3	1	0.264	2.82e-05	0.172	0.386	1.000	0.297
YAP1	SERPINE1	1	0.167	7.23e-05	0.088	0.233	1.000	0.184
CAV1	SRC	1	0.161	7.55e-05	0.088	0.228	1.000	0.357
SMAD1	SMAD4	1	0.159	1.03e-04	0.090	0.233	1.000	0.142
MAPK1	MYC	1	0.149	1.08e-04	0.091	0.211	1.000	0.321
PDPK1	AKT1	1	0.146	5.93e-04	0.061	0.222	1.000	0.159
MET	SRC	1	0.141	6.47e-04	0.053	0.213	1.000	0.357
AKT1	MTOR	1	0.128	7.16e-04	0.058	0.200	1.000	0.253
PIK3CA	PDPK1	1	0.201	0.001	0.103	0.325	1.000	0.141
BCL2L11	BAX	1	0.139	0.001	0.062	0.225	1.000	0.080
SMAD4	SERPINE1	1	0.134	0.002	0.055	0.213	1.000	0.184
MAPK1	JUN	1	0.132	0.005	0.057	0.207	1.000	0.100
ATM	CHEK2	1	0.082	0.005	0.026	0.131	1.000	0.488
MYC	PCNA	1	0.104	0.010	0.024	0.171	0.990	0.198
KDR	PIK3CA	1	0.112	0.012	0.030	0.180	1.000	0.147
NF2	YAP1	-1	-0.086	0.012	-0.147	-0.012	0.995	0.355
MTOR	EIF4EBP1	1	0.069	0.023	0.022	0.126	0.995	0.442

CDK1	RB1	1	0.102	0.032	0.030	0.181	0.995	0.089
MYC	CCNE1	1	0.057	0.045	0.001	0.107	0.975	0.611
MET	GAB2	1	0.058	0.046	0.003	0.100	0.985	0.539
ESR1	BCL2	1	0.062	0.047	0.008	0.119	0.990	0.685

B3. Repeated cross-validation summary

Model	mean_accuracy	mean_balanced_accuracy	mean_macro_auc	mean_macro_f1	ci_low_w_accuracy	ci_low_balanced_accuracy	ci_low_macro_auc	ci_low_macro_f1	ci_high_w_accuracy	ci_high_balanced_accuracy	ci_high_macro_auc	ci_high_macro_f1
Early_fusion	0.880	0.867	0.975	0.871	0.874	0.863	0.973	0.867	0.885	0.872	0.976	0.875
Network_augmented	0.876	0.861	0.972	0.865	0.872	0.855	0.971	0.860	0.879	0.868	0.973	0.870
Network_driver	0.688	0.679	0.875	0.667	0.682	0.668	0.871	0.657	0.695	0.689	0.879	0.678
RNA	0.874	0.862	0.974	0.862	0.870	0.858	0.973	0.858	0.877	0.867	0.976	0.866
RPPA	0.778	0.779	0.929	0.774	0.770	0.771	0.926	0.767	0.785	0.786	0.932	0.782

Confidence intervals were calculated from the ten repeat-level means. Full fold-level metrics are supplied as a machine-readable CSV.

B4. Top-five patient driver frequency by subtype

Subtype	Gene	Top-five frequency	N
LumA	ESR1	0.312	330
LumA	CHEK2	0.285	330
LumA	ATM	0.221	330
LumA	MAP2K1	0.221	330
LumA	AKT1	0.212	330
LumA	CHEK1	0.209	330
LumA	EGFR	0.209	330
LumA	MET	0.182	330
LumA	MTOR	0.182	330

LumA	MAPK1	0.179	330
LumA	NOTCH1	0.179	330
LumA	BCL2	0.167	330
LumB	ESR1	0.414	162
LumB	CHEK1	0.309	162
LumB	MTOR	0.253	162
LumB	CCNB1	0.235	162
LumB	ATM	0.222	162
LumB	CHEK2	0.198	162
LumB	MET	0.185	162
LumB	EGFR	0.154	162
LumB	SRC	0.142	162
LumB	RPS6KB1	0.142	162
LumB	AKT1	0.136	162
LumB	BCL2	0.136	162
Basal	NOTCH1	0.483	120
Basal	CCNB1	0.458	120
Basal	INPP4B	0.442	120
Basal	MET	0.342	120
Basal	SRC	0.300	120
Basal	AKT1	0.242	120
Basal	EGFR	0.200	120
Basal	CCNE1	0.200	120
Basal	KIT	0.192	120
Basal	MAP2K1	0.167	120
Basal	ATM	0.142	120
Basal	MTOR	0.125	120
Her2	ERBB2	0.607	61
Her2	EGFR	0.525	61
Her2	MET	0.311	61

Her2	AKT1	0.262	61
Her2	MTOR	0.246	61
Her2	INPP4B	0.213	61
Her2	CCNB1	0.197	61
Her2	NOTCH1	0.180	61
Her2	BCL2L11	0.164	61
Her2	MAP2K1	0.148	61
Her2	ATM	0.148	61
Her2	MAPK1	0.148	61

B5. Adjusted pathway survival models

Pathway	Hazard ratio	95% CI low	95% CI high	P	q
PI3K_AKT	0.789	0.609	1.022	0.073	0.449
AMPK	0.800	0.624	1.027	0.080	0.449
NFkB	0.806	0.620	1.047	0.106	0.449
Stress_MAPK	0.803	0.609	1.060	0.122	0.449
Notch	0.822	0.633	1.068	0.142	0.449
Global_driver_burden	0.809	0.602	1.086	0.158	0.449
Apoptosis	0.893	0.698	1.142	0.366	0.599
TGFb	0.886	0.676	1.162	0.381	0.599
MAPK	0.891	0.687	1.157	0.387	0.599
Hormone	0.886	0.662	1.187	0.418	0.599
Adhesion	0.900	0.698	1.162	0.419	0.599
DNA_damage	1.109	0.860	1.431	0.425	0.599
Cell_cycle	0.890	0.654	1.211	0.458	0.599
Hippo	0.925	0.723	1.183	0.533	0.647
SRC_STAT	0.940	0.734	1.205	0.627	0.710
RTK	1.047	0.824	1.331	0.704	0.748
WNT	0.997	0.765	1.299	0.980	0.980

Hazard ratios are per one-standard-deviation increase in pathway score and are adjusted for age, stage, and PAM50 subtype. No q value was <0.05.

B6. Reproducibility inventory

File

analysis_summary.json

run_real_tcga_brca_analysis.py

Table_1_cohort_characteristics.csv

Table_2_RNA_RPPA_concordance_all.csv

Table_3_structural_edges_all.csv

Table_3b_bootstrap_stable_edges.csv

Table_5_repeated_CV_fold_metrics.csv

Table_6_subtype_driver_contrasts.csv

Table_7_adjusted_pathway_survival.csv

Appendix_B_patient_top5_putative_drivers.csv

Appendix_B_patient_driver_score_matrix.tsv.gz

figures/

Contents

Machine-readable headline results and negative-control summaries

Executable Python workflow using the three input matrices

Cohort and subtype characteristics

All analyte-level concordance estimates

All tested structural edges

Stable edge set used for patient scores

All repeated cross-validation fold metrics

All subtype-specific score contrasts

Adjusted pathway Cox models

Top five candidate drivers for each tumour

Complete patient-by-node score matrix

Seven publication-resolution figures generated from observed results