
| RESEARCH ARTICLE**Cross-Cohort Analytical Validation of Mass-Spectrometry Proteomic and Phosphoproteomic Signalling Signatures in Molecularly Stratified Cancer Cohorts****Shaila Shahid***Department of Biochemistry, Bangladesh Medical University***Corresponding Author:** Shaila Shahid, **E-mail:** shaila07shahid@gmail.com

| ABSTRACT

Background: The functional implications of cancer genomes can be measured using mass-spectrometry proteomics, but for clinical translation, the ability to measure the same signaling across independent disease cohorts needs to be demonstrated. Methods: Harmonized CPTAC breast cancer data (n=122) were used for discovery, with ovarian cancer (n=83) and lung adenocarcinoma (n=110) for external validation. After cross-cohort filtering, 7,053 proteins and 3,822 phosphosites were analyzed. Three models (elastic-net regression) were used to predict the activity of MAPK/RTK, PI3K-AKT-mTOR, and cell-cycle activity, which were assessed by correlation, RMSE, R^2 , bootstrap confidence intervals, and AUC. Results: Pooled external correlations of MAPK/RTK, PI3K-AKT-mTOR, and cell cycle were 0.603, 0.561, and 0.782, respectively. External AUCs ranged from 0.810 to 0.956. However, integration was not always superior to the best single modality. Bootstrap-adjusted Rand index for the two stable states identified by the clustering was 0.974: signaling-low (n=78) and co-activated (n=114). LAMC2, GRB10-S104, RRM2, and CEP152-S1461 were some of the key contributors. Conclusions: Computational validity and development of targeted assays were supported by the validation of locked mass-spectrometry signatures, which signalled states across cancers, particularly those involving cell-cycle activity, with prospective clinical applications pending.

| KEYWORDS

Proteogenomics; Phosphoproteomics; Mass spectrometry; Signaling; Elastic net; Molecular stratification; Cancer biomarkers.

| ARTICLE INFORMATION**ACCEPTED:** August 20, 2024 **PUBLISHED:** December 28, 2024 **DOI:** <https://doi.org/10.61424/ijmhr.v2i1.1106>

1. Introduction

Genomic profiling has revolutionized the classification of cancer, but most cancer therapies are based on the biochemical state, which is not fully captured in the DNA sequence. Whether or not a genomic lesion is functionally expressed depends on protein abundance, post-translational modification, complex formation and pathway feedback. However, the Clinical Proteogenomic Tumor Analysis Consortium (CPTAC) has established highly characterized proteogenomic cohorts that relate genomic variation to RNA and protein measurements and phosphosite intensities for a variety of tumors (1-10). These analyses, now possible through pan-cancer harmonization, reveal that more than one thousand tumors can be analyzed on common functional states and that distinct oncogenic drivers can converge on shared functional states (1,2).

Phosphorylation is particularly informative, as it captures the dynamic regulation of kinase networks, and is also analytically challenging. Phosphosite matrices are sparse; identities of the sites must be explicitly stated; the same

Copyright: © 2024 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.

kinase can act on several substrates in different contexts. Signaling inference has become more systematic with the methods of PTM-SEA, KSEA, INKA, RoKAI, KSTAR and related methods (11-19). Inferred activities, however, are highly dependent on the context of a biological system and feature coverage and the substrate library used for the benchmarking (18,19). A clinically plausible measurement system should thus show that a signal acquired by one group of specimens is preserved when it is translated to new specimens and disease situations.

Typical analytical validation studies in clinical laboratory science involve precision, trueness, selectivity, linearity, reportable range, limits of detection and quantification, calibration verification, stability and interference testing studies (20-24). While the public processed omics matrices can provide many of the pieces of that puzzle, they cannot provide all the pieces because the raw instrument runs, replicate injections, calibrators, and standardized reference materials are not equally available to all public processed matrices. They can, however, ask a unique and essential question: can a measurement of the desired complexity and discriminatory power be reproduced at the level of biological ordering and state discrimination, in independently acquired cohorts? This is referred to as computational cross-cohort analytical validation.

We predicted that proteomic and phosphoproteomic signatures developed to predict orthogonal transcriptomic pathway activity in breast cancer will hold quantitative validity in ovarian high-grade serous carcinoma (HGSC) and lung adenocarcinoma (LCA). A priori, three pathways were selected for their pathway activity: MAPK/RTK signaling, PI3K-AKT-mTOR signaling, and cell-cycle activity. These axes play a key role in oncogenesis, targeted therapy, proliferation, and resistance and are depicted by complementary PROGENy and Hallmark transcriptional footprints (25-29). We also assessed whether a consistent increase in transportability was observed with the integration of protein and phosphosite features, whether locked scores distinguished between the high and low molecular state and whether external tumors assembled stable cross-cancer signaling classes.

2. Materials and methods

2.1 Study design and reporting framework

The analysis was a retrospective, computational, secondary analysis of de-identified open molecular data. A single discovery cohort and two external validation cohorts were used for the design. For all, feature selection and hyperparameter selection were limited to the breast-cancer discovery cohort. The ovarian and lung cohorts were not included in the selection of features, tuning of the complexity of the model, definition of coefficients, or in the selection of pathway targets. Analyses used a fixed random seed of 42. The full executable workflow as well as machine-readable tables of results are provided with the manuscript.

2.2 Data sources and cohort eligibility

The pan-cancer processed tumor proteome, phosphoproteome, and phenotype matrices were retrieved from the LinkedOmics distribution of the pan-cancer processed CPTAC resource (1). There were 122 breast cancers in the discovery cohort. External validation consisted of 83 ovarian high-grade serous carcinomas and 110 lung adenocarcinomas (total, $n = 315$). All tumours were eligible in the distributed matrices. The composite transcriptomic outcomes were not available for a single ovarian tumor, so 82 ovarian tumor were available for supervised validation. The main cohort publications outline the biospecimen collection, mass-spectrometry protocols, quality control and molecular results specific to the disease (4-6).

2.3 Molecular data and feature harmonization

The actual abundance values were used as processed values without any back transformation or retranslation of the absolute units. Protein features were identified using Ensemble gene identifiers. Phosphoproteomic features preserved the whole distributed identifier, including gene, protein isoform, residue, local peptide sequence and multiplicity field. Features had to be the same in all three cohorts and present in 70% or more of the tumors in each cohort. This "lite" approach minimized the number of features available on each platform and cohort prior to evaluating association with the outcome.

Non-finite values were coded as missing, missing values were replaced by the feature median and each feature was standardized to have a mean of zero and a variance of one within the cohort. The imputation and scaling were done separately within each cohort, so that distributional information in the validation cohort does not affect discovery coefficients. Due to the robustness required and that the analysis focussed on the relative sample ordering and not absolute concentration, median imputation was selected (30-33).

2.4 Orthogonal transcriptomic signaling outcomes

The response variables were built from transcriptomic phenotype scores provided with the pan-cancer resource as a means to avoid circular validation. The scores of these components were standardized within the cohort and then averaged and re-standardized. The PROGENy MAPK pathway was enriched with the MAPK/RTK pathway, along with PROGENy EGFR and HALLMARK_KRAS_SIGNALING_UP. Combined PROGENy PI3K, HALLMARK_PI3K_AKT_MTOR_SIGNALING and HALLMARK_MTORC1_SIGNALING. The combined pathways were Cell-cycle activity, HALLMARK_E2F_TARGETS, HALLMARK_G2M_CHECKPOINT and HALLMARK_MITOTIC_SPINDLE. While Hallmark sets group coherent biological programs (25-29), PROGENy summarizes perturbation-responsive transcriptional footprints. The composite outcomes were thus the molecular states rather than clinical endpoints.

2.5 Discovery feature selection and elastic-net signatures

Among components of the breast-cancer discovery cohort, features were sorted by the absolute Pearson correlation with each pathway outcome. In the cases of protein-only models, the top 120 proteins were retained, while 180 phosphosites were retained in the case of phosphoprotein-only models. The top 80 proteins and top 120 phosphosites were retained in the integrated models, where the modality prefixes indicate the source of the features. Extensive correlation among proteins and phosphosites made elastic-net linear regression the preferred method (34-37). For the discovery cohort, ten-fold cross-validation was performed at λ ratios of 0.1, 0.3, 0.5, 0.7 and 0.9 with 50 logarithmically spaced alpha values ranging from 0.001 to 1.0. The hyperparameters and coefficients that were selected were then frozen.

2.6 External validation endpoints

The main outcome was the correlation between the locked proteomic signature score and orthogonal transcriptomic pathway score in each external cohort (Spearman correlation). Secondary continuous endpoints were the R^2 and root-mean-square error (RMSE). Two-sided Spearman P-values were provided and 95% CIs were computed from 1000 nonparametric bootstrap re-samples. External Spearman correlations were Fisher z transformed, weighted by $n-3$, pooled, and then z reverted. The pooled estimate was considered descriptive as only two external cohorts were available, rather than a general random-effects meta-analysis.

2.7 Molecular-state discrimination

Each external cohort and pathway was classified as either having a low or high tumour score, based on the overlap with the 33th percentile and 67th percentile, respectively, of the transcriptomic outcome. The area under the receiver-operating-characteristic (ROC) curve was used to determine the locked continuous signature score; 1,000 bootstrap resamples with 95% confidence intervals were used. This extreme-tertile analysis was pre-specified as an endpoint to be interpreted as the molecular-stratification and was not presented as a diagnostic threshold to guide patient care.

2.8 Unsupervised cross-cancer signaling states

Predictions for these three pathways (MAPK/RTK, PI3K-AKT-mTOR, and cell cycle prediction) were standardized within each external cohort and then pooled. The average silhouette width was compared for K-means models ranging from 2 to 6 clusters (38). The selected two-cluster solution was evaluated after 500 iterations, in which 80% of the tumors were sampled without replacement, re-clustered, oriented according to the average activity of pathway, and compared with the full-sample solution using the adjusted Rand index (39).

2.9 Feature-stability analysis

A total of 100 bootstrap samples of the discovery cohort was drawn for each integrated signature. The feature pool was determined by discovery, the hyperparameters for the elastic-net used the previously determined values and the model was refitted. The selection frequency for each coefficient was the number of models in the bootstrap sample in which the coefficient was different from zero. Caution was used when interpreting the direction of coefficients as they may flip when there is correlated predictors, and biological interpretation was given weight to the stability of the chosen coefficient and not just the sign of the coefficient.

2.10 Software, ethics, and reproducibility

The analyses were conducted in Python with the packages Pandas, NumPy, SciPy, scikit-learn and Matplotlib. The script provided reproduces feature filtering, fixed fitting, locked cross-validated parameters, external prediction, bootstrapping, clustering, source data tables and figures. No new human subject participants were recruited and no new IRB review was needed for this study because only de-identified public data from the original reports of the CPTAC studies were used.

3. Results

3.1 Cohort composition and cross-cohort analytical feature space

The analysis included 315 tumors: 122 breast cancers in discovery and 193 tumors in external validation (83 ovarian high-grade serous carcinomas and 110 lung adenocarcinomas). The raw proteome features varied from 10,589 to 12,433, while the raw phosphoproteome features ranged from 37,334 to 61,705 features of phosphosites (Table 1; Figure 2). The missingness was significantly higher in the phosphoproteome (59.2%-65.2%) than the proteome (16.9%-20.9%), reflecting the added enrichment and site-localization restrictions when working with phosphoproteomic acquisition.

Identifiers intersecting exactly and coverage of 70% within the cohort resulted in 7,053 proteins and 3,822 phosphosites remaining. On this basis, the modeling space preferred was features that were measurable in more than one disease, and not features that were most informative in one disease. This limitation was in place prior to feature-outcome ranking.

Table 1. Cohorts, raw molecular depth, missingness, and shared filtered feature space.

Cohort	Modality	n	Raw features	Missing	Shared $\geq 70\%$
Breast cancer	Proteome	122	12,022	17.7%	7,053
Breast cancer	Phosphoproteome	122	49,871	65.2%	3,822
Ovarian HGSC	Proteome	83	10,589	20.9%	7,053
Ovarian HGSC	Phosphoproteome	83	37,334	59.2%	3,822
Lung adenocarcinoma	Proteome	110	12,433	16.9%	7,053
Lung adenocarcinoma	Phosphoproteome	110	61,705	64.1%	3,822

3.2 Locked integrated signatures transport across independent cancers

All three integrated signatures retained statistically strong associations with their orthogonal transcriptomic states in both external cohorts (Figure 4; Table 2). In ovarian cancer, MAPK/RTK validation yielded $p=0.684$ (95% CI 0.517-0.808; $R^2=0.515$), PI3K-AKT-mTOR yielded $p=0.589$ (0.417-0.721; $R^2=0.371$), and cell cycle yielded $p=0.703$ (0.550-0.803; $R^2=0.531$). In lung adenocarcinoma, corresponding correlations were 0.533 (0.390-0.658; $R^2=0.269$), 0.540 (0.349-0.715; $R^2=0.232$), and 0.828 (0.703-0.914; $R^2=0.618$). All external correlation P values were below 6×10^{-9} .

Fisher-z pooled external correlations were 0.603 (95% CI 0.503-0.686) for MAPK/RTK, 0.561 (0.455-0.652) for PI3K-AKT-mTOR, and 0.782 (0.719-0.832) for cell cycle. The median of the six external integrated correlations was 0.637.

These results established reproducible sample ordering across distinct tumor lineages, with cell-cycle state showing the greatest transportability.

Table 2. External validation performance of locked integrated proteomic-phosphoproteomic signatures.

Pathway	External cohort	n	Spearman ρ	95% CI	R ²	RMSE	P-value
Cell cycle	Lung adenocarcinoma	110	0.828	0.703-0.914	0.618	0.618	<0.001
Cell cycle	Ovarian HGSC	82	0.703	0.550-0.803	0.531	0.685	<0.001
MAPK/RTK	Lung adenocarcinoma	110	0.533	0.390-0.658	0.269	0.855	<0.001
MAPK/RTK	Ovarian HGSC	82	0.684	0.517-0.808	0.515	0.696	<0.001
PI3K-AKT-mTOR	Lung adenocarcinoma	110	0.540	0.349-0.715	0.232	0.876	<0.001
PI3K-AKT-mTOR	Ovarian HGSC	82	0.589	0.417-0.721	0.371	0.793	<0.001

3.3 Modality contributions are pathway and disease dependent

Integration was not uniformly superior to the best single modality (Figure 5), an important constraint against an overly simple multi-omics narrative. For ovarian MAPK/RTK activity, the integrated model ($\rho=0.684$) exceeded protein-only ($\rho=0.662$) and phosphoprotein-only ($\rho=0.527$) models. In lung MAPK/RTK, phosphoprotein-only performance was modestly higher ($\rho=0.558$) than integrated performance ($\rho=0.533$), while protein-only performance was lower ($\rho=0.444$).

For PI3K-AKT-mTOR, protein-only models were strongest in both ovarian cancer ($\rho=0.721$ versus 0.589 integrated) and lung adenocarcinoma ($\rho=0.574$ versus 0.540 integrated). For cell cycle, protein-only performance was highest in ovarian cancer ($\rho=0.768$), while phosphoprotein-only performance was highest in lung adenocarcinoma ($\rho=0.848$). The integrated cell-cycle model remained strong in both cohorts ($\rho=0.703$ and 0.828). These findings indicate that phosphosites add context-sensitive dynamic information, but abundance-level proteins may carry more portable information for some pathways.

3.4 Locked scores discriminate high and low molecular pathway states

Extreme-tertile classification converted the continuous external validation problem into molecular-state discrimination (Table 3; Figure 6). MAPK/RTK AUC was 0.927 (95% CI 0.848-0.982) in ovarian cancer and 0.877 (0.788-0.955) in lung adenocarcinoma. PI3K-AKT-mTOR AUC was 0.832 (0.719-0.931) and 0.810 (0.695-0.911), respectively. Cell-cycle AUC was 0.925 (0.849-0.977) and 0.956 (0.882-1.000), respectively. Across all six tests, the median AUC was 0.901. No classification threshold was optimized or proposed for clinical use.

Table 3. External discrimination of high versus low transcriptomic pathway states using locked integrated scores.

Pathway	External cohort	Extreme-tertile n	AUC	95% bootstrap CI
Cell cycle	Lung adenocarcinoma	74	0.956	0.882-1.000
Cell cycle	Ovarian HGSC	56	0.925	0.849-0.977
MAPK/RTK	Lung adenocarcinoma	74	0.877	0.788-0.955
MAPK/RTK	Ovarian HGSC	56	0.927	0.848-0.982
PI3K-AKT-mTOR	Lung adenocarcinoma	74	0.810	0.695-0.911
PI3K-AKT-mTOR	Ovarian HGSC	56	0.832	0.719-0.931

3.5 External tumors form stable signaling-low and co-activated states

Integrated pathway predictions were available for 192 external tumors. Among k=2-6 solutions, k=2 had the highest silhouette width (0.333; versus 0.297 for k=3 and ≤ 0.281 for k=4-6). A two-state solution was used to divide the sample into a signaling-low group (n=78) and a co-activated group (n=114). Within-cohort standardized scores for MAPK/RTK, PI3K-AKT-mTOR, and cell-cycle were, respectively, -0.379, -0.930, and -0.801 in the signaling-low state and 0.260, 0.636, and 0.548 in the co-activated state. In both cancers, both states were found in both cancers: lung adenocarcinoma (48 signaling-low and 62 co-activated) and ovarian cancer (30 signaling-low and 52 co-activated).

Cluster assignments were very reproducible when subjected to repeated 80% subsampling (median adjusted Rand index: 0.974; empirical interval: 0.665-1.000). Figure 7 illustrates heatmap of state reflectivity that suggests that there was no single dominant pathway.

3.6 Stable features connect the signatures to plausible signaling biology

Bootstrap refitting revealed small groups of features that were selected repeatedly (Figure 8). LPCAT2 protein, SOS1-S1178 and RPS6KA3 protein were also selected in 90%, 88% and 84% of the bootstrap models, respectively, for MAPK/RTK, LAMC2 protein and MRTFA-S385. LAMC2 is linked to programs of invasive epithelial cells, SOS1 directly links receptor tyrosine kinases to RAS, and RPS6KA3 is a kinase that is regulated by ERK, which provides biological coherence and does not imply causality between the coefficients.

The protein–protein interactions that were selected for high stability were CHDH protein and RRBP1-S615 (97% of the models); MYH10-S1975, GRB10-S104, OGT-S20, and DCAF8 also had high stability. GRB10 is an adaptor that is linked to feedback in insulin/PI3K signaling, while nutrient sensing occurs through OGT, which is linked to mTOR-dependent metabolism. For cell cycle, RRM2 protein was selected in every bootstrap model, CEP152-S1461 in 95%, HP1BP3-S156 in 92%, CAVIN3-S56 in 90%, and LIG1-S76 and PPM1G protein in 89%. The DNA synthesis/replication, centrosome biology and chromosome segregation are linked by RRM2, LIG1, CEP152 and SPC24, which is consistent with the remarkable external function of the cell-cycle signature.

4. Figures

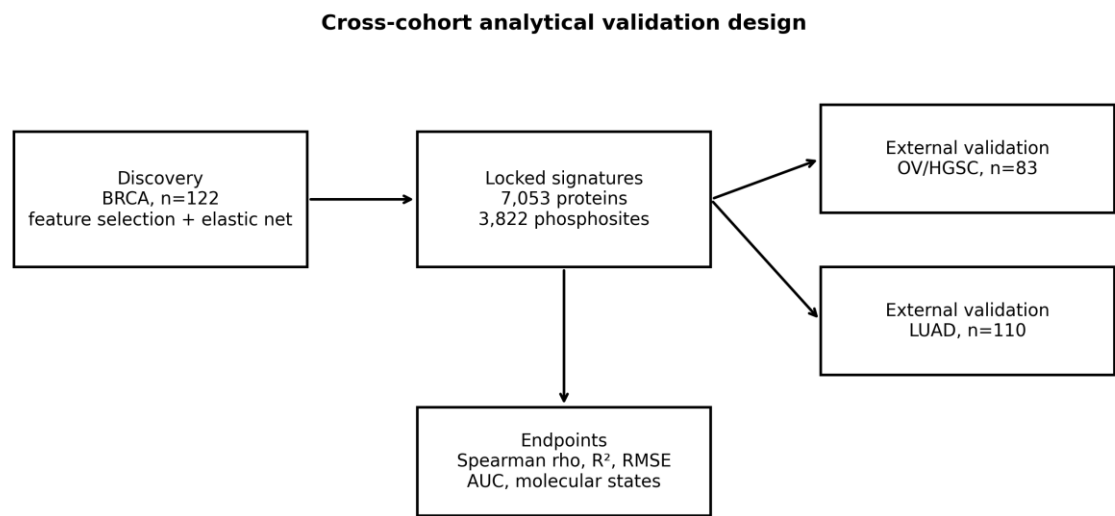


Figure 1. Study design. Discovery-only feature selection and elastic-net tuning were performed in CPTAC breast cancer. Coefficients were locked and evaluated without retuning in ovarian high-grade serous carcinoma and lung adenocarcinoma.

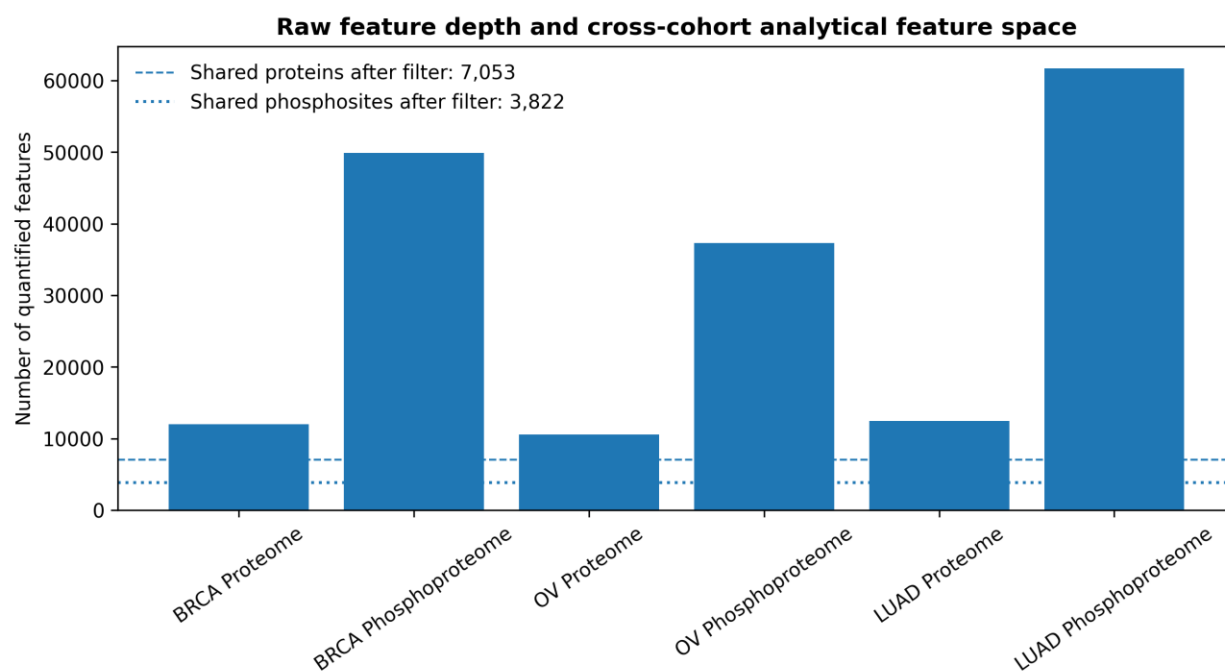
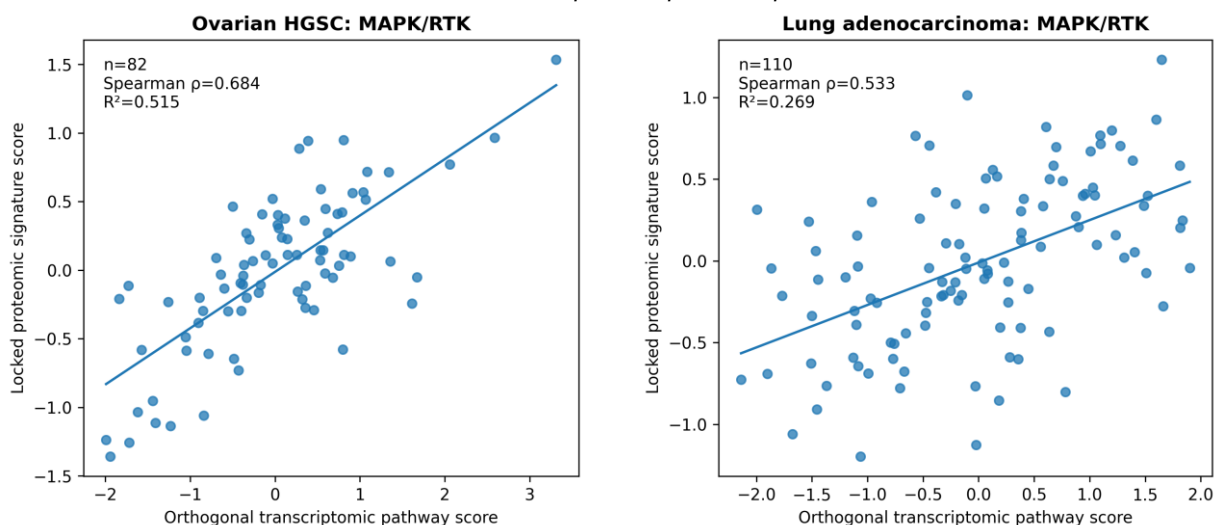


Figure 2. Raw feature depth and the common analytical feature space. Dashed lines show 7,053 proteins and 3,822 phosphosites retained after exact identifier intersection and at least 70% coverage in every cohort.

Figure 3. Observed orthogonal transcriptomic pathway states versus locked integrated mass-spectrometry signature scores in the two external cohorts. Each panel reports n , Spearman correlation, and R^2 .



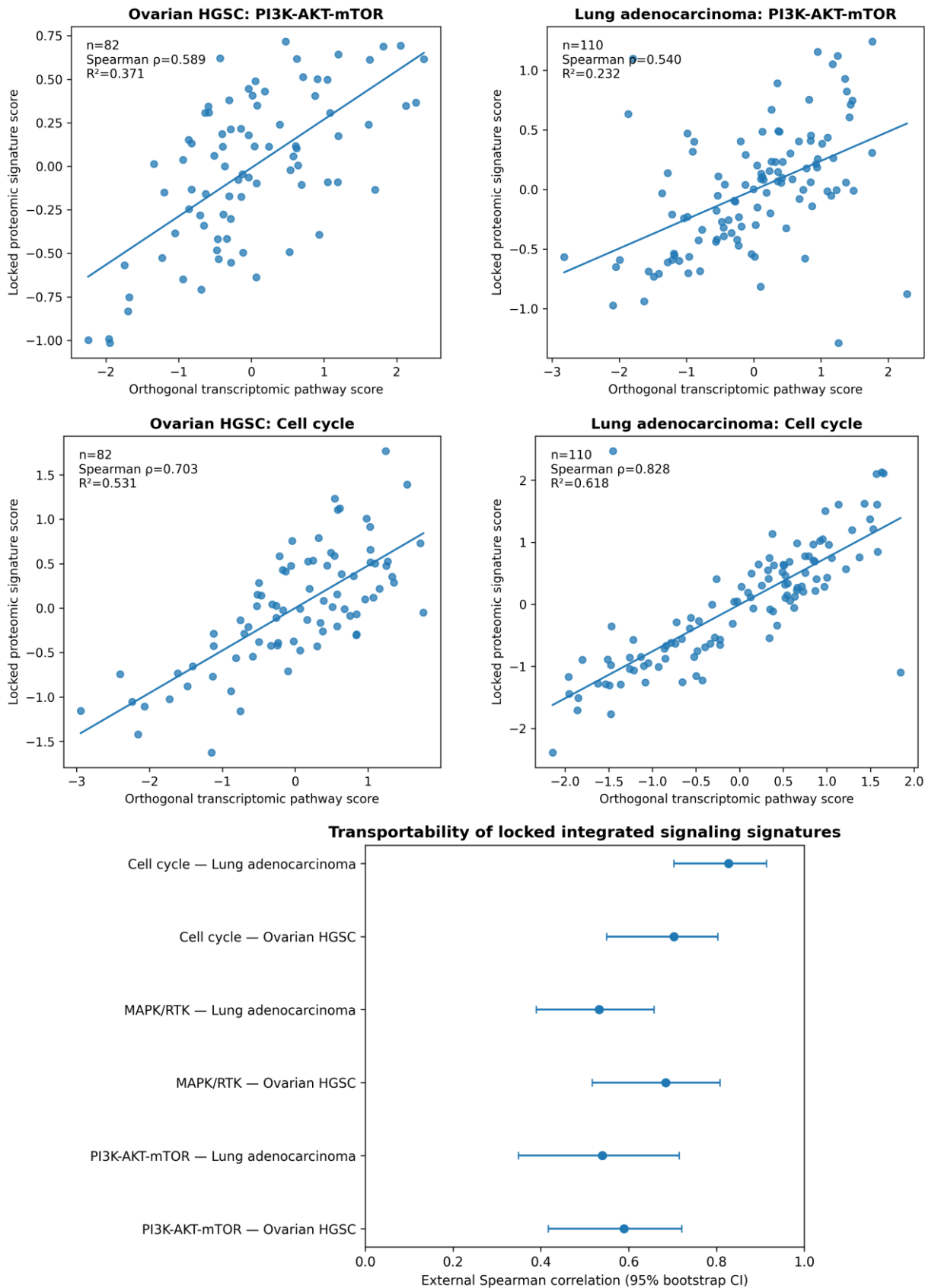


Figure 4. External transportability of the locked integrated signatures. Error bars are 95% nonparametric bootstrap confidence intervals.

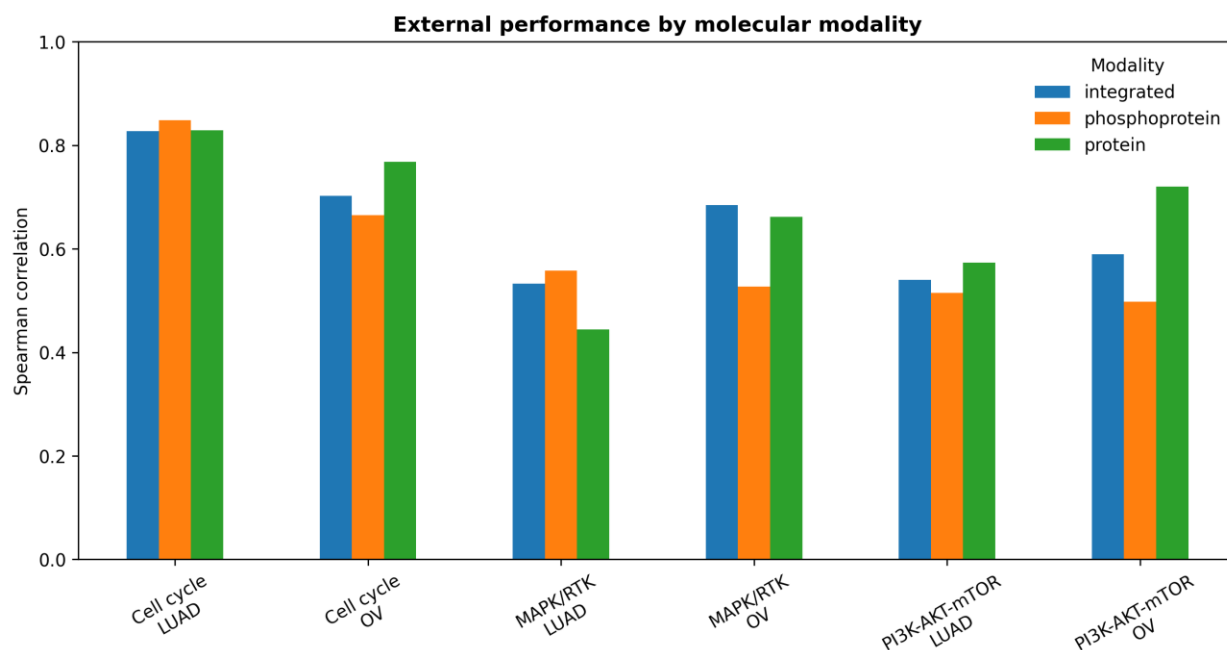


Figure 5. External Spearman correlations by protein-only, phosphoprotein-only, and integrated signatures. Multi-omic integration was beneficial in selected pathway-cohort combinations but was not uniformly superior.

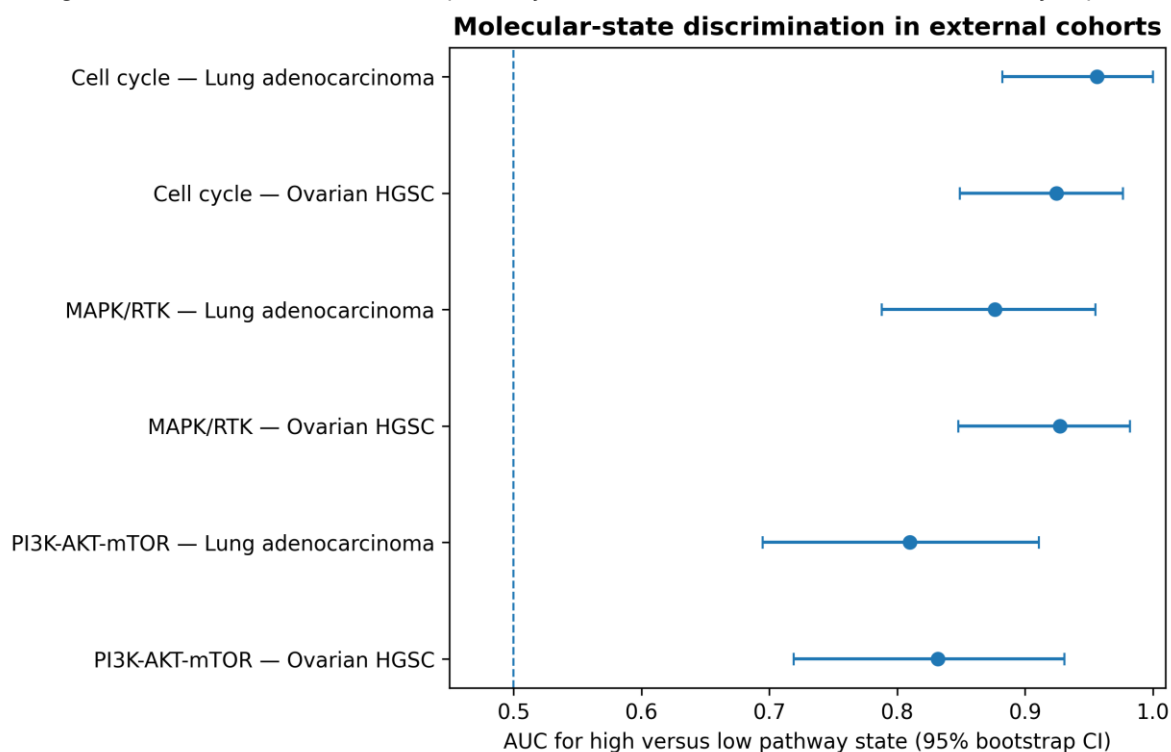


Figure 6. Discrimination of extreme-tertile high versus low transcriptomic pathway states by locked integrated scores. Error bars are 95% bootstrap confidence intervals; the dashed line denotes chance performance.

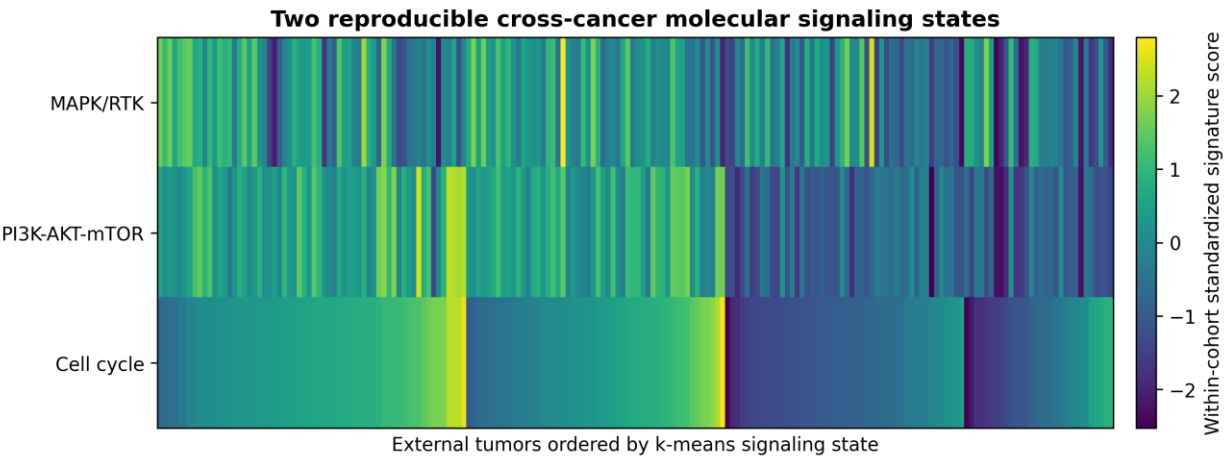
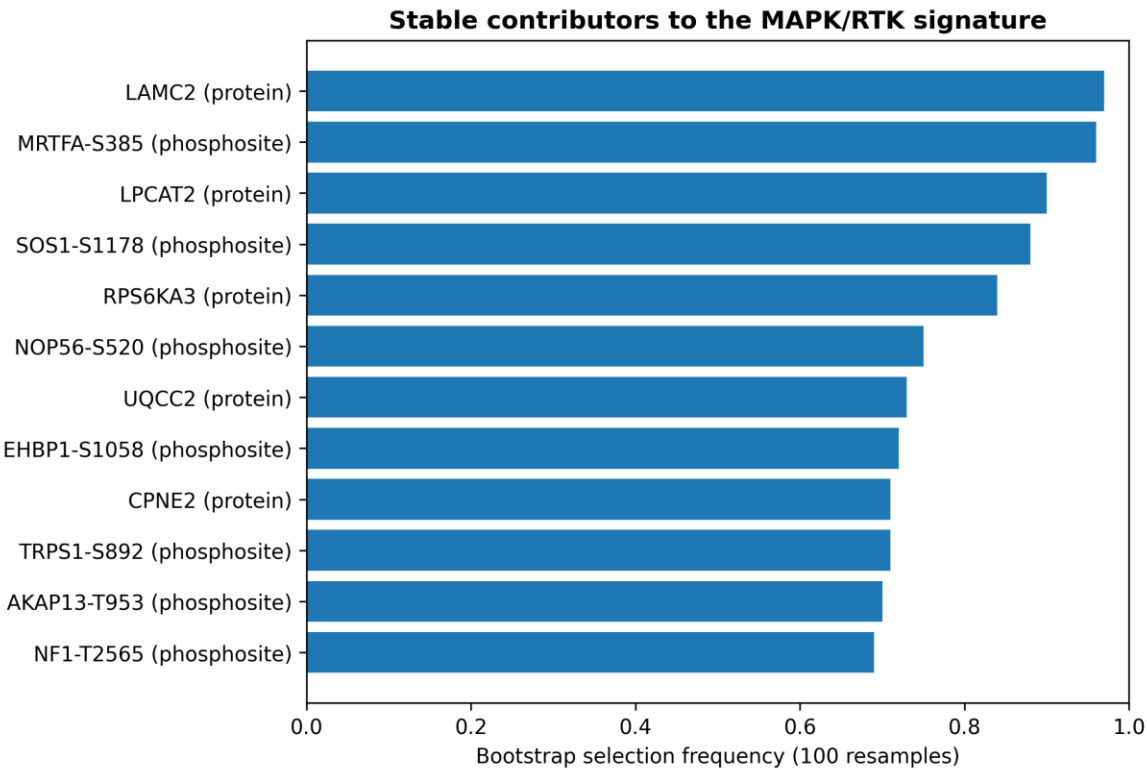
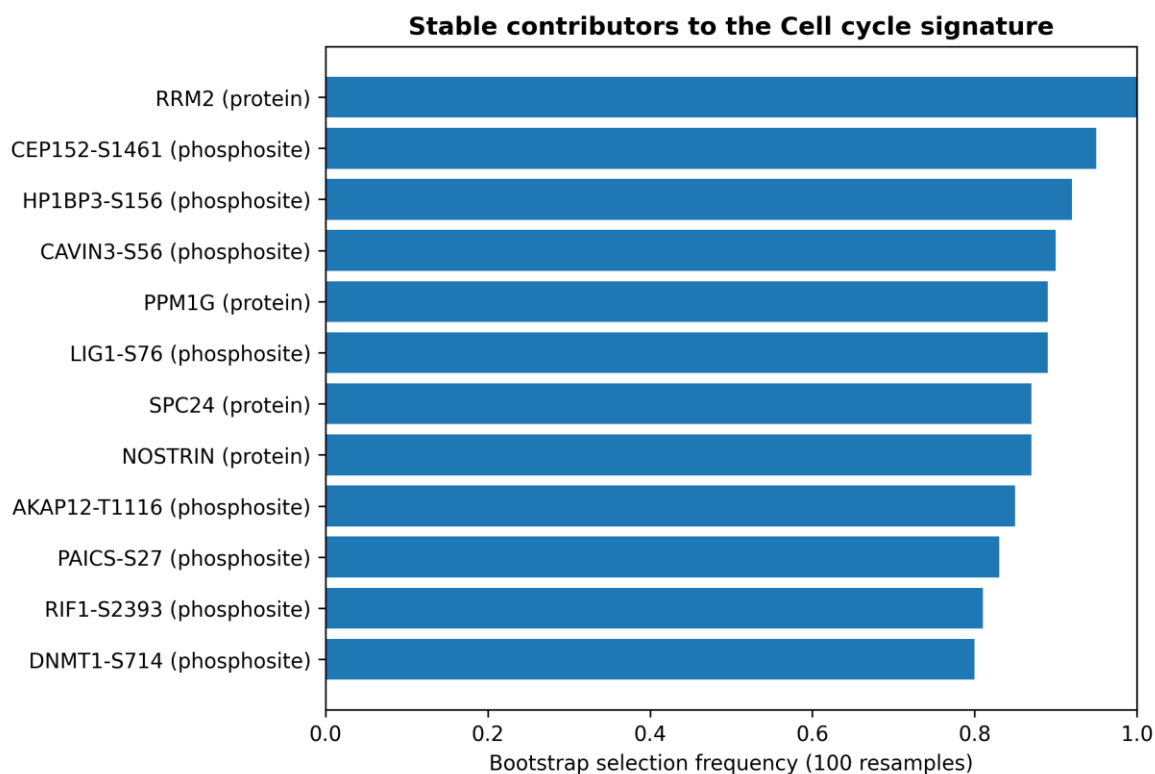
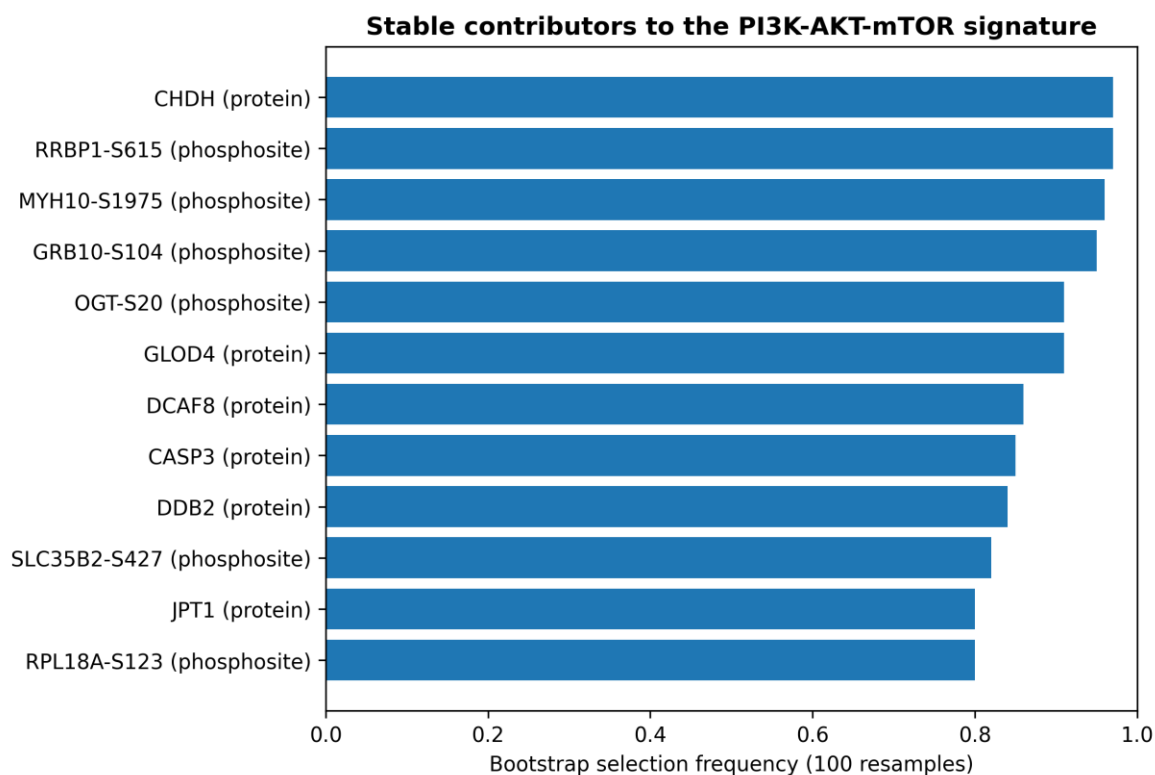


Figure 7. Heatmap of external integrated signaling scores after within-cohort standardization and ordering by the selected two-state k-means solution.

Figure 8. Bootstrap selection stability of leading contributors to the three integrated signatures. Selection frequency was calculated over 100 discovery-cohort bootstrap refits with the feature pool and hyperparameters fixed.





5. Discussion

The findings of this study show that multivariate measurements of mass-spectrometry proteomics and phosphoproteomics can maintain quantitative values in separate cohorts of cancer patients. Moderate-to-strong associations with independent states of transcriptomic pathways were preserved for ovarian high-grade serous carcinoma and lung adenocarcinoma, in a model developed in breast cancer. The best was the cell-cycle signaling

pathway, with a pooled external correlation of 0.782, and AUCs of 0.925 and 0.956. Significant lineage dependence was also observed with MAPK/RTK and PI3K-AKT-mTOR states, as these states were also reproducible.

The analysis goes beyond the within-cohort association in three aspects. First, it is to note that the validation cancers were not used to select features nor to tune the coefficients. Second, it was a totally orthogonal transcriptomic state, as it was not trained and validated on different transformations of the same proteomic matrix. Third, outcome ranking was preceded by a definition of the feature space of the analysis, which was based on cross-cohort measurability. These decisions were made with the intent to compromise a measure's "discovery-cohort fit" in favor of transportability - a key consideration for any measurement system that is planned for use across laboratories, cohorts, and clinical settings.

The result of the cell-cycle is biologically plausible. Coordinated DNA replication, mitosis, centrosome function, and chromosome segregation all play a role in proliferation programs, resulting in a wide and redundant molecular footprint. The conserved state across epithelial cancers is then encoded by stable features like RRM2, LIG1, CEP152, SPC24 and replication-associated phosphosites. In contrast, different receptors, mutations, paracrine signals and feedback loops in different lineages can activate MAPK and PI3K signaling. The fact that they performed moderately but significantly on the outside has special significance because they were not allowed to recalibrate as a disease-specific group.

Unfalsifiedness of the assumption of more omics layers better measurement is tempered by comparison between modalities. Protein-only models were superior in both cohorts in the validation models of PI3K-AKT-mTOR and ovarian cancer, and for cell cycle in lung adenocarcinoma. Protein-only models were superior in both cohorts in the validation models of PI3K-AKT-mTOR and ovarian cancer, while phosphoprotein-only models were superior in cell cycle in lung adenocarcinoma. If the second layer is sparse, noisier or subject to context-specific regulation, integration can cause a strong single-layer signal to be weakened. Thus, the selection of modality should be based on pathway, intended use in a particular context, and not just maximal molecule coverage.

The high versus low analyses serve as a meaningful transition between the continuous reproducibility and stratification. Lock scores have been found to distinguish molecular extremes across external tests without the need for retraining the coefficient for each cohort AUC values over 0.80. However, these tertiles did not represent clinically validated cut points, but were defined from the external molecular outcome. Prospective development would always involve fixed thresholds that are tied to reference materials, independent endpoints and a context of use, e.g., the enrichment of a targeted therapy, the measurement of residual pathway activity or the selection of combination therapies.

The two unsupervised external states indicate that the signatures represent a higher-order signaling structure. The co-activated state had scores that were both high for MAPK/RTK and PI3K-AKT-mTOR, and was found in both diseases, while the signaling-low state had low scores for both pathways. High bootstrap support suggests that the structure was not reliant on a limited number of tumors. Future studies should identify if the co-activated state is a good indicator of therapeutic sensitivity, aggressive behavior, or escape of compensatory pathways. Beyond that, the analysis of the present case cannot make such claims due to the fact that response to treatment and future outcomes were not part of the validation endpoint.

There must be a clear distinction between computational and laboratory assay validity in clinical translation. The current findings confirm the biological replicability of a multivariate measurement in are different cohorts. They do not develop precision and overall imprecision within runs or between runs, lot to lot stability, peptide recovery, matrix effects, interference, linearity, lower LOQ or traceability to certified reference materials (20-24). The pre-analytical ischemia, phosphatase activity, enrichment efficiency, site localization and missingness are significant when studying phosphoproteomics. A next step towards practice is to narrow down each pathway to a small set of

stable proteotypic peptides and phosphopeptides, then generate the internal standards and test performance in a formal fit-for-purpose protocol.

There are some limitations in this study. It could only handle processed relative-abundance matrices and it was incapable of performing an independent audit of the quality metrics of the instruments. Harmonization in CPTAC enhance comparability, but can overestimate transport to unrelated laboratories and acquisition platforms. The median imputation method relies on the assumption that a central estimate for the cohort is appropriate and that it is not able to address abundance-dependent missingness. Biologically important sites may be missed from feature filtering because they are not measured as frequently. The discovery cohort consisted of one cancer type and there were only two external cohorts available. Transcriptomic pathway scores are not direct kinase assays, treatment response biomarkers or treatment outcomes. Discovery performance was not considered as external validation. Last, stable coefficients detect predictors that can be reproduced, but do not establish a mechanistic causality.

These limitations notwithstanding, the study offers a rigorously set and repeatable basis for the development of assays. All of the above, from locked coefficients to independent cohorts of diseases, orthogonal pathway comparators, and feature-stability analysis, and openly executable code, are in keeping with recommendations for transitioning clinical proteomics from repeated discovery to explicit validation.

6. Conclusions

Proteomic and phosphosite signatures derived from mass spectrometry measurements quantitatively reproduced MAPK/RTK, PI3K-AKT-mTOR and cell-cycle molecular states in independent breast, ovarian and lung cancer datasets. High versus low molecular states were discriminated with consistently strong AUCs and cell-cycle activity was the most portable. Multi-omic integration was helpful but not always optimal, highlighting a need for pathway-specific assays. The findings provide a basis for the computational cross-cohort analytical validity, which is intended to inform targeted validation in prospective laboratory and clinical studies but should not be used as a substitute for such validation.

Highlights

- A breast-cancer discovery model was locked before testing in ovarian high-grade serous carcinoma and lung adenocarcinoma.
- External integrated-signature correlations ranged from 0.533 to 0.828 across MAPK/RTK, PI3K-AKT-mTOR, and cell-cycle states.
- High-versus-low molecular pathway states were discriminated with AUCs from 0.810 to 0.956.
- Cell-cycle signatures showed the strongest cross-cancer transportability, whereas the value of integrating proteome and phosphoproteome was pathway dependent.
- Unsupervised analysis identified stable signaling-low and co-activated cross-cancer states (median bootstrap adjusted Rand index, 0.974).

Declarations

Ethics approval and consent to participate: The analysis used de-identified public data. Ethics approvals and informed-consent procedures are described in the original CPTAC cohort publications. No new participant contact or specimen collection occurred.

Data availability: The processed CPTAC pan-cancer matrices used in this study are openly distributed through LinkedOmics/CPTAC resources. The accompanying reproducibility package contains the full analysis script, model outputs, source-data tables, and figures; raw public matrices are not redistributed because of file size and are referenced in the README.

Code availability: The accompanying Python script reproduces the reported feature filtering, locked model fitting, validation statistics, bootstrap confidence intervals, clustering, and figures. Random seed: 42.

Funding: This research received no external funding.

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

Acknowledgements: The authors acknowledge the patients, investigators, and staff of the National Cancer Institute Clinical Proteomic Tumor Analysis Consortium and the developers of LinkedOmics.

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] Badia-i-Mompel P, Vélez Santiago J, Braunger J, Geiss C, Dimitrov D, Müller-Dott S, et al. (2022) decoupleR: ensemble of computational methods to infer biological activities from omics data. Kuijjer ML, editor. *Bioinforma Adv.* 2022 Jan 10;2(1):vbac016. doi:10.1093/bioadv/vbac016
- [2] Beekhof R, Van Alphen C, Henneman AA, Knol JC, Pham TV, Rolfs F, et al. (2019) INKA, an integrative data analysis pipeline for phosphoproteomic inference of active kinases. *Mol Syst Biol.* 2019 Apr 1;15(4):MSB188250. doi:10.15252/msb.20188250
- [3] Bittremieux W, Tabb DL, Impens F, Staes A, Timmerman E, Martens L, et al. (2018) Quality control in mass spectrometry-based proteomics. *Mass Spectrom Rev.* 2018 Sep;37(5):697-711. doi:10.1002/mas.21544
- [4] Chiva C, Olivella R, Staes A, Mendes Maia T, Panse C, Stejskal K, et al. (2025) A Multiyear Longitudinal Harmonization Study of Quality Controls in Mass Spectrometry Proteomics Core Facilities. *J Proteome Res.* 2025 Feb 7;24(2):397-409. doi:10.1021/acs.jproteome.4c00359
- [5] Clark DJ, Dhanasekaran SM, Petralia F, Pan J, Song X, Hu Y, et al. (2019) Integrated Proteogenomic Characterization of Clear Cell Renal Cell Carcinoma. *Cell.* 2019 Oct;179(4):964-983.e31. doi:10.1016/j.cell.2019.10.007
- [6] Collins BC, Hunter CL, Liu Y, Schilling B, Rosenberger G, Bader SL, et al. (2017) Multi-laboratory assessment of reproducibility, qualitative and quantitative performance of SWATH-mass spectrometry. *Nat Commun.* 2017 Aug 21;8(1):291. doi:10.1038/s41467-017-00249-5
- [7] Crowl S, Jordan BT, Ahmed H, Ma CX, Naegle KM. (2022) KSTAR: An algorithm to predict patient-specific kinase activities from phosphoproteomic data. *Nat Commun.* 2022 Jul 25;13(1):4283. doi:10.1038/s41467-022-32017-5
- [8] Dou Y, Kawaler EA, Cui Zhou D, Gritsenko MA, Huang C, Blumenberg L, et al. (2020) Proteogenomic Characterization of Endometrial Carcinoma. *Cell.* 2020 Feb;180(4):729-748.e26. doi:10.1016/j.cell.2020.01.026
- [9] Frantzi M, Latosinska A, Kontostathi G, Mischak H. (2018) Clinical Proteomics: Closing the Gap from Discovery to Implementation. *PROTEOMICS.* 2018 Jul;18(14):1700463. doi:10.1002/pmic.201700463
- [10] Friedman J, Hastie T, Tibshirani R. (2010) Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw.* 2010;33(1). doi:10.18637/jss.v033.i01
- [11] Gillette MA, Satpathy S, Cao S, Dhanasekaran SM, Vasaikar SV, Krug K, et al. (2020) Proteogenomic Characterization Reveals Therapeutic Vulnerabilities in Lung Adenocarcinoma. *Cell.* 2020 Jul;182(1):200-225.e35. doi:10.1016/j.cell.2020.06.013
- [12] Hanahan D. (2022) Hallmarks of Cancer: New Dimensions. *Cancer Discov.* 2022 Jan 1;12(1):31-46. doi:10.1158/2159-8290.CD-21-1059
- [13] Hänzelmann S, Castelo R, Guinney J. (2013) GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics.* 2013 Dec;14(1):7. doi:10.1186/1471-2105-14-7
- [14] Hernandez-Armenta C, Ochoa D, Gonçalves E, Saez-Rodriguez J, Beltrao P. (2017) Benchmarking substrate-based kinase activity inference using phosphoproteomic data. Wren J, editor. *Bioinformatics.* 2017 Jun 15;33(12):1845-51. doi:10.1093/bioinformatics/btx082
- [15] Hirtz C, Bros P, Brede C, Lescuyer P, Maceski AM, Vialaret J, et al. (2018) Regulatory context and validation of assays for clinical mass spectrometry proteomics (cMSP) methods. *Crit Rev Clin Lab Sci.* 2018 Jul 4;55(5):346-58. doi:10.1080/10408363.2018.1470159
- [16] Holland CH, Tanevski J, Perales-Patón J, Gleixner J, Kumar MP, Mereu E, et al. (2020) Robustness and applicability of transcription factor and pathway analysis tools on single-cell RNA-seq data. *Genome Biol.* 2020 Dec;21(1):36. doi:10.1186/s13059-020-1949-z
- [17] Huang C, Chen L, Savage SR, Eguev RV, Dou Y, Li Y, et al. (2021) Proteogenomic insights into the biology and treatment of HPV-negative head and neck squamous cell carcinoma. *Cancer Cell.* 2021 Mar;39(3):361-379.e16. doi:10.1016/j.ccell.2020.12.007
- [18] Hubert L, Arabie P. (1985) Comparing partitions. *J Classif.* 1985 Dec;2(1):193-218. doi:10.1007/BF01908075
- [19] Kong W, Hui HWH, Peng H, Goh WWB. (2022) Dealing with missing values in proteomics data. *PROTEOMICS.* 2022 Dec;22(23-24):2200092. doi:10.1002/pmic.202200092
- [20] Krug K, Jaehnig EJ, Satpathy S, Blumenberg L, Karpova A, Anurag M, et al. (2020) Proteogenomic Landscape of Breast Cancer Tumorigenesis and Targeted Therapy. *Cell.* 2020 Nov;183(5):1436-1456.e31. doi:10.1016/j.cell.2020.10.036

- [21] Krug K, Mertins P, Zhang B, Hornbeck P, Raju R, Ahmad R, et al. (2019) A Curated Resource for Phosphosite-specific Signature Analysis. *Mol Cell Proteomics*. 2019 Mar;18(3):576-93. doi:10.1074/mcp.TIR118.000943
- [22] Li Y, Dou Y, Da Veiga Leprevost F, Geffen Y, Calinawan AP, Aguet F, et al. (2023) Proteogenomic data and resources for pan-cancer analysis. *Cancer Cell*. 2023 Aug;41(8):1397-406. doi:10.1016/j.ccell.2023.06.009
- [23] Li Y, Porta-Pardo E, Tokheim C, Bailey MH, Yaron TM, Stathias V, et al. (2023) Pan-cancer proteogenomics connects oncogenic drivers to functional states. *Cell*. 2023 Aug;186(18):3921-3944.e25. doi:10.1016/j.cell.2023.07.014
- [24] Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. (2015) The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst*. 2015 Dec;1(6):417-25. doi:10.1016/j.cels.2015.12.004
- [25] McDermott JE, Arshad OA, Petyuk VA, Fu Y, Gritsenko MA, Clauss TR, et al. (2020) Proteogenomic Characterization of Ovarian HGSC Implicates Mitotic Kinases, Replication Stress in Observed Chromosomal Instability. *Cell Rep Med*. 2020 Apr;1(1):100004. doi:10.1016/j.xcrm.2020.100004
- [26] Müller-Dott S, Jaehnig EJ, Munchic KP, Jiang W, Yaron-Barir TM, Savage SR, et al. (2025) Comprehensive evaluation of phosphoproteomic-based kinase activity inference. *Nat Commun*. 2025 May 22;16(1):4771. doi:10.1038/s41467-025-59779-y
- [27] O'Brien JJ, Gunawardena HP, Paulo JA, Chen X, Ibrahim JG, Gygi SP, et al. (2018) The effects of nonignorable missing data on label-free mass spectrometry proteomics experiments. *Ann Appl Stat*. 2018 Dec 1;12(4). doi:10.1214/18-AOAS1144
- [28] Ochoa D, Jarnuczak AF, Viéitez C, Gehre M, Soucheray M, Mateus A, et al. (2020) The functional landscape of the human phosphoproteome. *Nat Biotechnol*. 2020 Mar;38(3):365-73. doi:10.1038/s41587-019-0344-3
- [29] Piersma SR, Valles-Martí A, Rolfs F, Pham TV, Henneman AA, Jiménez CR. (2024) Inferring kinase activity from phosphoproteomic data: Tool comparison and recent applications. *Mass Spectrom Rev*. 2024 Jul;43(4):725-51. doi:10.1002/mas.21808
- [30] Poulos RC, Hains PG, Shah R, Lucas N, Xavier D, Manda SS, et al. (2020) Strategies to enable large-scale proteomics for reproducible research. *Nat Commun*. 2020 Jul 30;11(1):3793. doi:10.1038/s41467-020-17641-3
- [31] Rousseeuw PJ. (1987) Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math*. 1987 Nov;20:53-65. doi:10.1016/0377-0427(87)90125-7
- [32] Savage SR, Yi X, Lei JT, Wen B, Zhao H, Liao Y, et al. (2024) Pan-cancer proteogenomics expands the landscape of therapeutic targets. *Cell*. 2024 Aug;187(16):4389-4407.e15. doi:10.1016/j.cell.2024.05.039
- [33] Schubert M, Klinger B, Klünemann M, Sieber A, Uhlitz F, Sauer S, et al. (2018) Perturbation-response genes reveal signaling footprints in cancer gene expression. *Nat Commun*. 2018 Jan 2;9(1):20. doi:10.1038/s41467-017-02391-6
- [34] Stanfill BA, Nakayasu ES, Bramer LM, Thompson AM, Ansong CK, Clauss TR, et al. (2018) Quality Control Analysis in Real-time (QC-ART): A Tool for Real-time Quality Control Assessment of Mass Spectrometry-based Proteomics Data. *Mol Cell Proteomics*. 2018 Sep;17(9):1824-36. doi:10.1074/mcp.RA118.000648
- [35] Tibshirani R. (1996) Regression Shrinkage and Selection Via the Lasso. *J R Stat Soc Ser B Stat Methodol*. 1996 Jan 1;58(1):267-88. doi:10.1111/j.2517-6161.1996.tb02080.x
- [36] Vasaikar S, Huang C, Wang X, Petyuk VA, Savage SR, Wen B, et al. (2019) Proteogenomic Analysis of Human Colon Cancer Reveals New Therapeutic Opportunities. *Cell*. 2019 May;177(4):1035-1049.e19. doi:10.1016/j.cell.2019.03.030
- [37] Wiredja DD, Koyutürk M, Chance MR. (2017) The KSEA App: a web-based tool for kinase activity inference from quantitative phosphoproteomics. Valencia A, editor. *Bioinformatics*. 2017 Nov 1;33(21):3489-91. doi:10.1093/bioinformatics/btx415
- [38] Yılmaz S, Ayati M, Schlatter D, Çiçek AE, Chance MR, Koyutürk M. (2021) Robust inference of kinase activity using functional networks. *Nat Commun*. 2021 Feb 19;12(1):1177. doi:10.1038/s41467-021-21211-6
- [39] Zou H, Hastie T. (2005) Regularization and Variable Selection Via the Elastic Net. *J R Stat Soc Ser B Stat Methodol*. 2005 Apr 1;67(2):301-20. doi:10.1111/j.1467-9868.2005.00503.x

Supplementary materials

- Table S1: cohort-level raw feature depth, missingness, and shared-filtered feature counts.
- Table S2: sample-level observed and predicted pathway scores for every cohort, pathway, and modality.
- Table S3: complete elastic-net coefficient table and human-readable feature labels.
- Table S4: pooled external correlations.
- Table S5: discovery-bootstrap feature-selection stability.
- Table S6: external molecular-state assignments and standardized scores.
- Table S7: silhouette values for k=2-6.
- Reproducible analysis script and figure-generation script.