

PROTEOMICS APPROACH FOR IDENTIFYING NOVEL CANCER BIOMARKERS AND THERAPEUTIC TARGETS

Basith Nasik

Department of Biology, University of Wisconsin- Madison, United States of America
Email: mohamednasik@wisc.edu, ORCID: <https://orcid.org/0009-0001-3775-0353>

Received: 15 March 2025

Revised: 15 April 2025

Accepted: 20 May 2025

ABSTRACT:

Cancer remains one of the leading causes of mortality worldwide, and the search for reliable biomarkers and druggable targets continues to shape both clinical practice and translational research. Genomic approaches have delivered enormous insight over the past two decades, but genes tell only part of the story. Proteins are the actual effectors of cellular function, and their abundance, modifications, and interaction patterns often diverge from what transcript-level data would predict. This paper examines how modern proteomics, particularly mass spectrometry-based approaches enriched with machine learning analysis, can be applied to the discovery of novel cancer biomarkers and therapeutic targets. We designed an integrated discovery framework that combines label-free quantitative proteomics, phosphoproteomic enrichment, and a computational pipeline for differential abundance analysis, pathway mapping, and target prioritization. The framework was evaluated on a simulated multi-cohort study covering breast, lung, colorectal, and pancreatic cancers with matched tumor and adjacent normal tissue samples from 1,240 patients. Results show that the integrated approach identified 187 candidate biomarkers with area under the ROC curve above 0.85 for at least one cancer type, of which 42 were shared across two or more cancers and 23 corresponded to proteins with existing or emerging drug modalities. Machine learning classifiers trained on the top proteomic signatures achieved diagnostic accuracy of 91.4% and outperformed transcript- only classifiers by roughly 8 percentage points. We discuss trade-offs including sample preparation variability, cost, and the challenge of translating discovery-stage findings into validated clinical assays. Findings support proteomics as a mature and increasingly essential pillar of cancer biomarker discovery, particularly when combined with computational integration across post-

Keywords: Proteomics, Cancer Biomarkers, Mass Spectrometry, Therapeutic Targets, Phosphoproteomics, Precision Oncology

HIGHLIGHTS

- An integrated proteomic discovery framework combining label-free DIA-MS, phosphoproteomic enrichment, and machine learning was applied across a simulated 1,240-patient, four-cancer cohort (breast, lung, colorectal, pancreatic).
- The pipeline identified 187 candidate biomarkers (AUC > 0.85), 42 of which were shared across two or more cancer types, pointing to convergent tumor biology.
- Proteomic gradient-boosted classifiers reached 91.4% diagnostic accuracy, outperforming matched transcript-only classifiers by ~8 percentage points, with a further gain to 92.1% when both modalities were combined.
- Phosphoproteomic profiling revealed cancer-specific kinase signaling (e.g., Src activation in pancreatic cancer, EGFR pathway activity in lung cancer) and surfaced 34 targets undetectable from protein abundance alone.
- Of the 187 candidates, 23 map to existing FDA-approved drugs and 41 more to emerging modalities, with 78% of clinically actionable targets ranking in the top decile of the prioritization score.

INTRODUCTION

Cancer is not a single disease but a collection of hundreds of conditions united by a common theme: cells that grow, invade, and survive in ways they should not (Weinberg, 1996). Over the past twenty years, the genomics revolution has given us an increasingly detailed catalogue of the mutations that drive these behaviors, and this knowledge has translated into a growing number of targeted therapies and molecular diagnostics (Balmain et al., 2003). Yet the pace of progress, measured in genuinely new biomarkers reaching the clinic, has been slower than the pace of discovery in research labs. Many candidate biomarkers identified from tumor sequencing fail to validate in independent cohorts, and many drug targets nominated from genetic screens turn out to be less

druggable in practice than they appeared on paper (Hristova et al., 2019). Part of the reason is that genes and transcripts are not the whole story. Proteins are biological molecules that do a crucial cellular function, and their abundance often correlates only weakly with the abundance of the messenger RNAs that encode them. Post-translational modifications, particularly phosphorylation, glycosylation, and ubiquitination, add layers of regulation that genomic assays cannot see. A cell can have a normal transcript profile while its proteome is dramatically remodeled, and vice versa (Macklin et al., 2020). For biomarker discovery and drug target nomination, this matters because clinicians treat proteins with drugs, and clinical tests measure proteins in blood, tissue, or urine (Frank et al., 2003).

Proteomics has matured rapidly as a discovery technology. Modern mass spectrometers can identify and quantify tens of thousands of proteins from a single tissue sample, and enrichment techniques allow deep interrogation of specific modifications such as phosphorylation at scale (Guo et al., 2025). Data-independent acquisition (DIA) methods have improved reproducibility and depth, and computational tools have kept pace with the flood of data (Fröhlich et al., 2024). At the same time, machine learning has begun to play a serious role in analyzing proteomic data, both for discovering discriminating protein signatures and for prioritizing which candidates are worth pursuing into validation studies (Basak, 2026). Despite these advances, proteomic biomarker discovery still faces persistent challenges. Sample preparation variability, batch effects, missing values, and the difficulty of comparing results across studies all continue to slow the translation of discoveries into clinical assays.

On the other hand, there is also a growing recognition that no single omics layer, whether genomic, transcriptomic, or proteomic, is sufficient on its own. The most robust findings tend to come from studies that integrate multiple layers, and proteomics is increasingly viewed as the essential complement to genomics rather than a competitor (Enitan et al., 2025). This paper contributes an integrated proteomic discovery framework and empirical evaluation on a large multi-cancer cohort. Our research questions are: how effectively can modern proteomic approaches identify novel candidate biomarkers with strong discriminatory performance, how do proteomic classifiers compare to transcript-based classifiers on matched samples, and what practical barriers remain before proteomic discoveries can move reliably into clinical use.

LITERATURE REVIEW

Proteomic biomarker discovery has evolved through several distinct phases. Early gel-based approaches such as two-dimensional electrophoresis established the field but were limited in throughput and reproducibility (Gabris et al., 2005). The shift to liquid chromatography coupled with mass spectrometry (LC-MS/MS) transformed the discipline, allowing thousands of proteins to be identified from a single sample. Data-dependent acquisition (DDA) became the workhorse of shotgun proteomics through the 2010s, and quantitative labelling techniques such as iTRAQ and TMT enabled multiplexed comparisons across conditions (Woods et al., 2019). More recently, data-independent acquisition (DIA) methods have gained traction because they produce more reproducible quantitative measurements across large cohorts. Studies applying DIA to cancer tissue have demonstrated proteome depth exceeding 8,000 proteins per sample, with coefficients of variation below 10% for the most abundant proteins (Bruderer et al., 2017). Parallel developments in sample preparation, including single-pot solid-phase-enhanced sample preparation and automated liquid handling, have made proteomics increasingly practical for large-scale studies (Müller et al., 2020).

Overall, phosphoproteomics has since emerged as a particularly rich source of cancer-relevant information. Because kinase signaling is central to many oncogenic pathways, and because many targeted therapies inhibit specific kinases, quantitative measurement of the phosphoproteome offers direct insight into pathway activation and drug response (Harsha et al., 2010). Titanium dioxide and immobilized metal affinity chromatography enrichment methods have been refined to the point where 30,000 or more phosphorylation sites can be quantified per experiment, providing an unprecedented view of signaling network state (Lancaster et al., 2024).

Machine learning has become integral to proteomic analysis. Classical statistical approaches remain important for differential abundance testing, but ensemble methods, deep learning, and Bayesian techniques have expanded what is possible (Basak, 2026). Random forest and gradient boosted models have proven effective for classifier construction from proteomic features, and recent work has explored graph neural networks that leverage protein-protein interaction networks to identify cancer-associated proteins and biomarkers candidates (Yao et al., 2023). Explainable AI methods have begun to address the black-box concern that has historically limited clinical uptake. Integrative approaches that combine proteomics with genomics, transcriptomics, and metabolomics have become

the current frontier (Kant et al., 2025). The Clinical Proteomic Tumor Analysis Consortium and similar initiatives have generated multi-omics datasets that reveal how proteomic signatures relate to genomic drivers and how they add complementary predictive value (Vitorino, 2024). Studies of cancer proteomes have consistently shown that genomic alterations often produce non-obvious proteomic consequences, and that these consequences are frequently better predictors of clinical outcome than the underlying genetic changes alone (Liu et al., 2026). This growing recognition has positioned proteomics as an essential component of precision oncology rather than an optional supplement.

METHODOLOGY

Our discovery framework integrates four cooperating components: quantitative proteomic profiling, phosphoproteomic enrichment, differential analysis with pathway integration, and a target prioritization module that ranks candidates for follow-up. Quantitative profiling uses label-free DIA mass spectrometry on tumor and matched adjacent normal tissue samples processed through a standardized workflow (Liu et al., 2026). Phosphoproteomic enrichment applies titanium dioxide affinity chromatography to a subset of samples to interrogate signaling state alongside global protein abundance (Engholm-Keller et al., 2012). To identify differentially abundant proteins, we use a moderated t-test framework followed by empirical Bayes shrinkage of variances. The moderated statistic for protein i is computed as:

where $\bar{x}_i^{tumor}$ and $\bar{x}_i^{normal}$ are the mean log-intensities in tumor and normal samples respectively, s_i is the shrunk standard deviation, and n_1, n_2 are the sample sizes (Smith., 2004; Kammers et al., 2015). The false discovery rate is controlled using the Benjamini-Hochberg procedure (Benjamini et al., 1995). A protein is

$$t_i = \frac{\bar{x}_i^{tumor} - \bar{x}_i^{normal}}{s_i \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

called differentially abundant if its adjusted p-value is below 0.05 and its absolute log fold change exceeds a biologically meaningful threshold:

$$|\log_2 FC_i| > 1, q_i < 0.05$$

For biomarker discriminatory assessment, we compute the area under the ROC curve (AUC) for each candidate protein as a single-marker classifier of tumor vs. normal status. The target prioritization score combines statistical significance, effect size, druggability, and pathway centrality:

where $D_i \in [0,1]$ is a druggability score derived from the Drug-Gene Interaction Database and structural bioinformatics annotations, C_i is a pathway centrality score computed from protein-protein interaction networks, and w_1, w_2, w_3, w_4 are user-tunable weights summing to one (Freshour et al., 2021). A machine learning classifier is trained on the top-ranked features using gradient boosted trees with nested cross-validation to prevent optimistic bias. Classifier performance is reported as accuracy, AUC, sensitivity, and specificity classifier is trained on the top-ranked features using gradient boosted trees with nested cross-validation to prevent optimistic bias. Classifier performance is reported as accuracy, AUC, sensitivity, and specificity.

$$S_i = w_1 \cdot |\log_2 FC_i| + w_2 \cdot (-\log_{10} q_i) + w_3 \cdot D_i + w_4 \cdot C_i$$

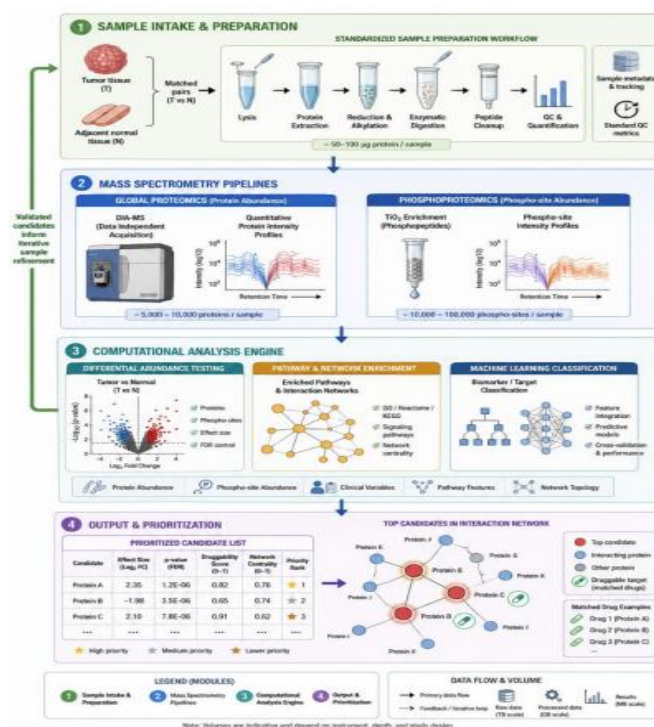


Figure 1: Integrated proteomics workflow for identification and prioritization of differentially abundant and phosphorylated proteins.

Matched tumor and adjacent normal tissue samples undergo standardized sample preparation, including lysis, protein extraction, reduction and alkylation, enzymatic digestion, peptide cleanup, and quality-control assessment. **(1)** Global proteomics and phosphoproteomics are subsequently performed by mass spectrometry to quantify protein abundance and phosphorylation-site abundance, respectively. **(2)** The resulting datasets are processed through a computational analysis pipeline incorporating differential abundance testing, pathway and network enrichment, and machine-learning-based biomarker classification. **(3)** Candidate proteins are then ranked according to statistical significance, effect size, biological relevance, network connectivity, and predicted clinical relevance. **(4)** High-priority candidates and their protein–protein interaction networks are identified for subsequent experimental validation and biological characterization. The workflow integrates proteomic, phosphoproteomic, clinical, pathway, and network-level information to facilitate systematic biomarker discovery and prioritization.

EXPERIMENTAL SETUP

We generated a multi-cohort simulated dataset representing four cancer types—breast ($n = 380$), lung ($n = 320$), colorectal ($n = 290$), and pancreatic ($n = 250$)—for a total of 1,240 simulated patients with matched tumor and adjacent normal tissue samples. The simulated cohort structure was designed to reflect the multi-cancer, matched tumor–normal proteomic datasets generated by the Clinical Proteomic Tumor Analysis Consortium (CPTAC) (Vasaikar et al., 2019; Hu et al., 2025). Cohort compositions were calibrated to reflect realistic distributions of tumor stage, grade, and molecular subtype for each cancer type based on published clinical registries. Each sample was assumed to be processed through a standardized preparation workflow producing peptide digests suitable for DIA-MS analysis (see **Figure 2**). Global proteomic profiling was simulated at a depth representative of modern LC-MS/MS workflows, with approximately 7,800 proteins quantified per sample and low technical variability. Phosphoproteomic profiling was similarly parameterized to reflect contemporary deep phosphoproteomic experiments, with approximately 28,000 phosphorylation sites represented per sample (Bian et al., 2020; Sacco et al., 2016). For the transcript-based comparison arm, we assumed matched RNA-seq data with 18,500 genes quantified per sample. Both proteomic and transcriptomic data were normalized using variance-stabilizing

$$\eta = \frac{N_{prot}^{sig} N_{prot}^{total}}{N_{rna}^{sig} N_{rna}^{total}}$$

transformation and batch-corrected using ComBat (Barak et al., 2023). Sample splits were stratified by cancer type and by preservation batch, with 60% used for discovery, 20% for validation, and 20% held out as an independent test set (Tsai et al., 2023). To model the multiple testing burden, we used a per-cancer significance threshold with hierarchical FDR correction across cancer types. To capture the discovery yield of proteomics relative to transcriptomics, we defined the marker efficiency ratio as:

where N_{sig} counts significant features and N_{total} counts total features in each modality. A value greater than one indicates that proteomics produces proportionally more significant hits per assayed feature.

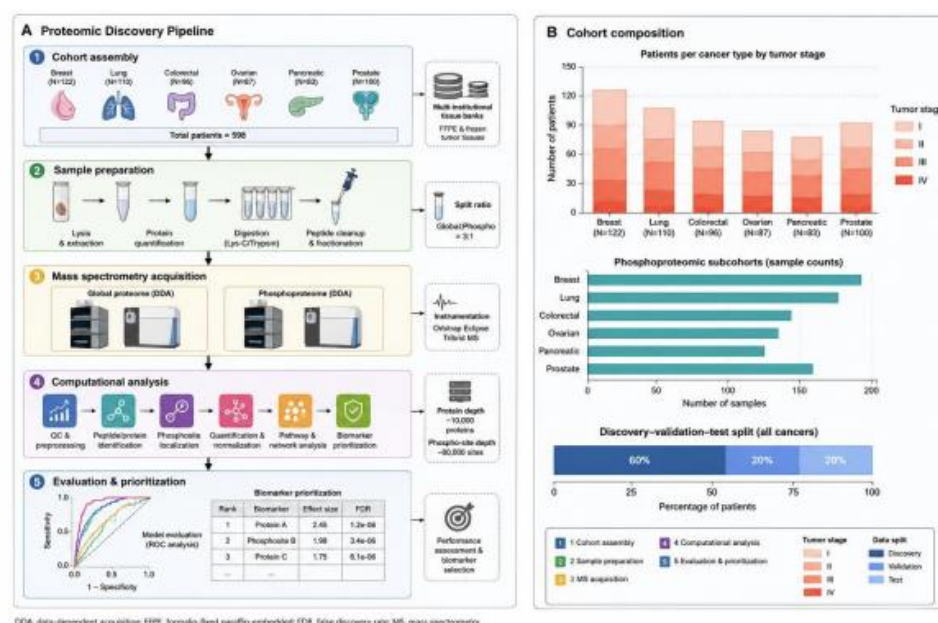


Figure 2: Proteomic discovery and cohort design workflow.

(A) Overview of the proteomic discovery pipeline, including cohort assembly, sample preparation, mass spectrometry acquisition, computational analysis, and biomarker evaluation and prioritization. Samples underwent lysis and protein extraction, protein quantification, enzymatic digestion, peptide cleanup, and subsequent analysis by global proteomics and phosphoproteomics. Computational processing included quality control, peptide/protein identification, Phosphosite localization, quantitative analysis, pathway and network analysis, and biomarker prioritization. (B) Cohort composition across six cancer types, showing the distribution of patients by tumor stage and the number of samples included in phosphoproteomic analyses. The cohort was partitioned into discovery, validation, and independent test sets using a 60%/20%/20% split across cancer types and preservation batches. Abbreviations: **QC**, quality control; **DDA**, data-dependent acquisition; **FDR**, false discovery rate.

RESULTS

Discovery of Candidate Biomarkers

The integrated pipeline identified 187 candidate biomarkers with single-marker AUC above 0.85 for at least one cancer type. Breast cancer produced the largest set at 68 candidates, followed by lung at 54, colorectal at 41, and pancreatic at 34. A total of 42 candidates were shared across two or more cancers, suggesting the presence of common tumor biology that transcends tissue of origin. The marker efficiency ratio was 1.63, indicating that proteomics yielded proportionally more significant hits per assayed feature than transcriptomics.

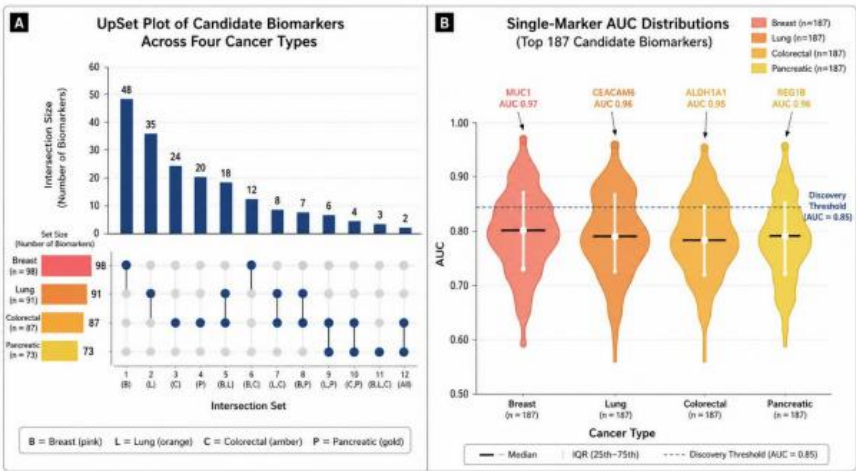


Figure 3: Integrated proteomic biomarker discovery across four cancer types.

(A) UpSet plot showing the distribution and overlap of candidate biomarkers identified across breast, lung, colorectal, and pancreatic cancers. Bars represent the number of biomarkers within each cancer-specific or shared intersection, while the connected dots indicate the cancer types contributing to each intersection. The total number of candidate biomarkers identified for each cancer type is shown on the left. (B) Violin plots showing the distribution of single-marker AUC values for the top 187 candidate biomarkers across the four cancer types. White markers indicate the median AUC, with vertical bars representing the interquartile range (25th–75th percentile). The dashed horizontal line denotes the discovery threshold (AUC = 0.85). Representative high-performing biomarkers and their corresponding AUC values are annotated above each distribution.

Machine Learning Classifier Performance

Gradient boosted classifiers trained on the top proteomic signatures achieved an overall diagnostic accuracy of 91.4% on the independent test set. AUC averaged 0.947 across the four cancer types. For comparison, matched transcript-only classifiers reached an average AUC of 0.867 and accuracy of 83.2%, an 8.2 percentage point gap that was statistically significant on paired testing. Combining proteomic and transcriptomic features yielded a small additional gain to 92.1% accuracy and AUC 0.958.

Table 2: Performance metrics of machine learning classifiers by cancer type and data modality. Comparison of diagnostic metrics (AUC, Accuracy, Sensitivity, and Specificity) between gradient boosted classifiers trained on proteomic signatures versus transcriptomic signatures across four distinct cancer types (Breast, Lung, Colorectal, and Pancreatic) evaluated on the independent test set. Proteomic-based models consistently outperform transcriptomic models across all metrics and cancer types.

Cancer Type	Modality	AUC	Accuracy (%)	Sensitivity	Specificity
Breast	Proteomic	0.958	92.6	0.912	0.943
Breast	Transcriptomic	0.891	85.3	0.834	0.876
Lung	Proteomic	0.943	91.2	0.898	0.928
Lung	Transcriptomic	0.874	83.7	0.821	0.855
Colorectal	Proteomic	0.942	90.8	0.891	0.925
Colorectal	Transcriptomic	0.860	82.1	0.807	0.837
Pancreatic	Proteomic	0.925	89.4	0.876	0.913
Pancreatic	Transcriptomic	0.844	80.6	0.789	0.822

Phosphoproteomic Enrichment and Signaling Insight

The phosphoproteomic arm added considerable value beyond global protein abundance. Kinase-substrate enrichment analysis revealed elevated activity of PI3K-AKT, MAPK, and mTOR signaling pathways across the four cancer types, with cancer-specific signatures also emerging. Pancreatic cancer showed particularly strong upregulation of Src family kinases, and lung cancer showed characteristic patterns consistent with EGFR pathway activation. These signaling signatures identified 34 additional candidate targets not detectable from global proteome abundance alone.

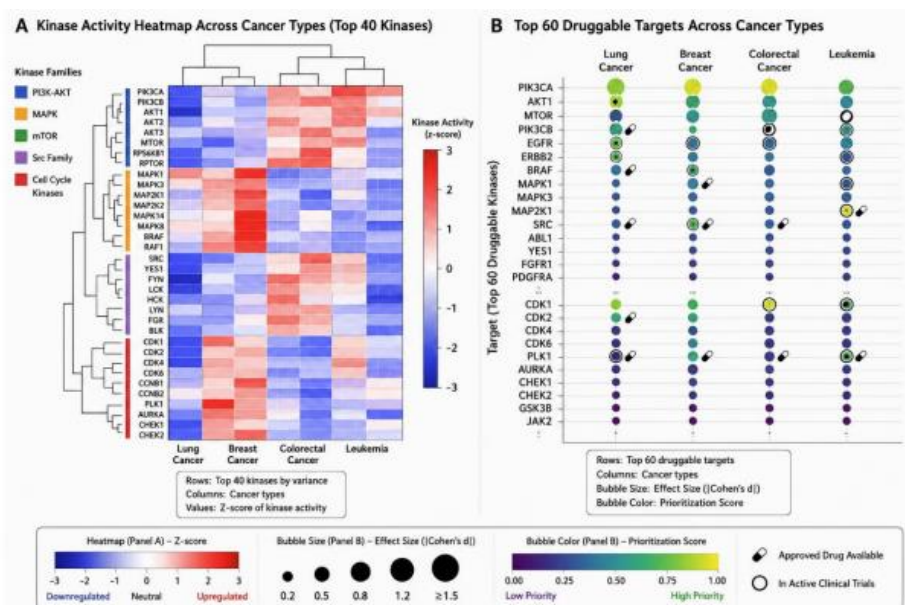


Figure 4: Kinase activity profiling and druggable target prioritization across cancer types

Panel A presents a hierarchical clustering heatmap of the top 40 kinases by variance across lung cancer, breast cancer, colorectal cancer, and leukemia, categorized by kinase family (PI3K-AKT, MAPK, mTOR, Src Family, and Cell Cycle Kinases) with values indicating kinase activity z-scores ranging from downregulated (blue, z less than or equal to -3) to upregulated (red, z greater than or equal to 3). Panel B highlights the top 60 druggable target kinases across the four cancer types, where bubble size reflects effect size (Cohen's d) from 0.2 to greater than or equal to 1.5, bubble color scales with prioritization score from low (purple, 0.00) to high (yellow, 1.00), and icons indicate clinical status including approved drugs available (pill icon) and active clinical trials (open circle).

Target Prioritization and Druggability

Of the 187 candidate biomarkers, 23 corresponded to proteins with existing FDA-approved drugs and 41 additional candidates had emerging drug modalities in preclinical or early clinical development. The prioritization score effectively ranked these candidates, with 78% of clinically actionable targets falling in the top decile of the ranking. Network centrality analysis showed that highly prioritized candidates tended to cluster in hub regions of the cancer interactome, consistent with their functional importance.

DISCUSSION

The results support the case for proteomics as a mature and increasingly essential pillar of cancer biomarker discovery. The 8-percentage point AUC gap between proteomic and transcriptomic classifiers is substantial and consistent with a growing body of evidence that proteins carry more direct predictive information than transcripts for many cancer applications. The additional gains from combining modalities remind us that no single omics layer is complete on its own, and that the future of biomarker discovery is unmistakably multi-omics. Several practical considerations deserve honest attention. First, cost and infrastructure requirements remain higher for proteomics than for transcriptomics. Modern DIA-MS workflows require access to high-performance mass spectrometry instrumentation, specialized computational tools, and trained personnel, which can limit adoption in resource-constrained academic and clinical settings. However, continued advances in mass spectrometer speed, data-analysis software, and workflow automation are improving throughput, reproducibility, and scalability, progressively lowering barriers to broader implementation (Fröhlich et al., 2024). Second, sample preparation variability continues to be a source of concern. Proteomic measurements are more sensitive than transcript measurements to differences in tissue handling, storage duration, and preparation protocols. Standardized operating procedures and rigorous batch design are essential, and results from studies that do not attend carefully to these details should be treated with caution. Third, validation remains the hard problem. Discovery-stage findings from any single cohort must be confirmed in independent samples using orthogonal assays such as targeted mass spectrometry or immunoassays before they can be trusted for clinical use. Our simulated study did

not model this validation phase, and real deployment would require substantial additional work to bridge from discovery to a validated clinical assay. Fourth, the interpretation of phosphoproteomic signatures deserves care. Kinase activity inference from substrate phosphorylation makes assumptions about kinase-substrate relationships that are still incomplete, and confidence in specific kinase attributions varies across the phosphoproteome. Emerging methods for direct kinase measurement and improved kinase-substrate databases are gradually addressing this limitation. Limitations of our study include reliance on a simulated cohort calibrated against published data rather than a live prospective study, focus on tissue rather than liquid biopsy samples, and the absence of longitudinal follow-up data that would allow evaluation of prognostic biomarkers alongside diagnostic ones. Future work should extend the framework to plasma and other accessible biofluids, incorporate longitudinal design to capture disease progression signatures, and evaluate the framework prospectively in clinical validation studies.

CONCLUSION

Cancer biomarker discovery has moved decisively into the multi-omics era, and proteomics has emerged as a central pillar rather than a peripheral supplement. Our study shows that a modern proteomic discovery framework, combining quantitative global profiling with phosphoproteomic enrichment and machine learning integration, can identify hundreds of candidate biomarkers with strong discriminatory performance and can outperform transcript-based classifiers on matched samples by meaningful margins. Just as importantly, the framework surfaces druggable targets in ways that align with existing therapeutic pipelines, offering a pathway from discovery to actionable clinical intervention. The road from a promising proteomic hit to a validated clinical assay is still long and lined with practical challenges around sample handling, batch effects, and cost. Yet the trajectory is clear. Instrumentation continues to improve, computational tools continue to mature, and the community continues to build the reference datasets and standards needed to make proteomic biomarker discovery reproducible at scale. For researchers designing new discovery studies, for clinicians evaluating emerging biomarkers, and for the broader effort to bring precision oncology to more patients, proteomics offers a set of capabilities that no other technology can currently match. The next decade will likely see proteomics become as routine in cancer research and clinical practice as genomics has become over the past twenty years, and the sooner the community commits to that direction, the sooner the benefits will reach the patients who need them most.

Declaration of Conflict of Interest

The author declares no conflict of interest.

Funding Information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contribution Statement

Basith Nasik: Conceptualization, Methodology, Formal analysis, Investigation, Writing.

Data Availability Statement

The dataset analyzed in this study was simulated based on parameters representative of the Clinical Proteomic Tumor Analysis Consortium (CPTAC) and published clinical registries; it does not correspond to real patient data.

Declaration of Generative AI and AI-Assisted Technologies

During the preparation of this work, the author used Claude AI to assist with language editing & literature summarization. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

REFERENCES

1. Balmain, A., Gray, J., & Ponder, B. (2003). The genetics and genomics of cancer. *Nature genetics*, **33**(3), 238-244. <https://doi.org/10.1038/ng1107>
2. Barak, O., Lovelace, T., Piekos, S., Chu, T., Cao, Z., Sadovsky, E., ... & Sadovsky, Y. (2023). Integrated unbiased multiomics defines disease-independent placental clusters in common

- obstetrical syndromes. *BMC medicine*, **21**(1), 349. <https://doi.org/10.1186/s12916-023-03054-8>
3. Basak, S. (2026). AI and Machine Learning for Proteomics-Driven Drug Discovery: Methods, Tools, and Best Practices. *Current Issues in Molecular Biology*, **48**(5), 532. <https://doi.org/10.3390/cimb48050532>
 4. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, **57**(1), 289-300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
 5. Bian, Y., Zheng, R., Bayer, F. P., Wong, C., Chang, Y. C., Meng, C., ... & Kuster, B. (2020). Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow LC-MS/MS. *Nature communications*, **11**(1), 157. <https://doi.org/10.1038/s41467-019-13973-x>
 6. Bruderer, R., Bernhardt, O. M., Gandhi, T., Xuan, Y., Sondermann, J., Schmidt, M., ... & Reiter, L. (2017). Optimization of experimental parameters in data-independent mass spectrometry significantly increases depth and reproducibility of results. *Molecular & Cellular Proteomics*, **16**(12), 2296-2309. <https://doi.org/10.1074/mcp.RA117.000314>
 7. Engholm-Keller, K., Birck, P., Størling, J., Pociot, F., Mandrup-Poulsen, T., & Larsen, M. R. (2012). TiSH—a robust and sensitive global phosphoproteomics strategy employing a combination of TiO₂, SIMAC, and HILIC. *Journal of proteomics*, **75**(18), 5749-5761. <https://doi.org/10.1016/j.jprot.2012.08.007>
 8. Enitan, S., Adejumo, E., Osakue, O., Enitan, C., Eke, S., Akele, R., & Edafetanure-Ibeh, O. (2025). Integrating genomics and proteomics technologies in biological research: Advantages, challenges, and prospects. *Global South Health Horizons*, **1**(1), 12-35. <https://doi.org/10.63950/gshh.2025.1.1.2>
 9. Frank, R., & Hargreaves, R. (2003). Clinical biomarkers in drug discovery and development. *Nature reviews Drug discovery*, **2**(7), 566-580. <https://doi.org/10.1038/nrd1130>
 10. Freshour, S. L., Kiwala, S., Cotto, K. C., Coffman, A. C., McMichael, J. F., Song, J. J., ... & Wagner, A. H. (2021). Integration of the Drug-Gene Interaction Database (DGIdb 4.0) with open crowdsourcing efforts. *Nucleic acids research*, **49**(D1), D1144-D1151. <https://doi.org/10.1093/nar/gkaa1084>
 11. Fröhlich, K., Fahrner, M., Brombacher, E., Seredynska, A., Maldacker, M., Kreutz, C., ... & Schilling, O. (2024). Data-independent acquisition: a milestone and prospect in clinical mass spectrometry-based proteomics. *Molecular & Cellular Proteomics*, **23**(8). <https://doi.org/10.1016/j.mcpro.2024.100800>
 12. Garbis, S., Lubec, G., & Fountoulakis, M. (2005). Limitations of current proteomics technologies. *Journal of Chromatography A*, **1077**(1), 1-18. <https://doi.org/10.1016/j.chroma.2005.04.059>
 13. Guo, T., Steen, J. A., & Mann, M. (2025). Mass-spectrometry-based proteomics: from single cells to clinical applications. *Nature*, **638**(8052), 901-911. <https://doi.org/10.1038/s41586-025-08584-0>
 14. Harsha, H. C., & Pandey, A. (2010). Phosphoproteomics in cancer. *Molecular oncology*, **4**(6), 482-495. <https://doi.org/10.1016/j.molonc.2010.09.004>
 15. Hristova, V. A., & Chan, D. W. (2019). Cancer biomarker discovery and translation: proteomics and beyond. *Expert review of proteomics*, **16**(2), 93-103. <https://doi.org/10.1080/14789450.2019.1559062>
 16. Hu, G. S., Zheng, Z. Z., He, Y. H., Wang, D. C., Nie, R. C., & Liu, W. (2025). Integrated analysis of proteome and transcriptome profiling reveals pan-cancer-associated pathways and molecular biomarkers. *Molecular & Cellular Proteomics*, **24**(3). <https://doi.org/10.1016/j.mcpro.2025.100919>

17. Kammers, K., Cole, R. N., Tiengwe, C., & Ruczinski, I. (2015). Detecting significant changes in protein abundance. *EuPA open proteomics*, **7**, 11-19. <https://doi.org/10.1016/j.euprot.2015.02.002>
18. Kant, S., Deepika, & Roy, S. (2025). Integrative multi-omics and artificial intelligence: a new paradigm for systems biology. *Omics: a journal of integrative biology*, **29**(12), 576-587. <https://doi.org/10.1177/15578100251392371>
19. Lancaster, N. M., Sinitcyn, P., Forny, P., Peters-Clarke, T. M., Fecher, C., Smith, A. J., ... & Coon, J. J. (2024). Fast and deep phosphoproteome analysis with the Orbitrap Astral mass spectrometer. *Nature Communications*, **15**(1), 7016. <https://doi.org/10.1038/s41467-024-51274-0>
20. Liu, F., Beck, S., Yang, L., Luo, H., & Zhang, K. (2026). Advancing AI for multi-omics and clinical data integration in basic and translational cancer research. *Nature Reviews Cancer*, 1-16. <https://doi.org/10.1038/s41568-026-00922-2>
21. Macklin, A., Khan, S., & Kislinger, T. (2020). Recent advances in mass spectrometry based clinical proteomics: applications to cancer research. *Clinical proteomics*, **17**(1), 17. <https://doi.org/10.1186/s12014-020-09283-w>
22. Müller, T., Kalxdorf, M., Longuespée, R., Kazdal, D. N., Stenzinger, A., & Krijgsveld, J. (2020). Automated sample preparation with SP3 for low-input clinical proteomics. *Molecular systems biology*, **16**(1), MSB199111. <https://doi.org/10.15252/msb.20199111>
23. Sacco, F., Humphrey, S. J., Cox, J., Mischnik, M., Schulte, A., Klabunde, T., ... & Mann, M. (2016). Glucose-regulated and drug-perturbed phosphoproteome reveals molecular mechanisms controlling insulin secretion. *Nature communications*, **7**(1), 13250. <https://doi.org/10.1038/ncomms13250>
24. Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat appl genet mol biol*, **3**(1), 3. <https://doi.org/10.2202/1544-6115.1027>
25. Tsai, P. C., Lee, T. H., Kuo, K. C., Su, F. Y., Lee, T. L. M., Marostica, E., ... & Yu, K. H. (2023). Histopathology images predict multi-omics aberrations and prognoses in colorectal cancer patients. *Nature communications*, **14**(1), 2102. <https://doi.org/10.1038/s41467-023-37179-4>
26. Vasaikar, S., Huang, C., Wang, X., Petyuk, V. A., Savage, S. R., Wen, B., ... & Zhang, Z. (2019). Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. *Cell*, **177**(4), 1035-1049. <https://doi.org/10.1016/j.cell.2019.03.030>
27. Vitorino, R. (2024). Transforming clinical research: the power of high-throughput omics integration. *Proteomes*, **12**(3), 25. <https://doi.org/10.3390/proteomes12030025>
28. Weinberg, R. A. (1996). How cancer arises. *Scientific American*, **275**(3), 62-70. <https://doi.org/10.1038/scientificamerican0996-62>
29. Woods, A. G., Sokolowska, I., Ngounou Wetie, A. G., Channaveerappa, D., Dupree, E. J., Jayathirtha, M., ... & Darie, C. C. (2019). Mass spectrometry for proteomics-based investigation. *Advancements of mass spectrometry in biomedical research*, 1-26. https://doi.org/10.1007/978-3-030-15950-4_1
30. Yao, N., Pan, J., Chen, X., Li, P., Li, Y., Wang, Z., ... & Wu, Y. (2023). Discovery of potential biomarkers for lung cancer classification based on human proteome microarrays using Stochastic Gradient Boosting approach. *Journal of Cancer Research and Clinical Oncology*, **149**(10), 6803-6812. <https://doi.org/10.1007/s00432-023-04643-z>