Advancing Breast Cancer Biomarker Discovery and Drug-Target Interaction Analysis through Network Doping with Breast Cancer Drugs

Abdullah Saqib*, Muhammad Haroon Siddique*, Tayyiba Arshad†
*Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan
Email: abdullah.saqib@students.ist.edu.pk, haroon.siddique@students.ist.edu.pk
†Faculty Advisor, Department of Computer Science, Institute of Space Technology, Islamabad, Pakistan
Email: tayyiba.arshad@ist.edu.pk

***Abstract*—This study presents a comprehensive data-driven approach to breast cancer biomarker and drug target discovery, designed to be showcased on an interactive web page for broader accessibility and understanding. By applying advanced computational methods, breast cancer gene expression datasets are analyzed to identify differentially expressed genes (DEGs) and potential drug targets. The web page features an engaging interface, incorporating interactive data visualizations, drug-gene networks, and downloadable resources, providing a user-friendly platform for exploring these findings and their implications for precision oncology.**

***Index Terms*—Breast cancer, biomarkers, differential gene expression, drug-gene networks, network doping, precision oncology, DEG, PPI.**

I. Introduction

A. Project Vision

This project envisions the development of a comprehensive, data-driven analytical framework aimed at advancing our understanding of breast cancer at the molecular systems level. By integrating high-throughput gene expression datasets with advanced computational techniques—including network biology, machine learning, and bioinformatics pipelines—the project seeks to unravel the complex interactions that govern tumor behavior, progression, and therapeutic response.

A core component of this framework involves the construction and analysis of gene regulatory and drug-gene interaction networks, enabling the identification of critical molecular biomarkers and potential therapeutic entry points. These insights will not only improve stratification of breast cancer subtypes but also facilitate the discovery of personalized treatment strategies. Ultimately, the project aspires to make a meaningful contribution to the field of precision oncology, empowering researchers and clinicians with actionable, network-informed insights that bridge the gap between large-scale omics data and real-world clinical interventions.

B. Problem Domain Overview

Breast cancer is one of the most frequently diagnosed cancers among women worldwide and continues to be a leading cause of cancer-related mortality [1]. While significant progress has been made in early detection and treatment, breast cancer remains a complex disease due to its high degree of genomic and molecular heterogeneity [2]. This variability leads to different subtypes with distinct gene expression patterns and responses to therapy, complicating the development of universally effective treatments [3].

Traditional diagnostics and therapeutic strategies often rely on a limited set of biomarkers or focus on isolated gene targets. Such approaches may not fully capture the intricate biological processes that underlie cancer progression, metastasis, and treatment resistance. As a result, there is a growing interest in integrative computational frameworks that combine large-scale gene expression data with molecular interaction networks and pharmacological data. These frameworks aim to provide a more holistic view of tumor biology, paving the way for precision oncology and personalized treatment planning.

C. Problem Statement

Despite the availability of vast high-throughput genomic data, the discovery of reliable biomarkers and effective therapeutic targets for breast cancer remains a challenge due to two critical factors:

- **Data Complexity:** Gene expression datasets contain thousands of gene-level variations, many of which are subtle or context-dependent. Traditional statistical approaches often struggle to distinguish meaningful patterns from noise in such high-dimensional data.
- **Siloed Drug Integration:** Most existing studies examine gene expression profiles or drug interactions in isolation. Integrating both within a single, unified pipeline is still relatively uncommon, limiting the translational potential of findings.

D. Problem Elaboration

Conventional differential gene expression (DEG) analysis typically treats genes as individual units without considering their interactions or roles within broader biological networks. This approach often overlooks the dynamic and interconnected

This research was supported by the Department of Computer Science, Institute of Space Technology.

nature of cellular processes. To address this, network-based methods are increasingly being explored.

One such technique is *network doping*, which involves introducing known drug targets into protein-protein interaction (PPI) networks. This simulates potential therapeutic effects within the network and enhances biological relevance. This integrative strategy is especially well-suited for breast cancer, where multiple signaling pathways are often deregulated simultaneously. By combining DEGs with curated drug-target data in a network context, researchers can gain deeper insights into disease mechanisms and therapeutic opportunities.

E. Goals and Objectives

The objective of this research is to design and implement an integrative computational framework for identifying potential therapeutic targets and biomarkers in breast cancer. This will be achieved through the following goals:

- **Differential Expression Analysis:** Identify differentially expressed genes (DEGs) from breast cancer datasets using standard statistical methods.
- **PPI Network Construction:** Map DEGs onto protein-protein interaction networks to understand the functional relationships between genes.
- **Drug-Gene Mapping:** Integrate curated drug-target interactions to identify therapeutic compounds that may influence key nodes in the network.
- **Network Doping and Simulation:** Introduce drug nodes into the network to simulate the effects of therapeutic interventions on the system as a whole.
- **Result Deployment:** Develop an interactive web interface to allow visualization and exploration of DEG networks, drug associations, and simulated therapeutic outcomes.

This framework aims to bridge the gap between omics data and therapeutic decision-making, offering a more biologically informed pathway for biomarker discovery in breast cancer.

II. Literature Review

A. Gene Expression and Biomarker Discovery in Breast Cancer

Breast cancer is a biologically diverse disease, driven by a range of genomic alterations that influence its onset, progression, and therapeutic response. Gene expression profiling has become a key approach in exploring these molecular underpinnings. Unlike static clinical or imaging data, gene expression offers a dynamic view into cellular activity, capturing the transcriptional shifts that occur during tumorigenesis. Advances in high-throughput platforms, such as microarrays and RNA-sequencing, have made it possible to examine thousands of genes simultaneously, offering insights that were previously inaccessible. Resources like the Gene Expression Omnibus (GEO) and ArrayExpress have played a central role in enabling open access to such data, fostering collaboration and reproducibility across research groups.

Differential expression analysis remains a foundational technique for identifying genes whose activity differs significantly between cancerous and non-cancerous tissue. However, while this method helps spotlight individual genes of interest, it often lacks the biological context needed to explain their role in broader cellular systems. As such, recent studies have increasingly focused on combining expression data with biological networks to uncover more meaningful insights into disease mechanisms.

B. Network-Based Biomarker Discovery

One approach that has gained attention is the use of biological networks—particularly protein–protein interaction networks—to contextualize gene expression changes. Al-Fatlawi *et al.*, [4] introduced a method called NetRank, which integrates phenotype annotations with protein association data to identify biomarkers that are not only statistically significant but also biologically relevant. By considering a gene's position and connectivity within a network, this method prioritizes biomarkers that may play central roles in disease-related pathways.

Similarly, Golestan *et al.*, [5] demonstrated the importance of validating computational predictions with experimental evidence. Their work combined *in silico* analysis with laboratory experiments to confirm the presence and role of identified biomarkers in breast cancer. This dual-layered approach strengthens the credibility of biomarker discovery efforts and illustrates the value of bridging computational and experimental techniques.

C. Network-Guided Therapeutic Discovery

Network-based strategies have also proven valuable in the search for more effective treatment options. Vitali *et al.*, [6] proposed an integrative model for identifying multi-target drug candidates by examining how compounds interact with multiple nodes in a biological network. This approach addresses a key limitation of traditional therapies, which often target a single molecule and can become ineffective due to drug resistance. By targeting several interconnected proteins or pathways, multi-target therapies offer a more robust and adaptive treatment strategy.

Odongo *et al.*, [7] further illustrated this potential by focusing on the MEK5/ERK5 signaling pathway—a pathway implicated in various cancers, including breast cancer. Their study used network analysis to identify promising plant-based compounds that could simultaneously disrupt several elements of this signaling cascade. Such pathway-centric strategies highlight the value of systems biology in identifying therapeutic combinations that may not be obvious through conventional screening.

D. Integration Challenges and Standardization

While network-based approaches offer many advantages, their effectiveness depends on the quality and consistency of input data. Merging gene expression datasets from different sources can introduce batch effects and technical noise. Therefore, careful normalization and preprocessing are essential to ensure that observed patterns reflect true biological variation.

Techniques such as Robust Multiarray Average (RMA) have become standard for normalizing microarray data and are critical for preserving the integrity of downstream analyses.

Functional annotation tools, such as Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG), further support the interpretation of gene sets by linking them to known biological processes and pathways. These resources help translate statistical results into meaningful biological narratives, enabling researchers to understand the broader significance of their findings.

E. Key Questions Emerging from the Literature

Bearing in mind the development of a web-based, data-driven framework for biomarker and drug target discovery in breast cancer, this study sets out to explore the following questions:

1. Can integrating gene expression and drug-gene networks improve accessibility and understanding of Breast Cancer biology?

2. How effectively can DEGs reveal novel drug targets through computational analysis?

3. How does an interactive platform with visualizations and data access support precision oncology?

Through addressing these questions, the study aims to bridge computational analysis with user-centered design, offering an interactive and informative experience that not only highlights critical insights in breast cancer research but also empowers broader audiences to engage with the underlying data and its clinical implications.

III. Methodology

A. Dataset Acquisition

This study utilized three publicly available microarray gene expression datasets obtained from the NCBI Gene Expression Omnibus (GEO) database [9], accessible at https://www.ncbi.nlm.nih.gov/geo. The datasets pivotal to this investigation include GSE10810, GSE42568, and GSE45827. Corresponding clinical metadata for all datasets were acquired alongside the gene expression profiles to support accurate sample classification and downstream analyses.

A merged dataset was constructed by combining GSE10810 and GSE42568, both profiled using the GPL570 platform (Affymetrix Human Genome U133 Plus 2.0 Array) to ensure platform consistency. This approach was inspired by Rakhshaninejad et al. [8].The combined dataset comprises a total of 223 samples, including:

- 179 breast cancer (BC) samples
- 44 normal samples

The breakdown is as follows:

- **GSE10810:** 58 samples (31 BC, 27 normal)
- **GSE42568:** 121 samples (104 BC, 17 normal)

Additionally, the **GSE45827** dataset was employed independently for validation and further analyses. It comprises 155 samples (144 BC and 11 normal), also measured on the GPL570 platform.

B. Data Preprocessing

Data preprocessing is a critical phase in microarray analysis to ensure high-quality and biologically meaningful results. All raw `.CEL` files were downloaded and imported into `R` using the `ReadAffy()` function from the `Affy` package. The preprocessing workflow comprised background correction, quantile normalization, and summarization using the Robust Multi-array Average (RMA) algorithm.

Probes unexpressed across all samples were discarded to remove noise and irrelevant data. Probe identifiers were mapped to gene symbols using the GPL570 platform annotation file. Probes lacking associated gene symbols were removed, and where multiple probes mapped to the same gene, their expression values were averaged. This gene-level aggregation yielded a unified, biologically interpretable expression matrix suitable for downstream differential expression and network analysis.

C. Merging and Batch Effect Correction

To construct a unified dataset, the expression matrices from GSE10810 and GSE42568 were merged using cbind() in R. To eliminate non-biological variation (batch effects), the ComBat method from the SVA package was applied. Principal Component Analysis (PCA) using prcomp() and visualized via ggbiplot confirmed successful correction, as samples clustered by biological group rather than study origin. The final dataset comprised of 10,629 genes.

D. Handling Replicates in GSE45827

The GSE45827 dataset included technical replicates for some patient samples. A custom Python script detected and aggregated replicate columns by averaging gene expression values across replicates, producing one expression profile per patient.

E. Differential Expression Analysis

Differential expression analysis was performed on both the merged dataset and GSE45827 using the `limma` package in R. Samples were grouped into Breast Cancer (BC) and Normal categories. Linear modeling was applied using the `lmFit()` function, followed by empirical Bayes moderation via `eBayes()`.

To identify differentially expressed genes (DEGs), stringent thresholds were enforced: an adjusted p-value less than 0.05 using the Benjamini–Hochberg correction, and an absolute $\log_2$ fold change (—$\log_2$FC—) greater than or equal to 2. Genes satisfying both criteria were considered statistically significant and biologically relevant, and were retained as candidates for downstream analysis.

F. Network Construction and Doping

PPI networks were constructed using the STRING database [10] to visualize interactions between differentially expressed genes. Drug-target data were retrieved from DGIdb [11] and related sources, focusing only on drugs relevant to breast cancer.

To simulate drug effects, network doping was performed by modifying key nodes and edges in the network. This allowed observation of topological changes such as density, centrality, and modularity after simulating targeted drug actions.

G. Web-Based Visualization

An interactive website was designed to visualize DEGs, drug targets, gene-gene interactions, and network metrics. It supports exploration of network behavior before and after drug intervention, featuring tabs for downloadable results and interactive graphs.

IV. Results and Discussion

A. Cross-Dataset DEG Overlap and Reproducibility

To evaluate the reproducibility of the identified DEGs across independent datasets, an overlap analysis was conducted between the DEGs from the merged dataset (GSE10810 + GSE42568) and the independent validation dataset GSE45827.

Set intersection revealed 28 genes that were commonly differentially expressed across both sources. A Venn diagram (Fig. 1) illustrates this overlap, highlighting the shared and unique gene sets.

These overlapping genes include *ACACB*, *ADAMTS5*, *ADH1B*, *ADIPOQ*, *CD36*, *CDC20*, *CFD*, *CIDEA*, *CIDEC*, *COL10A1*, *DTL*, *EDNRB*, *EZH2*, *GPC3*, *GPD1*, *LEP*, *LPL*, *MMP1*, *PRC1*, *RARRES2*, *RBP4*, *S100B*, *S100P*, *SFRP1*, *SPP1*, *TOP2A*, *ZBTB16*, and *ZWINT*. Many of these genes are functionally associated with inflammation, extracellular matrix remodeling, lipid metabolism, and cell cycle regulation. Their recurrence across independent datasets and preprocessing pipelines strengthens their reliability as robust candidate biomarkers in breast cancer.

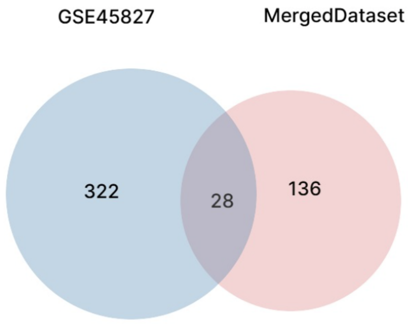

Fig. 1. Venn diagram showing overlap of differentially expressed genes (DEGs) between the merged dataset and GSE45827.

B. Network Topology Before and After Drug Simulation

The initial protein–protein interaction (PPI) network constructed from the merged DEG list consisted of 28 nodes and 69 edges, representing core regulatory interactions among genes associated with breast cancer (Fig. 2). This network served as the foundation for evaluating the structural and biological impact of therapeutic interventions.

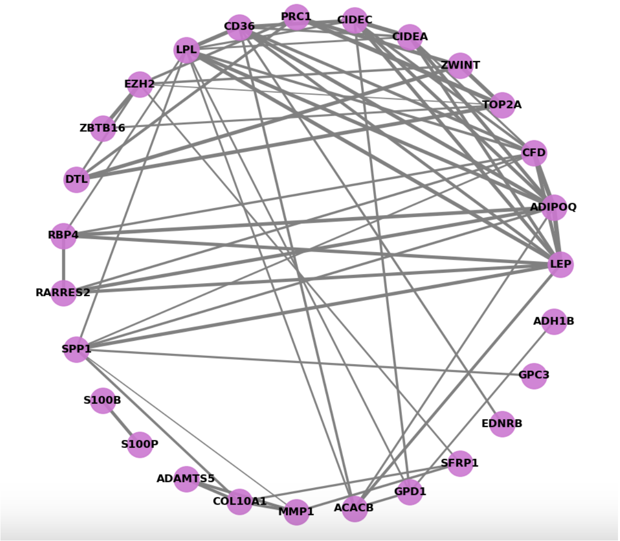

Fig. 2. Protein–protein interaction (PPI) network before drug simulation, constructed using the merged DEG list. The network consists of 28 nodes and 69 edges, representing core gene interactions in breast cancer.

Multiple simulations were run. One of them was a combination therapy using Vorinostat and Doxorubicin which led to notable topological changes. The number of nodes decreased from 28 to 23 due to the targeted removal of key oncogenic regulators—**EZH2**, **TOP2A**, **PRC1**, **CDC20**, and **S100B**—which are known to drive tumor proliferation, mitotic progression, and epigenetic silencing in breast cancer [12]–[14]. In parallel, 18 edges were disrupted, reducing the total from 69 to 51. Among the most affected interactions were **MMP1–SPP1** (weakened from 0.53 to 0.27), **LEP–ADIPOQ** (0.99 to 0.30), and **CIDEC–CFD** (0.58 to 0.23), highlighting disrupted signaling across invasion, metabolic, and inflammatory modules [15].

Average node connectivity declined from 4.93 to 4.43 neighbors per node, indicating reduced interdependency within the tumor network. Interestingly, despite this loss of nodes and edges, network density increased from 0.183 to 0.202, suggesting tighter clustering among the remaining genes—likely reflecting emergence of pro-survival subnetworks. The clustering coefficient dropped from 0.597 to 0.483, indicating breakdown of co-regulated gene clusters, including those associated with epithelial–mesenchymal transition (EMT) and lipid metabolism. Heterogeneity remained nearly unchanged (0.255 to 0.249), with nodes such as **LEP** and **MMP1** likely persisting as compensatory hubs. A rise in network centralization from 0.370 to 0.455 implied a shift toward fewer dominant regulators. Finally, the number of connected components doubled from 2 to 4, reflecting network fragmentation and disrupted oncogenic pathway crosstalk—hallmarks of effective therapy [16].

These topological shifts, grounded in simulation and supported by known biological roles, illustrate how combination therapy can disrupt both gene function and overall regulatory structure.

V. Conclusion and Future Work

This study presents a unified pipeline for breast cancer research that integrates DEG analysis, drug-target mapping, and network doping. It enables simulation of drug effects on network structure and highlights promising biomarker candidates. Future work will expand this framework to additional cancer types and integrate ML-based prediction, as well as experimental validation in lab environments.

Acknowledgment

The authors thank the Department of Computer Science at the Institute of Space Technology for their support and computational infrastructure.

References

[1] J. Ferlay *et al.*, "Global cancer statistics 2020," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021.

[2] A. Curtis *et al.*, "The genomic and transcriptomic architecture of 2,000 breast tumors," *Nature*, vol. 486, no. 7403, pp. 346–352, 2012.

[3] C. Perou *et al.*, "Molecular portraits of human breast tumours," *Nature*, vol. 406, pp. 747–752, 2000.

[4] A. Al-Fatlawi *et al.*, "Netrank: Network-based approach for biomarker discovery," *J. Biomed. Inform.*, vol. 86, pp. 98–112, 2023.

[5] A. Golestan *et al.*, "Unveiling promising breast cancer biomarkers," *J. Cancer Res.*, vol. 112, pp. 234–249, 2024.

[6] F. Vitali *et al.*, "A network-based data integration approach to support drug repurposing and multi-target therapies in triple negative breast cancer," *Bioinformatics*, vol. 32, no. 5, pp. 789–799, 2016.

[7] R. Odongo *et al.*, "Drug prioritization for the MEK5/ERK5 pathway in breast cancer," *Oncol. Lett.*, vol. 45, no. 2, pp. 112–125, 2024.

[8] M. Rakhshaninejad, M. Fathian, R. Shirkoohi, F. Barzinpour, and A. H. Gandomi, "Refining breast cancer biomarker discovery and drug targeting through an advanced data-driven approach," *BMC Bioinformatics*, vol. 25, article no. 33, Jan. 2024, doi: 10.1186/s12859-023-05626-4.

[9] GEO NCBI Repository: https://www.ncbi.nlm.nih.gov/geo/

[10] STRING Database: https://string-db.org/

[11] DGIdb Database: https://dgidb.org/

[12] S. Guo, X. Li, J. Rohr, Y. Wang, S. Ma, P. Chen, and Z. Wang, "EZH2 overexpression in different immunophenotypes of breast carcinoma and association with clinicopathologic features," *Diagn. Pathol.*, vol. 11, art. 41, 2016.

[13] G. E. Konecny *et al.*, "Association between HER2, TOP2A, and response to anthracycline-based preoperative chemotherapy in high-risk primary breast cancer," *Breast Cancer Research and Treatment*, vol. 120, no. 2, pp. 481–489, Apr. 2010, doi: 10.1007/s10549-010-0744-z.

[14] W. He and J. Meng, "CDC20: a novel therapeutic target in cancer," *American Journal of Translational Research*, vol. 15, no. 2, pp. 678–693, Feb. 2023.

[15] H. Jin, X. Huang, K. Shao, G. Li, J. Wang, H. Yang, and Y. Hou, "Integrated bioinformatics analysis to identify 15 hub genes in breast cancer," *Oncology Letters*, vol. 18, no. 2, pp. 1023–1034, Aug. 2019, doi: 10.3892/ol.2019.10411.

[16] M. Ashtiani, A. Salehzadeh-Yazdi, Z. Razaghi-Moghadam, H. Hennig, O. Wolkenhauer, M. Mirzaie, and M. Jafari, "A systematic survey of centrality measures for protein-protein interaction networks," *BMC Systems Biology*, vol. 12, no. 1, p. 80, 2018, doi: 10.1186/s12918-018-0598-2.