Thesis Report
on

Insights into the mechanisms of resistance to Axitinib in cancer cell lines

by

Khyati Patni (MT21304)
under the supervision of
Dr. Jaspreet Kaur Dhanjal

Submitted in partial fulfillment of the requirements for
the degree of Master of Technology
in
Computational Biology

Department of Computational Biology,
Indraprastha Institute of Information Technology - Delhi

Certificate

This is to certify that the thesis titled "**Insights into the mechanisms of resistance to Axitinib in cancer cell lines**" being submitted by Ms. Khyati Patni for the partial fulfillment of the requirements for the degree of Masters in Technology in Computational Biology at Indraprastha Institute of Information Technology, Delhi, is an authentic record of work carried out under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree. This work has not been submitted anywhere else for the reward of any other degree.

Dr. Jaspreet Kaur Dhanjal,
Assistant Professor,
Department of Computational Biology,
Indraprastha Institute of Information Technology, Delhi,
New Delhi 110020

Declaration

I submit this project entitled "**Insights into the mechanisms of resistance to Axitinib in cancer cell lines**" to the Department of Computational Biology, Indraprastha Institute of Information Technology (IIIT), Delhi. I declare that this is my original work carried out under the guidance of Dr. Jaspreet Kaur Dhanjal, Assistant Professor, Department of Computational Biology at IIIT Delhi.

Khyati Patni,
M.Tech Student (Dec 2021 - Jan 2024),
Department of Computational Biology,
Indraprastha Institute of Information Technology, Delhi,
New Delhi 110020

Acknowledgements

I extend my heartfelt gratitude to Dr. Jaspreet Kuar Dhanjal, Assistant Professor in the Department of Computational Biology at IIIT-Delhi, for her invaluable guidance and unwavering support during the entirety of my project. In the face of numerous challenges, her wealth of experience proved instrumental in helping me navigate obstacles, ultimately paving the way for the successful completion of my thesis. Her patience, kindness, and encouragement provided me with the strength and freedom to explore my ideas, ensuring I remained on the right path throughout this academic journey. Communication with her has always been comfortable, and I consider myself truly fortunate to have had such an exceptional guide.

A special note of thanks goes to my mentor, Ms. Samridhi Gupta, a Ph.D. student in the Translation Biology Lab. Her consistent support, guidance in aligning and channeling my work, and unwavering faith in my abilities served as a constant source of motivation. Akin to an elder sister and friend, significantly enriched my research experience.

I express my gratitude to all the faculty members and staff of the Department of Computational Biology at IIIT Delhi for their continuous support throughout my college journey. I also acknowledge the IT helpdesk and M.Tech. administration for their unwavering assistance whenever needed.

My deepest appreciation goes to my parents, brother, and friends for their unswerving support. I am also thankful to individuals such as Ms. Dhvani and Ms. Simarpreet Kaur, Mr. Prateek Paul, lab mates, and batchmates whose constant encouragement made the dissertation period more manageable. Their support has been a constant source of strength and inspiration throughout this academic endeavor.

Abstract

Along with the remarkable success stories in the evolutionary journey of targeted cancer treatment, there are examples where the targeted therapies failed to match the expectations. There have been multiple reports of resistance against such targeted therapies. Thus, it becomes important to understand the underlying evolution that contributes to the developed resistance in the tumor cells.

In this study, a comprehensive, high-throughput assessment was performed wherein machine-learning models have been implemented on the cellular expression omics (transcriptomics and proteomics) data of cell lines that are resistant and sensitive to FDA approved targeted drugs. Based on the predictive ability of the models, Axitinib was found to be a drug that showed high predictive accuracy. Thus, various algorithms rooted in explainable AI were implemented to find out the key molecular players that are responsible for mediating resistance. Further, the biological pathways wherein the derived resistance mediators are over-represented were found out and the cross-talk between them were analyzed to understand the affected/perturbed biological processes. Further, it was found out that there isn't one unique way to deploy resistance but there are multiple mechanisms via which resistance is enabled in the cells. These multiple forms of resistance and their mechanisms were further elucidated.

The findings thus gained from the work help us understand the limitations of using Axitinib for cancer treatment and the vulnerabilities present in the cells due to tumor evolution. This study can take us a step further in the domain of precision medicine enabling us to map those vulnerabilities/susceptibilities in the patient genome so as to pre-empt whether a patient will be responsive to Axitinib or not.

Keywords: Cancer, drug resistance, Axitinib, transcriptomics, proteomics, explainable AI, Machine learning.

Contents

Chapter 1. Introduction 9
Chapter 2. Methodology 12
2.1 Data Sourcing 12
2.1.1 Drug Response data 12
2.1.2 Omics Datasets 12
2.2 Automated Feature Engineering and Model Training 13
2.2.1 Feature Reduction 14
2.2.2 Model Training 14
2.3 Employing Algorithms to increase the Explainability of AI 15
2.4 Clustering for possible biological mechanisms mediating resistance 16
2.5 Pathway enrichment analysis of normalized data using iDEP 16
2.5.1. Data Input and Preprocessing 16
2.5.2. Differential gene expression and pathway analysis 17
Chapter 3. Results 19
3.1 Data sourcing and preprocessing 19
3.2 Feature engineering and model implementation automation 20
3.3 Explainable AI -feature importance analyses 23
3.4 Hierarchical agglomerative clustering 24
3.5 Differential gene analysis and pathway enrichment 27
3.5.1 Data Preprocessing 27
3.5.2 Differentially expressed genes 28
3.5.3 Pathway analysis 30
Chapter 4. Discussion 41
Chapter 5. Future directions 46

List of Figures

Figure 1 Data processing pipeline for creating transcriptomics and proteomics datasets. 13
Figure 2 A block diagram of the LIME framework..... 15
Figure 3 Imbalance Labels in FDA-approved therapies 19
Figure 4 Transcriptomics: top 7 drugs with their best performance model and metrics. 21
Figure 5 Proteomics: top 7 drugs with their best performance model and metrics. 22
Figure 6 Chemical structure of compound axitinib. 22
Figure 7 Explainable AI LIME : top 10 resistance features for SNU-407 cell line..... 24
Figure 8 Transcriptomics: Clustering evaluation and visualization 25
Figure 9 Proteomics: Clustering evaluation and visualization 25
Figure 10 Transcriptomics heatmap..... 26
Figure 11 Proteomics heatmap..... 27
Figure 12 Boxplot for transcriptomics data. 28
Figure 13 Boxplot for proteomics data 28
Figure 14 Summary plots for differential expression analysis using DESeq2 on Rnaseq Data. .. 29
Figure 15 Summary plots for differential expression analysis using DESeq2 on Proteomics 29
Figure 16 Network of Biological Processes GO terms for Transcriptomics. 35
Figure 17 Network of GO Biological Processes GO terms for proteomics..... 35
Figure 18 Hierarchical clustering tree of Molecular Function GO terms for Transcriptomics 36
Figure 19 Hierarchical clustering tree of Molecular Function GO terms for Proteomics. 36
Figure 20 Enriched KEGG pathways for the upregulated genes in the y1 cluster. 37
Figure 21 Enriched KEGG pathways for the upregulated genes in the y2 cluster. 37
Figure 22 KEGG extracellular matrix–receptor (ECM-receptor) interaction pathway 38
Figure 23 Figure: KEGG “DNA Replication” pathway 40

List of Tables

Table 1 Transcriptomics: Top 7 drugs with their best performance model in bold..... 20
Table 2 Proteomics: The top 7 drugs with their best performance model in bold..... 21
Table 3 GO Biological Process pathways obtained from transcriptomics analysis..... 30
Table 4 GO Molecular Function pathways obtained from transcriptomics analysis..... 31
Table 5 GO Biological Process pathways obtained from GAGE transcriptomics analysis. 32
Table 6 GO Biological Process pathways obtained from proteomics analysis. 33
Table 7 GO Molecular Function pathways obtained from proteomics analysis..... 34

Chapter 1. Introduction

One of the most debilitating illnesses that has afflicted human life for many years is cancer. In 2020, it was responsible for approximately 10 million deaths, accounting for roughly one in six deaths globally (WHO Cancer 2020). The primary approaches to treating cancer involve systemic strategies such as radiation, hormone treatments, chemotherapy, or surgical removal of the neoplastic part of the organ. However, the choice of treatment is influenced by factors such as the tumor's size, location and stage, the grade of the disease (including whether it has metastasized), and the overall physical health of the patient [1].

Over time, the modes of action of anticancer medications have evolved. Initially, cytotoxic drugs with general mechanisms were employed, targeting rapidly dividing cells through actions on DNA replication or cell division pathways. In contrast, modern approaches involve more focused mechanisms employed by targeted agents. These agents aim to inhibit the metastasis and proliferation of cancer cells by interacting with molecular targets involved in specific pathways. Targeted therapy can act on growth factors, cell surface antigens, receptors, or signal transduction pathways that regulate cell cycle progression, angiogenesis, cell death, metastasis, depending on the specific molecular targets [2].

Tyrosine kinase inhibitors (TKIs) represent an important category of targeted therapy. Tyrosine kinases, implicated in tumorigenesis and progression, have become crucial targets for drug discovery. TKIs prevent these kinases from phosphorylating tyrosine residues of their substrates, thus blocking the induction of downstream signaling pathways. They can be classified into four groups based on their mechanisms of action: ATP-competitive, non-active conformation binders, allosteric inhibitors, and covalent inhibitors [3]. In the last two decades, there has been significant progress in the development of potent and well-tolerated TKIs targeting receptors such as EGFR, ROS1, ALK, HER2, VEGFR, NTRK, RET, MEK, MET, PDGFR, FGFR, and KIT. This advancement has played a crucial role in advancing precision cancer medicine by tailoring treatment approaches to the specific genetic alterations found in individual patients [4]. The development of TKIs affecting angiogenesis promoting receptors has significantly ameliorated the outcomes of various types of cancers, including renal cell carcinoma, hepatocellular carcinoma, and colorectal carcinoma [5].

The process of creating new blood vessels, known as angiogenesis, is essential for both normal development and pathological processes like the growth of tumors, the healing of wounds, and the development of cardiovascular, inflammatory, ischemic, and infectious disorders. A few decades ago, it was discovered that tumors implanted into isolated perfused organs were unable to grow more than a few millimeters in diameter because the diffusion of oxygen and nutrients from

surrounding capillaries is insufficient to support cell function beyond 200 μm. Multicellular tumor clones require angiogenesis and vasculogenesis to attract new blood vessels to grow further. It is now generally acknowledged that oncogene and tumor suppressor gene mutations both cause tumors to become angiogenic [6]. Tumor tissues synthesize and release angiogenic growth factors, such as vascular endothelial growth factor (VEGF), acidic and basic fibroblast growth factors (aFGF, bFGF), and platelet-derived endothelial cell growth factor (PD-ECGF) [7], in response to hypoxia, a condition in which tumor cells have been deprived of oxygen [8].

Numerous signal transduction pathways are triggered to encourage the activation of endothelial cells when these angiogenic growth factors attach to their corresponding unique tyrosine kinase receptors on the endothelial cells of pre-existing blood vessels. The original vessels then experience distinctive morphological changes, such as diameter enlargement, basement membrane degradation, a thinning of the endothelial cell lining, an increase in endothelial number, a decrease in pericyte number, and pericyte detachment, all of which ultimately aid in the growth and migration of tumor cells [9], [10].

VEGFR, a tyrosine kinase crucial in angiogenesis, is activated by the placental growth factor and the VEGF family, including VEGF-A/B/C/D/E. VEGF-A, the primary regulator, binds to VEGFR-2, stimulating endothelial cell proliferation, vascular permeability, and migration. VEGFR-targeted multi-targeted TKIs have been established as effective anti-tumor agents against various solid tumors. Activation of VEGFR leads to signaling cascades involving PI3K, MAPK, Src, and PLCγ pathways. VEGFs contribute to cancer metastasis through mechanisms such as angiogenesis stimulation, formation of disorganized vessels, promotion of inflammation, alteration of hypoxia, remodeling of tumor vessels, modification of the microenvironment, and support for metastatic regrowth in distant organs. In summary, VEGF-induced cancer metastasis involves multiple intricate steps in the metastatic signaling [11].

There are many TKIs that inhibit angiogenesis, such as sorafenib, sunitinib, axitinib and pazopanib with varying binding capabilities to angiogenic kinases, and have been recently approved as treatment for patients in advanced tumor stages. Despite their efficacy, drug resistance in cancer patients that are supplemented with anti-angiogenic medications is a crucial clinical problem. Tumors have been observed to grow resistant during the treatment duration and exhibit intrinsic resistance [12].

The diminished sensitivity of cancer cells is a complex occurrence influenced by genetic, epigenetic, and microenvironmental elements. Diverse mechanisms, such as drug inactivation, decreased internal drug retention due to lower intake or heightened removal, modifications in drug targets, initiation of alternative pathways promoting cell survival, control over DNA repair, cell cycle, and cell demise, as well as tumor adaptability and interactions with the tumor microenvironment (TME), have been identified and extensively researched [13].

Though many common drug resistance mechanisms are already known for cancer, and numerous individual drugs have been experimentally tested on tumors to determine why they fail, this process is often costly and requires elaborate experimentations to elucidate the complete biological mechanism mediating the resistance. Hence, there is a need to find alternative approaches to find these answers in a much effective and efficient way. Highly specialized techniques are currently generating vast amounts of 'omics' data, including metabolomics, proteomics, transcriptomics, genomics, epigenetics, and pharmacogenomics. The latest developments in tools and methodologies rooted in artificial intelligence (AI), along with the developments in computation abilities and high throughput data management, enable the use of these large data sources to understand the mechanisms of drug resistance [1]. The integration of computational and biological approaches, supported by experimental literature, can offer rapid insights into individual drugs for drug repurposing, combination therapies, and personalized medicine.

In this study, attempts have been made to study the intrinsic characteristics of various cancer cell lines that make them resistant to one of the targeted anticancer drugs, called axitinib. The omics-data of various cancer cell lines and their response to treatment with different targeted drugs was retrieved. Axitinib was one of the drugs for which a machine-learning based model, trained on this dataset, was able to better classify the cancer cells into resistant and sensitive groups. The features involved in prediction were further studied using advanced algorithms like LIME. The features contributing to the resistance in cancer cells were further used to cluster the resistant cell lines to find if at all there exist multiple mechanisms for making the cancer cell lines resistant to Axitinib treatment. Moving further with the hypothesis of multiple resistance mechanisms in the dataset, a comprehensive analysis was carried out to elucidate the different biological pathways behind resistance response.

Chapter 2. Methodology

In the pursuit of unraveling the intricacies of resistance mechanisms in various cell lines, a systematic methodology was crafted. The details about the data used in the study and the techniques and approaches employed to answer the core objectives of the thesis work are discussed in this chapter.

2.1 Data Sourcing

2.1.1 Drug Response data

This study embarked on a comprehensive data collection process started by harnessing the drug response data from the Genomics of Drug Sensitivity in Cancer (GDSC) database https://www.cancerrxgene.org/, a reputable source providing drug response data for 1,000 human cancer cell lines. The use of the GDSC Bulk Data Download feature Jan 2023 facilitated the efficient acquisition of data for 580 drugs, each accompanied by critical features such as TCGA classification, tissue type, tissue subtype, IC_{50}, max and min concentration, etc. To align with the focus of our research, GDSC1 data was selectively retained only for drugs lacking corresponding GDSC2 data, and among the two GDSC2 datasets containing a single compound with the same screening process, the dataset with a greater number of cell lines was chosen. This deliberate choice directed our efforts predominantly toward the latest GDSC2 version.

The selected drug response indicators for our analysis were based on half maximal inhibitory concentration (IC_{50}), AUC, and Max Concentration. IC_{50} represents the concentration causing a 50% reduction in cell viability, while AUC signifies the area under the fitted dosage-response curve. The data labeling strategy involved categorizing samples as resistant (labeled as 0) or sensitive (labeled as 1) based on the IC_{50} values and maximum concentration inducing resistance for a specific drug shown in the AUC graph. Further refinement of the dataset focused on FDA-approved targeted therapies, narrowing our investigation for clinical relevance. The FDA-approved targeted therapies list was sourced from the National Cancer Institute (NCI) https://www.cancer.gov/about-cancer/treatment/types/targeted-therapies/approved-drug-list, pre-processed and response data associated with these drugs for different cell lines was extracted from the available GDSC data. To ensure the modeling efficacy, drugs with highly imbalanced labels (one label less than 5% of overall samples) were removed from the dataset.

2.1.2 Omics Datasets

Further enriching our omics datasets, we relied on the same resource, Cell Model Passport, intricately interlinked with GDSC. Within this repository, we accessed both proteomics and transcriptomics data for human cancer cell lines. The proteomics data, termed ProCan-

DepMapSanger, furnished insights into protein intensity levels across the myriad of cell lines, was derived from 6,864 mass spectrometry runs encompassing 949 cell lines and estimating the quantification of close to 8,498 proteins. Data was already log2-transformed [14].

Parallelly, RNA-seq data collected from the Broad Institute and the Wellcome Sanger Institute [15]were processed using the iRAP pipeline (BAM files - Illumina HiSeq 2000) [16]. The primary datasets collected from each institute are accessible for download with the corresponding read counts and FPKM (fragments per kilobase million) values. The combined dataset consists of read counts, TPM (transcripts per million) and FPKM values. For the analysis, TPM data was considered and was further transformed into log2 of TPM values.

The final transcriptomics and proteomics datasets were made separately by combining omics data with drug response through automation and were utilized for subsequent machine learning analysis framed as a binary classification problem. Figure 1 summarizes the data processing pipeline.

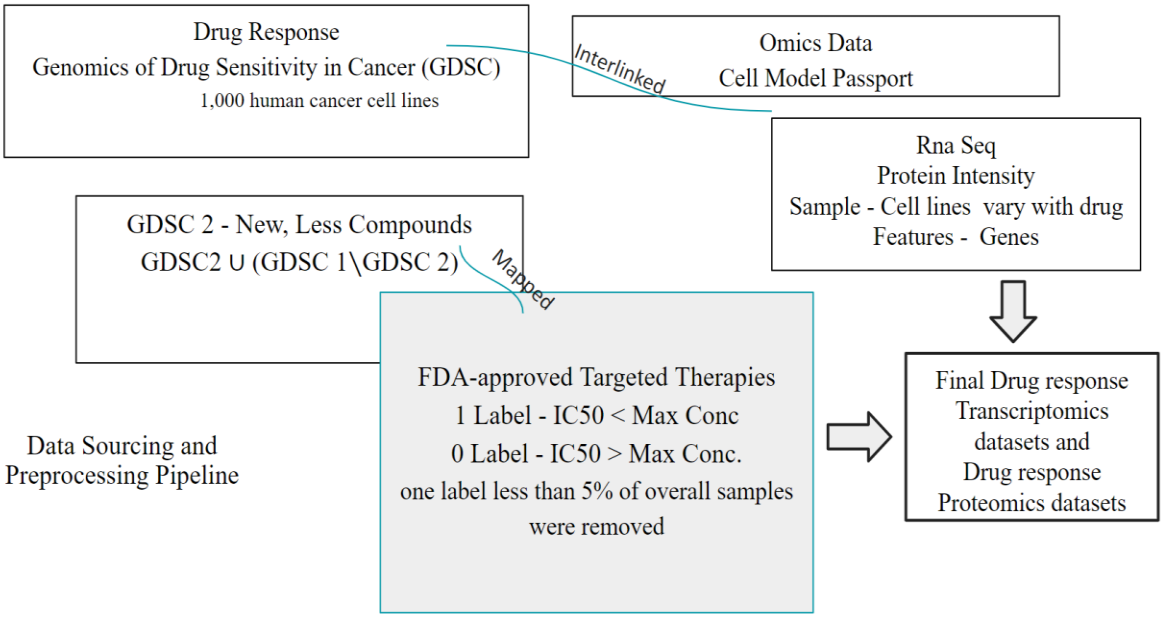


Figure 1. Data processing pipeline for creating transcriptomics and proteomics datasets.

2.2 Automated Feature Engineering and Model Training

In addressing the complex challenge of identifying the drugs whose resistance or sensitivity can be explained by machine learning algorithms, the shortlisted FDA targeted drugs and omics data integrated datasets were subjected to a constructed automated feature engineering and machine learning (ML) pipeline. Recognizing the redundancy often present in high-dimensional data in modern machine learning problems, we implemented robust statistical techniques to enhance model performance through feature selection.

2.2.1 Feature Reduction

High-dimensional datasets often contain redundant or noisy data, necessitating the extraction of relevant information. Feature selection becomes a critical step in enhancing model efficiency, interpretability, and performance metrics such as accuracy, precision, or recall. Traditional methods such as Singular Value Decomposition and Principal Component Analysis (PCA) are unsupervised and may not account for the specific relationships between feature values and target classes. Additionally, the principal components created in PCA may not have a clear and direct interpretation of the original feature space. Each principal component is a mixture of all the original features, and understanding the contribution of each original feature to a particular principal component can be challenging.

To address these limitations, we adopted the Boruta algorithm - a wrapper method around the random forest classification algorithm, known for its ability to capture important features related to an outcome variable. This algorithm employs a multistep process to assess feature importance. Initially, the procedure involves duplicating the dataset and shuffling values within each column to create shadow features. Subsequently, a classifier, such as a Random Forest Classifier, is trained on this augmented dataset to assess the importance of each feature through metrics like Mean Decrease Accuracy or Mean Decrease Impurity. The algorithm then evaluates the real features by comparing their Z-scores to the maximum Z-score among their corresponding shadow features. If a real feature consistently outperforms its shadow counterparts, the algorithm records this as a hit in a vector. This process iterates for a predefined number of cycles, resulting in a table of hits. Throughout each iteration, the algorithm validates feature importance by scrutinizing Z-scores from shuffled copies against those of the original features, using a binomial distribution to determine statistical significance. Essentially, this method enhances robustness by assessing feature importance relative to randomly shuffled copies, offering a comprehensive validation approach [17].

2.2.2 Model Training

Following the reduction of features through the Boruta algorithm, the next step involved a meticulous comparison of machine learning classifiers to identify the drug whose resistance or sensitivity in different cell lines could be optimally explained by the ML algorithms. This assessment was done separately on drug transcriptomics and proteomics datasets. The ensemble of classifiers employed for this comparison included Logistic Regression, Linear Support Vector Classification (LinearSVC), other classifiers such as Random Forest, Gradient Boosting, AdaBoost, Decision Tree, K-Nearest Neighbors (KNeighbors), XGBoost, Gaussian Naive Bayes (GaussianNB), and Multi-layer Perceptron Classifier (MLP) were implemented. The modeling phase was executed with an effective stratified 5-fold cross-validation strategy, ensuring balanced labels and mitigating the risk of overfitting. Stratified sampling is crucial in the context of imbalanced datasets, maintaining the distribution of class labels across folds to provide a more robust evaluation of classifier performance. Performance measures, including Matthews

Correlation Coefficient (MCC), Area Under the Receiver Operating Characteristic curve (AUC-ROC), precision and recall for both labels, accuracy, and F-measure, were computed for each machine learning classifier. These metrics collectively offer a comprehensive evaluation of classifier performance across various dimensions.

Furthermore, to identify the drug and machine learning model with the most robust performance on imbalanced datasets, special emphasis was placed on MCC, AUC-ROC, and F1 scores. Matthews Correlation Coefficient is particularly valuable in handling imbalanced datasets as it considers true and false positives and negatives. AUC-ROC provides insights into the classifier's ability to discriminate between classes, and F1 score balances precision and recall, making it suitable for imbalanced scenarios. This technical methodology ensures a rigorous and thorough assessment of drug performance, laying the groundwork for informed decisions regarding the selection of drugs for further analysis and potential clinical applications.

2.3 Employing Algorithms to increase the Explainability of AI

In critical fields such as understanding molecular mechanisms in cancer, binary responses may fall short, necessitating a focus on "how" or "where" and “why” events occur. Explainability can be described as the extent to which a human can comprehend the rationale behind cellular decisions and the resulting phenotypes or in general the extent to which a human can intuitively presume the model’s result [18]. Consequently, following the identification of the drug whose labels could be best predicted, and the associated model, Local Interpretable Model-agnostic Explanations (LIME) algorithm was applied to discover the crucial features influencing drug sensitivity or resistance in each instance of cancer cell lines.

LIME, as an instance-based explainer, generates simulated data points around an instance through random perturbation. It offers explanations by fitting a weighted sparse linear model over predicted responses from these perturbed points. LIME's explanations are locally faithful to an instance, regardless of the classifier type, creating an explanation for a single prediction by any machine learning (ML) model. This is achieved by learning a simpler interpretable model, such as a linear classifier, and determining feature importance through a form of feature selection method such as forward/ backward selection, lasso etc. Figure 2 illustrates the block diagram of the LIME framework [19].

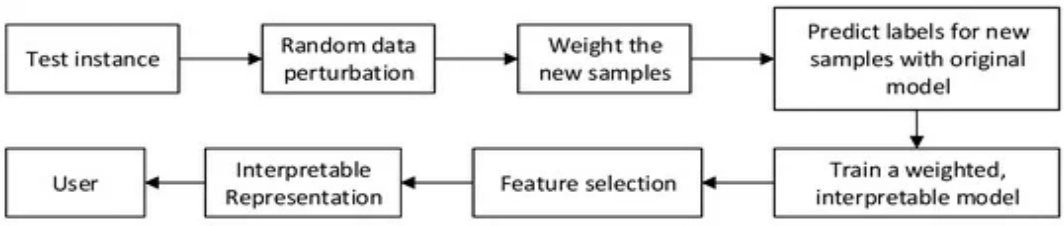


Figure 2. A block diagram of the LIME framework

Employing a robust five-fold stratified k-fold cross-validation strategy, the best model was trained and evaluated iteratively on distinct subsets of the top drug dataset to ensure model robustness and generalization. Within each fold, the LimeTabularExplainer was initialized using the training data, on both proteomics and transcriptomics datasets of the same drug separately. This explainer elucidated the model's decision-making process by generating explanations for individual instances in the test set. For each cell line instance, the LIME algorithm identified the top ten feature explanations deemed most influential in predicting the given drug sensitivity and resistivity of drug. It is crucial to note that at last lime explanations were obtained for both training and test data, providing a comprehensive understanding of feature importance across all the instances. This nuanced approach enhances the transparency and interpretability of our predictive modeling framework, shedding light on the intricate molecular determinants of drug response in cancer cell lines.

2.4 Clustering for possible biological mechanisms mediating resistance

In our comprehensive investigation of resistance mechanisms, we prioritized the identification of crucial features for each instance, in this case cell line, focusing exclusively on resistance cell lines in both transcriptomics and proteomics datasets. Utilizing the Lime method, we pooled all the important genes for resistant cell lines from the Lime explanation and their respective expression data and subsequently applied agglomerative hierarchical iterative clustering with "Euclidean" distance and ward linkage. Through the silhouette index, the number of optimal clusters was determined. Next, the cell lines common between the proteomics and transcriptomics clusters to were assessed and retained. This facilitated a concise and in-depth exploration of resistance mediators prevailing in different clusters.

2.5 Pathway enrichment analysis of normalized data using iDEP

In our quest to discern resistance mechanisms across distinct clusters of proteomics and transcriptomics data, we employed the iDEP tool [20], an integrated web application facilitating the understanding of genes/molecular players that are differentially expressed and recognizing the pathways wherein those molecular players are enriched. This tool integrates the 63 R/Bioconductor packages, and a comprehensive backend annotation, ontology and pathway database covering various plant and animal species, a total of 220 species.

2.5.1. Data Input and Preprocessing

To conduct the enrichment analysis seamlessly, the genes whose expression showed strong correlation (the threshold for adequate correlation was taken to be values greater than 0.8) with the genes derived from the LIME analyses were further. This ensures an adequate pathway enrichment or identification of over-represented pathways, thus resulting in improved analyses. The iDEP tool accepts various types of normalized expression data, including gene-level read count data, FPKM,

TPM, RPKM, and other varieties of normalized expression data (proteomics, DNA microarray etc). For the analysis, the data has to be structured with each row representing the expression level of a gene, where the first row serves as sample names, and each column contains data on a sample.

To accurately represent a given gene's expression (TPM for transcriptomics, protein intensity for proteomics) in a given cluster, central tendency of gene expression was computed, and two metrics were utilized: mean and weighted mean values. The mean expression was calculated by computing the average of the expression values of a particular feature across cells within a given cluster. The weighted mean was calculated by first considering all cells within the derived clusters. To normalize the expression values, the formula (expression - mean) / standard deviation was applied. Since the normalized values are continuous data and for further computation, discrete data is required, the expression values were approximated to the closest integer and thus transformed into class intervals. The mean of these classes represented the normalized expression. The frequency of occurrence (fi) for each derived class interval was computed, and the product of the mean of each class interval and its frequency was calculated. The average of these products resulted in the weighted mean normalized expression. This weighted mean (c) was then algebraically derived as c * standard deviation + mean.

Following the generation of these two representations of central tendency, statistical tests such as Kruskal-Wallis, and t-test were employed on mean and weighted mean to assess whether the central tendency values exhibited any statistical difference, where the null hypothesis posited no difference between groups, while the alternative hypothesis suggested a significant difference.

2.5.2. Differential gene expression and pathway analysis

In the field of bioinformatics, differentially expressed genes and pathway enrichment are crucial techniques for understanding biological phenomena. For the analyses, a fold change greater than 2 and a False Discovery Rate (FDR) cutoff of 0.1 was applied to derive differentially expressed genes in the derived clusters (DEGs) using DESeq2. FDR is used where multiple null hypothesis testing is being done. This is used to regulate the amount of expected proportion of incorrectly rejected null hypotheses that can be tolerated. In this context, controlling for an FDR of 0.1 ensures that a maximum of 10% of the identified genes are false positives.
Fold change is a measure used in the analysis of expression data from RNA-Seq experiments to quantify changes in the expression level of a gene between two groups. These parameters were carefully chosen based on the specifics of the study and the standards in the field.

Pathway Enrichment analyses were then performed on the derived DEGs. One of the key methods used in this process is Gene Set Enrichment Analyses (GSEA), also known as pathway enrichment analysis or functional enrichment analysis. GSEA is utilized to find the pathways that get affected due to the differentially expressed genes. In the backend of the toll, there is a database that has a list of all the known pathways present across multiple species. It also has a list of all the genes that

are known to the scientific community so far that participates in the corresponding pathway. On uploading the differentially expressed genes, these genes are mapped to the database of pathways with all the genes present in pathway and then it analyses the pathways where multiple DEGs are present thus concluding the pathways which are enriched due to the input DEGs. Since multiple genes are linked to a single biological pathway, it is the additive change in expression within gene sets that contribute to the difference in phenotypic expression [21]. In GSEA, Gene Ontology (GO) enriched Biological Processes and Molecular Functions were obtained. Enriched KEGG pathways were obtained using the ShinyGO tool by pooling all the upregulated genes from the different clusters separately. Pathway enrichment analysis helps interpret high throughput molecular data and generate hypotheses about the underlying biological phenomena of experiments. A GO network and hierarchical clustering tree to summarize such correlation, internally it first measures the distance among the terms by the percentage of overlapped genes and then this distance is used to construct a hierarchical clustering tree and a network of GO terms [20]. The parameters for the analysis were set as follows: the FDR cutoff for pathway Oval Cutoff was set at 0.2, the minimum size for the gene set (minSetSize) was set at 5, and the maximum size for the gene set (maxSetSize) was set at 2000. These parameters ensure that the analysis is both robust and manageable, allowing for meaningful insights to be drawn from the data.

Chapter 3. Results

The following sections summarize the results obtained from each of the steps comprising the methodology discussed in the previous chapter.

3.1 Data sourcing and preprocessing

The final datasets for transcriptomics consist of samples represented by cell lines, with 36,000 genes serving as features, and their corresponding expression levels in the cell lines. In parallel, the proteomics datasets comprise 8,498 protein-encoding genes along with their protein intensity values. The mapping of GDSC and FDA-approved datasets, specifically targeting FDA-approved therapies, resulted in the identification of 44 unique drugs. Among these, 11 drugs were sourced from GDSC 1, while the remaining were from GDSC 2. The number of cell lines varied for each drug, ranging from 225 for Tretinoin to 968 for Palbociclib, as illustrated in Figure 3. The integration of RNAseq and protein intensity data, along with response data labeled as 0 and 1, resulted in the creation of drug transcriptomics and drug proteomics datasets. These datasets were adjusted to include only those cell lines for which both proteomics and transcriptomics data were available for the respective drugs.

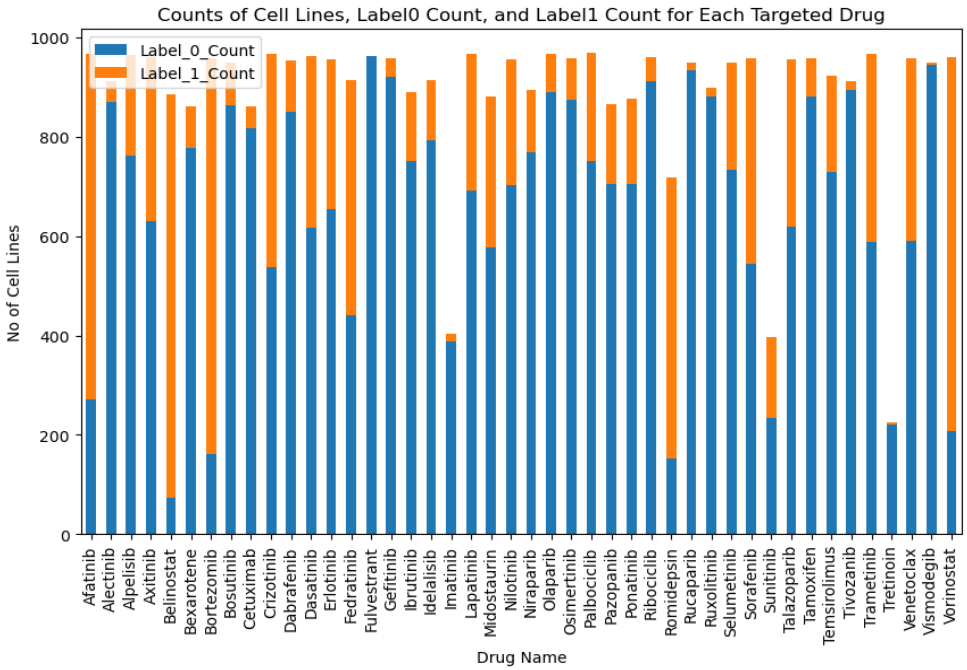


Figure 3. Imbalance in the number of resistant and sensitive cell lines for each of the FDA-approved drug

To address data imbalance shown in Figure 3, nine drugs with extremely imbalanced samples, where one label represented less than 5% of the overall samples, were removed. These drugs are

'Tretinoin', 'Gefitinib', 'Vismodegib', 'Rucaparib', 'Fulvestrant', 'Ruxolitinib', 'Imatinib', 'Alectinib', and 'Tivozanib'. The exclusion of these drugs aimed to ensure a more balanced representation in the final datasets for subsequent analyses.

3.2 Feature engineering and model implementation automation

In this study, an automated feature selection technique, Boruta, was employed during the feature engineering and model implementation phase. This method effectively identified relevant features associated with the outcome variable under specific conditions. Leveraging this technique, we evaluated ten diverse models, ranging from logistic regression to advanced approaches like XGBoost, utilizing RNA-seq and protein intensity data for each drug. The selection of the top-performing drug was based on the maximization of MCC, ensuring a balanced accuracy metric. In both transcriptomics and proteomics, Axitinib, Crizotinib, and Palbociclib emerged as the top drugs, demonstrating superior predictive capabilities. For the purpose of elucidating resistance mechanisms, Axitinib was singled out as the best-performing common drug across both datasets. Notably, gradient boosting proved to be the optimal model for Axitinib in transcriptomics data, while the XGBoost classifier excelled in axitinib proteomics data. Rigorous random state and hyperparameter tuning were meticulously performed to achieve the enhanced MCC, ensuring optimized model performance and accurate predictions. Detailed results, including the top drugs and their corresponding models for transcriptomics (Table 1, Figure 4) and proteomics (Table 2, Figure 5) are summarized below.

Table 1. Transcriptomics: Top 7 drugs with their best performance model in bold

Model	MCC	AUC_ROC	Accuracy	F1	Precision_0	Precision_1	Recall_0	Recall_1	Drug name
Gradient Boosting	**0.61**	**0.9**	**0.94**	**0.61**	**0.96**	**0.7**	**0.97**	**0.53**	**Bosutinib**
AdaBoost	0.61	0.88	0.94	0.64	0.96	0.7	0.98	0.59	Bosutinib
Gradient Boosting	**0.61**	**0.88**	**0.82**	**0.74**	**0.85**	**0.77**	**0.89**	**0.69**	**Axitinib**
XGBClassifier	0.59	0.87	0.82	0.72	0.84	0.77	0.89	0.68	Axitinib
KNeighbors	**0.59**	**0.86**	**0.81**	**0.75**	**0.83**	**0.76**	**0.85**	**0.74**	**Trametinib**
XGBClassifier	**0.58**	**0.93**	**0.95**	**0.57**	**0.95**	**0.81**	**0.99**	**0.45**	**Olaparib**
XGBClassifier	0.58	0.87	0.94	0.6	0.95	0.76	0.98	0.5	Bosutinib
Random Forest	0.58	0.88	0.81	0.73	0.85	0.77	0.89	0.67	Axitinib
XGBClassifier	**0.58**	**0.85**	**0.79**	**0.75**	**0.79**	**0.81**	**0.87**	**0.7**	**Crizotinib**
Gradient Boosting	0.58	0.88	0.8	0.73	0.82	0.78	0.88	0.7	Trametinib
Gradient Boosting	**0.57**	**0.87**	**0.86**	**0.65**	**0.89**	**0.74**	**0.94**	**0.57**	**Palbociclib**

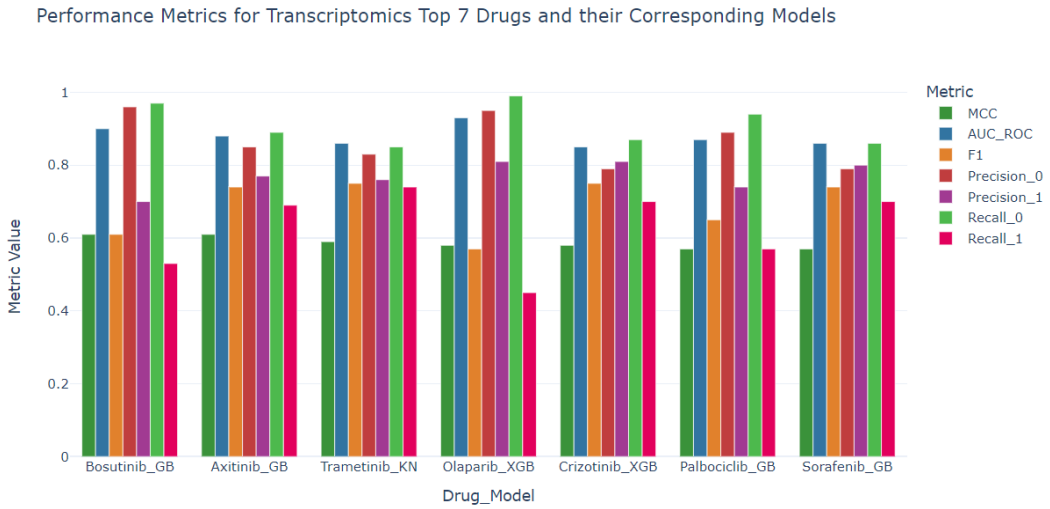


Figure 4. Transcriptomics: Top 7 drugs with their best performance model and metrics.

Table 2 Proteomics: The top 7 drugs with their best performance model in bold.

Drug_Model	**MCC**	**AUC_ROC**	**F1**	**Precision_0**	**Precision_1**	**Recall_0**	**Recall_1**	**Drugname**
XGBClassifier	**0.6**	**0.87**	**0.73**	**0.85**	**0.77**	**0.9**	**0.68**	**Axitinib**
Random Forest	0.58	0.87	0.71	0.83	0.78	0.9	0.67	Axitinib
Gradient Boosting	0.58	0.87	0.71	0.84	0.77	0.9	0.66	Axitinib
XGBClassifier	**0.58**	**0.91**	**0.62**	**0.93**	**0.72**	**0.97**	**0.55**	**Niraparib**
Gradient Boosting	**0.57**	**0.85**	**0.74**	**0.79**	**0.8**	**0.87**	**0.68**	**Sorafenib**
Gradient Boosting	0.57	0.91	0.6	0.93	0.74	0.97	0.52	Niraparib
GaussianNB	0.56	0.82	0.71	0.85	0.73	0.86	0.7	Axitinib
MLPClassifier	0.56	0.86	0.69	0.82	0.74	0.87	0.64	Axitinib
Random Forest	0.56	0.85	0.74	0.77	0.8	0.88	0.67	Sorafenib
GaussianNB	0.56	0.83	0.74	0.79	0.79	0.86	0.69	Sorafenib
XGBClassifier	0.56	0.85	0.73	0.78	0.8	0.87	0.67	Sorafenib
LogisticRegression	0.55	0.84	0.69	0.83	0.75	0.88	0.65	Axitinib
AdaBoostClassifier	0.54	0.83	0.69	0.83	0.73	0.87	0.66	Axitinib
LogisticRegression	**0.54**	**0.85**	**0.62**	**0.88**	**0.71**	**0.93**	**0.56**	**Palbociclib**
Random Forest	0.54	0.85	0.61	0.87	0.73	0.95	0.52	Palbociclib
Random Forest	**0.54**	**0.83**	**0.71**	**0.75**	**0.8**	**0.87**	**0.62**	**Crizotinib**
KNeighbors	0.53	0.77	0.61	0.87	0.72	0.94	0.53	Palbociclib

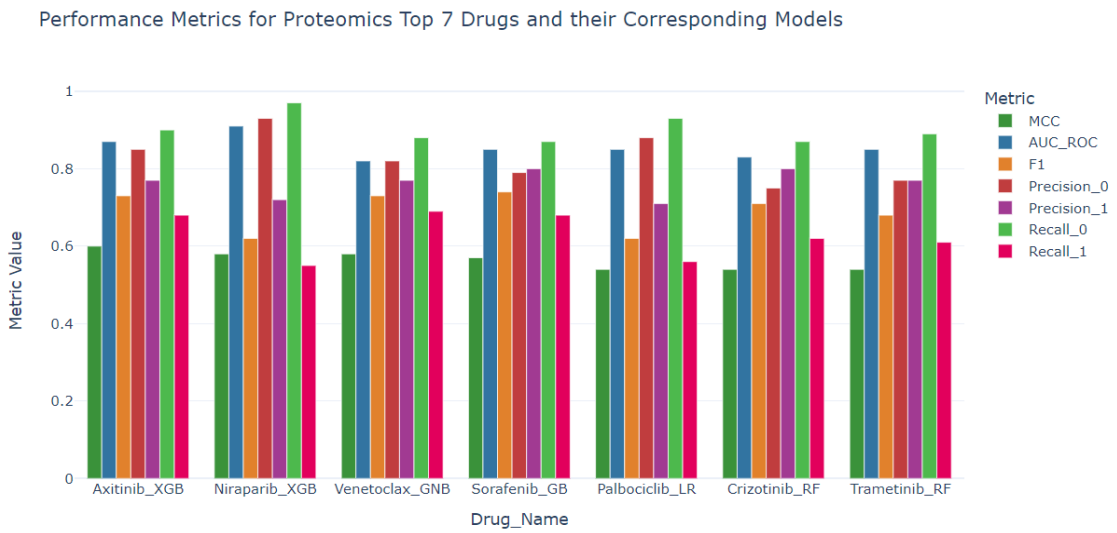


Figure 5. Proteomics: Top 7 drugs with their best performance model and metrics.

Axitinib, marketed as Inlyta, is a small molecule tyrosine kinase inhibitor developed by Pfizer. This compound is an indazole derivative with the molecular formula C22H18N4OS and a molecular weight of 386.47 Da. The chemical structure of axitinib is shown in Figure 6 [22].

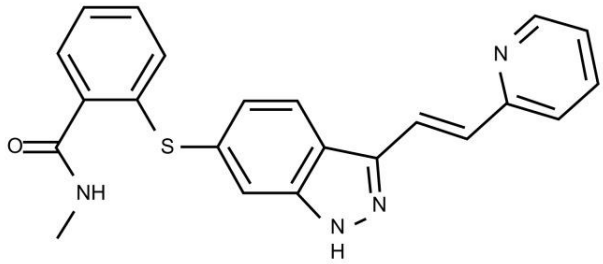


Figure 6. Chemical structure of axitinib.

Axitinib exhibits inhibition towards VEGFR1, VEGFR2, and VEGFR3. In contrast to numerous tyrosine kinase inhibitors (TKIs), axitinib demonstrates limited activity on additional receptors like KIT and platelet-derived growth factor receptor (PDGFR). Axitinib stands out as a potent VEGFR inhibitor with half maximal inhibitory concentrations (IC_{50}) for VEGF family receptors being 10 times lower than other TKIs. In *in vivo* studies, axitinib has shown dose-dependent inhibition of angiogenesis and tumor growth in mice, including a xenograft model of human tumors. Notably, axitinib induces regression in tumor vasculature, characterized by the loss of endothelial sprouts and fenestration, along with a reduction in vessel density [23]. It has also been demonstrated to bind to the BCR-ABL fusion protein (although in a different conformation than VEGF binding), thereby selectively blocking the drug-resistant T315I mutant isoform [24]

In the Axitinib transcriptomics dataset, we observed a comprehensive set of 954 cell lines, with 137 genes meticulously selected through Boruta. Among these, 625 cell lines exhibited resistance to Axitinib, contrasting with 329 sensitive cell lines. Similarly, the Axitinib proteomics dataset comprised 937 cell lines and featured a subset of 100 Boruta-selected genes. In this context, 617 cell lines displayed resistance, while 320 were classified as sensitive.

A noteworthy trend emerged in both datasets, with a predominant presence of resistance cell lines. This imbalance is reflected in the precision scores, where Precision 0 (indicating the prediction of resistivity) is 0.85, and Precision 1 (for sensitivity prediction) attains a score of 0.77. The recall scores further emphasize the models' effectiveness, particularly with a Recall 0 (resistance prediction) of 0.9 and Recall 1 (sensitive prediction) of 0.7, the same in both Table 2 and Table 3.

Despite the potential for reduced balanced Matthews Correlation Coefficient (MCC) due to the high number of resistance cell lines, the precision and recall scores suggest a robust capability of the prediction models in identifying resistance instances. This imbalance in the dataset, coupled with the high predictive scores for resistance, may enhance the models' effectiveness in dissecting the mechanisms underlying resistance cell lines in both transcriptomics and proteomics contexts.

3.3 Explainable AI -feature importance analyses

LIME analysis provides feature importance for each instance or cell lines in a binary classification task. It displays prediction probabilities for every feature, indicating whether it contributes to determining class 0 or class 1.

The graphical representation of LIME outcome with transcriptomics data for one of the cell lines, SNU-407, which is resistant to Axitinib in shown in figure 7. The left panel of the figure has a bar graph showing the prediction probabilities for class 0 and class 1, 0.93 and 0.07 respectively, indicating a strong inclination towards class 0. On the right is a heatmap displaying the top 10 features contributing to the prediction of this particular cell line such as PPA1, PTPN1, SQSTM1, CAT, BID, ATP1A1, LMNA, F5, SUCLG2, and NAPG. Features with positive LIME values (PPA1, CAT, BID, ATP1A1, F5) are colored in orange and contribute to the incorrect prediction of class 1. On the other hand, features with negative LIME values (PTPN1, SQSTM1, LMNA, SUCLG2, NAPG) are colored in blue and contribute to the correct prediction of class 0. Despite the presence of features that support class 1 (orange bars), the overall probability strongly favors class 0 due to the more significant contribution of features supporting this class (blue bars).

This analysis is specific to the SNU-407 cell line and provides insights into how the model behaves for this particular data point. The positive values for certain genes may be indicative of the complex interplay between different features and their impact on the model’s decision, even if the overall trend suggests a class 0 prediction for this specific instance, thus to identify the important

mediators in resistant cell lines using LIME, all the top 10 features across all resistant instances were considered for further analysis.

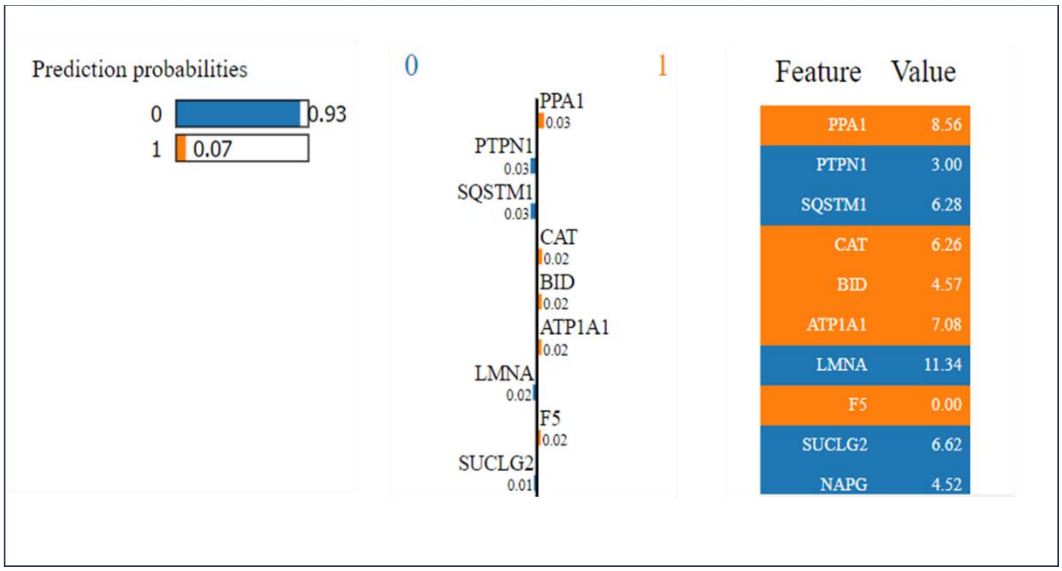
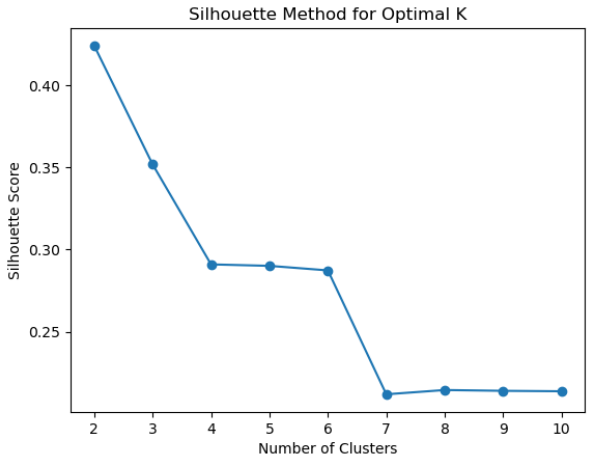


Figure 7. LIME prediction probabilities and top 10 features for SNU-407 cell line.

3.4 Hierarchical agglomerative clustering

Clustering was done separately for transcriptomics and proteomics data. After pooling all the important genes for each resistance cell line and applying hierarchical agglomerative clustering, through the silhouette index, the optimal cluster was determined as 2 with associated scores 0.42 and 0.40 in transcriptomics and proteomics, respectively as shown in figure 8a and figure 9a. The clusters were denoted as T_y1, comprising 520 cell lines and T_y2 comprising 105 cell lines for RNA-seq data and P_y1 comprising of 535 cell lines and P_y2 comprising of 82 cell lines for protein intensity data. T-SNE visualization for transcriptomics data and proteomics clustering is presented in Figure 8b and Figure 9b, where purple and red represent two distinct clusters. Purple corresponds to the T_y1 cluster, and red corresponds to the T_y2 cluster. Interestingly when assessing cell line similarity between proteomics and transcriptomics clusters, 508 common cell lines were uncovered between T_y1 and P_y1 and 79 common cell lines between T_y2 and P_y2. These results indicate that essential features from Lime, derived from proteomics and transcriptomic-resistant cell lines, are very similar. Consequently, their expression leads to the merging of the same cell lines into identical clusters in both datasets. This observation underscores the efficiency of clustering algorithms, Lime explainable AI, and other processes employed to reach this step. This difference in clusters may account for variations in gene expression, as illustrated in the heatmap figure 10 and figure 11 for transcriptomics and proteomics respectively where the clusters are separated by a black line, with the smaller section representing y2 and the larger section representing y1.

a. Silhouette score graph



b. T-SNE visualization

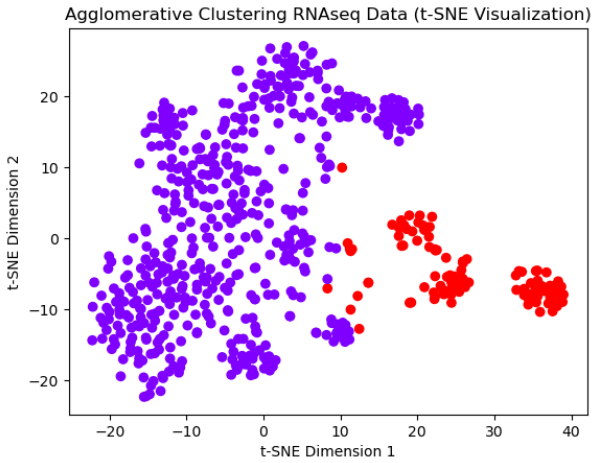
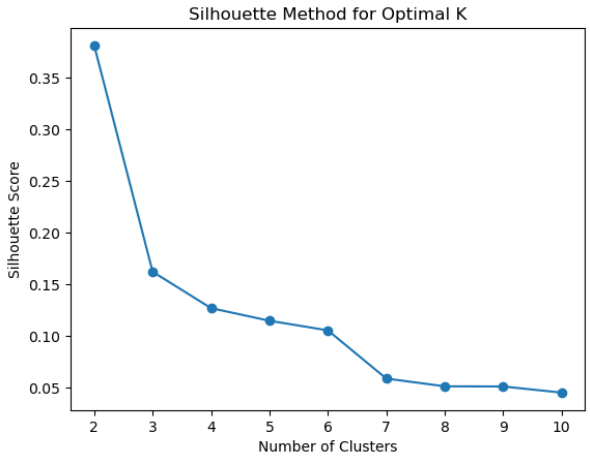


Figure 8. Transcriptomics: Clustering evaluation and visualization

a. Silhouette score graph



b. T-SNE visualization

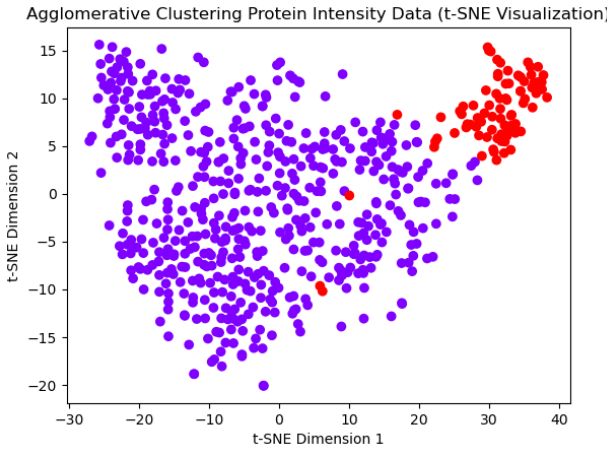


Figure 9. Proteomics: Clustering evaluation and visualization.

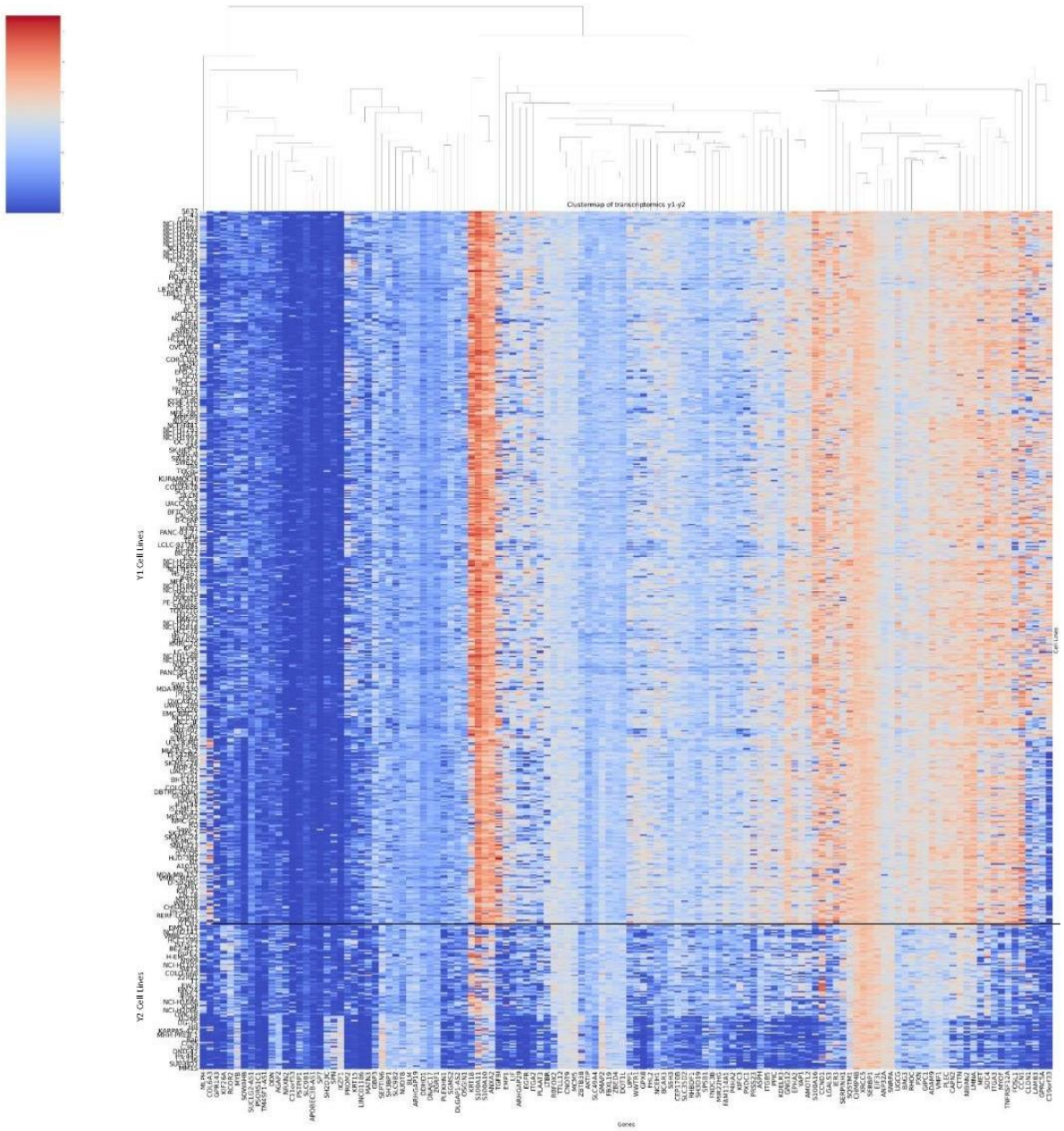


Figure 10. Transcriptomics: Heatmap showing the difference in expression of important genes in the two clusters derived from the pool of resistant cell lines (T_y1 and T_y2)

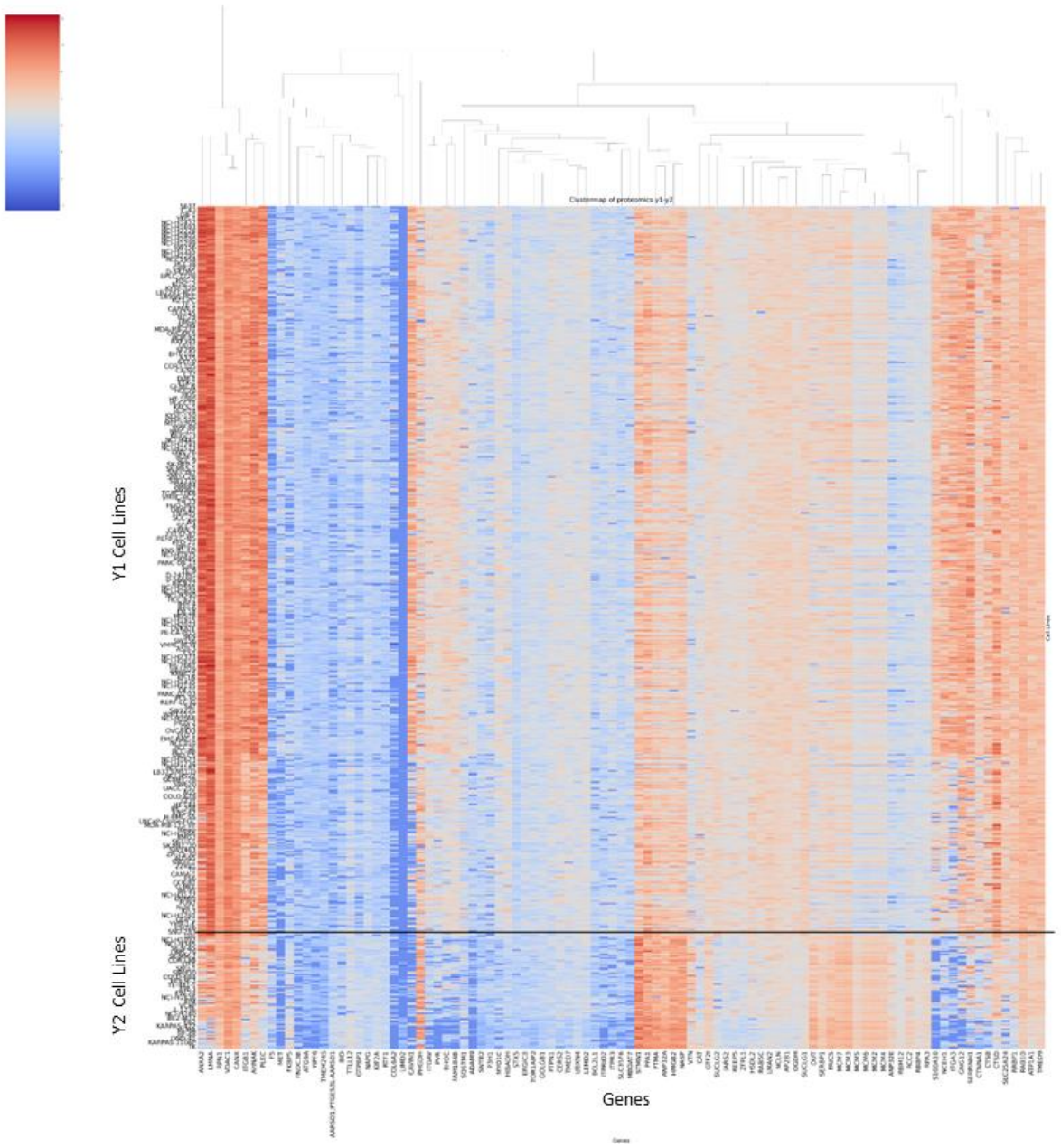


Figure 11. Proteomics: Heatmap showing the difference in expression intensity of important proteins in the two clusters derived from the pool of resistant cell lines (P_y1 and P_y2)

3.5 Differential gene analysis and pathway enrichment

3.5.1 Data Preprocessing

To accurately represent the gene expression of a cluster, central tendency measures chiefly mean and weighted mean values were calculated for each gene across samples. Weighted mean was computed to identify outliers if any present. Statistical tests such as Kruskal-Wallis and ANOVA confirmed that there is no significant difference between mean and weighted mean value, thus

confirming that there were no outliers. The above metrics to represent gene expression in each cluster were efficiently utilized to calculate the DEG between two groups. Box plots (figure 12 and figure 13) show the same observation in a graphical way.

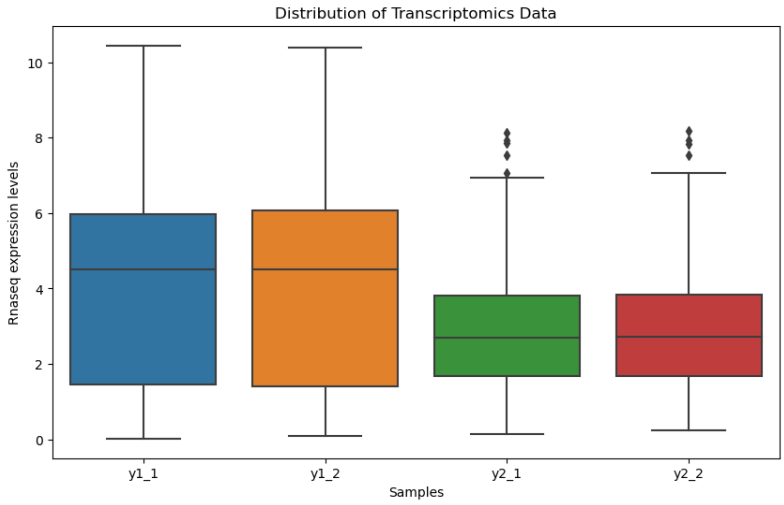


Figure 12. Boxplot for transcriptomics data.

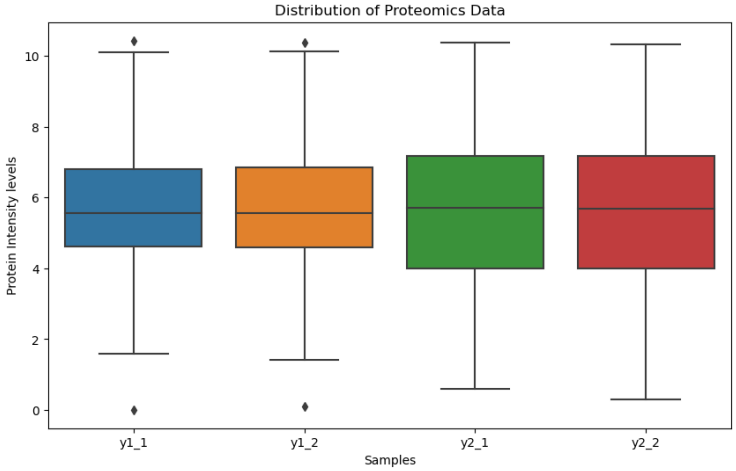


Figure 13. Boxplot for proteomics data

3.5.2 Differentially expressed genes

RNA seq and protein intensity data derived from the above processing of central tendency of the cell lines of y1 and y2 clusters were compared to derive differentially expressed genes (DEGs) using the DESeq R package. A total of 199 important genes were identified from LIME and features that showed a strong correlation with them (>0.8) in transcriptomics. Differential gene expression analysis revealed that 107 genes were upregulated in the T_y1 cluster, while 53 genes were downregulated i.e. upregulated in T_y2. The remaining 39 genes were deemed statistically insignificant. Similarly, in proteomics data, a total of 175 important proteins were considered using

the features derived from LIME and the features showing a high correlation with those features. Differential expression analysis on protein intensity revealed 43 highly expressed and 42 under expressed proteins in the P_y1 cluster. The difference in the expression of remaining 89 proteins was deemed statistically insignificant. The volcano plot and the MA plot are shown in figure 14 and figure 15 for RNA-seq and protein intensity data respectively. Both analyses were conducted with a False Discovery Rate (FDR) cutoff of 0.1. The minimum fold change was set at 2. A fold change of 2 indicates that the expression of the gene in y1 is doubled than that of y2.

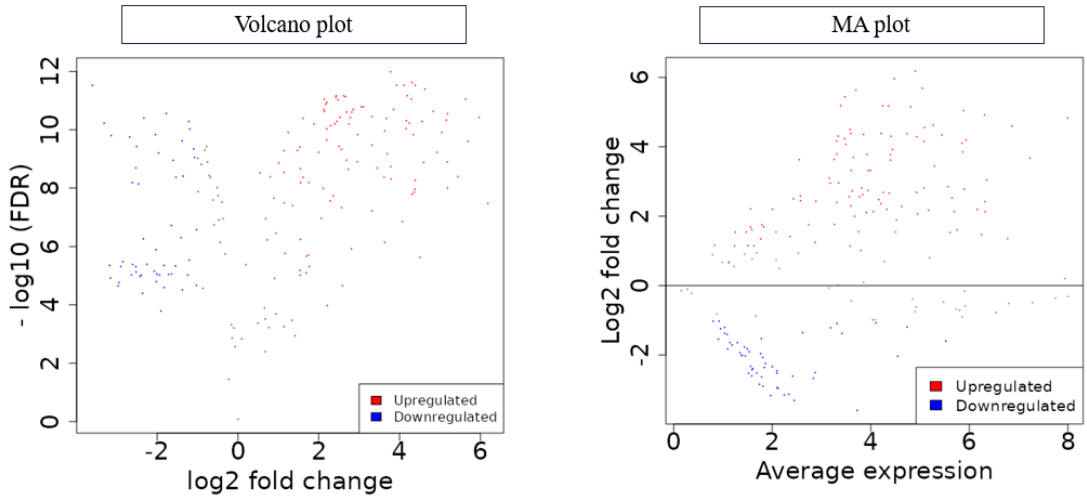


Figure 14. Summary plots for differential expression analysis using DESeq2 on RNA-seq data.

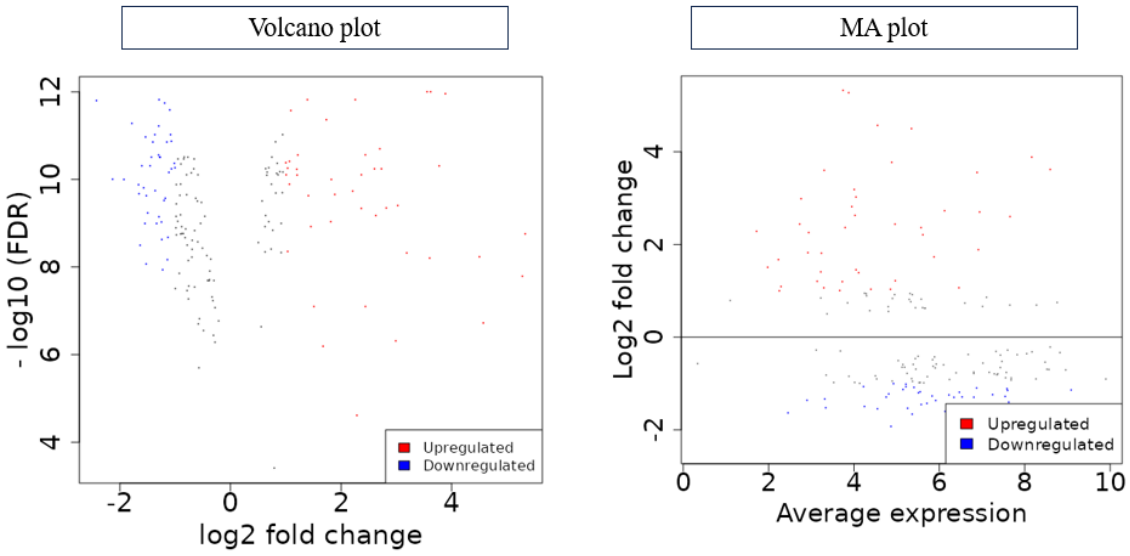


Figure 15. Summary plots for differential expression analysis using DESeq2 on protein intensity data

3.5.3 Pathway analysis

The Gene Ontology (GO) Biological Processes enriched post-differential expression gene (DEG) analyses were computed for both transcriptomics and proteomics, as shown in tables 3 and 6, respectively. The upregulated genes of y1 in both datasets were found associated with extracellular matrix organization, cytoskeletal organization (especially the regulation of actin filament-based processes), cell junctions, and cell migration and locomotion. The GO Molecular Function, listed in tables 4 and 7 for transcriptomics and proteomics respectively, provides similar insights. The molecular functions enriched in y1 included actin-binding, cadherin binding, and binding and interaction with Integrin, Fibronectin, Laminin, and other components of the cellular extracellular architecture along with some other molecular functions related to transcriptional coregulator activity. Additionally, GAGE (generally applicable gene set enrichment) GO terms, shown in Table 5, highlighted terms like vasculature development and blood vessel development which are important process for angiogenesis.

Examining these GO terms revealed that a crucial factor influencing TKI resistance is the tumor microenvironment (TME). The TME is essential for both initiating and sustaining tumorigenesis, playing a significant role in the resistance to anti-angiogenic TKIs. Various components within the TME, both cellular and acellular, contribute to tumor growth by facilitating nutrient supply, aiding immune cell infiltration, and regulating the production and restructuring of the extracellular matrix (ECM) and cytoskeleton. Mechanical irregularities within the TME, such as compressive forces and matrix stiffness, impact cancer cells at molecular, cellular, and tissue levels [25].

The downregulated genes in y1 exhibited significant enrichment in pathways associated with the immune system. Transcriptomics data indicate enrichment in pathways related to leukocytes, specifically lymphocytes, involving cell-cell adhesion, activation, and proliferation. Additionally, GTPase activator activity and actin-binding pathways are implicated. From the proteomics perspective, the downregulated genes were prominently associated with MCM 2-7 genes, highlighting pathways such as DNA replication origin binding, helicase activity, ATPase activity, and other cell cycle-related processes, as detailed in the corresponding tables.

Table 3. GO biological processes obtained from transcriptomics analysis.

	Direction	**adj.Pval**	**nGenes**	**Pathways**
1	Down regulated	3.14E-10	22	Cell activation
2	Down regulated	3.14E-10	21	Leukocyte activation
3	Down regulated	2.21E-09	14	T cell activation
4	Down regulated	2.21E-09	16	Lymphocyte activation
5	Down regulated	6.05E-09	18	Immune effector process
6	Down regulated	6.05E-09	29	Immune system process
7	Down regulated	2.32E-08	25	Immune response

8	Down regulated	1.12E-07	15	Leukocyte mediated immunity
9	Down regulated	1.43E-07	14	Cell activation involved in immune response
10	Down regulated	1.46E-07	11	Leukocyte cell-cell adhesion
11	Down regulated	2.05E-07	20	Regulation of immune system process
12	Down regulated	2.05E-07	10	Mononuclear cell proliferation
13	Down regulated	2.05E-07	9	T cell proliferation
14	Down regulated	2.05E-07	10	Lymphocyte proliferation
15	Down regulated	3.91E-07	10	Leukocyte proliferation
16	Up regulated	4.07E-09	30	Cell adhesion
17	Up regulated	4.07E-09	30	Biological adhesion
18	Up regulated	5.89E-09	34	Tissue development
19	Up regulated	2.07E-08	38	Anatomical structure morphogenesis
20	Up regulated	7.16E-07	29	Locomotion
21	Up regulated	7.54E-07	31	Movement of cell or subcellular component
22	Up regulated	8.19E-07	27	Cell motility
23	Up regulated	8.19E-07	27	Localization of cell
24	Up regulated	1.60E-06	25	Cell migration
25	Up regulated	2.69E-06	8	Contractile actin filament bundle assembly
26	Up regulated	2.69E-06	17	Cell junction organization
27	Up regulated	2.69E-06	8	Stress fiber assembly
28	Up regulated	6.30E-06	13	Extracellular matrix organization
29	Up regulated	6.30E-06	13	Extracellular structure organization
30	Up regulated	6.30E-06	18	Supramolecular fiber organization

Table 4. GO molecular functions obtained from transcriptomics analysis.

	Direction	**adj.Pval**	**nGenes**	**Pathways**
1	Down regulated	3.55E-08	12	GTPase regulator activity
2	Down regulated	3.55E-08	12	Nucleoside-triphosphatase regulator activity
3	Down regulated	1.56E-07	11	GTPase activator activity
4	Down regulated	6.82E-05	13	Enzyme regulator activity
5	Down regulated	0.001824646503	14	Molecular function regulator
6	Down regulated	0.006518859415	6	Actin binding
7	Down regulated	0.00960404853	10	Protein-containing complex binding
8	Up regulated	6.20E-09	18	Cell adhesion molecule binding
9	Up regulated	2.65E-06	12	Cadherin binding
10	Up regulated	4.17E-06	7	Virus receptor activity
11	Up regulated	4.17E-06	6	Extracellular matrix binding

12	Up regulated	4.17E-06	7	Exogenous protein binding
13	Up regulated	8.94E-05	7	Integrin binding
14	Up regulated	0.001042325074	3	Cadherin binding involved in cell-cell adhesion
15	Up regulated	0.001079628825	3	Proline-rich region binding
16	Up regulated	0.001555081068	4	Cell adhesion mediator activity
17	Up regulated	0.001771621313	12	Structural molecule activity
18	Up regulated	0.002571549559	22	Identical protein binding
19	Up regulated	0.002571549559	3	Laminin binding
20	Up regulated	0.002847786856	4	Collagen binding
21	Up regulated	0.002847786856	4	Myosin binding
22	Up regulated	0.003701533451	5	Growth factor binding

Table 5. GO biological processes obtained from GAGE transcriptomics analysis.

	Direction	**Pathways**	**statistic**	**nGenes**	**adj.Pval**
1	Down	Adaptive immune response	-4.0073	10	0.15
2	Down	Regulation of immune response	-3.3036	20	0.15
3	Down	Immune effector process	-3.2327	23	0.15
4	Up	Embryo development	3.7994	16	0.054
5	Up	Tube development	3.7716	18	0.054
6	Up	Circulatory system development	3.7001	18	0.054
7	Up	Animal organ morphogenesis	3.5577	13	0.078
8	Up	Epithelium development	3.2022	25	0.078
9	Up	Tissue development	3.1763	40	0.078
10	Up	Vasculature development	3.1703	13	0.089
11	Up	Regulation of cellular protein localization	3.1399	12	0.1
12	Up	Tube morphogenesis	3.0684	14	0.089
13	Up	Epithelial cell differentiation	3.0017	19	0.089
14	Up	Anatomical structure formation involved in morphogenesis	2.9575	23	0.089
15	Up	Skin development	2.951	13	0.11
16	Up	Blood vessel development	2.8741	11	0.13
17	Up	Anatomical structure morphogenesis	2.8705	48	0.089
18	Up	Embryonic morphogenesis	2.8678	10	0.14
19	Up	Tissue morphogenesis	2.6751	13	0.16
20	Up	Regulation of protein localization	2.5988	14	0.17
21	Up	Blood vessel morphogenesis	2.5795	10	0.19
22	Up	System process	2.491	23	0.17

Table 6. GO biological processes obtained from proteomics analysis.

	Direction	adj.Pval	nGenes	Pathways
1	Down regulated	2.75E-14	6	Pre-replicative complex assembly involved in nuclear cell cycle DNA replication
2	Down regulated	2.75E-14	6	Pre-replicative complex assembly
3	Down regulated	1.03E-12	6	Double-strand break repair via break-induced replication
4	Down regulated	1.03E-12	7	Nuclear DNA replication
5	Down regulated	1.03E-12	7	Cell cycle DNA replication
6	Down regulated	3.53E-12	11	DNA recombination
7	Down regulated	3.81E-12	9	Double-strand break repair via homologous recombination
8	Down regulated	3.81E-12	9	Recombinational repair
9	Down regulated	1.72E-11	10	DNA replication
10	Down regulated	2.96E-11	10	protein-DNA complex subunit organization
11	Down regulated	5.24E-11	6	DNA replication initiation
12	Down regulated	9.92E-11	15	Cell cycle process
13	Down regulated	9.92E-11	16	Chromosome organization
14	Down regulated	1.10E-10	17	Cell cycle
15	Down regulated	4.83E-10	10	DNA conformation change
16	Up regulated	2.67E-06	27	Localization
17	Up regulated	2.03E-05	13	Regulation of cellular component movement
18	Up regulated	4.70E-05	12	Regulation of cell motility
19	Up regulated	4.97E-05	19	Regulation of localization
20	Up regulated	4.97E-05	12	Regulation of locomotion
21	Up regulated	0.0001343882912	3	Membrane raft assembly
22	Up regulated	0.000166256878	9	Positive regulation of locomotion
23	Up regulated	0.000166256878	9	Positive regulation of cellular component movement
24	Up regulated	0.000166256878	9	Positive regulation of cell motility
25	Up regulated	0.0005680384403	10	Regulation of cell migration
26	Up regulated	0.0006007767898	7	Regulation of actin filament-based process
27	Up regulated	0.0006567758589	8	Positive regulation of cell migration
28	Up regulated	0.0006567758589	3	Membrane raft organization
29	Up regulated	0.0006941107103	3	Negative regulation of rho protein signal transduction
30	Up regulated	0.0008555273429	11	Positive regulation of cellular component organization

Table 7. GO molecular functions obtained from proteomics analysis.

	Direction	adj.Pval	nGenes	Pathways
1	Down regulated	1.39E-10	6	DNA replication origin binding
2	Down regulated	7.32E-10	8	Single-stranded DNA binding
3	Down regulated	2.66E-08	38	Binding
4	Down regulated	2.66E-08	8	Catalytic activity, acting on DNA
5	Down regulated	3.84E-08	6	DNA helicase activity
6	Down regulated	2.95E-07	6	DNA-dependent ATPase activity
7	Down regulated	2.13E-06	6	Helicase activity
8	Down regulated	3.45E-06	28	Protein binding
9	Down regulated	1.69E-05	6	Histone binding
10	Down regulated	2.52E-05	4	Damaged DNA binding
11	Down regulated	0.0002296901896	2	Supercoiled DNA binding
12	Down regulated	0.0002546857074	8	Pyrophosphatase activity
13	Down regulated	0.0002546857074	8	Hydrolase activity, acting on acid anhydrides
14	Down regulated	0.0002841970702	6	ATPase activity
15	Down regulated	0.0004448978288	9	Double-stranded DNA binding
16	Up regulated	3.26E-06	15	Protein-containing complex binding
17	Up regulated	5.29E-06	5	Extracellular matrix binding
18	Up regulated	7.81E-06	5	Collagen binding
19	Up regulated	0.0002569514455	7	Actin binding
20	Up regulated	0.0002569514455	2	C-X3-C chemokine binding
21	Up regulated	0.0002569514455	6	Cell adhesion molecule binding
22	Up regulated	0.0003849308734	3	Fibronectin binding
23	Up regulated	0.0003849308734	3	Laminin binding
24	Up regulated	0.000827630873	9	Cytoskeletal protein binding
25	Up regulated	0.001460771621	4	Integrin binding
26	Up regulated	0.001610792768	4	Protease binding
27	Up regulated	0.002078806372	2	Phosphate ion binding
28	Up regulated	0.004236239184	24	Protein binding
29	Up regulated	0.004522969509	4	Actin filament binding
30	Up regulated	0.00523285341	2	Ankyrin binding

As many GO terms are related or redundant, two plots were created to summarize the pathways with their degree of up and down-regulation based on adjusted p-value given in a network of GO terms (figure 16 and figure 17) and in a hierarchical clustering tree (figure 18 and figure 19). In

the network, nodes with a darker color signify more significantly enriched gene sets, while larger nodes denote larger gene sets; thicker edges indicate a higher degree of overlap in shared genes between connected sets. In the hierarchical clustering tree, gene sets closer share more genes, with dot sizes corresponding to adjusted p-values; closer sets exhibit a higher degree of shared genes. In both representations red dots signify upregulated pathways in y1, while green dots represent upregulated pathways in y2.

Figure 16. Network of biological processes GO terms for transcriptomics.

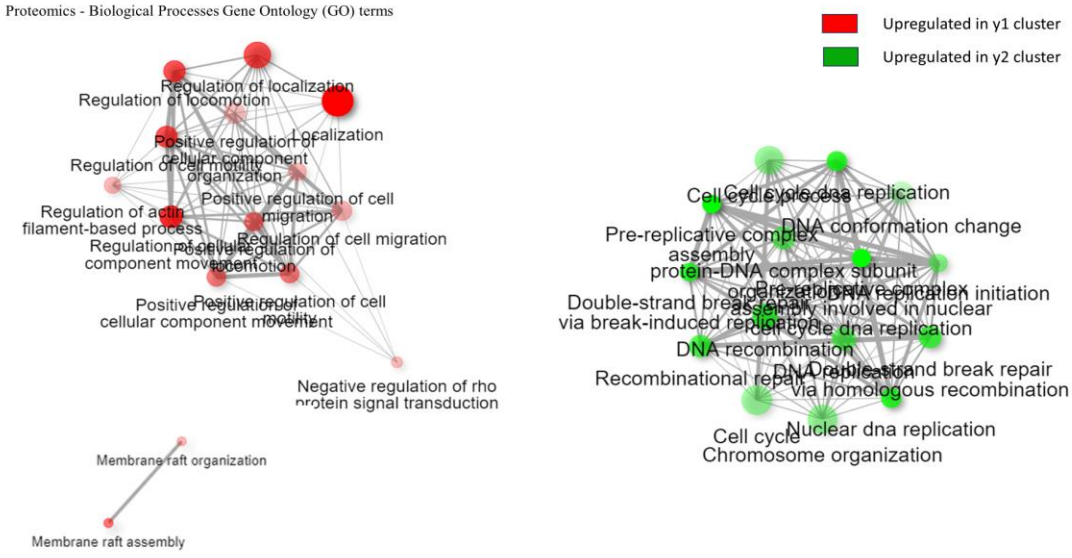


Figure 17. Network of biological processes GO terms for proteomics.

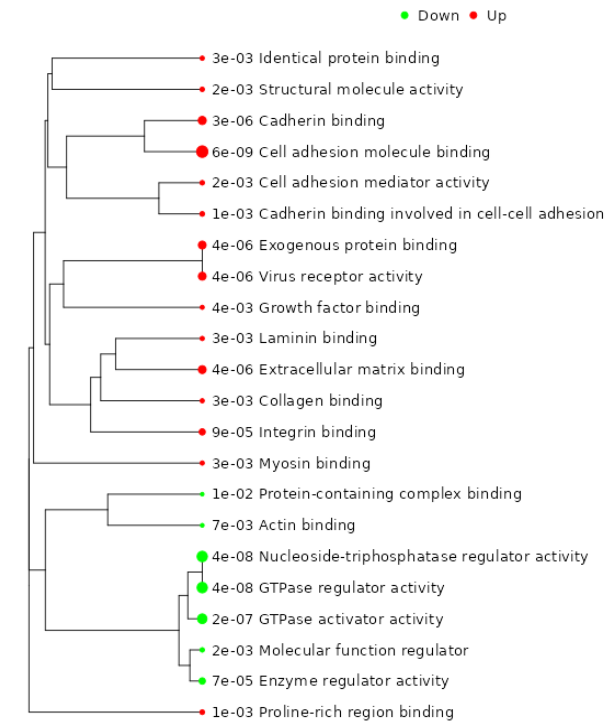


Figure 18. Hierarchical clustering tree of molecular function GO terms for transcriptomics

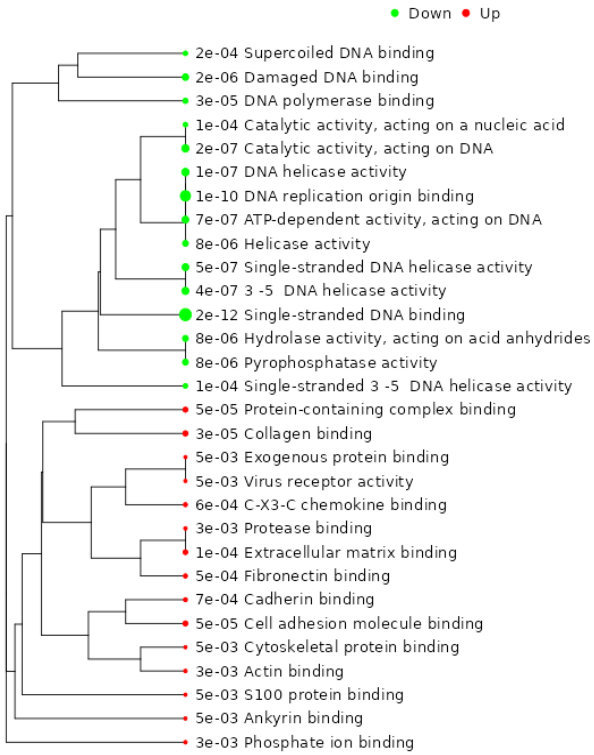


Figure 19. Hierarchical clustering tree of molecular function GO terms for proteomics.

Further ShinyGO was utilized [26] to identify enriched KEGG pathways [27] associated with upregulated proteomics and transcriptomics genes in the y1 and y2 clusters separately. The analysis confirmed the overrepresentation of genes from the y1 cluster in pathways such as Regulation of Actin Cytoskeleton, Focal Adhesion, PI3K-Akt Signaling Pathway, ECM-Receptor Interaction, Proteoglycans in Cancer, Tight Junction, and Hippo Signaling Pathway, as depicted in the accompanying figure 20.

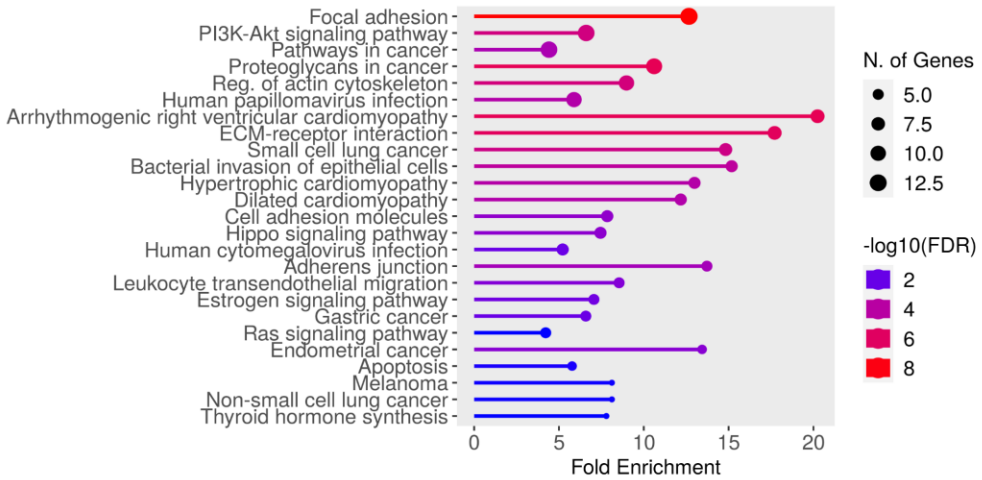


Figure 20. Enriched KEGG pathways for the upregulated genes in the y1 cluster (Transcriptomics).

In the y2 cluster, we identified KEGG upregulated pathways, including DNA Replication, Mismatch Repair, Nucleotide Excision Repair, and Cell Cycle, as depicted in the figure 21. These findings aligned with the synchronized Gene Ontology (GO).

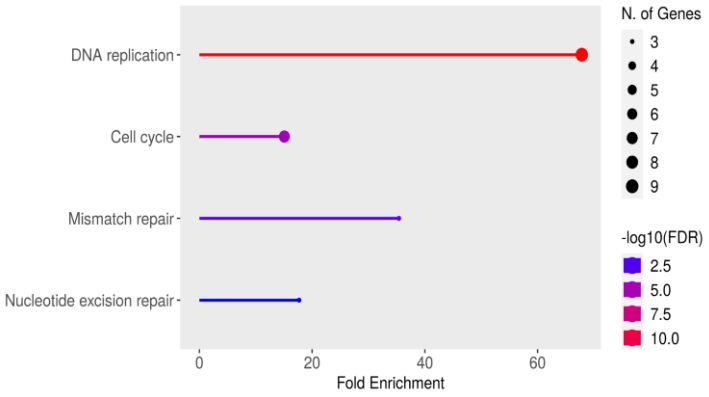


Figure 21. Enriched KEGG pathways for the upregulated genes in the y2 cluster.

The y1 cluster in the depicted figure 20 exhibited significant enrichment in the "extracellular matrix–receptor (ECM-receptor) interaction pathway" according to KEGG pathways. All genes within this pathway, including integrins, collagen, and laminins, showed significant upregulation as shown in red color in figure 22, while vitronectin was downregulated as shown in green colored

box. The extracellular matrix (ECM) is a complex blend of macromolecules crucial for tissue and organ development, maintaining cell and tissue structure, and regulating cellular activities like adhesion, migration, and apoptosis. Cellular interactions with the ECM involve transmembrane molecules, primarily integrins and potentially proteoglycans, influencing various cellular functions. Integrins, as glycosylated heterodimeric adhesion receptors, play a key role in serving as mechanoreceptors and forming a physical link between the ECM and the cytoskeleton. This pathway's significance is underscored by research highlighting the ECM's role in tumorigenesis and drug resistance [28].

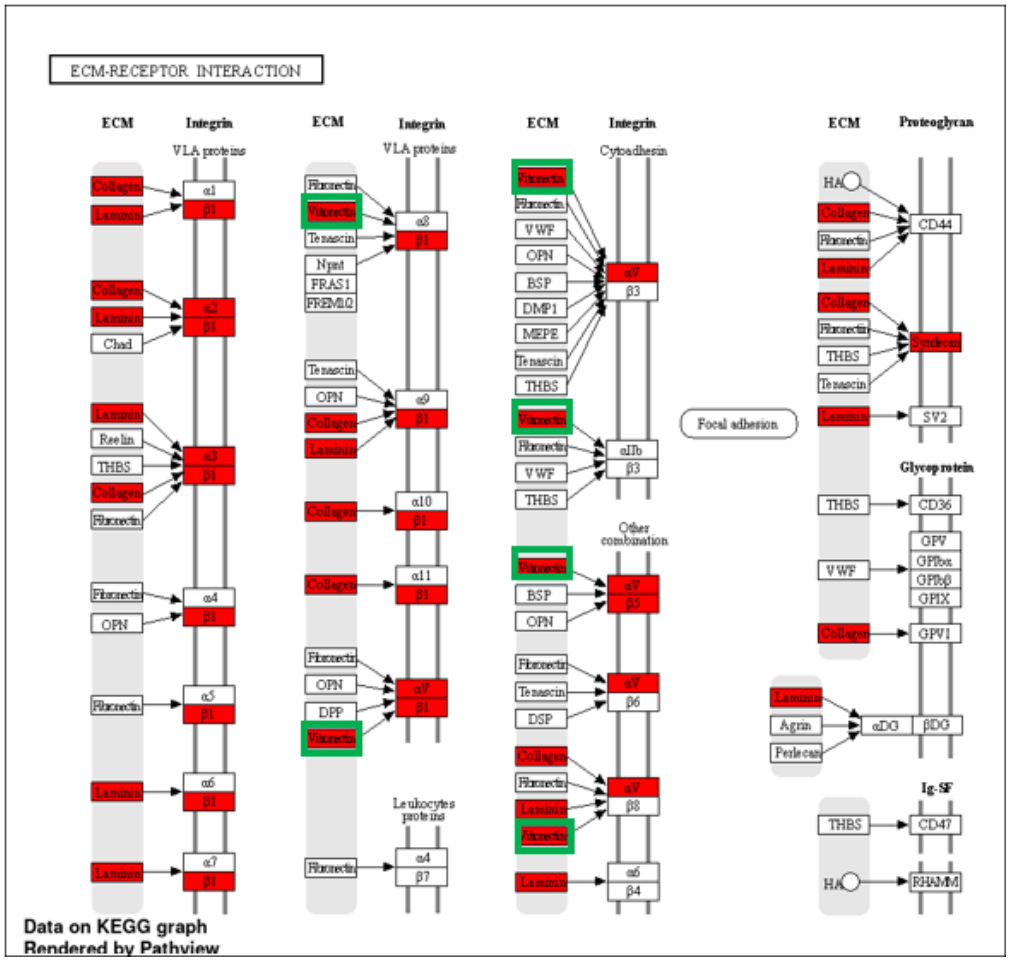


Figure 22. KEGG extracellular matrix–receptor (ECM-receptor) interaction pathway.

Other than ECM, KEGG pathways related to the Regulation of Actin Cytoskeleton and Focal Adhesion were also enriched in y1 as shown in figure 20. In the context of cancer progression, particularly during metastasis, a complex series of interconnected cellular events becomes crucial. These events involve the regulation of actin filaments, extracellular matrix (ECM), and focal adhesion, and are pivotal for processes such as invasion, migration, immune evasion, extravasation, and angiogenesis. The metastatic process relies heavily on the dynamic reorganization of cytoskeletal components, including actin, myosin, and intermediate filaments, as

well as associated proteins. Localized invasion occurs at the interface between tumor and stromal cells, where interactions lead to the exchange of enzymes and cytokines. This modulates the local ECM and promotes cell migration. Cytoskeletal reorganization results in the formation of membrane structures like ruffles, lamellipodia, filopodia, and invadopodia. Lamellipodia, transient actin-rich protrusions at the cell's leading edge, drive cell locomotion through integrin-mediated adhesion, contraction, and detachment. Filopodia, finger-like protrusions, support slow migration and act as sensory organelles. Invadopodia, specialized actin-based protrusions, facilitate invasion through the basement membrane [29]. The intricate regulation of cytoskeletal organization is governed by Rho GTPases, with Rac controlling lamellipodia and membrane ruffles, while Cdc42 and Rho regulate filopodia and actin-myosin contractility, respectively. Dysregulation of cytoskeletal dynamics is implicated in tumor metastasis [30], [31].

Moreover, cell attachment to ECM components induces integrin clustering, leading to the formation of focal adhesions (FAs). Adaptor/scaffold proteins within FAs, such as talin, paxillin, tensin, p130Cas, and α-actinin, establish robust connections to the actin cytoskeleton, firmly anchoring cells to the ECM. This linkage facilitates the generation of tension, altering cell morphology and promoting cell migration. FAs also recruit signaling proteins, including focal adhesion kinase (FAK) and Src, acting as molecular switches transmitting signals to pathways controlling proliferation, survival, and migration. FAK and Src play pivotal roles in integrin-mediated signaling cascades by phosphorylating integrin-associated proteins, influencing cellular behavior in processes such as cell migration and tumorigenesis [32].

The complex interplay between cytoskeletal dynamics, ECM interactions, and focal adhesion signaling is central to understanding and potentially controlling the metastatic progression of cancer cells. Simultaneously, the resulting cellular locomotion enables tumor cells to evade drug exposure, leading to drug resistance. ECM remodeling and disruption in focal adhesion are downstream effects of perceived hypoxia. In response to oxygen deficit, epithelial cells undergo dedifferentiation to form mesenchymal cells with a stem cell-like phenotype. This process causes cells to detach from the basement membrane and induces "wound healing" or stiffening of the extracellular matrix by increasing crosslinking within the ground matrix, ultimately replacing the pre-existing epithelial cell structure [33].

Another upregulated pathway PI3K/Akt signaling pathway has huge role in tyrosine kinase drug Resistance. This pathway chiefly comprises of threonine/serine protein kinase B (PKB) and phosphatidylinositol-3-kinase (PI3K). PI3K are a crucial family of lipid enzymes which are activated when 3-hydroxyl group is phosphorylated of the phosphatidylinositol. When EGFR (epidermal growth factor receptor) is bound to Axitinib, the downstream signaling has shown to promote cellular survival and further proliferation. This can be due to a mutation in the oncogenes, PI3KCA gene that encodes for the subunit p110, or in tumor suppressor genes, such as phosphatase and tensin homolog (PTEN) which are the major mutations causing dysregulation of this pathway [34], [35].

The MET, EGFR, and AXL receptors are known to promote the PI3K/Akt signalling pathway in various ways. EGFR, a receptor tyrosine kinase, initiates several signalling cascades when activated, including the PI3K/Akt pathway. In some cancers, AXL forms a dimer with EGFR and phosphorylates it, leading to the activation of phospholipase Cg (PLCg) and protein kinase C (PKC). This sequence of events activates mTOR, which can result in resistance to PI3Kα inhibitors [36]. On the other hand, MET has the ability to circumvent the inhibited EGFR phosphorylation kinase pathway. It gets amplified via the ERBB3-P13K/AKT and MAPK-ERK1/2 pathways. The amplification of c-MET enhances downstream signal transduction through a bypass mechanism, which helps evade cell death induced by EGFR-TKIs. This intricate interplay between EGFR and MET underscores their role in promoting drug resistance [37].

Further it has been demonstrated that axitinib induces increased abundance of phosphorylated Protein Kinase B or pAKT and this increment in pAkt increases Glut-1 translocation from the plasma membrane thereby causing a 2-fold increment in glycolysis rates which has been shown to promote drug resistance in Pancreatic adenocarcinoma [38].

The y2 cluster depicted in the figure 23 exhibited a significant enrichment in the DNA Replication pathway, with all genes within this pathway, including those encoding the minichromosome maintenance protein complex (MCM) 2 to MCM 7, RPA3, RFA2/4, and PCNA, being notably upregulated as DEGs. The MCM complex, comprised of MCM2–7, functions as a crucial DNA helicase essential for genomic DNA replication. Research has demonstrated that various MCMs play a pivotal role in promoting cell proliferation both in vitro and in vivo, particularly in specific cancer cell lines such as those associated with gastrointestinal, lung, brain, renal, breast, ovarian, and blood cancers. The link between MCMs and increased proliferation in cancer cells is primarily attributed to their capacity to enhance DNA replication and confer resistance to chemotherapy by modulating cell cycle dynamics and managing DNA replication stress [39].

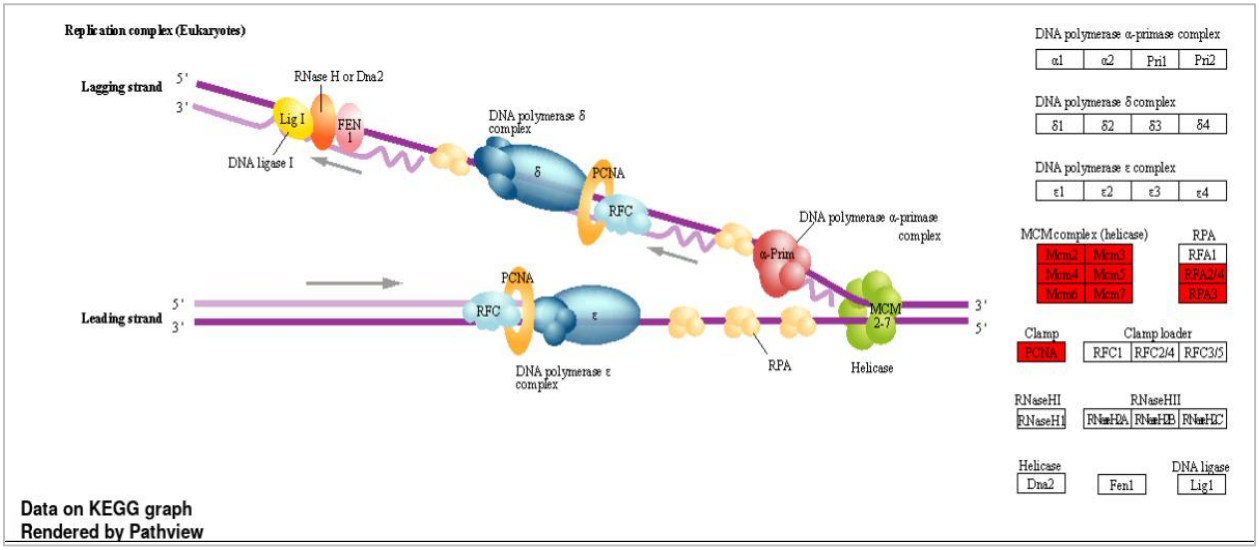


Figure 23 Figure: KEGG "DNA Replication" pathway

Chapter 4. Discussion

Cancer remains a prominent global cause of mortality, claiming nearly 10 million lives in 2020. As estimated by WHO, the primary culprits in cancer-related deaths included lung (1.80 million), colon and rectum (916,000), liver (830,000), stomach (769,000), and breast (685,000) cancers. Targeted therapy, gaining traction for its reduced toxicity and enhanced specificity, encompasses approaches like anti-angiogenetic, which inhibits new blood vessel formation, inducing hypoxia and tumor cell death. However, tumors often develop resistance mechanisms, necessitating a detailed study of drug resistance.

In this study, we employed a comprehensive approach integrating transcriptomics, proteomics, and drug response data across various cancer cell lines to unravel axitinib drug resistance mechanisms. Utilizing automated feature engineering, we streamlined relevant features and implemented ten machine-learning models with stratified cross-validation on diverse drugs. Axitinib consistently emerged as a top predictive drug with an MCC score of 0.60, and an AUC-ROC of approximately 0.88. Gradient boost and XGBoost models proved optimal for transcriptomics and proteomics data, respectively. Precision and recall scores for predicting resistance were higher, approximately 0.85 and 0.9, respectively, compared to those for sensitive cell lines in both proteomics and transcriptomics datasets which was favorable for this analysis to study resistance mechanism. It was also observed that axitinib exhibited an imbalance, with resistance cell lines outnumbering sensitive ones by double in both datasets. This underscored the urgency of dissecting axitinib's resistance mechanisms. Employing Lime explainable AI, genes or proteins pivotal in predicting resistant cell lines were identified, and taking those features hierarchical clustering was applied leading to the identification of two optimal clusters in both Transcriptomics and proteomics datasets (T_y1, T_y2 and P_y1, P_y2 respectively) with a silhouette index of approximately 0.41 in both datasets. Subsequent analysis revealed that larger transcriptomics clusters (T_y1) largely overlapped with larger proteomics clusters (P_y1), similarly for smaller clusters. This alignment prompted a focused comparison between the two clusters and increased the confidence in the systematic pipeline used for the analysis.

Since a lot of genes/proteins (high correlations within the data) were removed for developing prediction models, a lot of biological information was filtered. To account for the information loss in the previous steps, genes/proteins with highly correlated expression values in reference to the identified features (LIME result) were added before the differential gene expression and pathway enrichment analyses were employed to elucidate mechanisms cohesively.

Further features (+ the correlated genes) associated with the cell lines common among the T_y1 and P_y1, and T_y2 and P_y2 were only used for the enrichment analysis. These identifies genes/proteins were predominantly associated with enhanced cell migration, locomotion, blood

vessel development, cytoskeleton regulation, extracellular matrix interactions, protein bindings, and signaling pathways like Hippo and PI3K/AKT.

Given axitinib's role as a tyrosine kinase inhibitor targeting VEGFR 1–3, c-KIT, and PDGFR, the identified mechanisms aligned well with its primary action, inhibiting angiogenesis. Notably, the revealed pathways were largely related to cell migration, a critical aspect in tumor progression, underscoring the importance of cytoskeleton regulation in facilitating nutrient and oxygen supply through blood vessels in response to hypoxia.

Based on the genes and pathways observed, as well as supporting literature, one proposed resistance mechanism for cell lines in y1 clusters involves mechanosensor and its interaction with ECM and Cytoskeleton. The tumor microenvironment (TME) plays a crucial role in initiating and sustaining angiogenesis and tumorigenesis. As the number of both cancerous and non-cancerous cells in the TME increases, mechanical forces come into play, influencing physiological and pathological processes in cancers [40]. Mechanosensing, defined as a cell's ability to perceive mechanical cues from its microenvironment, encompasses the detection of force, stress, strain, substrate stiffness, topography, and adhesiveness [41]. Cells engage in the detection and transmission of diverse mechanical signals through the activation of surface mechanosensors, including integrins, G protein-coupled receptors (GPCR), transient receptor potential (TRP) ion channels, and YAP/TAZ. Notably, these mechanosensors were highly expressed in the y1 cluster, comprising ITGA2, ITGA3, ITGB5, ITGAV, ITGB, GPRC5A, GNG12, LIF (a receptor within the cytokine receptor family, including GPCRs), TGFBI, YAP/TAZ, and the ankyrin binding pathway. In response to mechanosensors, the cytoskeleton specially actin filament undergoes substantial alterations. The cytoskeleton, connecting cells intricately to their external environment, interprets external signals that govern complex behaviors like lamellipodia formation, invasion, and migration [41]. Physical forces act on the cytoskeleton through mechanosensors, triggering various pathways, including the GTP-binding protein RhoA, the Hippo pathway, focal adhesion, and PI3K-AKT pathways [41], [42]. Interestingly, these pathways were also observed to be upregulated in the y1 cluster, with the exception of RhoC being upregulated instead of RhoA. RhoC, akin to RhoA and RhoB isoforms, plays a role in diverse cellular processes such as organizing cytoskeletal components, cell division, and intracellular trafficking. Notably, RhoC upregulation has already been identified in malignant pancreatic ductal carcinoma, inflammatory breast cancer tumors, and highly metastatic melanoma [43].

Mechanosensors also contribute to the stiffness of the Extracellular Matrix (ECM), which in turn plays an important role in the organization of the cytoskeleton in the Tumor Microenvironment (TME). The ECM, composed of proteoglycans (PGs), glycoproteins (GAGs), and fibrous proteins like collagen, fibronectin (FN), and laminin, is not merely a physical support but an active component of living tissue. It facilitates cell-cell communication, cell proliferation, and metastasis [44]. In the context of various KEGG pathways such as tight junction, Adherens junction, and Proteoglycans in cancer, upregulation of collagen COL6A3, fibronectin (FN) binding pathway,

laminin LAMB3 and LAMB2 genes, and integrins were clearly observed in the y1 cluster. Tumors are often stiffer than their surrounding normal tissue, primarily due to the amount of ECM components present. This ECM stiffness can lead to intracellular contractions and an increase in the stiffness of an actin cytoskeleton, favoring cancer migration. Matrix stiffening can activate TGF-β signaling, which was also found to be highly upregulated in the y1 cluster. Activation of TGF-β signaling mediates the epithelial-mesenchymal transition (EMT), leading to a more aggressive phenotype that promotes cancer metastasis. EMT contributes to cancer progression through signaling pathways, with the actin cytoskeleton playing a predominant role [40]. Cancer-Associated Fibroblasts (CAFs) are major contributors to ECM stiffness. They interact with almost all cells within the TME and regulate cancer cell cytoskeletons indirectly by mediating ECM stiffness. Hypoxia and TGF-β are key inducers for CAFs in regulating TME stiffness, influencing tumor and stromal cell properties such as proliferation and motility [45].

YAP (yes-associated protein) and TAZ, also known as WWTR1, which were highly overexpressed along with their transcription factor TEAD1, are regulated by ECM stiffness and control cell fate decisions. Mesenchymal stem cells differentiate into bone cells by activating YAP/TAZ when cultured on a stiff matrix, mimicking the natural bone environment. Conversely, on a soft matrix, they differentiate into fat cells by inactivating YAP/TAZ3. The actin cytoskeleton is a key mediator of the regulation of Hippo-YAP signaling by means of a variety of biochemical and mechanical cues. Expression of several regulators and components of the actomyosin cytoskeleton, including myosin IIB, myosin regulatory light chain 2, and filamin A, are also enhanced by YAP [46], [47]. MYO1C and MYOF were found be to upregulated. Research suggests that, unlike the sole blockade of VEGF signaling by Axitinib or anti-VEGF antibody, the combination of YAP1-knock-down in CAFs with VEGF signaling inhibitors could enhance the suppression of tumor angiogenesis. This indicates that YAP1 serves as a significant pro-angiogenic factor in the VEGF-independent pathway [48].

GAGE ((Generally Applicable Gene-set Enrichment) GO analysis further revealed some additional pathways such as Vasculature development, Tube development, Blood vessel morphogenesis, and Blood vessel development, which hint at a resistance mechanism through intussusceptive angiogenesis. This is a dynamic intravascular process that can significantly alter the structure of the microcirculation. Several high-impact publications have shown that in the absence of VEGF stimulation, blood vessels can divide into new vessels without the need for endothelial proliferation. Intussusceptive angiogenesis is a relatively quick process where existing endothelial cells migrate and are remodeled by increasing in volume and becoming thinner. As existing endothelial cells are incorporated, the tumor can overcome the need for VEGF-mediated endothelial proliferation. It is believed that factors such as blood flow dynamics and shear stress influence this process, leading to interactions between endothelial cells (related to cell adhesion) and endothelial-pericyte, cytoskeletal rearrangements, and adaptation of gap junction complexes. It is also hypothesized that the inhibition of sprouting angiogenesis may stimulate intussusceptive angiogenesis [49].

While looking at individual genes, it was found that several gene-encoded receptor tyrosine kinases other than VEGF, such as AXL, EGFR, EPHA2, and MET, were also highly upregulated in the y1 cluster. Some tumors depend on a single deregulated oncogene for their growth and/or survival, a phenomenon known as oncogene addiction, which explains the effectiveness of targeted therapies. EGFR is a tyrosine kinase that regulates cellular homeostasis and stimulates cancer growth. EGFR gene mutations and protein overexpression, both of which activate downstream pathways, and VEGF and EGFR pathways are closely related, sharing common downstream signaling pathways [50]. Both can trigger PI3K/AKT and RAS/RAF/ERK pathways which were also found to be upregulated in KEGG pathways. This has been also seen for tyrosine kinase MET and AXL as mentioned in the results section. These pathways enhance cell proliferation, motility, and angiogenesis. The expression of the EGFR protein and the activation of the EGFR pathway cause VEGF production and VEGF Receptor activity, promoting angiogenesis through the upregulation of hypoxia-dependent Hypoxia-Inducible Factor (HIF)-α expression [51]. The expression of HIF-1α is known to trigger another receptor tyrosine kinase MET gene. This gene encodes for a protein named MET, which is vital for cell survival, growth, and migration. When MET is activated, it can promote downstream signaling which lead to tumor invasion and metastasis, making the cancer cells more aggressive and resistant to therapy. One of the previous studies shows that using a combination of axitinib and crizotinib did not show a clinical advantage for patients with advanced RCC, implying that MET expression could be the primary cause of axitinib resistance [52]. Another tyrosine kinase receptor AXL abnormal expression has been associated with clinicopathological characteristics and a poor prognosis in cancer patients. There is growing evidence that AXL plays a role in the development and progression of cancer, as well as in resistance to anti-angiogenesis drugs [53]. One study showed that after prolonged pre-treatment with sunitinib, a drug similar to axitinib, the level of AXL protein increased in patient samples. Additionally, some other groups have reported that the activation of AXL and MET kinases led to the stimulation of their downstream signaling pathways and promoted EMT signaling [54].

Though a lot of information could be derived from the analysis of the genes found to be contributing to the resistance to axitinib in cell lines in y1 cluster, the same could not be applied to the y2 cluster consisting of 79 cell lines that represented a distinct subgroup. The LIME features of these cell lines were found to be enriched in different set of pathways. To get insights into the observed differences, firstly the tissue of origin of these cell lines was analyzed. Interestingly, the cell lines in this y2 cluster were predominantly associated with connective tissue or a combination of connective and epithelial tissues, with a prevalence of blood-type cell lines. This unique composition suggested that y2 operates under different resistance mechanisms compared to the y1 cluster. Notably, the majority of y2 cell lines exhibit low resistance to axitinib as reported in the drug response data based on IC_{50} analysis. The important genes here were found to be associated with immune activation, heightened leukocyte activation, and proliferation, particularly in lymphocytes. These changes extend to increased cell-cell adhesion and regulatory activities within the immune system, potentially contributing to the prevalence of blood-type cancers or other

mechanisms yet to be elucidated. Additionally, there is an upregulation of genes related to GTPase activator proteins and nucleoside-triphosphatase regulator activities, both linked to actin binding and cell locomotion. The intricate interplay of immune activation and GTPase regulatory functions may play a role in the observed resistance to axitinib in y2 cell lines. Though it was clear that the mechanisms elucidated for the cell lines clustered in T_y1 group were not directly applicable to the T_y2 grouped cell lines. Further analysis is warranted to explore the questions that still remain unanswered.

Chapter 5. Future directions

This study elucidated the resistance mechanisms associated with axitinib across diverse cancer cell lines. Recognizing the need for a more nuanced understanding, iterative or nested clustering has to be integrated allowing for the exploration of finer sub-clusters based to different resistance mechanisms.

To bolster the credibility of identified mechanisms, genetic mutation data, copy number variations, and other omics data can also be integrated into the analysis. This multi-faceted approach will provide a more holistic view of the molecular landscape underlying resistance.

Moreover, the methodology can be extended to investigate resistance mechanisms associated with other anti-cancer drugs. Employing a similar or enhanced pipeline will enable the categorization of drugs based on shared or distinct resistance mechanisms. This not only contributes to a broader understanding of drug resistance but also opens avenues for exploring potential synergies between drugs with complementary actions.

The insights gained from such studies could pave the way for precision medicine approaches, where tailored treatment strategies can be devised based on the identified resistance mechanisms. This comprehensive and integrative research framework holds promise for advancing our understanding of cancer drug resistance and optimizing therapeutic interventions for improved patient outcomes.

References

[1] Bueschbell, B., Caniceiro, A. B., Suzano, P. M., Machuqueiro, M., Rosário-Ferreira, N., & Moreira, I. S. (2022). Network biology and artificial intelligence drive the understanding of the multidrug resistance phenotype in cancer. *Drug Resistance Updates*, *60*, 100811.

[2] Shuel, S. L. (2022). Targeted cancer therapies: Clinical pearls for primary care. *Canadian Family Physician*, *68*(7), 515.

[3] Sierra, J. R., Cepero, V., & Giordano, S. (2010). Molecular mechanisms of acquired resistance to tyrosine kinase targeted therapy. *Molecular cancer*, *9*, 1-13.

[4] Huang, L., Jiang, S., & Shi, Y. (2020). Tyrosine kinase inhibitors for solid tumors in the past 20 years (2001–2020). *Journal of hematology & oncology*, *13*, 1-23.

[5] Qin, S., Li, A., Yi, M., Yu, S., Zhang, M., & Wu, K. (2019). Recent advances on anti-angiogenesis receptor tyrosine kinase inhibitors in cancer therapy. *Journal of hematology & oncology*, *12*, 1-11.

[6] Niu, G., & Chen, X. (2010). Vascular endothelial growth factor as an anti-angiogenic target for cancer therapy. *Current drug targets*, *11*(8), 1000-1017.

[7] Gupta, M. K., & Qin, R. Y. (2003). Mechanism and its regulation of tumor-induced angiogenesis. *World journal of gastroenterology: WJG*, *9*(6), 1144.

[8] Emami Nejad, A., Najafgholian, S., Rostami, A., Sistani, A., Shojaeifar, S., Esparvarinha, M., ... & Manian, M. (2021). The role of hypoxia in the tumor microenvironment and development of cancer stem cell: a novel approach to developing treatment. *Cancer Cell International*, *21*(1), 1-26.

[9] Hofer, E., & Schweighofer, B. (2007). Signal transduction induced in endothelial cells by growth factor receptors involved in angiogenesis. *Thrombosis and haemostasis*, *97*(03), 355-363.

[10] Tsuji-Tamura, K., & Ogawa, M. (2018). Morphology regulation in vascular endothelial cells. *Inflammation and regeneration*, *38*, 1-13.

[11] Yang, Y., & Cao, Y. (2022, November). The impact of VEGF on cancer metastasis and systemic disease. In *Seminars in Cancer Biology* (Vol. 86, pp. 251-261). Academic Press.

[12] Gotink, K. J., & Verheul, H. M. (2010). Anti-angiogenic tyrosine kinase inhibitors: what is their mechanism of action?. *Angiogenesis*, *13*, 1-14.

[13] Rotow, J., & Bivona, T. G. (2017). Understanding and targeting resistance mechanisms in NSCLC. *Nature Reviews Cancer*, *17*(11), 637-658.

[14] Gonçalves, E., Poulos, R. C., Cai, Z., Barthorpe, S., Manda, S. S., Lucas, N., ... & Reddel, R. R. (2022). Pan-cancer proteomic map of 949 human cell lines. *Cancer Cell*, *40*(8), 835-849.

[15] Garcia-Alonso, L., Iorio, F., Matchan, A., Fonseca, N., Jaaks, P., Falcone, F., ... & Saez-Rodriguez, J. Transcription factor activities enhance markers of drug response in cancer.

[16] Fonseca, N. A., Petryszak, R., Marioni, J. C., & Brazma, A. (2014). iRAP-an integrated RNA-seq Analysis Pipeline. *bioRxiv*, 005991.

[17] Kursa, M. B., Jankowski, A., & Rudnicki, W. R. (2010). Boruta–a system for feature selection. *Fundamenta Informaticae*, *101*(4), 271-285.

[18] Hakkoum, H., Idri, A., & Abnane, I. (2020). Artificial neural networks interpretation using LIME for breast cancer diagnosis. In *Trends and Innovations in Information Systems and Technologies: Volume 3 8* (pp. 15-24). Springer International Publishing.

[19] Zafar, M. R., & Khan, N. (2021). Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, *3*(3), 525-541.

[20] Ge, S. X., Son, E. W., & Yao, R. (2018). iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. *BMC bioinformatics*, *19*(1), 1-24.

[21] Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S. and Mesirov, J.P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, *102*(43), 15545-15550.

[22] Gross-Goupil, M., Françlois, L., Quivy, A., & Ravaud, A. (2013). Axitinib: a review of its safety and efficacy in the treatment of adults with advanced renal cell carcinoma. *Clinical Medicine Insights: Oncology*, *7*, CMO-S10594.

[23] Bellesoeur, A., Carton, E., Alexandre, J., Goldwasser, F., & Huillard, O. (2017). Axitinib in the treatment of renal cell carcinoma: design, development, and place in therapy. *Drug design, development and therapy*, 2801-2811.

[24] Okabe, S., Tauchi, T., Tanaka, Y., Sakuta, J., & Ohyashiki, K. (2015). Anti-leukemic activity of axitinib against cells harboring the BCR-ABL T315I point mutation. *Journal of Hematology & Oncology*, *8*, 1-4.

[25] Khalaf, K., Hana, D., Chou, J. T. T., Singh, C., Mackiewicz, A., & Kaczmarek, M. (2021). Aspects of the tumor microenvironment involved in immune resistance and drug resistance. *Frontiers in immunology*, *12*, 656364.

[26] Ge, S. X., Jung, D., & Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. *Bioinformatics*, *36*(8), 2628-2629.

[27] Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, *28*(1), 27-30.

[28] Bao, Y., Wang, L., Shi, L., Yun, F., Liu, X., Chen, Y., ... & Jia, Y. (2019). Transcriptome profiling revealed multiple genes and ECM-receptor interaction pathways that may be associated with breast cancer. *Cellular & molecular biology letters*, *24*, 1-20.

[29] Vasilaki, D., Bakopoulou, A., Tsouknidas, A., Johnstone, E., & Michalakis, K. (2021). Biophysical interactions between components of the tumor microenvironment promote metastasis. *Biophysical Reviews*, *13*(3), 339-357.

[30] Biber, G., Ben-Shmuel, A., Sabag, B., & Barda-Saad, M. (2020). Actin regulators in cancer progression and metastases: From structure and function to cytoskeletal dynamics. *International Review of Cell and Molecular Biology*, *356*, 131-196.

[31] Sit, S. T., & Manser, E. (2011). Rho GTPases and their role in organizing the actin cytoskeleton. *Journal of cell science*, *124*(5), 679-683.

[32] Nagano, M., Hoshino, D., Koshikawa, N., Akizawa, T., & Seiki, M. (2012). Turnover of focal adhesions and cancer cell migration. *International journal of cell biology*, *2012*.

[33] Fife, C. M., McCarroll, J. A., & Kavallaris, M. (2014). Movers and shakers: cell cytoskeleton in cancer metastasis. *British journal of pharmacology*, *171*(24), 5507-5523.

[34] Samuels, Y., & Ericson, K. (2006). Oncogenic PI3K and its role in cancer. *Current opinion in oncology*, *18*(1), 77-82.

[35] Song, M. S., Salmena, L., & Pandolfi, P. P. (2012). The functions and regulation of the PTEN tumour suppressor. *Nature reviews Molecular cell biology*, *13*(5), 283-296.

[36] Elkabets, M., Pazarentzos, E., Juric, D., Sheng, Q., Pelossof, R. A., Brook, S., ... & Baselga, J. (2015). AXL mediates resistance to PI3Kα inhibition by activating the EGFR/PKC/mTOR axis in head and neck and esophageal squamous cell carcinomas. *Cancer cell*, *27*(4), 533-546.

[37] Zhang, Z., Yang, S., & Wang, Q. (2019). Impact of MET alterations on targeted therapy with EGFR-tyrosine kinase inhibitors for EGFR-mutant lung cancer. *Biomarker Research*, *7*(1), 1-7.

[38] Hudson, C. D., Hagemann, T., Mather, S. J., & Avril, N. (2014). Resistance to the tyrosine kinase inhibitor axitinib is associated with increased glucose metabolism in pancreatic adenocarcinoma. *Cell death & disease*, *5*(4), e1160-e1160.

[39] Yu, S., Wang, G., Shi, Y., Xu, H., Zheng, Y., & Chen, Y. (2020). MCMs in cancer: prognostic potential and mechanisms. *Analytical Cellular Pathology*, *2020*.

[40] Jiang, Y., Zhang, H., Wang, J., Liu, Y., Luo, T., & Hua, H. (2022). Targeting extracellular matrix stiffness and mechanotransducers to improve cancer therapy. *Journal of hematology & oncology*, *15*(1), 34.

[41] Li, X., & Wang, J. (2020). Mechanical tumor microenvironment and transduction: cytoskeleton mediates cancer cell invasion and metastasis. *International journal of biological sciences*, *16*(12), 2014.

[42] Reif, S., Lang, A., Lindquist, J.N., Yata, Y., Gäbele, E., Scanga, A., Brenner, D.A. and Rippe, R.A. (2003). The role of focal adhesion kinase-phosphatidylinositol 3-kinase-akt signaling in hepatic stellate cell proliferation and type I collagen expression. *Journal of Biological Chemistry*, *278*(10), 8083-8090.

[43] Lou, Y., Jiang, Y., Liang, Z., Liu, B., Li, T., & Zhang, D. (2021). Role of RhoC in cancer cell migration. *Cancer Cell International*, *21*(1), 1-16.

[44] Liao, X., Li, X., & Liu, R. (2023). Extracellular-matrix mechanics regulate cellular metabolism: A ninja warrior behind mechano-chemo signaling crosstalk. *Reviews in Endocrine and Metabolic Disorders*, *24*(2), 207-220.

[45] Gargalionis, A. N., Papavassiliou, K. A., Basdra, E. K., & Papavassiliou, A. G. (2022). mTOR signaling components in tumor mechanobiology. *International Journal of Molecular Sciences*, *23*(3), 1825.

[46] Rahaman, S. G., Mahanty, M., Mukherjee, P., Dutta, B., & Rahaman, S. O. (2023). Mechanosensing and Mechanosignal Transduction in Atherosclerosis. *Current Atherosclerosis Reports*, *25*(10), 711-721.

[47] Pobbati, A. V., & Hong, W. (2020). A combat with the YAP/TAZ-TEAD oncoproteins for cancer therapy. *Theranostics*, *10*(8), 3622.

[48] Du, Y. E., Tu, G., Yang, G., Li, G., Yang, D., Lang, L., ... & Hou, Y. (2017). MiR-205/YAP1 in activated fibroblasts of breast tumor promotes VEGF-independent angiogenesis through STAT3 signaling. *Theranostics*, *7*(16), 3972.

[49] Abdullah, S. E., & Perez-Soler, R. (2012). Mechanisms of resistance to vascular endothelial growth factor blockade. *Cancer*, *118*(14), 3455-3467.

[50] Tabernero, J. (2007). The role of VEGF and EGFR inhibition: implications for combining anti–VEGF and anti–EGFR agents. *Molecular cancer research*, *5*(3), 203-220.

[51] Farooq, M., Bhat, G. R., Besina, S., Thakur, N., Zahoor, S., Rather, R. A., ... & Afroze, D. (2023). Expression of HIF-1α and markers of angiogenesis and metabolic adaptation in molecular subtypes of breast cancer. *Translational Medicine Communications*, *8*(1), 1-13.

[52] Fernandes, M., Jamme, P., Cortot, A. B., Kherrouche, Z., & Tulasne, D. (2021). When the MET receptor kicks in to resist targeted therapies. *Oncogene*, *40*(24), 4061-4078.

[53] Tang, Y., Zang, H., Wen, Q., & Fan, S. (2023). AXL in cancer: a modulator of drug resistance and therapeutic target. *Journal of Experimental & Clinical Cancer Research*, *42*(1), 148.

[54] Zhou, L., Liu, X. D., Sun, M., Zhang, X., German, P., Bai, S., ... & Jonasch, E. (2016). Targeting MET and AXL overcomes resistance to sunitinib therapy in renal cell carcinoma. *Oncogene*, *35*(21), 2687-2697.