IDENTIFICATION OF POTENTIAL BREAST CANCER DRUG THROUGH AN INTEGRATIVE APPROACH OF SINGLE-CELL RNA SEQUENCING DATA ANALYSIS AND MACHINE LEARNING

DISSERTATION SUBMITTED IN FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF DEGREE OF

MASTER OF SCIENCE

IN

BIOTECHNOLOGY

THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

By

SARVBHAUM SHUKLA

302101023

Under the supervision of:

Dr. Mohd. Imran Siddiqi
Sr. Principal Scientist, Professor AcSIR
Biochemistry and Structural Biology Division (CSIR-CDRI)

Internal guide:

Dr. Vikas Handa
Assistant Professor
Department of Biotechnology
Thapar Institute of Engineering and Technology

CERTIFICATE

This is to certify that **Mr. Sarvbhaum Shukla**, final year student at the Thapar Institute of Engineering and Technology (Thapar University), Patiala has successfully completed the dissertation report titled "**Identification of Potential Breast Cancer Drug through an Integrative Approach of Single-Cell RNA Sequencing Data Analysis and Machine Learning**" under the supervision of Dr. Mohd. Imran Siddiqi, Sr. Principal Scientist, CDRI, and under the guidance of internal guide Dr. Vikas Handa, Assistant Professor, Thapar University, in partial fulfilment of requirement for the award of degree of **Master of Science in Biotechnology**. The matter embodied in this dissertation report has not been previously submitted in part or full for the award of any degree at any other university or institution and any such material which has been obtained from other sources has been duly acknowledged in the dissertation.

Imran Siddiqi

Dr. Mohd. Imran Siddiqi
Guide
Senior Principal Scientist
Professor, AcSIR
Coordinator/PI-Center for Bioinformatics & Computational Biology
Biochemistry and Structural Biology Division
CSIR-CDRI

25.08.23

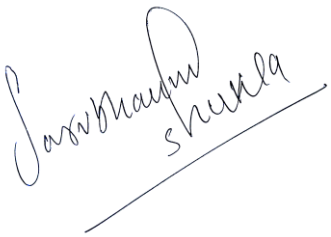
Dr. Vikas Handa
Internal Guide
Assistant Professor
Thapar University

DECLARATION

I, **Sarvbhaum Shukla**, a student of M.Sc. Biotechnology, have completed my six months dissertation work embodied in the thesis entitled, **"Identification of Potential Breast Cancer Drug through an Integrative Approach of Single-Cell RNA Sequencing Data Analysis and Machine Learning"** was carried out by me under the supervision of Dr. Mohd. Imran Siddiqi, Sr. Principal Scientist, Biochemistry & Structural Biology, Central Drug Research Institute, in partial fulfilment of the requirements for the award of M.Sc. Degree in Biotechnology. This work has not been submitted in part or full, to this or any other university or institution, for any degree or diploma. I, hereby, affirm that the work has been done by me in all aspects. I have sincerely prepared this project report and the results reported in this study are genuine and authentic.

Place: CDRI, Lucknow

Date: 14th July, 2023



Sarvbhaum Shukla

Acknowledgements

The completion of my master's thesis on the topic " Identification of Potential Breast Cancer Drug through an Integrative Approach of Single-Cell RNA Sequencing Data Analysis and Machine Learning" would not have been possible without the support and guidance of several individuals and institution. I would like to take this opportunity to express my sincere gratitude to all of them.

Firstly, I would like to express my heartfelt gratitude to my thesis guide, **Dr. Mohd. Imran Siddiqi**, Senior Principal Scientist, Professor, AcSIR, Biochemistry and Structural Biology Division, CSIR-CDRI and internal guide **Dr. Vikas Handa**, assistant Professor, Thapar Institute of engineering &Technology, for their constant support, guidance, and encouragement throughout the entire research period. Their invaluable feedback, constructive criticism, and expert insights have helped me to gain a deeper understanding of the subject matter and to refine my research work to a great extent.

I would like to thank **Dr. Aman C. Kaushik**, Project Scientist at Bioinformatic division, CDRI, who has been a constant support of encouragement and has provided valuable insights throughout my endeavour. Big thanks to **Dr. Gaurav Srivastava**, ICMR-RA, CDRI who has always helped me in his calm and composed manner and has been my go-to person since day one of my dissertation. I would really like to thank the research scholars (PhD-JRF) **Shubham Talware, Shivangi Yadav, and Saumya Singh**, who were always there to answer my frequent and annoying doubts, helped me understand the vast world of research and showered me with their experiences and invaluable feedbacks on my project topic. My heartful thanks to all the lab members of Bioinformatics division, CDRI, their humour and intellect made my lab experience smooth and unforgettable.

I would also like to extend my gratitude to Thapar University for providing me with the opportunity to work on this project, and Central Drug Research Institute for providing me with the resources. I am also grateful to my six constant friends, **Tathagata, Sahibleen, Ananaya, Dhruv, Aanchal and Avneet** who despite being miles away, always encouraged me to be productive and boosted my moral.

Lastly, I would like to express my heartfelt gratitude to my parents **Mr. Rajeev Lochan Shukla** and **Mrs. Sushma Shukla** for always being there for me, my brother **Yasharth Shukla** for

being my all-time mentor and encouraging me to achieve greater heights. A special mention to **Ayushika Saraswat** who has been a constant source of love, encouragement, and support throughout my academic journey. Their unwavering support, patience, and understanding have been instrumental in helping me to stay focused and motivated throughout this challenging research process.

I am truly grateful to everyone who has supported me in this research endeavour, and I acknowledge their contributions wholeheartedly. Their support and guidance have been instrumental in shaping me as a researcher and academic, and I will always cherish their contribution towards my success.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADME: Absorption, Distribution, Metabolism, and Excretion

ANN: Aritificial neural network

ASN: Aspargine

ASP: Aspartic acid

AUC: Area under the Receiver Operating Characteristic Curve

BRCA1: Breast cancer gene 1

BRCA2: Breast cancer gene 2

Bulk-RNA-Seq: Bulk RNA Sequencing

CCLE: Cancer Cell Encyclopedia

ChEMBL: a manually curated database of bioactive molecules with drug-like properties

CNN: convolutional neural networks

DDRC: Drug dose response curves

DNA: Deoxyribonucleic Acid

EF: Enhancement factor

EGFR: Epidermal growth factor receptor

ERBB2: Erb-B2 Receptor Tyrosine Kinase 2

FDA: Food and Drugs Act

GBRT: Gradient Boosted Regression Trees

GDSC: Genomics of Drug Sensitivity in Cancer

GLU: Glutamic acid

GROMACS: Groningen Machine for Chemical Simulations

IC50: Half-maximal inhibitory concentration

IDE: Integrated Development Environment

LEU: Leucine

Log: logarithmic

LogP: the log of the partition coefficient of a solute between octanol and water, at near infinite dilution

MACCS: Molecular ACCess Systems

MD: Molecular Dynamics

MET: Methionine

nM: Nano molar

PAINS: Pan Assay Interference Compounds

PDB: Protein data bank

PDBID: Protein data bank identifier

PHE: Phenylalanine

PRO: Proline

Pset: Pharmacoset

RF: Random Forest

RNA: Ribonucleic Acid

RNN: recurrent neural networks

ScDNA-Seq: Single cell DNA sequencing

ScRNA-Seq: Single-Cell RNA Sequencing

SEED: an integer value to ensure that the pseudo-random generation are reproducible.

Seq: Sequencing

SMART-Seq: Switch Mechanism at the 5' End of RNA Templates

SMILES: Simplified Molecular input line entry system

SVM: Support vector machine

TNBC: Triple Negative Breast Cancer

UMAP: Uniform Manifold Approximation and Projection

Abstract

This abstract presents a study that aimed to identify potential drugs for breast cancer treatment by analyzing large datasets and employing computational methods. Initially, breast cancer data from two datasets, Cancer Cell Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC), were combined and tested for drug sensitivity using a drug-dose response curve. The drugs showing sensitivity in all eleven breast cancer cell lines were shortlisted for further evaluation. ScRNA-Seq analysis identified five drugs overlapping with the drug-sensitivity data, namely Afatinib, Bortezomib, Gemcitabine, Navitoclax, and Trametinib, which were selected for further investigation. Afatinib was evaluated based on its IC50 value and consistency across eleven cell lines. The target of Afatinib was identified from the datasets and subjected to machine learning using a Python script, which provided predictions of similar molecules in the ChEMBL database. Docking analysis was performed using the PDB ID 4G5J as a reference to predict the interactions of the target. The docking score for Afatinib was compared to a resultant compound, ChEMBL233325, with a high score of -9.4. The test compound ChEMBL233324 was selected for MD simulation as it consisted the same residue as our control compound. Overall, this study employed computational methods and analysis techniques to identify potential drugs for breast cancer treatment, providing valuable insights for further research and development.

Keywords: *Breast cancer, drug-dose response, ScRNA-Seq analysis, Machine Learning, Docking analysis, MD simulation*

INTRODUCTION

1 Introduction

Breast Cancer occurs in both men and women. However, being the most common among women, it reported approx—2.3 million cases globally in 2020 alone (Bray *et al*., 2021). Breast cancer occurs when abnormal cells in the breast grow and divide uncontrollably, forming a mass or lump. The symptoms of breast cancer range from a bump in the breast, a shift in the form of the breast, skin dimpling, liquid from the nipple, and a newly inverted nipple to a yellow or scaly hair patch.

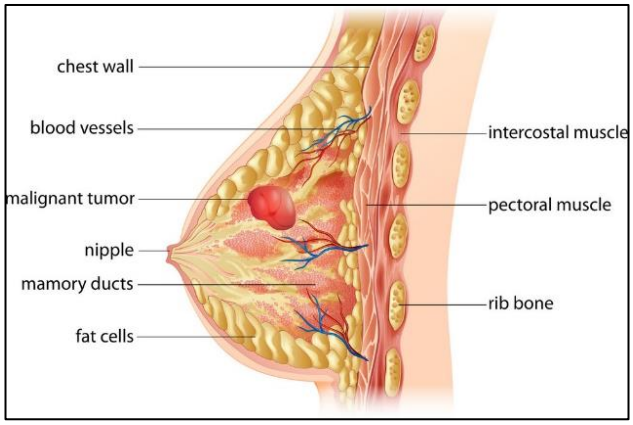


Figure 1. Typical Breast Tumor

The outcomes of breast cancer greatly depend on the form of cancer, the disorder's magnitude, and the person's age[1]. Risk factors for breast cancer are depression, increasing age, absence of physical activity, smoking, frequent consumption of liquor, hormone substitution treatment during menopause, ionizing radiation, having kids late or not at all, elderly era, and family history such as inherited mutations in specific genes [1]. Genes most well-known to trigger the onset of breast cancer are BRCA1 & BRCA2, which produce proteins to suppress tumor growth in the breast and other tissues. However, due to mutational changes, these genes increase the risk of developing breast cancer or different types of cancer. We can acknowledge that Breast cancer is a complex and heterogeneous disease, presenting significant challenges in its diagnosis and treatment. On the other hand, the advent of high-throughput technologies and the availability of large-scale genomics datasets have opened new ways and approaches for understanding the underlying mechanisms of breast cancer and identifying potential therapeutic strategies. In recent years, bioinformatics approaches have played a crucial role in deciphering the intricate molecular landscape of breast cancer and have become instrumental in drug discovery and development.

The first step in drug discovery is the evaluation of drug response profiles across various cancer cell lines. The Genomics of Drug Sensitivity in Cancer (GDSC) [2] and Cancer Cell Line Encyclopedia (CCLE) [3] databases provide valuable resources for accessing drug response data across various cancer cell lines, including breast cancer. These databases integrate comprehensive genomic and pharmacological information, enabling researchers to investigate drug sensitivity patterns and identify candidates for further investigation.

In this study, we leveraged the wealth of information available in the GDSC and CCLE databases to generate drug dose-response curves (Rplots) specifically for breast cancer cell lines. Using the robust statistical computing environment RStudio [4] we analyzed the drug response data and identified drugs exhibiting promising efficacy against breast cancer. These drugs were potential candidates for further exploration in our drug discovery pipeline. We performed single-cell RNA sequencing (ScRNA-Seq) [5] analysis on breast cancer cell lines to gain deeper insights into the molecular mechanisms underlying drug response. Single-cell RNA-seq offers the ability to capture gene expression patterns at the single-cell level, identifying cellular heterogeneity and characterizing distinct subpopulations within a tumor.

Contrary to the Single-cell RNA sequencing Analysis, a more usual approach is Bulk-RNA sequencing analysis [6], where pooled cell populations are analyzed to measure the average expression of individual genes across millions of cells to get a general idea of differences in the gene expression between samples. Bulk RNA Sequencing analysis incorporates the usage of many cells, is less sensitive, and needs to be more precise due to the average gene expression it provides. ScRNA-Seq analysis, on the other hand, is more sensitive in comparison and more precise as it provides gene expression levels for individual cells. Since the ScRNA-Seq analysis method is superior and preferable to Bulk-RNA-Seq Analysis, it is used in this study for better quality results.

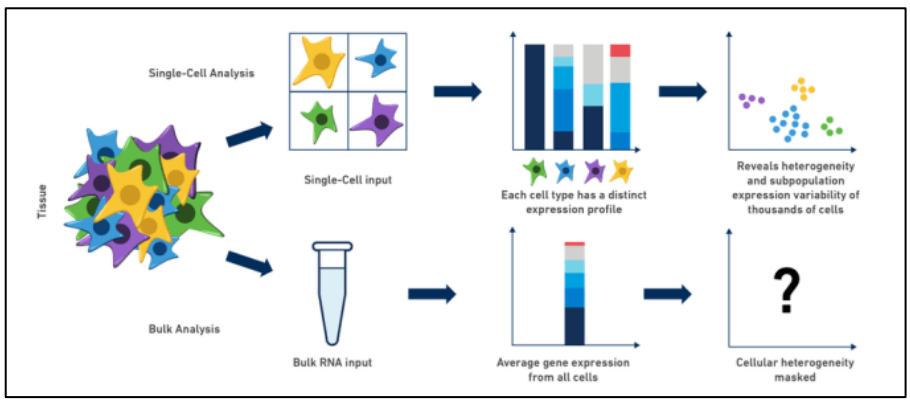


Figure 2. Bulk RNA-Seq analysis vs. Single-cell RNA-Seq analysis

By analyzing the ScRNA-Seq data, we aimed to identify specific cellular subsets or molecular signatures associated with drug response in breast cancer. Based on the ScRNA-Seq results, we shortlisted drugs that demonstrated the most promising efficacy in targeting relevant cellular subpopulations or gene expression patterns associated with breast cancer. To further optimize these drug candidates, we employed machine learning techniques. Machine learning algorithms have great potential in identifying novel drug-like compounds with desirable properties. In our study, we utilized three machine learning algorithms to generate derivatives of the shortlisted drugs to enhance their efficacy and specificity for breast cancer treatment.

(RF) Random Forest [7] is a popular machine learning algorithm used for classification and regression tasks. It belongs to the ensemble learning methods, which combine multiple decision trees to make predictions. In Random Forest, a collection of decision trees is built on randomly selected subsets of the training data, and the final prediction is made by aggregating the predictions of individual trees. The algorithm has gained popularity due to its ability to handle large datasets, high-dimensional features, and complex interactions between variables. It has been successfully applied in various domains, including finance, healthcare, and image recognition.

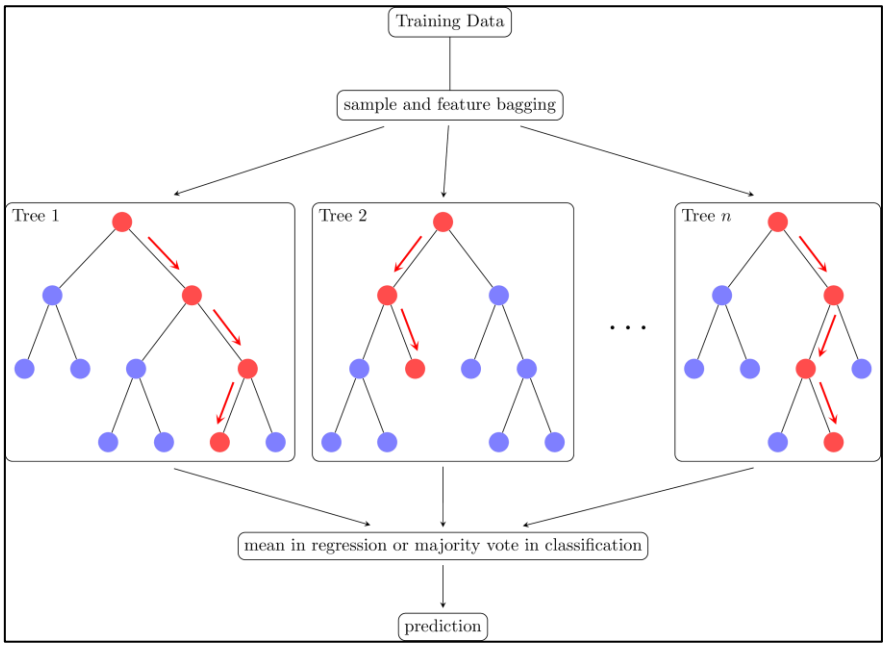


Figure 3. A flow chart of Random Forest Algorithm

(SVM) Support Vector Machine [8] is a powerful supervised learning algorithm used for both classification and regression tasks. SVM aims to find the best hyperplane that separates the

data points belonging to different classes with the largest margin. It works by mapping the input data into a high-dimensional feature space and finding an optimal hyperplane that maximally separates the classes. SVM is known for its ability to handle high-dimensional data and its robustness against overfitting. It has been widely used in areas such as image classification, text categorization, and bioinformatics.

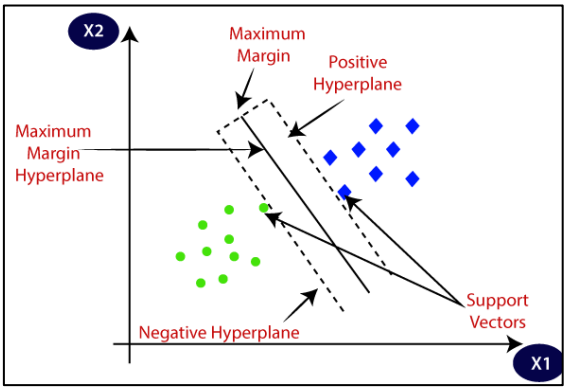


Figure 4. A representation of Support Vector Machine (SVM) Algorithm

(ANN) Artificial Neural Network [9] is a computational model inspired by the structure and function of biological neural networks. It consists of interconnected nodes, called neurons, organized in layers. Each neuron takes input signals, applies a non-linear activation function, and passes the output to the next layer. ANNs are capable of learning complex patterns and relationships in the data through a process called training. This training involves adjusting the weights and biases of the neurons to minimize the difference between the predicted and actual outputs. ANNs have shown remarkable success in various applications, including speech recognition, image processing, and natural language processing.

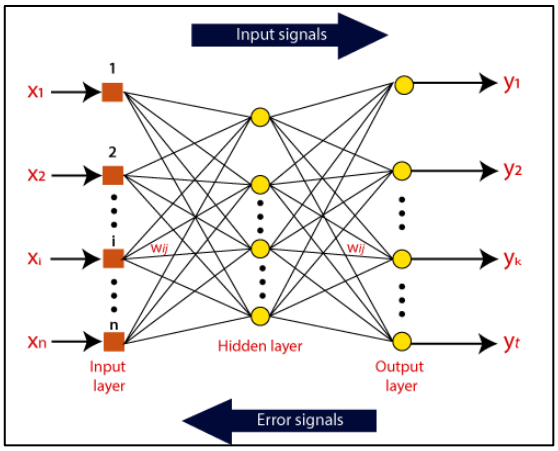


Figure 5. Artificial Neural Network Algorithm

The machine learning-derived drug derivatives were subjected to molecular docking [10] studies against the ChEMBL library. Molecular Docking provides insights into the binding interactions between small molecules and their target proteins, enabling the prediction of their potential binding affinities. We sought to identify the most favorable drug-target interactions through molecular Docking and prioritize the drug derivatives with the highest binding affinities.

We performed molecular dynamics (MD) simulations [11] to validate the potential therapeutic candidates. MD simulations allow for exploring the dynamic behavior of drug molecules and their interactions with target proteins over time. By simulating the drug-target complexes, we aimed to gain a comprehensive understanding of their stability, conformational changes, and binding kinetics. The MD simulations provided crucial information for evaluating the proposed Drug and its potential efficacy and safety.

This study integrates bioinformatics approaches with experimental and computational techniques to identify and propose the best Drug (s) possible for breast cancer treatment. By leveraging drug dose responses, single-cell RNA-seq analysis, machine learning-driven derivative generation, molecular Docking, and MD simulations, we aim to identify a promising drug candidate with enhanced efficacy and target specificity. The findings from this research hold the potential to advance breast cancer therapeutics significantly, improving patient outcomes and contributing to the ongoing efforts to combat this devastating disease.

LITERATURE REVIEW

2 Literature review

Breast cancer is a complex and heterogeneous disease with diverse molecular subtypes, necessitating the development of targeted and personalized therapeutic approaches. Extensive research has been undertaken to unravel the underlying mechanisms of breast cancer biology, drug response, and treatment resistance. By reviewing the existing literature, we can identify the progress made, the challenges encountered, and the opportunities for further investigation.

The literature review will also encompass the utilization of machine learning algorithms in the context of breast cancer drug discovery. By examining studies that have utilized machine learning for predictive modeling, drug candidate optimization, and therapeutic target identification, we can assess the potential of these computational approaches in enhancing personalized treatment strategies for breast cancer patients.

Additionally, the review will explore the field of molecular docking, which enables the investigation of drug-protein interactions, and MD simulations, which provide insights into the stability and dynamics of drug molecules and target proteins. By analyzing relevant studies, we can understand the role of these computational techniques in evaluating the binding affinity, conformational changes, and binding kinetics of potential drug candidates.

Through their analysis, [12] the authors made several important findings that shed light on the clonal evolution of breast cancer. Firstly, they observed extensive intra-tumor heterogeneity, with the presence of multiple co-existing clones within each tumor sample. This heterogeneity suggests diverse subpopulations of cells with distinct genomic alterations and evolutionary trajectories. Secondly, the authors identified common genetic alterations, such as copy number variations and somatic mutations, that were shared among multiple clones, indicating the presence of clonal expansions. These clonal expansions likely contribute to tumor growth and development. Additionally, the authors discovered subclones with unique genomic alterations, highlighting the ongoing diversification and evolution of breast cancer cells within the tumor.

[12] The authors discuss the integration of deep learning algorithms with single-cell data to improve drug response prediction. Deep learning methods, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have shown great promise in analyzing complex biological data and extracting informative features. By training deep

learning models on large-scale single-cell datasets, researchers can uncover patterns and signatures associated with drug response, leading to more accurate predictions. The authors provide an overview of various single-cell techniques that enable the measurement of cellular responses to drugs with high resolution. These techniques include single-cell RNA sequencing (ScRNA-Seq), single-cell proteomics, and single-cell imaging assays. By capturing molecular profiles and functional readouts of individual cells, these techniques enable the characterization of cellular states and heterogeneity, which are critical for predicting drug response accurately.

[13] The authors proposed an ensemble method that integrates matrix completion and ridge regression to improve drug-response prediction. Matrix completion is used to fill in missing drug-response measurements, utilizing known responses to infer and complete the missing values. Ridge regression, a regularization technique, is then applied to train predictive models using the completed matrix. The ensemble method combines the predictions from multiple matrix completion and ridge regression models, resulting in a more robust and accurate drug-response prediction model. The authors evaluate the performance of their proposed ensemble method using publicly available cancer cell line datasets with associated drug-response measurements. They compare the predictive accuracy of their method with other state-of-the-art approaches for drug-response prediction. The evaluation includes metrics such as mean squared error and Pearson correlation coefficient to assess the accuracy and robustness of the predictions. The authors present the results of their experiments, demonstrating that the proposed ensemble method outperforms other methods in predicting anticancer drug response. The combination of matrix completion and ridge regression effectively addresses the challenges of missing data and high dimensionality, leading to improved accuracy and reliability in drug-response prediction. The authors conclude that their ensemble method has the potential to enhance personalized medicine by providing more accurate predictions of drug response, facilitating informed treatment decisions.

[14] The authors employed ScRNA-Seq technology to perform transcriptomic profiling at the single-cell level, allowing them to capture the heterogeneity of tumor and immune cell populations within primary breast cancer samples. By isolating individual cells and sequencing their RNA, the researchers obtained detailed gene expression profiles, enabling comprehensive analysis of the cellular composition and functional states of the tumor and immune microenvironment. Through their analysis, [14] made several important findings that contribute to our understanding of primary breast cancer. Firstly, they identified distinct cell clusters

corresponding to different cell types, including tumor cells, immune cells, stromal cells, and endothelial cells. This comprehensive cellular profiling provides insights into the cellular heterogeneity within primary breast tumors. Secondly, the authors discovered unique gene expression signatures associated with different cell types and states. This information sheds light on the molecular characteristics and functional diversity of tumor and immune cells in breast cancer.

[15] The authors employed ScRNA-Seq technology to profile the transcriptomes of individual immune cells within breast tumor samples. By analyzing gene expression patterns at the single-cell level, they identified distinct immune cell populations and characterized their phenotypic and functional diversity. The researchers used a computational method called Biscuit to identify cell populations that differed in co-expression patterns, providing insights into the heterogeneity of immune cells within the tumor microenvironment. Through their analysis, the authors made several important findings that enhance our understanding of immune phenotypes in the breast tumor microenvironment. Firstly, they observed a phenotypic expansion of intra-tumoral immune cells, indicating a broad range of functional states and activation levels within the immune cell populations. This diversity suggests complex interactions between tumor cells and the immune system. Secondly, the authors identified continuous activation and differentiation trajectories of T cells, revealing the dynamic nature of T cell responses within the tumor microenvironment. Furthermore, they discovered that combinatorial environmental inputs and T cell receptor (TCR) usage shape the phenotypic diversity of T cells, underscoring the impact of both intrinsic and extrinsic factors on T cell phenotypes.

[1] Here the authors describe their integrated approach, which involves several steps. Firstly, they analyze gene expression data from breast cancer cell lines to identify differentially expressed genes between cancer and normal cell lines. Then, they construct a protein-protein interaction network and prioritize genes based on their network properties and connectivity. They further analyze the enriched biological pathways and functional modules using gene set enrichment analysis and module identification algorithms. Finally, they validate the potential biomarkers using independent breast cancer datasets. Through their analysis, the authors identified several novel biomarkers associated with breast cancer. They discovered differentially expressed genes that play critical roles in breast cancer progression and identified key biological pathways and functional modules involved in the disease. The integration of network-based approaches allowed for the identification of hub genes and potential regulatory

relationships, providing further insights into the molecular mechanisms underlying breast cancer development. The study by [1] has important implications for breast cancer research and clinical applications. The identified biomarkers have the potential to improve diagnosis, prognosis, and treatment selection in breast cancer patients. The integrated systems biology and network-based approach provide a comprehensive framework for biomarker discovery, facilitating a deeper understanding of the molecular basis of breast cancer and enabling the development of personalized therapeutic strategies.

[16] The authors employed ScRNA-Seq technology to profile gene expression in TNBC samples at the single-cell level. By analyzing the transcriptomes of individual cells, they identified distinct TNBC subtypes based on their gene expression profiles. Subsequently, they systematically investigated the association between these subtypes and patient prognosis using survival analysis. The authors developed a computational framework to identify subtype-specific prognosis signatures, which were based on gene expression patterns and their correlation with patient outcomes. Through their analysis, [16] made several important findings that enhance our understanding of TNBC subtypes and their prognostic implications. Firstly, they identified multiple TNBC subtypes characterized by distinct gene expression profiles. These subtypes exhibited differences in their prognosis and clinical outcomes. Secondly, the authors systematically identified subtype-specific prognosis signatures by pinpointing genes whose expression patterns were associated with patient survival. These signatures provide potential biomarkers for predicting patient prognosis and may have implications for personalized treatment approaches in TNBC. The study by [16] has important implications for the stratification and management of TNBC patients. The identified TNBC subtypes and associated prognosis signatures provide insights into the heterogeneity and clinical outcomes within this subtype of breast cancer. These findings have the potential to contribute to the development of personalized treatment strategies and the identification of therapeutic targets specific to TNBC subtypes.

[17] The authors describe the methodology behind Digitaldlsorter, which utilizes deep learning techniques to deconvolute ScRNA-Seq data. The model is trained using a large dataset of ScRNA-Seq profiles with known cell type annotations. Digitaldlsorter learns to recognize cell type-specific gene expression patterns and can subsequently classify and quantify the abundance of different cell types within a mixed cell population based on gene expression data. Through their analysis, [17] demonstrate the effectiveness of Digitaldlsorter in accurately

deconvoluting gene expression data obtained from ScRNA-Seq experiments. The deep learning model successfully identifies and quantifies the abundance of distinct cell types within a mixed population. The authors showcase the utility of Digitaldlsorter by applying it to different ScRNA-Seq datasets, highlighting its ability to provide valuable insights into cellular composition and gene expression dynamics. This study has important implications for the analysis of ScRNA-Seq data and the study of cellular heterogeneity. Digitaldlsorter offers a powerful tool for deconvoluting complex gene expression data, enabling researchers to identify and quantify the abundance of different cell types within a sample. This information can aid in understanding cellular dynamics, identifying novel cell populations, and investigating cell type-specific gene expression patterns in various biological processes and diseases.

[5] The authors discuss different ScRNA-Seq technologies, including droplet-based methods (e.g., Drop-seq, 10x Genomics), microwell-based methods (e.g., SMART-seq) , and combinatorial indexing methods (e.g., sci-RNA-seq, Perturb-seq). They describe the key steps involved in ScRNA-Seq, such as cell isolation, reverse transcription, library preparation, and sequencing. The authors highlight the strengths and limitations of each technology and provide insights into choosing the appropriate method based on the research objectives. [5] discuss the computational challenges in analyzing ScRNA-Seq data and provide an overview of the common analytical steps. These steps include quality control, read alignment, gene quantification, normalization, dimensionality reduction, clustering, and differential expression analysis. The authors describe the algorithms and tools commonly used for each step and discuss considerations for data interpretation and downstream analysis. They also touch upon emerging computational methods for addressing specific challenges in ScRNA-Seq analysis, such as batch effect correction and trajectory inference. The authors highlight the broad applications of ScRNA-Seq technologies and their impact on various fields, including developmental biology, cancer research, immunology, neuroscience, and regenerative medicine. They emphasize the importance of computational data analysis in extracting meaningful biological insights from ScRNA-Seq datasets. The ability to study gene expression profiles at the single-cell level offers unprecedented resolution and has the potential to uncover novel cell types, regulatory networks, and therapeutic targets.

[18] The authors describe their methodology, which involves integrating gene expression data and pathway cross-talk information to identify potential drug target pathways. They first analyze gene expression profiles of breast cancer subtypes to identify differentially expressed

genes and pathways. Next, they analyze the cross-talk between pathways to identify potential targetable interactions. Finally, they validate their findings using drug sensitivity data and assess the clinical relevance of the identified drug target pathways. Through their analysis, [18] made several important findings. They identified subtype-specific differentially expressed genes and pathways, highlighting the molecular differences between breast cancer subtypes. By analyzing pathway cross-talk, they identified potential drug target pathways that are involved in pathway interactions specific to each subtype. The authors provide evidence of the clinical relevance of the identified drug target pathways by correlating them with drug sensitivity and patient survival data.

[19] The authors describe their methodology, which involves integrating and analyzing publicly available breast cancer datasets to identify genes associated with breast cancer progression and prognosis. They utilize bioinformatics tools and algorithms to perform gene expression analysis, identify differentially expressed genes, and assess their potential biological functions and pathways. The authors also discuss the potential therapeutic implications of the identified genes and explore their interactions with existing drugs or molecular targets. Through their analysis, [19] made several important findings. They identified genes that are differentially expressed in breast cancer samples compared to normal tissues and explored their potential roles in breast cancer progression and prognosis. The authors highlight the potential therapeutic relevance of the identified genes by discussing their interactions with existing drugs or molecular targets. Additionally, they provide insights into the biological pathways and processes associated with breast cancer based on the analysis of gene expression patterns. The study has important implications for breast cancer research and personalized medicine. The identified genes associated with breast cancer progression and prognosis provide potential biomarkers for disease prognosis, risk stratification, and treatment response prediction. Moreover, the exploration of potential therapeutic targets and their interactions with existing drugs opens avenues for developing novel therapeutic strategies and personalized treatment approaches in breast cancer.

[20] The authors describe the methodology used to develop the GBRT model. They collected experimental solubility data for Decitabine and utilized various molecular descriptors and physicochemical properties as input features. The GBRT model was trained using the collected data, and its predictive performance was assessed. Additionally, the authors utilized the model to optimize the solubility of Decitabine by identifying key molecular features and conditions

that enhance its solubility. Through their analysis, [20] made several important findings. They developed a GBRT model capable of accurately predicting the solubility of Decitabine based on its molecular descriptors and physicochemical properties. The model demonstrated robustness and high predictive performance. Additionally, the authors utilized the model to optimize the solubility of Decitabine by identifying specific molecular features and conditions that can be manipulated to enhance solubility. The study has important implications for the formulation and development of Decitabine as an anti-cancer drug. The GBRT model provides a novel mathematical tool for predicting and optimizing the solubility of Decitabine, which can aid in the formulation of more effective drug delivery systems. The insights gained from the model can help researchers and pharmaceutical companies design strategies to enhance the solubility and bioavailability of Decitabine, ultimately improving its therapeutic efficacy. The model demonstrates high predictive performance and offers insights into the molecular features and conditions that influence solubility. This research has implications for enhancing the formulation and delivery of Decitabine, potentially improving its therapeutic effectiveness in the treatment of myelodysplastic syndromes.

[21] The advent of single-cell technologies has revolutionized our ability to study the complexities of cellular heterogeneity and responses to drugs at a high-resolution level. Analyzing the vast amount of data generated from single-cell studies requires advanced computational approaches. In this research paper titled "Trends and Potential of Machine Learning and Deep Learning in Drug Study at Single-Cell Level," [21] explore the trends and potential applications of machine learning and deep learning techniques in drug studies at the single-cell level. The authors provide an overview of the methodologies employed in single-cell drug studies using machine learning and deep learning approaches. They discuss the key steps involved, including data preprocessing, feature selection, model construction, and evaluation. The authors highlight the advantages of machine learning and deep learning techniques in handling the high-dimensional and complex nature of single-cell data, as well as their ability to capture intricate patterns and make predictions. Through their analysis, [21] identify several important trends and potential applications of machine learning and deep learning in drug studies at the single-cell level. They discuss the utilization of these techniques in drug response prediction, drug discovery, drug toxicity assessment, and personalized medicine. The authors highlight the ability of machine learning and deep learning models to leverage single-cell data to identify drug targets, understand drug mechanisms, and optimize treatment strategies. The study by both the authors has important implications for advancing

drug study approaches at the single-cell level. Machine learning and deep learning techniques offer powerful tools for analyzing and interpreting the complex and high-dimensional single-cell data, enabling the discovery of novel drug targets, the prediction of drug responses, and the optimization of therapeutic interventions. The application of these techniques in drug studies has the potential to accelerate the development of personalized medicine and improve patient outcomes. The findings highlight the power of these computational approaches in analyzing complex single-cell data and their implications for drug discovery, drug response prediction, and personalized medicine.

[22] The authors discuss the progress made in applying single-cell sequencing to various aspects of cancer research. They highlight the use of single-cell RNA sequencing (ScRNA-Seq) in characterizing cellular heterogeneity, identifying cell types and states within tumors, and elucidating signaling pathways and regulatory networks. They also discuss the application of single-cell DNA sequencing (ScDNA-seq) and single-cell epigenomic profiling in studying clonal evolution, genetic alterations, and epigenetic changes in cancer cells. [22] discuss the prospects and future directions of single-cell sequencing in cancer research. They highlight the potential of integrating multi-omics approaches, such as simultaneous analysis of transcriptome and epigenome, to provide a comprehensive understanding of tumor biology. The authors also discuss the challenges and opportunities in leveraging single-cell sequencing for clinical applications, including the identification of novel biomarkers, monitoring treatment responses, and guiding personalized therapies. The study by [22]underscores the significance of single-cell sequencing technologies in advancing cancer research. The ability to capture the molecular and functional heterogeneity of tumor cells at the single-cell level provides valuable insights into tumor biology and has implications for precision medicine. Single-cell sequencing holds promise for identifying novel therapeutic targets, optimizing treatment strategies, and improving patient outcomes in cancer.

Overall, this literature review aims to provide a comprehensive synthesis of the current knowledge, gaps, and opportunities in the field of breast cancer research, drug response profiling, ScRNA-Seq analysis, machine learning, molecular docking, and MD simulations. By critically analyzing the existing literature, this thesis aims to contribute to the advancement of knowledge and the development of more effective therapeutic strategies for breast cancer patients.

PROBLEM STATEMENT

3 Problem statement

Breast cancer is a complex and heterogeneous disease requiring practical therapeutic approaches tailored to its diverse molecular subtypes and variable treatment responses. While significant advancements have been made in understanding breast cancer biology, a critical need remains to identify novel drugs and therapeutic targets specific to breast cancer subtypes. Moreover, integrating multiple data sources and employing advanced computational techniques are crucial for unraveling the underlying mechanisms of drug response and resistance in breast cancer.

Despite extensive drug response data in databases such as GDSC and CCLE, identifying optimal drug candidates for breast cancer treatment remains a challenge. The molecular heterogeneity within breast cancer tumors poses further obstacles in predicting drug response patterns and identifying effective therapeutic targets.

Single-cell RNA sequencing (ScRNA-Seq) analysis offers an unprecedented opportunity to explore breast cancer's cellular heterogeneity and molecular landscape. However, there needs to be more knowledge regarding translating ScRNA-Seq data analysis into identifying clinically relevant drug targets. Integrating ScRNA-Seq data with drug response profiles holds promise in identifying specific cellular subsets and molecular signatures associated with drug sensitivity or resistance, enabling the development of personalized treatment approaches.

The potential of machine learning-driven derivative generation and molecular docking techniques in breast cancer drug discovery remains underexplored. These computational approaches provide avenues for optimizing drug candidates and enhancing their efficacy and specificity for breast cancer treatment. However, their application in the context of breast cancer requires further investigation and validation.

Furthermore, molecular docking studies provide valuable insights into the binding interactions between drug candidates and target proteins. However, their predictive power must be assessed and expanded to incorporate factors such as protein flexibility and solvent effects. Molecular dynamics (MD) simulations offer a dynamic view of drug-protein interactions, enabling the evaluation of stability, conformational changes, and binding kinetics.

Therefore, this thesis addresses the overarching problem of identifying and optimizing drug candidates for breast cancer treatment by integrating diverse datasets, including drug response profiles, ScRNA-Seq analysis, machine learning-driven derivative generation, and molecular Docking. The thesis aims to bridge the gaps in knowledge by leveraging these methodologies to uncover potential therapeutic targets, enhance drug efficacy, and advance personalized treatment strategies for breast cancer patients.

AIMS & OBJECTIVES

4 Aims and Objectives

4.1 Aim

This thesis aims to investigate and identify potential drug candidate(s) for breast cancer treatment by integrating diverse datasets, including drug response profiles, single-cell RNA sequencing (ScRNA-Seq) analysis, Machine Learning-driven derivative generation, Molecular Docking, and Molecular Dynamics (MD) simulations.

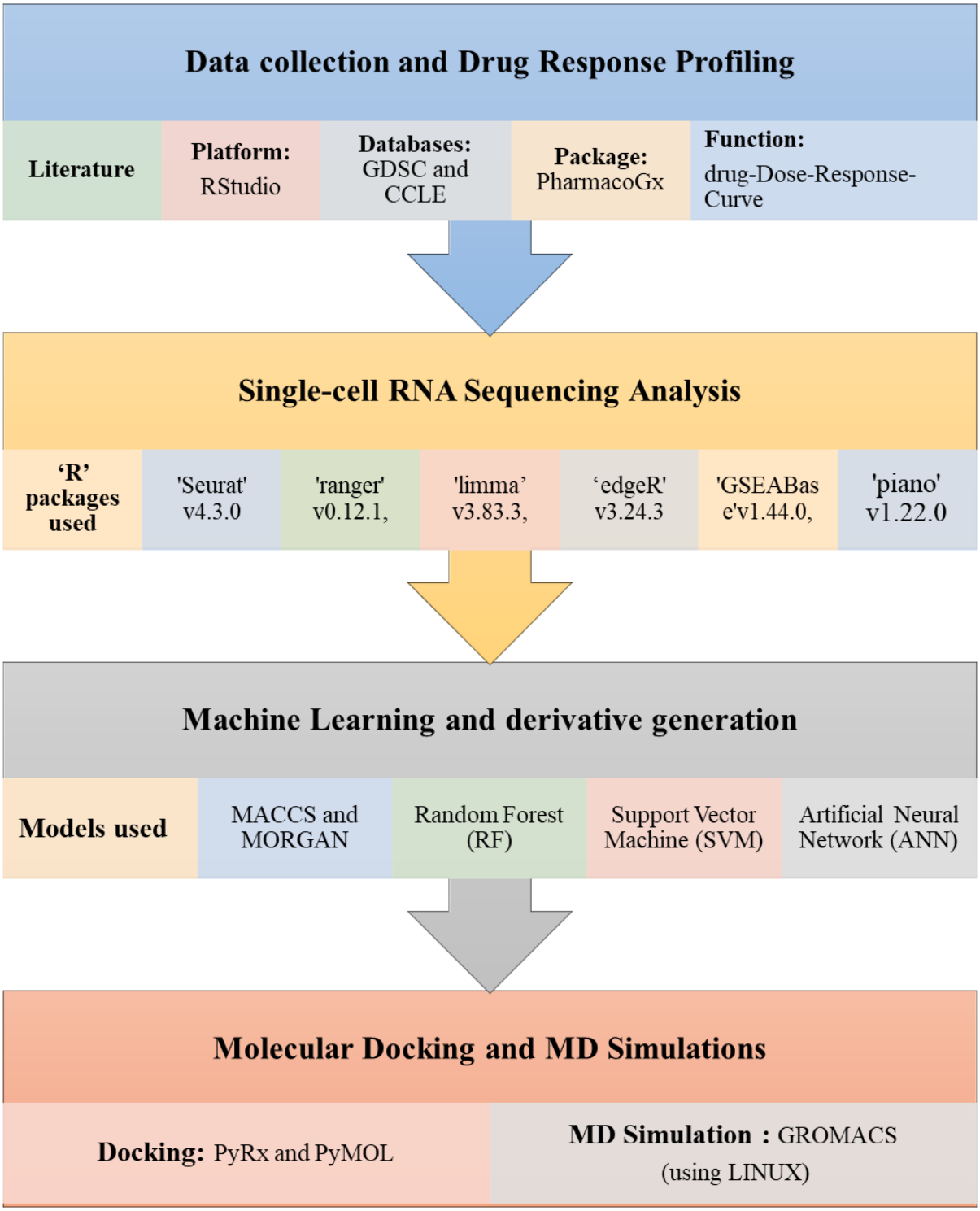
4.2 Objectives

1. **Data collection and Drug response profiling:** To analyze drug response profiles from the GDSC and CCLE databases for breast cancer cell lines.
2. **Single-cell RNA Sequencing Analysis:** Perform ScRNA-Seq analysis on breast cancer cell lines to capture gene expression profiles at the single-cell level.
3. **Machine Learning and Derivative Generation:** Develop predictive models using machine learning algorithms to generate derivatives of the shortlisted drugs.
4. **Molecular Docking and MD Simulations:** Perform molecular docking studies to explore the binding interactions between the generated drug derivatives and target proteins. Conduct molecular dynamics (MD) simulations to evaluate the stability, conformational changes, and binding kinetics of the proposed drug candidates.

MATERIALS AND METHODS

5 Materials and methods

5.1 Workflow



5.2 Data extraction from freely accessible public repositories

The data for this study was obtained from freely accessible public repositories, namely, CCLE [3] and GDSC [2]. These two cloud-based repositories were used to curate Breast Cancer cell lines. GDSC and CCLE pharmacosets were converted to data frames to generate datasets containing cancer cell line data and treatment data to acquire information on the anti-cancer drugs applied. To overcome any overlapping, breast cancer drugs and breast cancer cell lines were shortlisted after intersecting both databases. A total of 125 drugs were sorted out for further evaluation. Subsequently, cell line–drug sensitivity analysis of 11 breast cancer cell lines (listed in Table 1) was performed for the 125 sorted drugs.

Table 1. List of breast cancer cell lines

S.no	Cell lines from GDSC	Cell lines overlapping with CCLE	Characteristic disease
1.	HCC1187	HCC1187	Ductal carcinoma
2.	HCC1806	HCC1806	Squamous cell breast carcinoma
3.	ZR-75-30	ZR-75-30	Ductal carcinoma
4.	COLO 824	-	Breast carcinoma
5.	MDA-MB-330	-	Invasive lobular carcinoma
6.	MFM-223	-	Ductal carcinoma
7.	MRK-nu-1	-	Breast carcinoma
8.	OCUB-M	-	Breast carcinoma
9.	Evsa-T	-	Invasive ductal carcinoma
10.	MDA-MB-361	MDA-MB-361	Breast adenocarcinoma
11.	UACC-812	UACC-812	Ductal carcinoma

5.3 Data evaluation and assessment

'RStudio'[4], an integrated development environment (IDE) for the R programming language[23], was used as the platform for data analysis. Massive pharmacogenomics data comprising breast cancer cell lines and drugs derived from GDSC and CCLE were analyzed using the "PharmacoGx" package [24]. The "pharmacoGx" package analyzes pharmacological and molecular data and stores the generated information as R objects. In recent studies, drug-related data has become more prominent and has risen exponentially. Therefore, this package is an effective tool for exploring new research avenues for meta-analysis of pharmacogenomics data. Through this study, we want to explore the prevailing relationship between various breast cancer cell lines and their drug responses, correlating these diverse cell lines and their responses with the ScRNA-Seq analysis data to propose a potential drug candidate. The pharmacoset (Pset) function showcases the information regarding the analyzed data: annotations, molecular details, and treatment response data.

5.4 Modeling the Sensitivity Data

The 'drugDoseResponseCurve' function was used under the library of the 'pharmacoGx' package to plot the drug-dose response curve (DDRCs) for the 11 cell lines against the 125 drugs. In order to complete this function, statistics or parameters were obtained from PSet objects. Required parameters to execute the operation were: a list comprising PSets (GDSC or CCLE), drug name (the name of the Drug against which the cell line's response is desired), cell-line name, and the plot (actual/fitted/both) desired. Executing the command line comprising the mentioned parameters plots the DDRCs for a given cell-drug combination allowing quick visualization and comprehension of the result for each cell line.

5.5 Prediction of Drug Sensitivity

The next step after generating the DDRCs is shortlisting the drugs that showed direct responses against the cell lines. Cell lines were tested against drugs to observe the response to increasing drug concentrations, due to which their IC50 values were calculated using IC50_published or IC50_recomputed as parameters in the DDRCs function itself. The IC50 value of a drug represents the Drug's inhibitory concentration at which it effectively reduces the biological process by half. 37 drugs out of 125 were sorted for further evaluation.

5.6 Drug screening from single-cell RNA-Seq data

The data acquired, processed, and evaluated till now was subjected to Single-cell RNA sequencing data analysis to analyze the data in hand at the single-cell level. The analysis incorporates various software packages to perform the required tasks. 'Seurat' v4.3.0 [25] was used to carry out the ScRNA-Seq analysis, leading to the identification and characterization of cell populations and their gene expression patterns. Another package named 'ranger' v0.12.1 [26] helped us in feature selection, identification, and selection of essential genes in order to distinguish between cell populations. Following this, to perform differential expression analysis, two more packages were deployed: 'limma' v3.83.3 [27] and 'edgeR' v3.24.3 [28]. These mentioned packages are widely deployed and have been proven helpful in RNA sequencing analysis and determining the differential expression between cell populations based on statistical tests provided. The 'GSEABase' package v1.44.0 [29] was incorporated to perform gene set enrichment analysis, helping us identify the biological pathways differentially expressed between the cell populations. The last package to be utilized in this analysis was 'piano' v1.22.0 [30], which performed pathway-level analysis, thereby allowing us to identify the biological processes and signaling pathways that were either enriched or depleted in specific cell populations. The collective usage of the packages mentioned above provided a comprehensive and systematic approach to the ScRNA-Seq data analysis in identifying the biologically meaningful differences between the cell populations.

The single-cell data analysis was conducted using R packages, which included Seurat, to identify the cell types and then compare the transcriptional responses to drug perturbations. In the final stage, the in vitro drug screening was conducted using the cancer cell lines, and drug response data was integrated with the RNA-sequencing and single-cell RNA sequencing data to predict the drug responses computationally. Further investigation of the cell line composition and drug response heterogeneity was carried out using these data.

In this study, we used 37 different cancer cell lines. We generated 2D representations of the single-cell expression profiles using Seurat v4.3.0. The single-cell counts data underwent initial normalization and log transformation using the NormalizeData function. Subsequently, cell-wise normalization was performed using the ScaleData function. For UMAP [31] embeddings, we utilized the RunUMAP Seurat function, applying default parameters for "min.dist". The resulting two-dimensional UMAP representations enabled the analysis of individual cell

expression profiles within each cancer cell line, facilitating the identification of potential clusters exhibiting similar gene expression patterns. This approach offers valuable insights into the heterogeneity of cancer cell populations and holds the potential to guide the development of novel targeted therapies. For cell cycle analysis, the authors utilized the Seurat function CellCycleScoring, which employs gene lists specific to the S and G2M phases to classify cell cycle stages. The change in the proportion of cells in each phase between treatment and control conditions was estimated for each cell line using an R function. Aggregate scores, representing how each compound altered the composition of the cell cycle, were computed by calculating weighted averages across cell lines based on the change in the proportion of cells in each phase. The weights were determined by the measured drug sensitivity of the cell lines, bounded between 0 and 1. After this stage, 5 out of 37 drugs were sorted for further evaluation.

5.7 Similarities Search using Machine Learning

The Afatinib drug, sorted from 5 shortlisted drugs, was taken as the query molecule for similarities search. Afatinib was used for similarity search along with ERBB2 and EGFR as the targets. The ligand-based virtual screening analysis was performed using our script in Python against the ChEMBL[32] library of compounds.

5.8 Pre-processing of the bioactivity compounds

Data from the ChEMBL library of compounds was analyzed to filter out the bioactivity compounds for breast cancer. Initially, we converted the datatype of standard value from "object" to "float," then deleted entries with missing molecule structure entries and kept only entries with the standard unit (nM), and deleted duplicate molecules. We kept molecules with canonical SMILES and removed all molecules without canonical SMILES for further evaluations, then converted IC50 to pIC50 (log value) to allow IC50 to the negative logarithmic scale, which is essential -log10(IC50).

5.9 Labeling of active compounds with Lipinski's Rule of five

We labeled those compounds having values of less than 1000 nM (considered to be active) while those greater than 10,000 nM (considered to be inactive) and those values between 1,000 and 10,000 nM (referred to as intermediate) for further evaluation. Then we did Lipinski [33]

calculation to evaluate the drug-likeness of compounds. The drug-likeness is based on Absorption, Distribution, Metabolism, and Excretion (ADME), also known as the pharmacokinetic profile. Lipinski analyzed all orally active FDA-approved drugs [34] in the formulation of what is to be known as the Rule of Five, which stated the following: molecular weight should be < 500 Dalton, octanol-water partition coefficient (LogP) should be < 5, hydrogen bond donors should be < 5, and hydrogen bond acceptors < 10. The total DataFrame Shape for our study observed was (2127, 11). Upon normalization, the recalculated DataFrame shape was (1822,11). We found the total number of compounds in the unfiltered data set (1818), the total number of compounds in the filtered data set (1337), and the total number of compounds not compliant with the Rule of five (481). We also did PAINS (Pan Assay Interference Compounds) analysis [35] and obtained 89 compounds with pan assay interference compounds and 1248 without pan assay interference compounds.

5.10 Ligand-based screening using machine learning

We generated MACCS and Morgan fingerprints [36] for the Afatinib and then generated MACCS and Morgan fingerprints for all molecules in the processed dataset. In our findings, the number of found unwanted substructures was 89, and the number of compounds without unwanted substructures was 1248. We calculated the Experimental EF for the 5 % of the ranked dataset (tanimoto_maccs) to be 7.0% and the experimental EF for the 5% of the ranked dataset (tanimoto_morgan) to be 6.7%. Random EF for 5% of the ranked dataset was 5.0%, and optimal EF for 5% of the ranked dataset was 8.7%. Next, we calculated the Tanimoto similarity [37] between the query molecule (Afatinib) and all molecules in the processed dataset (using MACCS and Morgan fingerprints) and found 1337 compounds converted where fingerprint length per compound was 2048 and Tanimoto similarity (0.20) and distance matrix (0.80)

Then, we ran the clustering procedure for the entire dataset and found the total number of clusters was 243, where the number of clusters with only 1 compound was 125, the number of clusters with >5 compounds was 60, the number of clusters with >25 compounds was 8 and number of clusters with >100 compounds was 0.

We selected a total of 951 molecules, and the total number of compounds was 1818. First, we performed the step of data preparation or data labeling, where a column for activity with a pIC50 of > 9.0 was added, and we found 766 active compounds, while the number of inactive

compounds was 571. Molecule encoding was done using the MACCS method, and three classical machine learning approaches were applied to classify our molecules named Random Forest (RF) [7], Support Vector Machine (SVM) [8], and Artificial Neural Network (ANN) [9]. Then we performed the performance of models where we fit classical machine learning models on a train-test split of the data. Splitting the data was reused for the two other classical models. We used test (x) and train (x) for the respective fingerprint splitting and test (y) and train (y) for the respective label splits. Random forest classifier was applied where the set model parameter for random forest estimators was 100, number of trees to grow criterion (entropy), and number cost function to be optimized for a split. We observed sensitivity for RF, SVM, and ANN.

5.11 Docking Analysis

Structure-based virtual screening was performed through the AutoDock Vina44 [38]] tool compiled in PyRx16 [39]. PyRx is a software tool designed for computer-aided drug discovery (CADD) and virtual screening. It provides a user-friendly interface to facilitate the process of molecular docking, virtual screening, and molecular dynamics simulations. PyRx integrates various open-source tools and libraries, such as AutoDock, Vina, and Open Babel, to perform these tasks efficiently. With PyRx, researchers can explore the interactions between small molecules and target proteins, predict their binding affinity, and identify potential drug candidates. It has been widely used in both academic and industrial settings for drug discovery and optimization. We retrieved the Crystal structure of EGFR Kinase in complex with BIBW2992 (PDBID: 4G5J) [40] from Protein Data Bank [41]. The Complex PDBID 4G5J retrieved from the PDB database was cleaned of all other bonded atoms and saved as a separate molecule (macromolecule), similarly the ligand bound to the complex was removed from the total structure and saved as a separate molecule (ligand). The macromolecule and ligand were added in PyRx (this ligand served as control). The rest of the derivates were added and the grid box was built with center (x, y, z) = (60.31, 10.17, -23.24) and of dimensions (x, y, z) = (50, 50, 50). The docking simulation was then done with an exhaustiveness of 8. After completion of Docking, top hits were analyzed by PyMOL molecular visualizer [42] (The PyMOL Molecular Graphics System, Version 2.0 Schrödinger, LLC.) PyMOL is a powerful and versatile molecular visualization tool used for the analysis and representation of three-dimensional (3D) molecular structures. It enables researchers to visualize and manipulate complex biomolecules, such as proteins, nucleic acids, and small molecules, in an interactive

graphical environment. PyMOL provides a range of features, including molecular rendering, surface generation, and measurement tools, which aid in the understanding of molecular structures and their functions. It also supports scripting and customization, allowing users to automate tasks and create high-quality figures for presentations and publications. PyMOL is widely used in structural biology, biochemistry, and drug discovery research. 9 different orientations were observed for each derivative. The protein – ligand interaction was visualized in PyMOL and the ligand(s) fitting into the active site of the protein with best orientation, was shortlisted for further evaluation. The compound showing bond with the same residue as the control was taken up for MD simulations

5.12 Molecular Dynamics (MD) Simulation

Molecular Dynamics (MD) Simulation was performed using GROMACS [43] via LINUX OS. There are several stages to a Molecular Dynamics simulation.

- Generating Topology:
    - Protein topology was prepared with pdb2gmx
    - Ligand topology was prepared using external tools
    - CHARM27 force field was used which was obtained from Mackerall lab website[44].
    - CGenFF was accessed which is the official CHARMM force field server
    - Hydrogen atoms were added to the pre-retrieved protein complex PDBID: 4G5J
    - Topology was built with CGenFF
- Box and solvate were defined
- Addition of ions:
    - We then had a solvated system that contained a charged protein
- Energy Minimization was performed
- Equilibration
- Production MD
- Analysis

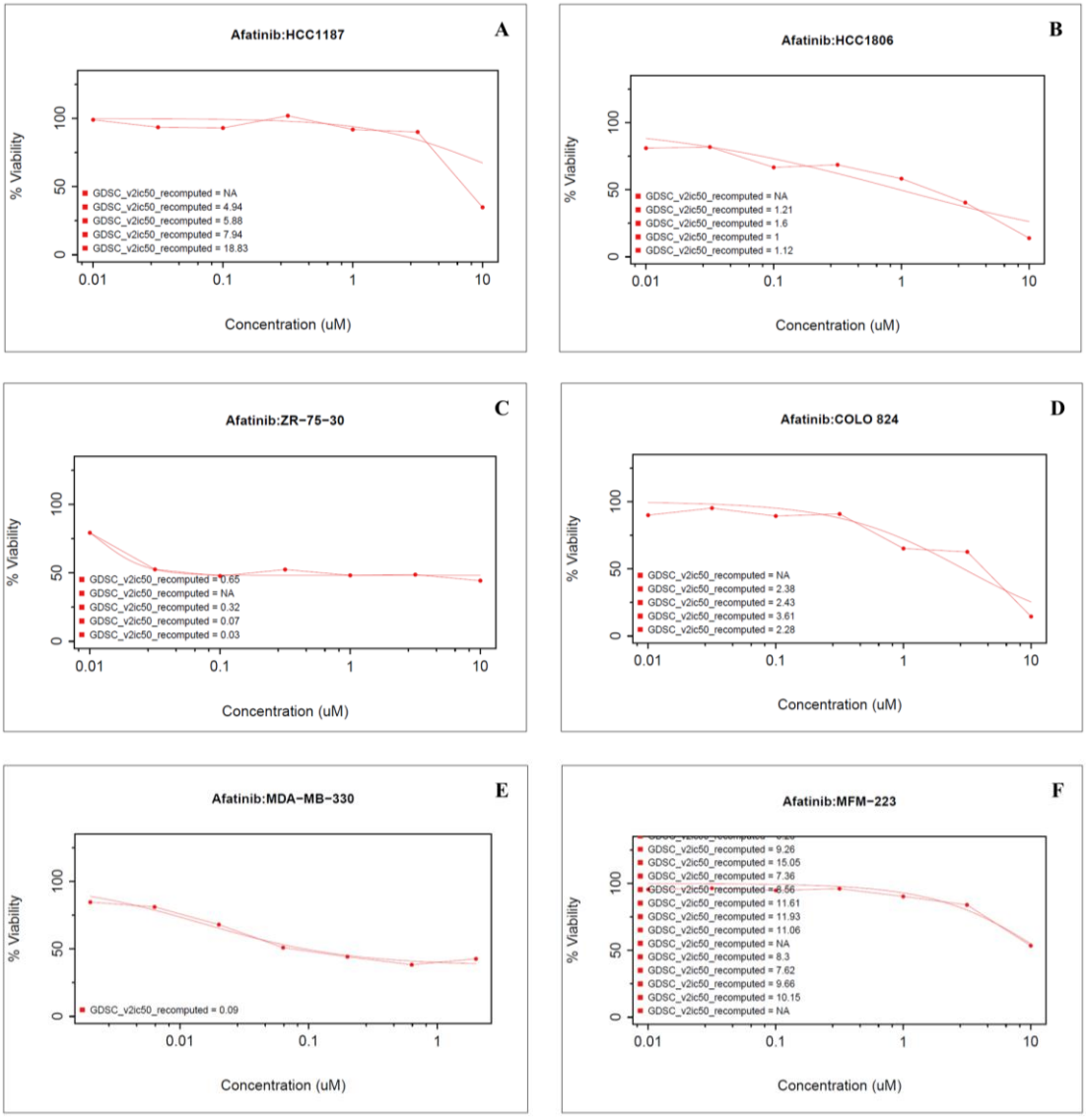
All the steps were followed to perform Molecular dynamics simulation of 1 control compound and 1 test compound. Minimization was done for 10 nano-seconds. Observations were recorded.

RESULTS

6 Results

Prediction of Drug Sensitivity

The PharmacoGx package was successfully deployed to record the drug sensitivity profiles against the breast cancer cell lines. The profiles were generated in the form of drug-dose response curves plotted for concentration (micromole) vs viability (%) with each graph showcasing the IC50_recomputed values. Figure 6. (A-K) contains the drug-dose response curves of the drug 'Afatinib' against the 11 breast cancer cell lines.



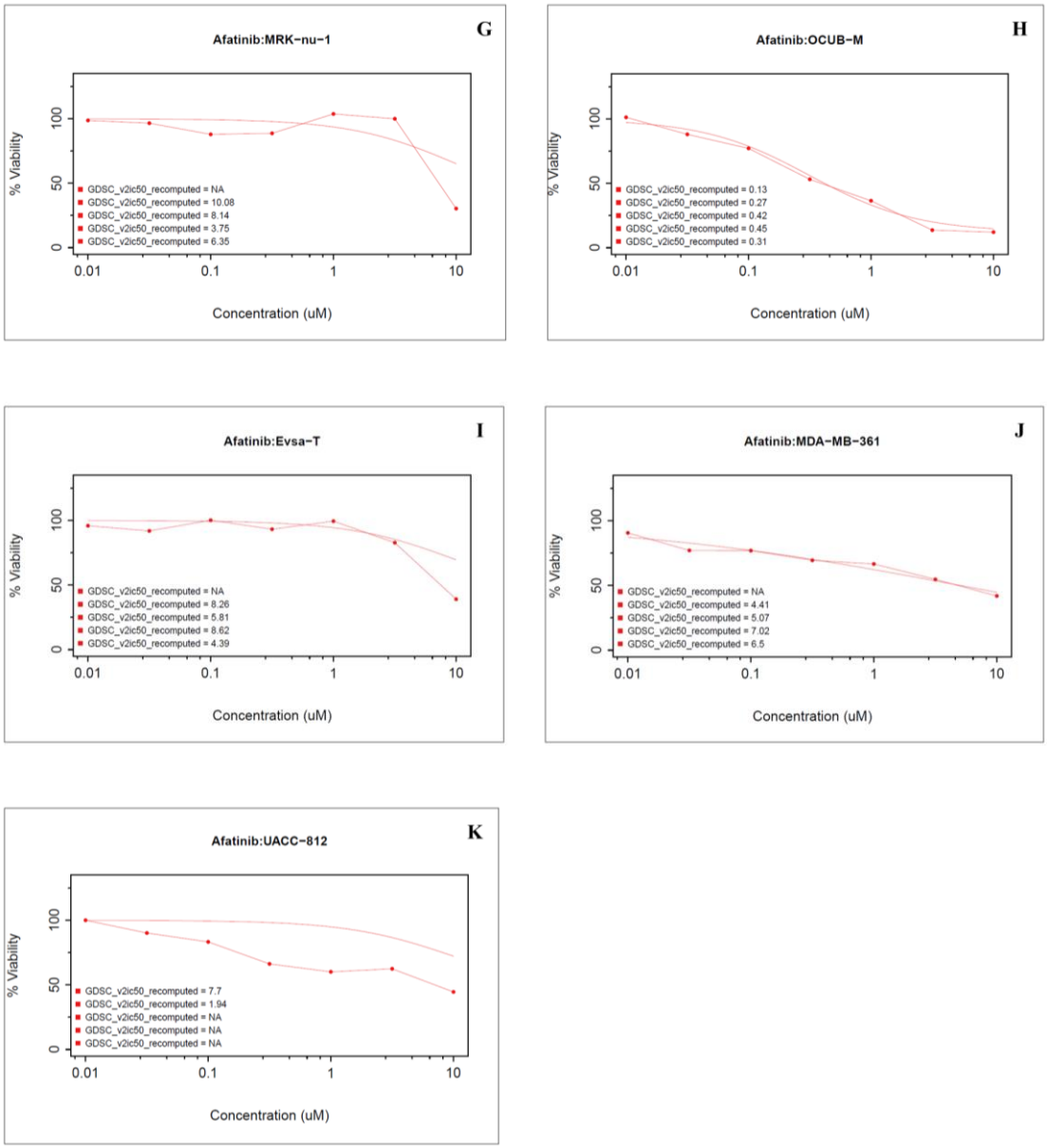


Figure 6.Drug-Dose-Response-Curves of Afatinib drug against 11 breast cancer cell lines (A) HCC1187, The Afatinib drug shows high potency and efficacy at initial concentration and maintains the same response throughout the concentration until 10 micromole where the decline can be seen. (B)HCC1806, the potency and efficacy at initial concentration is very positive however with increasing concentrations they both decline. (C)ZR-75-30, after the response at initial concentration the drug remains in a steady state with increasing concentrations. (D)COLO824, the drug shows decline in potency and efficacy for this cell line with increasing concentrations. (E)MDA-MB-330, After initial concentrations response, no increase in potency or efficacy is seen. (F)MFM-223, potency of the drug declines rapidly after concentration of 1micromole is administered. (G)MRK-nu-1, a fall in the potency after the initial concentration is interrupted after concentration of 1 micromole. (H)OCUB, Highest potency initially but rapid decline in potency with increasing concentrations. (I)Evsa-T, Graph of this cell line can be seen as fluctuating with sudden increase and decrease in potency as the concentration increases. (J) MDA-MB-361, the cell line shows negative response as with increasing of dosage the potency declines. (K)UACC-812, From initial concentration to final concentration a gradual decrease in potency can be seen.

Both the the red lines represent the biological activity of the drug against the respective cell line. The plain line signifies the actual response while the line joining the dots represents fitted

curve meaning the one fitted by logl logistic regression. The IC50 _recomputed values were recorded for each drug's response against each cell line. A record of the values are shown in the Table 2. The drug dose response curves for drugs other than Afatinib are shown in figures 13-16 under Annexure 10.2

Table 2. Record of drugs and the IC50_recomputed values

	Afatinib	**Bortezomib**	**Gemcitabine**	**Navitoclax**	**Trametinib**
HCC1187	18.83	0	NA	NA	NA
HCC1806	1.12	0.01	0	2.61	0.04
ZR-75-30	0.03	0.03	NA	12.63	NA
COLO 824	2.28	0	41.93	0.24	4.37
MDA-MB-330	0.09	NA	40.73	0.04	NA
MFM-223	NA	0	20307.79	7.55	11.94
MRK-nu-1	6.35	0.01	0.43	7.46	NA
OCUB-M	0.31	0	0.69	1.05	NA
Evsa-T	4.39	0	13.17	5.54	NA
MDA-MB-361	6.5	NA	NA	5.33	NA
UACC-812	1.94	0.01	78.56	2.66	NA

Drugs 'Afatinib' and 'Navitoclax' showed better IC50 _recomputed values compared to others

6.1 ScRNA-Seq Analysis

The ScRNA-Seq analysis results can be seen in the figures 7,8 and 9.

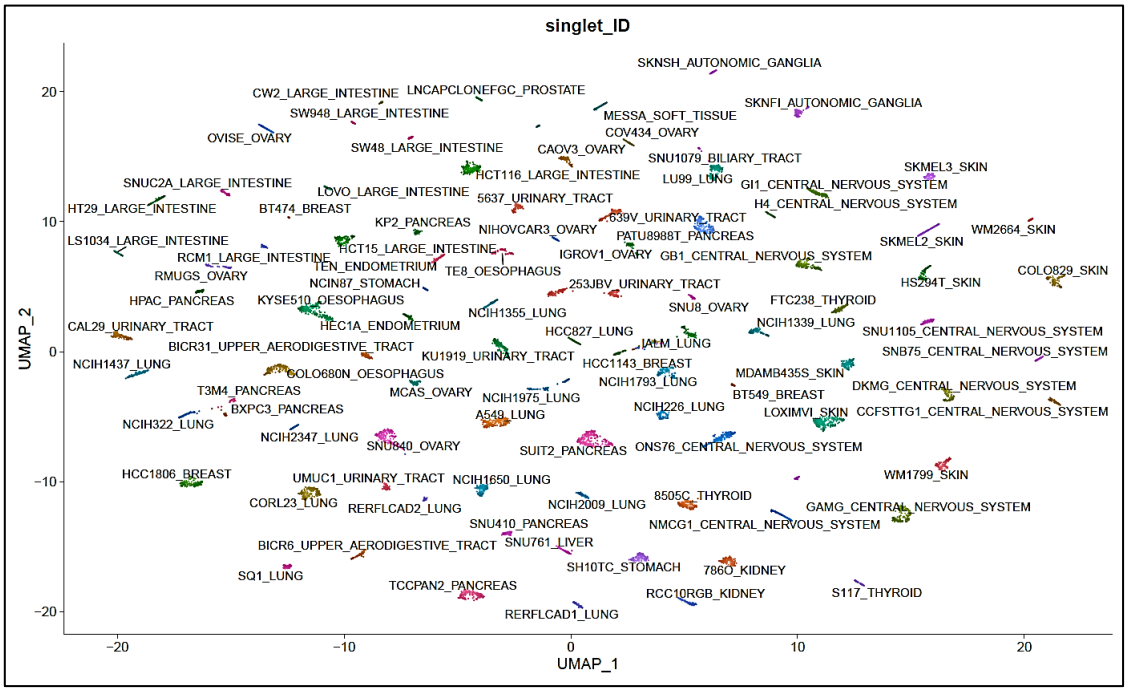


Figure 7. UMAP of Afatinib Drug (UMAP embedding visualized with color-coded 'singlet_ID' metadata column, representing the predicted cell line.)

EGFR is a crucial gene involved in maintaining genome integrity and preventing cancer formation; mutations in EGFR can lead to alterations in cellular pathways. These alterations can include disruption of the cell cycle, inhibition of apoptosis, and inhibition of cellular senescence.

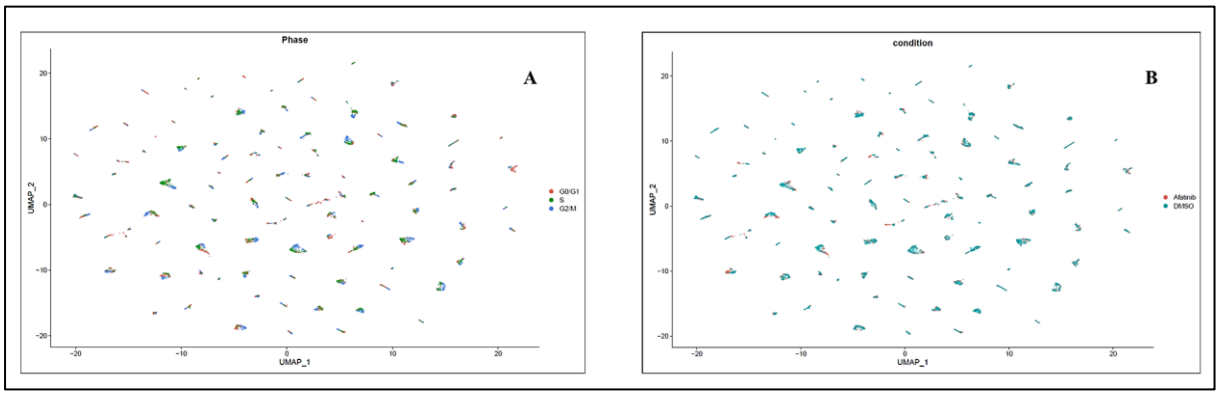


Figure 8. (A) Cell Cycle of Afatinib Drug (Cell-cycle phase distribution, categorized by cell line and EGFR status, depicted as a fraction of cells for each condition.)

(B) Condition of Afatinib Drug (UMAP visualization of cells treated with either DMSO control (green) or Afatinib (red) across a diverse set of cell lines.)

The drugs found from the ScRNA-Seq analysis data and common among the GDSC and CCLE pharmacosets were shortlisted for further use and are mentioned in the Table 3.

Table 3. List of Drugs shortlisted by ScRNA-Seq analysis data

S.no	Drugs	synonyms	Rplots	Target	Target_ Pathway	FDA approved
1.	Afatinib	BIBW2992, Tovok, Gilotrif	11/11	ERBB2, EGFR	EGFR signaling	FALSE
2.	Bortezomib	PS-341, LDP-341, Velcade	11/11	Proteasome	Protein stability and degradation	TRUE
3.	Gemcitabine	Gemzar, LY-188011///Gemzar, LY-188011///	11/11	Pyrimidine antimetabol-ite	DNA replication	TRUE
4.	Navitoclax	ABT-263, ABT263, ABT 263	11/11	BCL2, BCL-XL, BCL-W	Apoptosis regulation	FALSE
5.	Trametinib	GSK1120212, Mekinist	11/11	MEK1, MEK2	ERK MAPK signaling	FALSE

6.2 Similarities search using Machine Learning

Similarity search was done to retrieve Afatinib derivates for target Erb-B2 Receptor Tyrosine Kinase 2 (ERBB2), Epidermal Growth Factor Receptor (EGFR).

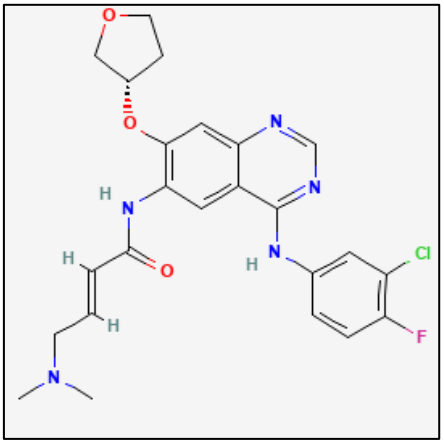


Figure 9. Chemical structure of Afatinib

We conducted a comprehensive analysis to determine the similarity between the query molecule Afatinib and all molecules in the dataset using MACCS and Morgan fingerprints. We calculated the Dice similarity for these comparisons and analyzed Tanimoto and Dice similarities for the two fingerprint types.

We retrieved the most similar molecules to Afatinib based on the obtained similarity values. Additionally, we gathered information on the bioactivities associated with these molecules to assess their potential relevance. The results of this analysis are presented in the figure below, where we plotted enrichment plots. We set a pIC50 (log p-value) cutoff of 9.0 to discriminate between active and inactive molecules. Furthermore, we evaluated the enrichment factors (EF) for the MACCS and Morgan fingerprints using Tanimoto similarity elaborated in the supplementary data

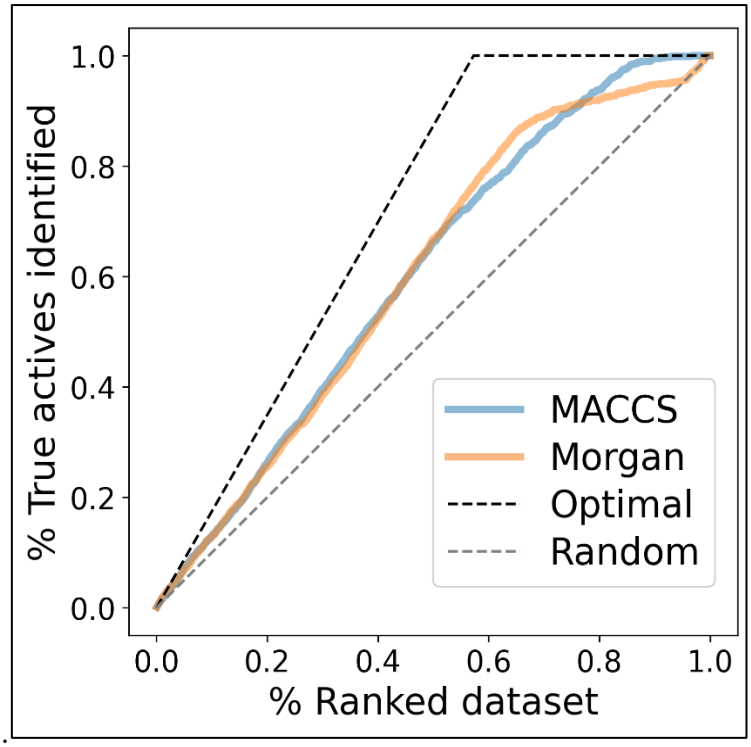


Figure 10. Enrichment graph obtained from MACCS and MORGAN methods

6.3 Ligand-based screening using machine learning

In our study focused on ERBB2 and Breast cancer, we employed three different classification models: Random Forest (RF), Support Vector Machine (SVM), and Artificial Neural Network (ANN). The RF model exhibited a sensitivity of **0.95**, a specificity of **0.77**, and an AUC of **0.93**. For the SVM model, we used an RBF kernel with a C value of 1 and a gamma value of 0.1 and set the probability to True. The SVM model achieved a sensitivity of **0.94**, a specificity of **0.75**, and an AUC of **0.92**.

We also utilized an ANN classifier, configuring it with hidden layer sizes of 5 and 3 and a random state of SEED. The ANN model demonstrated a sensitivity of **0.89**, a specificity of **0.75**, and an AUC of **0.91**

These results indicate that all three models exhibited strong performance in classifying compounds related to ERBB2 and Breast cancer. The RF and SVM models demonstrated high sensitivity and specificity and excellent AUC values. The ANN model also achieved a notable sensitivity, albeit with a slightly lower specificity. The cross-validation experiments further

validated the performance of the models, providing additional confidence in their predictive capabilities when using the Morgan fingerprint encoding approach.

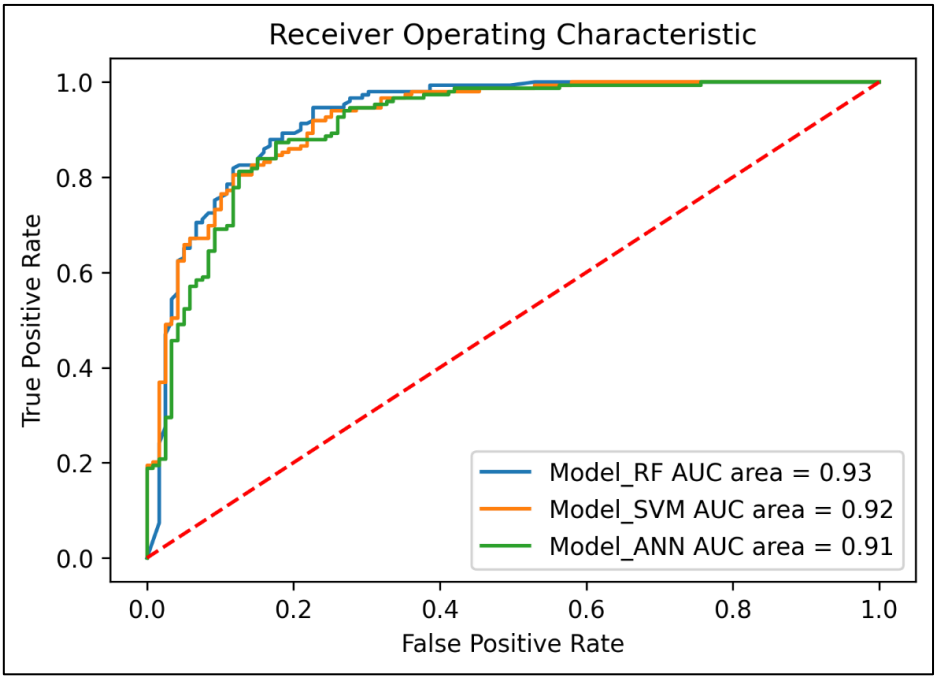
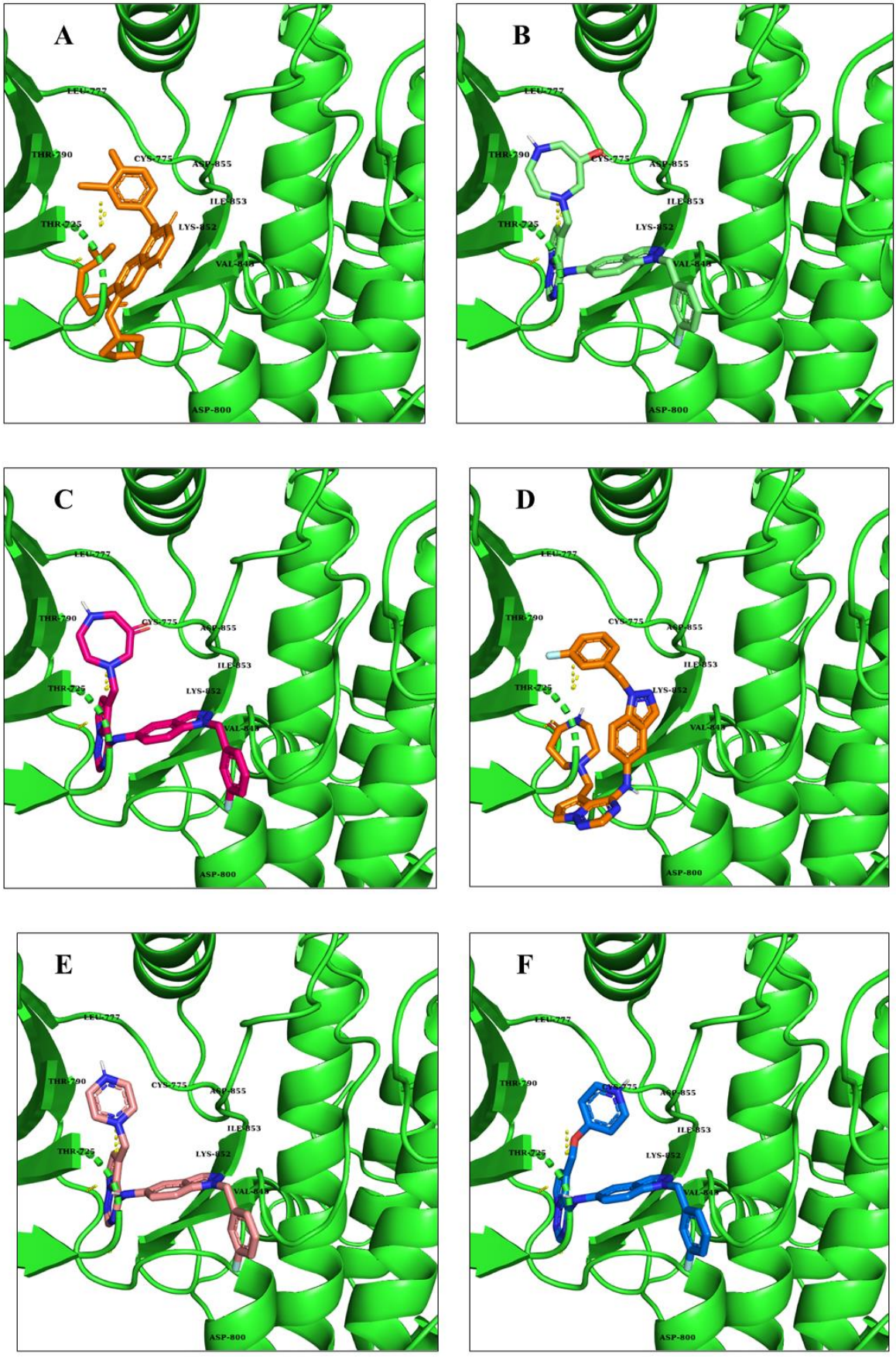


Figure 11. The ROC plot for true positive rate vs false positive rate provided the results

6.4 Docking Analysis

Docking was conducted using PyRx. Visualization was done using PyMOL. The initially retrieved complex PDB ID 4G5J was cleaned and protein (control) and ligand (control) were derived from it. PyMOL was used majorly to visualize the interactions, to locate the active site, find the number of bonded amino acid residues with the control as well as the test compounds. 10 orientations of protein-ligand complexes (including control) were shortlisted based on the docking scores (shown in figure 12, listed in Table 4).


A
LEU-777
THR-790
CYS-775
ASP-855
ILE-853
THR-725
LYS-852
VAL-848
ASP-800
B
LEU-777
THR-790
CYS-775
ASP-855
ILE-853
THR-725
LYS-852
VAL-848
ASP-800
C
LEU-777
THR-790
CYS-775
ASP-855
ILE-853
THR-725
LYS-852
VAL-848
ASP-800
D
LEU-777
THR-790
CYS-775
ASP-855
ILE-853
THR-725
LYS-852
VAL-848
ASP-800
E
LEU-777
THR-790
CYS-775
ASP-855
ILE-853
THR-725
LYS-852
VAL-848
ASP-800
F
LEU-777
THR-790
CYS-775
ASP-855
ILE-853
THR-725
LYS-852
VAL-848
ASP-800

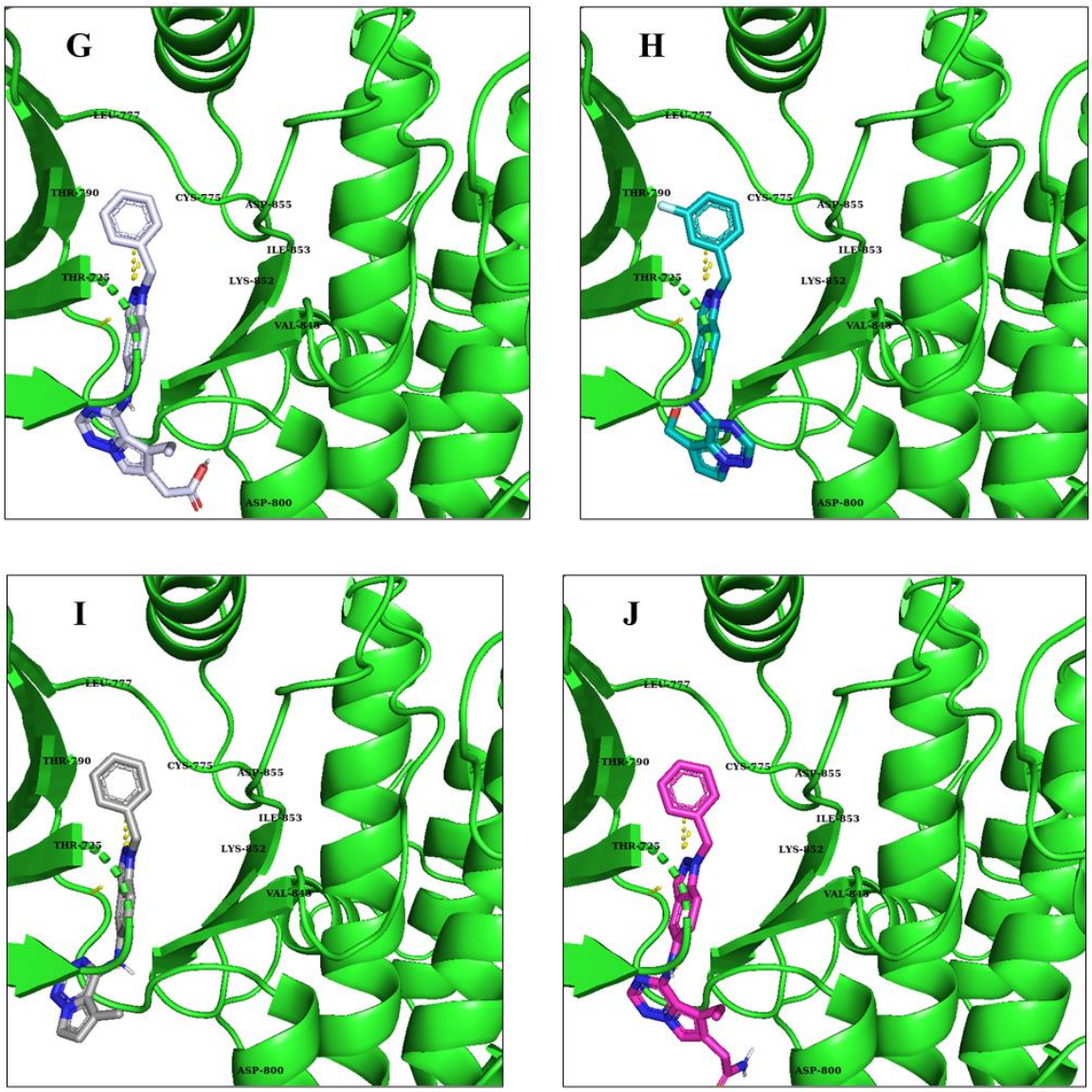


Figure 12. Docking configurations (A) Control compound: Afatinib (control) The green coloured macromolecule depicted in the figure is the protein structure. The Orange coloured small molecule is the ligand docked in the active site of protein, bonded to an amino acid residue ASP-855. **(B) ChEMBL233325** This figure depicts the Protein (Green) and the derivate ligand (lime-bluish) docked in the active site of protein, creating bond with an amino acid residue ASN-842. **(C) ChEMBL234580** Image shows Protein (Green) and another derivate ligand (pink) docked into the active site of protein, showing bonding with amino acid residues THR-854, ASN-842. **(D) ChEMBL233324** The shown figure depicts Protein structure (green) and 3rd derivate ligand (orange) docked into the active site of the protein, showing bonds with amino acid residues ASP-855, MET-793, LEU-718. **(E) ChEMBL231875** In the figure is seen protein (green) and 4th top derivative ligand (bluish-pink) docked into the active site of the protein, with bonds with amino acid residues MET-793, ASN-842. **(F) ChEMBL245869** Shown in the figure are protein (green) and 5th derivate ligand (blue) docked into the active site of protein, bonded with amino acid residues MET-793, ASN-842, GLU-762. **(G) ChEMBL372692** Visible in the figure are protein (green) and 6th derivate ligand (white) docked into the active site of protein, bonded with amino acidic residues PRO-794. **(H) ChEMBL246276** In the figure are protein (green) and 7th derivate ligand (blue) docked into the active site of protein, bonded with amino acidic residue ASP-800. **(I) ChEMBL429827** In the figure are protein (green) and 8th derivate ligand (white) docked into the active site of the protein, bonded with amino acidic residue MET-793. **(J) ChEMBL194160** Seen in the figure are protein (green) and 9th derivate ligand (magenta) docked into the active site of the protein, bonded with amino acid residues PRO-794, PHE-795

Table 4. Docking results of Afatinib Drug

S.no	Compounds (Results)	Docking (Score)	Interactions
			H-Bonding
1.	Afatinib (control)	-7.4	**ASP-855**
2.	ChEMBL233325	-9.4	ASN-842
3.	ChEMBL234580	-9.3	THR-854, ASN-842
4.	ChEMBL233324	-9.2	**ASP-855**, MET-793, LEU-718
5.	ChEMBL231875	-9.1	MET-793, ASN-842
6.	ChEMBL245869	-9.1	MET-793, ASN-842, GLU-762
7.	ChEMBL372692	-8.9	PRO-794
8.	ChEMBL246276	-8.7	ASP-800
9.	ChEMBL429827	-8.7	MET-793
10.	ChEMBL194160	-8.6	PRO-794, PHE-795

The Molecular Docking results provided us with compounds better than our control compound, however, the 9 compounds in Table 4 (from 2-10) were shortlisted as the best compounds with excellent docking scores. The control compound (Afatinib control) generated a docking score of -7.4 and showed presence of amino acid ASP-855 in the H-Bonding. The Test compound (ChEMBL233324) generated a higher docking score of -9.2 with ASP-855, MET-793 and LEU-718 present as amino acid residues in the H-bonding.

6.5 MD ANALYSIS

Following graphs were generated upon completion of the Molecular Dynamics (MD) Simulations.

- GYRATE
- RMSD LIGAND
- RMSD PROTEIN
- RMSF
- H-BONDS
- SASA

GYRATE

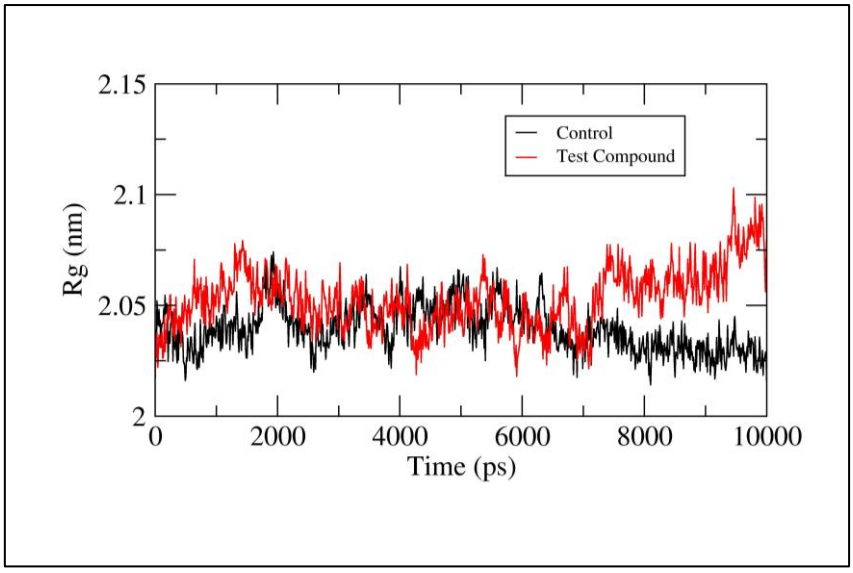


Figure 13. GYRATE GRAPH

The Rg graph signifies the compactness of the protein and was determined for protein backbone. The Rg of complex was found to be ranging between 2nm -2.15nm respectively. Fluctuation was observed after 7000 ps.

RMSD LIGAND

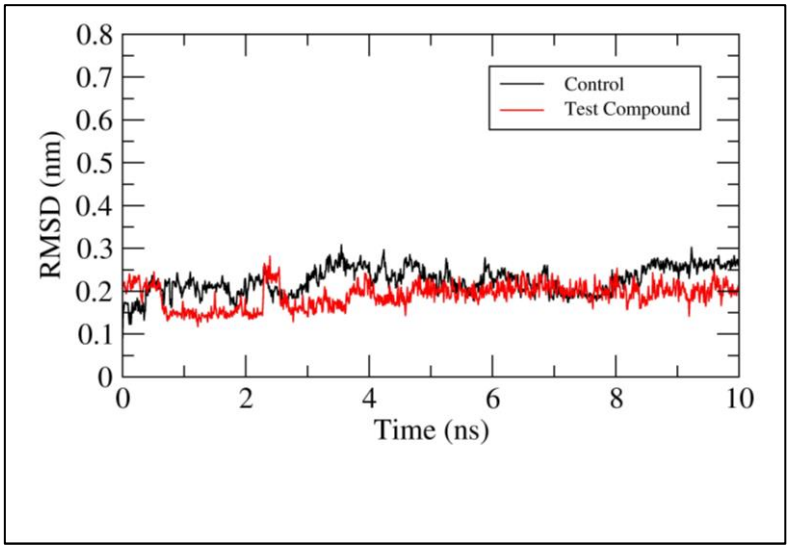


Figure 14. RMSD LIGAND GRAPH.

The combined rmsd of protein-ligand complex during simulation remained stable throughout the simulation with a time frame of 10ns. Fluctuation observed initially at 2.2 ns in the Test Compound compared to the Control compound is 'minor' or miniscule and hence, can be neglected. The average rmsd was completed from range 0 nm – 0.8 nm.

RMSD PROTEIN

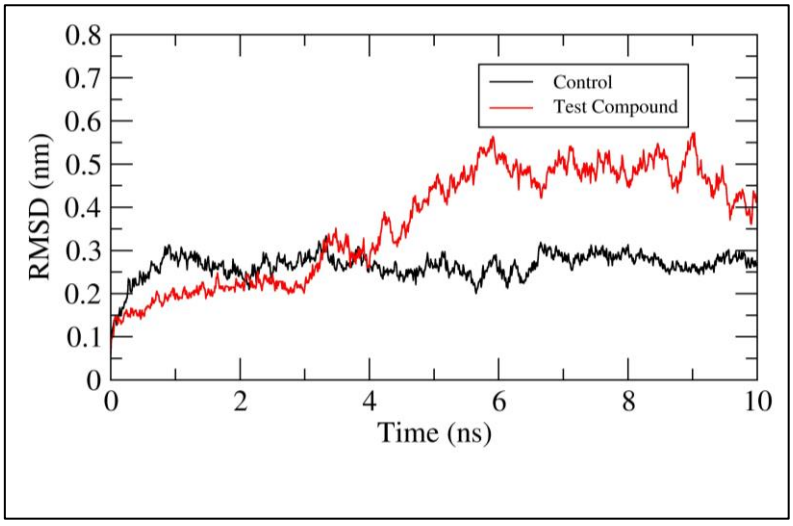


Figure 15. RMSD PROTEIN GRAPH

The combined rmsd of protein-ligand complex initially remained stable from 0 ns - 4 ns, interestingly fluctuation was observed till 9 ns. After 9 ns the rmsd was seen to be regaining stability. The average rmsd was completed from range 0nm – 0.8 nm.

RMSF

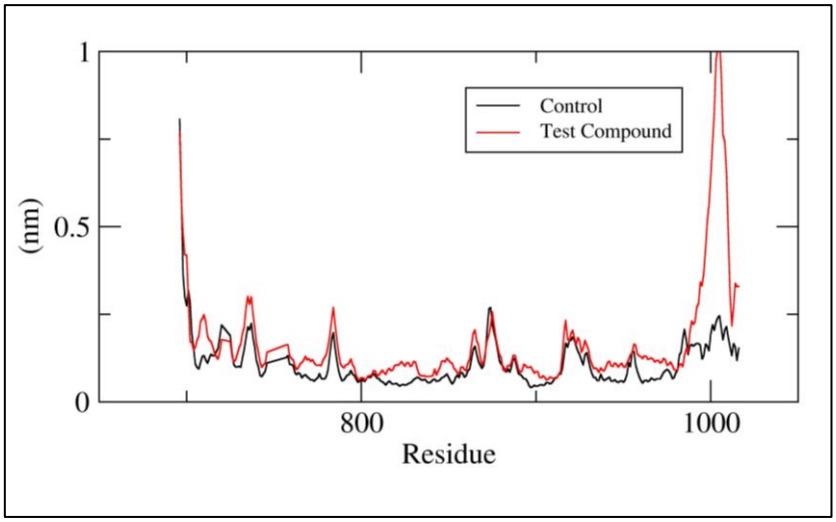


Figure 16. Root Mean Square Fluctuation Graph

RMSF graph demonstrates the residue of protein backbone fluctuations. The residue was found to be stable until 950 residues. After that fluctuation was observed between the residues 950-1000. The fluctuating residues observed out of the pocket of active site.

H-BOND

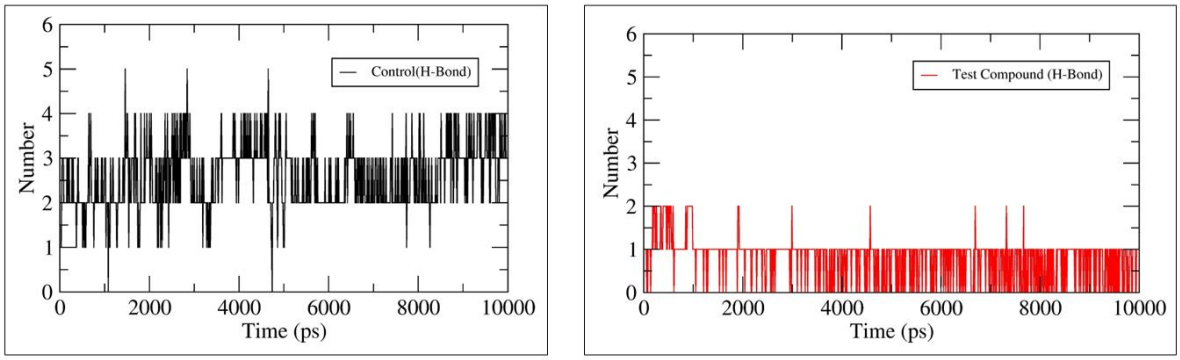
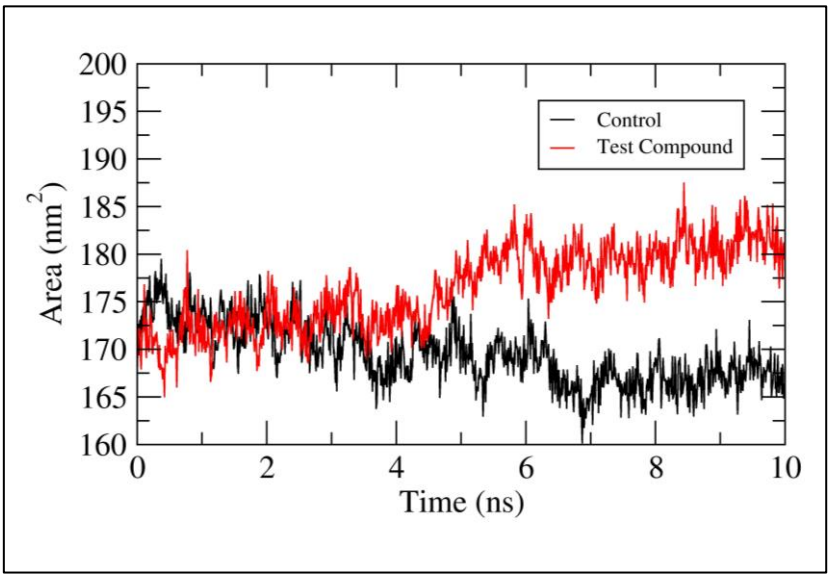


Figure 17. H-Bond. Control compound (Black), Test Compound (Red)

Throughout the MD Simulation studies the H-Bonds were notably formed. The average range was 0 – 6. The number of H-Bonds formed in control compound was 3 and whereas in the Test compound 1 H-Bond was formed.

SASA (Solvent accessible surface area)



The SASA graph represents the compactness of the complex. The variation of the protein-ligand complex as shown by SASA during 10 ns of simulation. The colour black represents control while the red colour represents the Test compound. The compactness of the protein-ligand complex was initially stable till 4ns. From range between 5 ns – 9.5 ns fluctuations in the test compound can be seen.

DISCUSSIONS

7 Discussion

Initially, Breast cancer data was pooled from two large datasets, Cancer Cell Encyclopedia (CCLE) and Genomics of Drug Sensitivity in Cancer (GDSC), in the form of large pharmacists. The breast cancer cell lines and the data acquired from the datasets were combined to avoid overlapping and tested for drug sensitivity using a drug-dose response curve. This enabled us to plot and compare the drug-response plots.

There were 125 drugs for breast cancer among the 190 drugs in the datasets. After shortlisting them for breast cancer, we plotted the drug-dose response curves for 125 drugs against the retrieved 11 breast cancer cell lines.

We shortlisted those drugs for further evaluation that showed sensitivity for all eleven cell lines, and any that showed less than 11 (<11) were neglected and not taken up for further evaluation. Using this method as a means of filtering, we were able to shortlist 37 drugs out of the 125 breast cancer drugs that showed response for all eleven cell lines.

ScRNA-Seq analysis further revealed five drugs that overlapped with our drug-sensitivity data. The five drugs, Afatinib, Bortezomib, Gemcitabine, Navitoclax, and Trametinib, were selected for further evaluation of the 37 drugs. The Drug Afatinib was further evaluated based on its IC50 value and consistency across the seven cell lines.

The target of the Drug Afatinib was identified from the datasets and further taken up in machine learning, performed by a Python script of our own; the system was trained using the training data set to provide possible predictions of similar molecules in the ChEMBL database. The information generated from machine learning was used in Docking, where the PDB ID – 4G5J ("crystal structure of **EGFR** kinase in complex with BIBW2992) was used as a reference.

Docking was performed via PyRx and visualized through PyMOL using standard principles to predict the possible interactions of our target.The docking score for our control 'Afatinib' was **-7.4**, and that of our top resultant compound, ChEMBL233325, was **-9.4.** Despite the higher docking scores than the control compound, only 1 test compound showed presence of similar amino acid residue with that of the control compound. ChEMBL233324 compound and control consisted of ASP-855. The other 4 compounds of our top 5 had excellent docking scores but

lacked the common residue, so instead of taking the usual approach and shortlisting 5 compounds for MD simulations, we just shortlisted 1 compound as our test compound, namely ChEMBL233324 and took it further for Molecular Dynamics (MD) Simulations.

MD Simulations were performed for 10 ns using GROMACS and the results were recorded. In the MD simulation, we have found that the test-compound has shown almost similar pattern of fluctuation in the trajectories, which reflects that we can carry forward the test compound for further *in vitro*/*in vivo* validation.

CONCLUSION

8 Conclusion

In conclusion, our study aimed to identify potential drugs for breast cancer by analyzing large datasets and employing a series of computational methods. Initially, we combined breast cancer data from the CCLE and GDSC datasets, avoiding overlapping, and conducted drug sensitivity testing using a drug-dose response curve. This allowed us to plot and compare drug-response plots for 125 breast cancer drugs against 11 breast cancer cell lines. Through this filtering process, we identified 37 drugs that exhibited sensitivity across all eleven cell lines.

Subsequently, ScRNA-Seq analysis further narrowed down our selection to five drugs that overlapped with our drug-sensitivity data. These drugs, namely Afatinib, Bortezomib, Gemcitabine, Navitoclax, and Trametinib, were chosen for further evaluation. Afatinib was assessed based on its IC50 value and consistency across seven cell lines. To gain insights into the target of Afatinib, machine learning was employed using a Python script trained with the available data, generating predictions for similar molecules in the ChEMBL database.

To investigate the possible interactions of our target, we performed docking using PyRx and visualized the results on PyMOL. The docking score for our control compound, Afatinib, was -7.4, while our resultant compound, ChEMBL233325, obtained a docking score of -9.4. These findings indicated that ChEMBL233325 displayed a stronger predicted interaction with the target compared to Afatinib. However, the test compound ChEMBL233324 with a docking score of -9.2 and having the same amino acid residue as control compound was taken up for MD Simulations. The MD simulations revealed similar patterns in the fluctuation of trajectories in the Test Compound and therefore can be further taken ahead for *in vitro/ in vivo* evaluation.

In summary, our comprehensive approach encompassed data analysis, drug sensitivity testing, machine learning, docking, and MD simulations to identify potential drugs for breast cancer. The results obtained thus far provide a promising starting point for further investigation and potential development of novel therapeutic candidates in the fight against breast cancer.

9 Plagiarism Report

Sarvbhaum Shukla 302101023

ORIGINALITY REPORT

18%	12%	12%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	James M. McFarland, Brenton R. Paolella, Allison Warren, Kathryn Geiger-Schuller et al. "Multiplexed single-cell profiling of post-perturbation transcriptional responses to define cancer vulnerabilities and therapeutic mechanism of action", Cold Spring Harbor Laboratory, 2019 Publication	1%
2	link.springer.com Internet Source	1%
3	www.tandfonline.com Internet Source	<1%
4	Aamir Mehmood, Sadia Nawab, Yifan Jin, Hesham Hassan, Aman Chandra Kaushik, Dong-Qing Wei. "Ranking Breast Cancer Drugs and Biomarkers Identification Using Machine Learning and Pharmacogenomics", ACS Pharmacology & Translational Science, 2023 Publication	<1%
5	Submitted to Al Akhawayn University in Ifrane Student Paper	<1%

Dr. M.I. Siddiqi
Senior Principal Scientist
Biochemistry and Structural Biology Division
C.S.I.R. Central Drug Research Institute
Lucknow- 226031

BIBLIOGRAPHY

10 Bibliography

1. Khan, A., Rehman, Z., Hashmi, H. F., Khan, A. A., Junaid, M., Sayaf, A. M., Ali, S. S., Hassan, F. U., Heng, W., & Wei, D.-Q. (2020). An Integrated Systems Biology and Network-Based Approaches to Identify Novel Biomarkers in Breast Cancer Cell Lines Using Gene Expression Data. *Interdisciplinary Sciences, Computational Life Sciences*, *12*(2), 155–168. https://doi.org/10.1007/s12539-020-00360-0

2. Yang, W., Lightfoot, H., Bignell, G., Behan, F., Cokelear, T., Haber, D., Engelman, J., Stratton, M., Benes, C., McDermott, U., & Garnett, M. (2016). Genomics of Drug Sensitivity in Cancer (GDSC): A resource for biomarker discovery in cancer cells. *European Journal of Cancer*, *S1*(69), S82. https://doi.org/10.1016/S0959-8049(16)32839-8

3. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., … Garraway, L. A. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, *483*(7391), Article 7391. https://doi.org/10.1038/nature11003

4. Loraine, A. E., Blakley, I. C., Jagadeesan, S., Harper, J., Miller, G., & Firon, N. (2015). Analysis and Visualization of RNA-Seq Expression Data Using RStudio, Bioconductor, and Integrated Genome Browser. In J. M. Alonso & A. N. Stepanova (Eds.), *Plant Functional Genomics: Methods and Protocols* (pp. 481–501). Springer. https://doi.org/10.1007/978-1-4939-2444-8_24

5. Chen, G., Ning, B., & Shi, T. (2019). Single-Cell RNA-Seq Technologies and Related Computational Data Analysis. *Frontiers in Genetics*, *10*, 317. https://doi.org/10.3389/fgene.2019.00317

6. Thind, A. S., Monga, I., Thakur, P. K., Kumari, P., Dindhoria, K., Krzak, M., Ranson, M., & Ashford, B. (2021). Demystifying emerging bulk RNA-Seq applications: The application and utility of bioinformatic methodology. *Briefings in Bioinformatics*, *22*(6), bbab259. https://doi.org/10.1093/bib/bbab259

7. Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

8. Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

9. Kouamo, S., & Tangha, C. (2016). Fingerprint Recognition with Artificial Neural Networks: Application to E-Learning. *Journal of Intelligent Learning Systems and Applications*, *08*(02), 39–49. https://doi.org/10.4236/jilsa.2016.82004

10. Morris, G. M., & Lim-Wilby, M. (2008). Molecular Docking. In A. Kukol (Ed.), *Molecular Modeling of Proteins* (pp. 365–382). Humana Press. https://doi.org/10.1007/978-1-59745-177-2_19

11. Hansson, T., Oostenbrink, C., & van Gunsteren, W. (2002). Molecular dynamics simulations. *Current Opinion in Structural Biology*, *12*(2), 190–196. https://doi.org/10.1016/S0959-440X(02)00308-1

12. Wu, Z., Lawrence, P. J., Ma, A., Zhu, J., Xu, D., & Ma, Q. (2020). Single-Cell Techniques and Deep Learning in Predicting Drug Response. *Trends in Pharmacological Sciences*, *41*(12), 1050–1065. https://doi.org/10.1016/j.tips.2020.10.004

13. Liu, C., Wei, D., Xiang, J., Ren, F., Huang, L., Lang, J., Tian, G., Li, Y., & Yang, J. (2020). An Improved Anticancer Drug-Response Prediction Based on an Ensemble Method Integrating Matrix Completion and Ridge Regression. *Molecular Therapy. Nucleic Acids*, *21*, 676–686. https://doi.org/10.1016/j.omtn.2020.07.003

14. Chung, W., Eum, H. H., Lee, H.-O., Lee, K.-M., Lee, H.-B., Kim, K.-T., Ryu, H. S., Kim, S., Lee, J. E., Park, Y. H., Kan, Z., Han, W., & Park, W.-Y. (2017). Single-cell RNA-seq enables comprehensive tumor and immune cell profiling in primary breast cancer. *Nature Communications*, *8*, 15081. https://doi.org/10.1038/ncomms15081

15. Azizi, E., Carr, A. J., Plitas, G., Cornish, A. E., Konopacki, C., Prabhakaran, S., Nainys, J., Wu, K., Kiseliovas, V., Setty, M., Choi, K., Fromme, R. M., Dao, P., McKenney, P. T., Wasti, R. C., Kadaveru, K., Mazutis, L., Rudensky, A. Y., & Pe'er, D. (2018). Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell*, *174*(5), 1293-1308.e36. https://doi.org/10.1016/j.cell.2018.05.060

16. Xing, K., Zhang, B., Wang, Z., Zhang, Y., Chai, T., Geng, J., Qin, X., Chen, X. S., Zhang, X., & Xu, C. (2023). Systemically Identifying Triple-Negative Breast Cancer Subtype-Specific Prognosis Signatures, Based on Single-Cell RNA-Seq Data. *Cells*, *12*(3), 367. https://doi.org/10.3390/cells12030367

17. Torroja, C., & Sanchez-Cabo, F. (2019). Digitaldlsorter: Deep-Learning on scRNA-Seq to Deconvolute Gene Expression Data. *Frontiers in Genetics*, *10*, 978. https://doi.org/10.3389/fgene.2019.00978

18. Cava, C., Bertoli, G., & Castiglioni, I. (2018). In silico identification of drug target pathways in breast cancer subtypes using pathway cross-talk inhibition. *Journal of Translational Medicine*, *16*(1), 154. https://doi.org/10.1186/s12967-018-1535-2

19. Uchida, S., & Sugino, T. (2022). In Silico Identification of Genes Associated with Breast Cancer Progression and Prognosis and Novel Therapeutic Targets. *Biomedicines*, *10*(11), 2995. https://doi.org/10.3390/biomedicines10112995

20. Abdelbasset, W. K., Elsayed, S. H., Alshehri, S., Huwaimel, B., Alobaida, A., Alsubaiyel, A. M., Alqahtani, A. A., El Hamd, M. A., Venkatesan, K., AboRas, K. M., & Abourehab, M. A. S. (2022). Development of GBRT Model as a Novel and Robust Mathematical Model to Predict and Optimize the Solubility of Decitabine as an Anti-Cancer Drug. *Molecules (Basel, Switzerland)*, *27*(17), 5676. https://doi.org/10.3390/molecules27175676

21. Qi, R., & Zou, Q. (2023). Trends and Potential of Machine Learning and Deep Learning in Drug Study at Single-Cell Level. *Research (Washington, D.C.)*, *6*, 0050. https://doi.org/10.34133/research.0050

22. Zhang, X., Marjani, S. L., Hu, Z., Weissman, S. M., Pan, X., & Wu, S. (2016). Single-Cell Sequencing for Precise Cancer Research: Progress and Prospects. *Cancer Research*, *76*(6), 1305–1312. https://doi.org/10.1158/0008-5472.CAN-15-1907

23. Cribari-Neto, F., & Zarkos, S. G. (1999). R: Yet another econometric programming environment. *Journal of Applied Econometrics*, *14*(3), 319–329. https://doi.org/10.1002/(SICI)1099-1255(199905/06)14:3<319::AID-JAE533>3.0.CO;2-Q

24. Smirnov, P., Safikhani, Z., El-Hachem, N., Wang, D., She, A., Olsen, C., Freeman, M., Selby, H., Gendoo, D. M. A., Grossmann, P., Beck, A. H., Aerts, H. J. W. L., Lupien, M., Goldenberg, A., & Haibe-Kains, B. (2016). PharmacoGx: An R package for analysis of large pharmacogenomic datasets. *Bioinformatics*, *32*(8), 1244–1246. https://doi.org/10.1093/bioinformatics/btv723

25. Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. *Nature Biotechnology*, *33*(5), Article 5. https://doi.org/10.1038/nbt.3192

26. Wright, M. N., & Ziegler, A. (2017). ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *Journal of Statistical Software*, *77*(1). https://doi.org/10.18637/jss.v077.i01

27. Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., & Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, *43*(7), e47. https://doi.org/10.1093/nar/gkv007

28. Robinson, M. D., McCarthy, D. J., & Smyth, G. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, *26*(1), 139–140. https://doi.org/10.1093/bioinformatics/btp616

29. Innis, S. E., Reinaltt, K., Civelek, M., & Anderson, W. D. (2021). GSEAplot: A Package for Customizing Gene Set Enrichment Analysis in R. *Journal of Computational Biology*, *28*(6), 629–631. https://doi.org/10.1089/cmb.2020.0426

30. Väremo, L., Nielsen, J., & Nookaew, I. (2013). Enriching the gene set analysis of genome-wide data by incorporating directionality of gene expression and combining statistical hypotheses and methods. *Nucleic Acids Research*, *41*(8), 4378–4391. https://doi.org/10.1093/nar/gkt111

31. McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (arXiv:1802.03426). arXiv. https://doi.org/10.48550/arXiv.1802.03426

32. Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., & Overington, J. P. (2012).

ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, *40*(D1), D1100–D1107. https://doi.org/10.1093/nar/gkr777

33. Pollastri, M. P. (2010). Overview on the Rule of Five. *Current Protocols in Pharmacology*, *49*(1), 9.12.1-9.12.8. https://doi.org/10.1002/0471141755.ph0912s49

34. Ciociola, A. A., Cohen, L. B., Kulkarni, P., Kefalas, C., Buchman, A., Burke, C., Cain, T., Connor, J., Ehrenpreis, E. D., Fang, J., Fass, R., Karlstadt, R., Pambianco, D., Phillips, J., Pochapin, M., Pockros, P., Schoenfeld, P., Vuppalanchi, R., & Gastroenterology, the F.-R. M. C. of the A. C. of. (2014). How Drugs are Developed and Approved by the FDA: Current Process and Future Directions. *Official Journal of the American College of Gastroenterology | ACG*, *109*(5), 620. https://doi.org/10.1038/ajg.2013.407

35. Pouliot, M., & Jeanmart, S. (2016). Pan Assay Interference Compounds (PAINS) and Other Promiscuous Compounds in Antifungal Research. *Journal of Medicinal Chemistry*, *59*(2), 497–503. https://doi.org/10.1021/acs.jmedchem.5b00361

36. Zhang, X., Mao, J., Wei, M., Qi, Y., & Zhang, J. Z. H. (2022). HergSPred: Accurate Classification of hERG Blockers/Nonblockers with Machine-Learning Models. *Journal of Chemical Information and Modeling*, *62*(8), 1830–1839. https://doi.org/10.1021/acs.jcim.2c00256

37. Bajusz, D., Rácz, A., & Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, *7*(1), 20. https://doi.org/10.1186/s13321-015-0069-3

38. Eberhardt, J., Santos-Martins, D., Tillack, A. F., & Forli, S. (2021). AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *Journal of Chemical Information and Modeling*, *61*(8), 3891–3898. https://doi.org/10.1021/acs.jcim.1c00203

39. Dallakyan, S., & Olson, A. J. (2015). Small-molecule library screening by docking with PyRx. *Methods in Molecular Biology (Clifton, N.J.)*, *1263*, 243–250. https://doi.org/10.1007/978-1-4939-2269-7_19

40. Bank, R. P. D. (n.d.). *RCSB PDB - 4G5J: Crystal structure of EGFR kinase in complex with BIBW2992*. Retrieved July 7, 2023, from https://www.rcsb.org/structure/4G5J

41. Burley, S. K., Berman, H. M., Kleywegt, G. J., Markley, J. L., Nakamura, H., & Velankar, S. (2017). Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. In A. Wlodawer, Z. Dauter, & M. Jaskolski (Eds.), *Protein Crystallography: Methods and Protocols* (pp. 627–641). Springer. https://doi.org/10.1007/978-1-4939-7000-1_26

42. *Pymol: An Open-Source Molecular Graphics Tool – ScienceOpen*. (n.d.). Retrieved July 10, 2023, from https://www.scienceopen.com/document?vid=4362f9a2-0b29-433f-aa65-51db01f4962f

43. Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B., & Lindahl, E. (2015). GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, *1–2*, 19–25. https://doi.org/10.1016/j.softx.2015.06.001

44. Vanommeslaeghe, K., Hatcher, E., Acharya, C., Kundu, S., Zhong, S., Shim, J., Darian, E., Guvench, O., Lopes, P., Vorobyov, I., & Mackerell, A. D. (2009). CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *Journal of Computational Chemistry*, NA-NA. https://doi.org/10.1002/jcc.21367