Study of dysregulated genes during Squamous Cell Carcinoma for identification of potential biomarkers and therapeutics

THESIS

Submitted to the Delhi Technological University for the award of the degree of

DOCTOR OF PHILOSOPHY

Submitted by

Jaishree Meena

Guide

Prof. Yasha Hasija

Department of Biotechnology

Delhi Technological University, Delhi

Department of Biotechnology

Delhi Technological University

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-110042,

INDIA

JUNE 2023

Study of dysregulated genes during Squamous Cell Carcinoma for identification of potential biomarkers and therapeutics

THESIS

Submitted to the Delhi Technological University for the award of the degree of

DOCTOR OF PHILOSOPHY

Submitted by

Jaishree Meena

Guide

Prof. Yasha Hasija

Department of Biotechnology

Delhi Technological University, Delhi

Department of Biotechnology

Delhi Technological University

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-110042,

INDIA

JUNE 2023

"In the embrace of gratitude, I find solace and strength. With each breath, I offer heartfelt thanks to the Almighty, for the abundant blessings that grace my life's journey. Through every triumph and challenge, I am humbled by the blessings bestowed upon me, like gentle whispers of divine love. With a grateful heart, I cherish the tapestry of blessings, weaving moments of joy, grace, and growth."

Dedicated to

My Parents and

Beloved Son,

Krishna

DECLARATION

I hereby declare that the thesis entitled ***"Study of dysregulated genes during Squamous Cell Carcinoma for identification of potential biomarkers and therapeutics"*** submitted by me for the award of the degree of "***Doctor of Philosophy"*** to **Delhi Technological University (Formerly DCE), Delhi** is a record of *bona fide* work carried out by me under the guidance of Prof. Yasha Hasija.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this Institute or any other Institute or University.

Name: Jaishree Meena
Reg No: 2K18/PHDBT/10
Department of Biotechnology
Delhi Technological University (DTU)
Shahbad Daulatpur, Bawana Road, Delhi-110042
Place: Delhi
Date: 20/06/2023

CERTIFICATE

This is to certify that the thesis entitled ***"Study of dysregulated genes during Squamous Cell Carcinoma for identification of potential biomarkers and therapeutics"*** submitted by **Mrs. Jaishree Meena** to **Delhi Technological University (Formerly DCE), Delhi**, for the award of the degree of ***"Doctor of Philosophy"*** in Biotechnology is a record of *bona fide* work carried out by her. Jaishree Meena has worked under my guidance and supervision and has fulfilled the requirements for the submission of this thesis, which to our knowledge, has reached the requisite standards.

The results contained in this thesis are original and have not been submitted to any other university or institute for the award of any degree or diploma.

Prof. Yasha Hasija, PhD
Supervisor
Department of Biotechnology
Delhi Technological University
Place: Delhi
Date: 20/06/2023

Prof. Pravir Kumar
Head of the Department and DRC Chairman
Department of Biotechnology
Delhi Technological University
Place: Delhi
Date: 20/06/2023

Acknowledgements

In a world bursting with limitless possibilities and infinite wonders, I find myself immersed in an ocean of gratitude, compelled to express my deepest appreciation to those who have played an integral role in shaping my remarkable journey. With every step I take, my path is illuminated by the guiding light of extraordinary individuals who have bestowed upon me their wisdom, support, and unwavering faith in my abilities.

At the forefront of my acknowledgments, I extend my profound gratitude to ***Prof. Jai Prakash Saini****, the Honourable Vice Chancellor of Delhi Technological University, Delhi, for bestowing upon me the invaluable opportunity to embark on this extraordinary endeavor within the esteemed halls of this institution. It is with immense appreciation that I recognize* ***Prof. Yogesh Singh****, the former Vice Chancellor of Delhi Technological University, Delhi, for his visionary leadership and for providing the foundation that has nurtured my growth and development.*

Within the intricate maze of academia, I have been blessed to encounter a mentor of unparalleled caliber. ***Prof. Pravir Kumar****, DRC Chairman and Head, Department of Biotechnology, an epitome of knowledge and compassion, has not only provided me with an environment conducive to exploration but has also offered steadfast support and guidance in navigating the complexities of my research. His unwavering dedication to his students and the seamless facilitation of our academic pursuits are truly commendable.*

At the heart of my journey stands a radiant source of inspiration, ***Prof. Yasha Hasija****, my esteemed supervisor. Her unwavering dedication to excellence, coupled with her profound insight and nurturing guidance, has transformed the trajectory of my academic pursuits. Through each interaction, Professor Hasija has unveiled new perspectives, urging me to push the boundaries of my intellect and unleash my creativity. Her mentorship has been the cornerstone of my growth as a researcher and scholar, and for this, I am eternally grateful.*

No creative endeavor is complete without the collective wisdom of an exceptional group of individuals. To my esteemed External DRC and SRC experts, ***Dr. Neel Sarovar Bhavesh*** *(International Centre for Genetic Engineering and Biotechnology, New Delhi, India),* ***Dr. Md Imtaiyaz Hassan*** *(Center for Interdisciplinary Research in Basic Sciences, Jamia Millia Islamia, Delhi),* ***Prof. Sonika Bhatnagar*** *(Department of Biological Sciences and Engineering, NSUT, Delhi),* ***Prof. A.K. Dubey*** *(Department of Biological Sciences and Engineering, NSUT,*

Delhi), ***Prof. Jai Gopal Sharma*** *(Former Head of the Department, Department of Biotechnology, Delhi Technological University), and* ***Prof. Nitin Kumar Puri*** *(Department of Applied Physics, Delhi Technological University), I extend my heartfelt thanks. Their invaluable guidance, insightful perspectives, and stimulating discussions have not only shaped my work but have also broadened the horizons of my understanding.*

I am filled with heartfelt appreciation for my fellow lab mates, whose unwavering support and encouragement have served as an essential pillar throughout my research. ***Rajkumar Chakraborty****,* ***Neha Kumari, Nakul Tanwar****,* ***Khushi Yadav****, and* ***Priya Rai****, your assistance has been invaluable, and I am immensely grateful for the collaborative spirit we have fostered. To my incredible friends,* ***Niharika Gupta, Sunil Kumar, Sidharth Sharma, Rohan Gupta, Shweta Gulia****, and* ***Ritu Dhankas****, who have consistently stood by my side, infusing positivity and hope into my journey, words cannot adequately express my gratitude.*

I extend my sincere thanks to ***Mr. Chhail Bihari****,* ***Mr. Jitender Kumar****,* ***Mrs. Saumya Maurice****, and* ***Mr. Lalit Kumar****, whose unwavering support has made significant contributions to the success of my research endeavors. Moreover, I am grateful to the senior management and the dedicated technical staff of DTU for their invaluable assistance in facilitating my research work.*

To my beloved brother, ***Hari Shanker Meena****, and my dear sisters,* ***Sapna Meena****,* ***Shivani Meena****,* ***Sushila Meena****,* ***Anita Meena****, and* ***Kailash Meena****, along with my adored nephews and nieces, your unwavering presence during both triumphs and tribulations has been a constant source of motivation. Your unwavering belief in me and steadfast support have played an instrumental role in my journey.*

In a heartfelt tribute, I dedicate this thesis to my remarkable parents, ***Mr. Radhey Shyam Meena*** *and* ***Mrs. Panna Devi****. Their boundless love and unwavering encouragement have been the foundation of my strength, propelling me forward in pursuit of knowledge. I am forever grateful for their consistent presence in my life and the sacrifices they have made to nurture my educational endeavors. Their unwavering commitment to my dreams has left an indelible mark on my heart, and I carry their legacy with immense pride.*

Furthermore, I dedicate this thesis to my cherished son, ***Krishna Jai Kumar****, whose infectious laughter and radiant smile illuminate my world. In his innocent mischief, I find the strength and motivation to overcome challenges and embrace the unknown. He personifies hope, reminding me of the profound purpose that fuels my academic journey.*

To all those who have played a role, whether through direct guidance or indirect influence, I extend my heartfelt appreciation. Your unwavering support and invaluable contributions have shaped the person I have become. I am humbled by the collective wisdom, encouragement, and guidance that have accompanied me on this transformative path. Together, we have woven a tapestry of shared experiences, and I am forever grateful for the profound impact each of you has had on my journey.

Jaishree Meena

ABSTRACT

ABSTRACT

The aim of this study was to investigate dysregulated genes during Squamous Cell Carcinoma (SCC) and identify potential biomarkers and therapeutics for the disease. Three specific objectives were pursued: (1) to identify key genes and pathways involved in the progression of Cutaneous SCC from AK, (2) to investigate the impact of somatic non-synonymous mutations on BTK protein and their potential influence on FDA-approved therapies for SCC, and (3) to find chemical perturbations associated with identified biomarkers for the correction of SCC.

The first objective focused on identifying key genes and pathways involved in the progression of cutaneous SCC from AK. To accomplish this, the study applied an e**X**plainable **A**rtificial **I**ntelligence (XAI) approach to the XGBoost classification model. XAI techniques, such as SHAP barplot and SHAP summary plots, were utilized to establish interpretability by linking the model outputs to relevant genes. By analyzing the model's predictions, significant genes associated with SCC progression were identified. This approach not only provided insights into the genes involved in the disease but also demonstrated the potential of XAI methods in identifying biomarkers [1].

The second objective aimed to investigate the impact of somatic non-synonymous mutations on Bruton's tyrosine kinase (BTK) protein and their potential influence on FDA-approved therapies for SCC. A literature survey was conducted to identify FDA-approved drugs for skin cancer, leading to the discovery of Ibrutinib, a BTK inhibitor. Although there has been limited research on the role of BTK protein in SCC, the study chose to focus on it to address the existing research gap. Molecular dynamics (MD) simulations were performed to analyze the effects of individual amino acid mutations on the stability of the BTK protein. The findings indicated that these mutations may contribute to the prognosis of SCC by rendering the protein unstable. Additionally, the interaction between the BTK protein and its mutants with Ibrutinib was

examined, revealing that the mutants exhibited comparable binding to Ibrutinib as the wild-type protein. This observation highlighted the potential efficacy of Ibrutinib-based therapy in targeting these mutations for SCC treatment [2].

The third objective aimed to find chemical perturbations associated with the identified biomarkers for the correction of SCC. To achieve this, gene expression profiles of individuals diagnosed with SCC, healthy individuals, and those with AK were rigorously compared. Several dysregulated genes that exhibited significant differential expression were identified. These dysregulated genes were found to be involved in crucial biological processes closely associated with SCC progression, such as cellular disassembly, regulation of protein catabolism, and extracellular matrix disassembly. Additionally, important biological pathways, including WNT signaling and regulation of the actin cytoskeleton, were found to play a role in SCC progression. To further augment the research outcomes, the Drug Gene Budger tool was utilized to investigate potential therapeutic interventions. Analysis using this tool revealed the notable effectiveness of certain drugs, such as Doxorubicin, Dasatinib, and Tretinoin, in rectifying the abnormal expression patterns of the identified dysregulated genes associated with SCC.

This study contributes to the identification of potential biomarkers and therapeutics for SCC through a comprehensive approach. The utilization of XAI techniques facilitated the identification of significant genes associated with SCC progression. The investigation of somatic non-synonymous mutations in the BTK protein provided insights into its stability and potential therapeutic targeting with Ibrutinib. The exploration of chemical perturbations associated with dysregulated genes shed light on potential treatment options for SCC. Collectively, these findings have implications for precision medicine and innovative drug discovery strategies in the field of SCC.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

SCC	Squamous Cell Carcinoma
BCC	Basal Cell Carcinoma
AK	Actinic Keratosis
ML	Machine Learning
AI	Artificial Intelligence
XAI	eXplainable Artificial Intelligence
SHAP	SHapley Additive exPlanations
LIME	Local Interpretable Model-Agnostic Explanations
KNN	K-Nearest Neighbors
SVM	Support Vector Machines
XGBoost	Extreme Gradient Boosting
PCA	Principle Component Analysis
AUC-ROC	Area Under the Receiver Operating Characteristic Curve
CONAN	CONtact ANalysis
DGB	Drug Gene Budger
GO	Gene Ontology
BP	Biological Processes
MF	Molecular Function
CC	Cellular Components
PAMR1	Peptidase Domain Containing Associated with Muscle Regeneration-1
CTSC	Cathepsin-C
PHYHIP	Phytanoyl-CoA 2-Hydroxylase Interacting Protein
CD24	Cluster of Differentiation-24

WNT5A	WNT family member-5A
RAB3B	RAS-associated binding family member-3B
WIF1	WNT-Inhibitory Factor-1
TNNC1	Troponin-C1
PARK7	Parkinson disease protein-7
MMP14	Matrix Metalloprotease-14
ARHGEF4	Rho guanine nucleotide exchange factor 4
CFL1	Cofilin-1
HNRNPM	heterogeneous nuclear ribonucleoprotein M
RPS13	Ribosomal Protein S-3
GTSE1	G2 And S-Phase Expressed-1
CHTOP	Chromatin Target Of PRMT1
EDNRB	Endothelin Receptor Type-B
DNAJC8	DnaJ heat shock protein family (Hsp40) member C-8
S100A11	S100 calcium binding protein A-11
TFG	Tropomyosin-receptor kinase fused gene
GAPDH	Glyceraldehyde-3-Phosphate Dehydrogenase
RPS3A	Ribosomal Protein S-3A

Chapter 1

Overview of the Thesis

Chapter 1. Overview of the Thesis

I. Introduction

The integration of computational methodologies and biological data is of utmost importance in the healthcare domain, wherein bioinformatics assumes a central position. Through the analysis, management, and interpretation of vast biological datasets that encompass genomics, transcriptomics, proteomics, and metabolomics, researchers acquire knowledge regarding diverse facets of health and disease [3], [4]. The acquisition of such datasets is crucial in order to gain a comprehensive understanding of the molecular mechanisms that underlie various diseases and to facilitate the development of targeted therapeutic interventions [5]. The advent of e**X**plainable **A**rtificial **I**ntelligence (XAI) and Machine Learning (ML) has significantly bolstered the capabilities of bioinformatics in extracting meaningful insights from intricate biological data, thereby transforming the healthcare industry by facilitating comprehensive analysis of genomic and clinical datasets. ML algorithms accelerate the identification and management of various diseases by revealing patterns, relationships, and predictive models. The analysis of genomic data enables the use of algorithms to predict the probability of disease onset, thereby enabling prompt intervention and tailored preventive measures. Also, ML models utilize electronic health records (EHRs) and clinical data to assist precise diagnoses and anticipate patient outcomes for healthcare professionals [6], [7].

The amalgamation of bioinformatics and ML has become an essential component in the healthcare industry, specifically in the realm of drug discovery and development. Docking and simulation techniques are extensively utilized by researchers to comprehend the interaction between drug molecules and target proteins, as well as to evaluate their binding affinity [8]. The methodologies employed entail the simulation of the docking mechanism between a small molecule and a target protein, with the objective of ascertaining the most favorable binding

configuration and potential energy. By utilizing sophisticated computational algorithms, researchers are able to effectively examine extensive compound libraries in order to identify potential drug candidates, resulting in a notable acceleration of the drug discovery process [9].

The application of comprehensive drug databases has brought about a significant transformation in the field of drug repurposing, which involves the identification of alternative therapeutic indications for drugs that were initially developed for different purposes [10]. Professionals in the field of bioinformatics utilize information on drug structures, targets, and clinical outcomes to identify possible candidates for repurposing. ML algorithms are utilized to forecast new therapeutic indications for already existing drugs by examining databases and taking into account various factors such as target interactions, chemical similarity, and disease pathways. This methodology results in significant savings of both time and resources in the drug development process [11], [12].

In the advancement of personalized medicine, bioinformatics tools, and techniques have been indispensable, involving the customization of therapeutic approaches according to the distinct genomic, environmental, and lifestyle attributes of individual patients. Bioinformatics software is utilized to analyze genomic data, detecting genetic variations that impact a patient's reaction to medication or vulnerability to a particular condition [13], [14]. This information is utilized by medical practitioners to make well-informed decisions regarding the selection of treatment and to tailor drug dosages for individual patients, leading to enhanced effectiveness and minimized adverse reactions [15].

The fusion of bioinformatics with other omics technologies, namely proteomics and metabolomics, amplifies our comprehension of disease mechanisms [16], [17]. As a result, significant biomarkers are recognized, facilitating the detection, anticipation, and surveillance of diseases. ML algorithms are utilized by researchers to scrutinize vast omics datasets in order

to detect accurate biomarkers that are linked to particular diseases or treatment results. Biomarkers play a crucial role in enabling the timely identification of diseases, tailoring treatment options to individual patients, and tracking the development of diseases. The domain of bioinformatics and computational biology is in a state of perpetual evolution, as it assimilates a wide range of data sources and devises sophisticated algorithms to enhance patient outcomes and transform the domains of disease diagnosis, treatment, and prevention [18], [19].

The study of dysregulated genes during SCC relied on the integration of bioinformatics and computational methodologies. The primary aim of this study is to identify precise biomarkers and therapeutics for SCC. To achieve this goal, the study is divided into three main objectives:

II. Aim

Study of dysregulated genes during Squamous Cell Carcinoma for identification of potential biomarkers and therapeutics.

III. Objectives

1. **To identify key genes and pathways involved in the progression of Cutaneous Squamous Cell Carcinoma from Actinic Keratosis.**

XAI has garnered increasing attention in recent years for its potential to identify biomarkers associated with diverse conditions, such as cancer. The present objective was centered on the utilization of XAI to discern biomarkers associated with (Actinic Keratosis) AK to SCC. A two-phase methodology was utilized, which involved the development of a classification model based on ML using XGBoost, followed by the application of XAI techniques to establish interpretability by linking the model outputs to pertinent genes. This study reported the outcomes of our research, encompassing the identification of the most significant genes that contribute to the precision of the model and their plausible implications in the development of cancer (Figure 1.1).

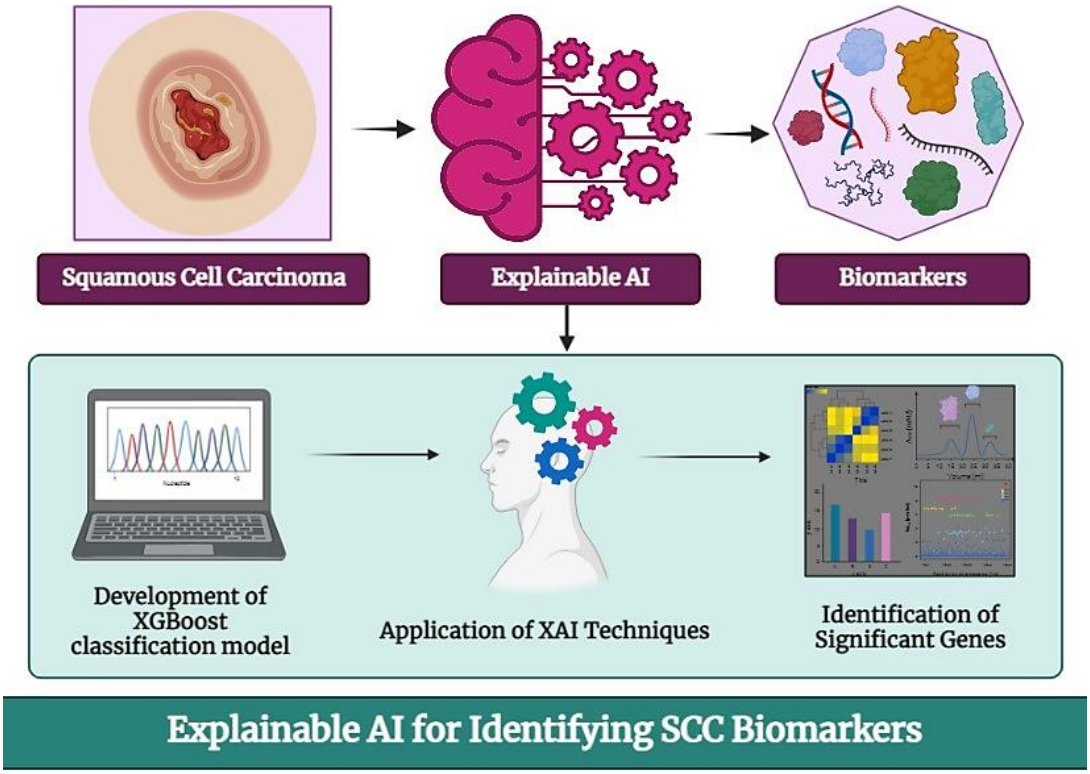


Figure 1.1: Workflow of the proposed study showing the applicability of XAI in SCC biomarkers identification.

In the beginning, a classification model based on XGBoost was developed by applying ML approach. The XGBoost algorithm is widely recognized for its exceptional predictive performance, rendering it a highly suitable option for our research. The model was trained utilizing a dataset comprising pertinent genes associated with SCC. In order to guarantee the interpretability of the XGBoost model, we applied XAI techniques. The SHAP (SHapley Additive exPlanations) barplot and SHAP summary plots were employed to elucidate the predictions of the model. The plots yielded valuable insights into the individual gene contributions toward the classification outcomes, thus establishing a correlation between the model output and genes associated with SCC. The findings of our analysis indicated that the precision of the XGBoost classification model remained stable both prior to and subsequent to the implementation of SHAP values. This discovery suggested that it is possible to attain the interpretability of ML models without compromising their efficacy. The observation

underscored the capacity of XAI methods to augment the interpretability of models while maintaining their precision. The genes that were identified in this study exhibit potential as prospective targets for the management of SCC. The present research strongly supported the implementation of XAI in the identification of biomarkers for predictive and prognostic purposes in the biomedical domain [1].

2. To investigate the impact of somatic non-synonymous mutations on BTK protein and their potential influence on FDA-approved therapies for SCC.

To achieve this objective, we began with the genes implicated in the progression of SCC based on our previous study. As these genes are still being studied and there are no known FDA-approved drugs that target them, we conducted a literature survey to identify FDA-approved drugs for skin cancer. This survey led to the discovery of Ibrutinib, a Bruton's tyrosine kinase (BTK) inhibitor. Although there have been few studies on the role of BTK protein in SCC, we chose to focus on BTK in order to close the extant research gap. By investigating the impact of somatic non-synonymous mutations on BTK protein and their potential influence on Ibrutinib for SCC, we obtained a better understanding of this drug as a potential therapeutic target for this cancer.

The development and progression of SCC are significantly impacted by protein mutations that play a pivotal role in the disease. This study was directed to examine the effects of individual amino acid mutations in the BTK protein. The adverse effects of selected deleterious mutations in the BTK protein on protein stability were analyzed using molecular dynamics (MD) simulations. The results suggested that by making the protein unstable, these mutations may affect SCC prognosis. Additionally, a study was done to see how the BTK protein, and its mutations interacted with Ibrutinib, a medication created especially for the treatment of SCC. Despite the adverse impact of mutations on protein structure, it was noted that the mutants

exhibited comparable binding to Ibrutinib as the wild-type protein. The outcomes of this objective suggested that regardless of the adverse effects of missense mutations on protein function, particularly in SCC, Ibrutinib-based therapy remains a viable option for targeting these mutations with efficacy (Figure 1.2).

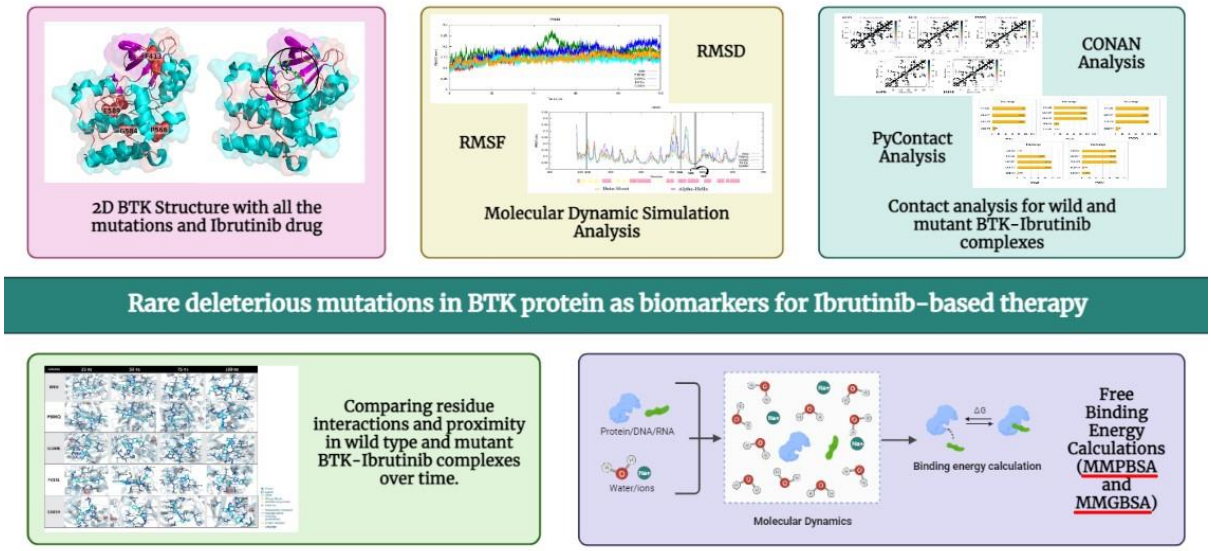


Figure 1.2: Workflow of the proposed study showing RMSD and RMSF analysis to find the impact of mutations on BTK protein and their potential influence on the Ibrutinib drug for SCC therapy.

Seven computational tools were used in this study to evaluate the impact of individual amino acid mutations. Insights into the differences in protein and mutant dynamics were obtained through the use of Molecular Dynamics (MD) simulations and trajectory analysis techniques such as Root Mean Square Deviation (RMSD), Root Mean Square Fluctuation (RMSF), Principal Component Analysis (PCA), and Contact analysis. Furthermore, we utilized Molecular docking, Molecular Mechanics-Generalized Born Surface Area (MM-GBSA), Molecular Mechanics-Poisson Boltzmann Surface Area (MM-PBSA), and interaction analysis to evaluate the free binding energy and its decomposition for each protein-drug complex for both the wild-type protein and its mutants. This study provided evidence that the stability and function of the BTK protein are negatively affected by single amino acid

mutations in SCC. However, it is noteworthy that Ibrutinib-based therapy continues to exhibit efficacy against these mutations, thereby highlighting their potential as biomarkers for targeted treatment with Ibrutinib in SCC [2].

3. To find chemical perturbations associated with identified biomarkers for correction of Squamous Cell Carcinoma.

The main objective of this study was to identify chemical perturbations that are linked to biomarkers/dysregulated genes for the treatment of SCC, with a significant emphasis on their potential use in drug discovery and repurposing. This study identified crucially dysregulated genes that are associated with SCC and assessed their potential as therapeutic targets. Through a rigorous comparison of gene expression profiles among individuals diagnosed with SCC, healthy individuals, and those with AK, we successfully identified several genes that exhibit significant differential expression. These dysregulated genes have been identified to have involvement in crucial biological processes that are closely associated with the progression of SCC. These processes include cellular disassembly, regulation of protein catabolism, and extracellular matrix disassembly. Also, we have found several biological pathways, such as WNT signaling, regulation of actin cytoskeleton, etc., which are crucial in the progression of SCC.

In order to augment our research outcomes, we utilized Drug Gene Budger (DGB), a tool designed to investigate potential therapeutic interventions. The analysis revealed the notable effectiveness of certain drugs, such as Doxorubicin, Dasatinib, Tretinoin, etc., in rectifying the atypical expression patterns of identified dysregulated genes associated with SCC. This study presents promising opportunities for precise therapeutic interventions in personalized treatment, thereby facilitating innovative drug development and repurposing strategies in the field of cancer (Figure 1.3).

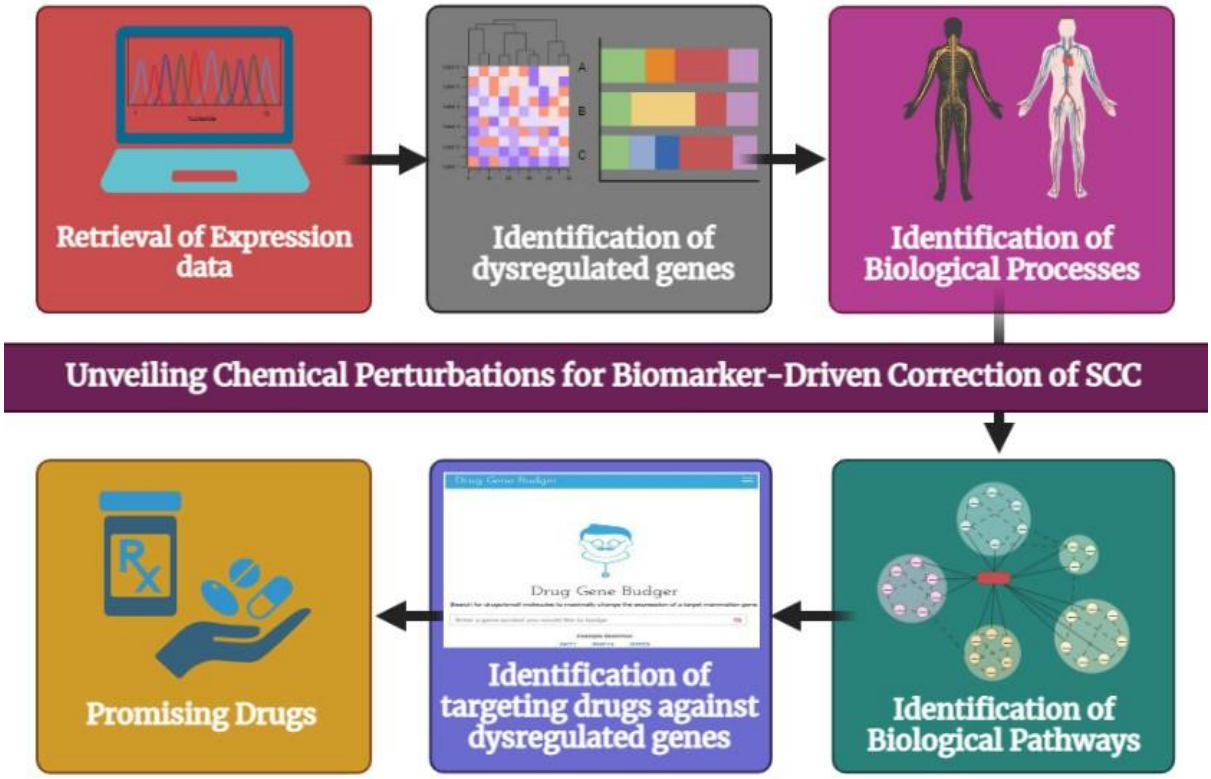


Figure 1.3: Workflow of the proposed study showing the identification of chemical perturbations for biomarker-driven correction of SCC.

Chapter 2

Review of Literature

Chapter 2. Review of literature

I. Introduction

"The two areas that are changing are information technology and medical technology. Those are the things that the world will be very different 20 years from now than it is today."

Bill Gates

Bill Gates's quote emphasizes the transformative impact of information technology and medical technology on the global landscape within the next two decades. These two areas have the potential to revolutionize various aspects of society. Information technology advancements, such as AI and data analytics, are already reshaping industries and opening new avenues for innovation. Medical technology breakthroughs, including genomics and personalized medicine, are paving the way for precise diagnostics and targeted therapies. When it comes to SCC, these advancements have already contributed to more accurate diagnoses, personalized treatment approaches, and ongoing research advancements. As we embrace these technologies further, we can anticipate even greater progress in the fight against cancer and other complex diseases.

Non-melanoma skin cancer (NMSC) is the fifth most prevalent cancer in men and women worldwide. About 1 million cases of NMSCs are diagnosed worldwide, with SCC being the second most prevalent NMSC, accounting for approximately 20% of all skin malignancies [20], [21]. SCC is characterized by the uncontrolled proliferation of aberrant squamous cells, which are slender and flat cells that make up the skin's outer layer and the lining of the body's hollow organs. Numerous studies have indicated that SCC is likely to develop from AK or from Bowen's disease, also known as Carcinoma-in-situ [22]. The chance of developing SCC exhibits variability contingent upon the cancer's anatomical site and also upon an individual's age and gender, among other risk factors. Adults over the age of 65 experience

the highest rates of diagnosis, making them the most likely group to experience it. Men are more likely than women to acquire SCC, and also individuals with a pale complexion and green or blue eyes are at greater risk [23]. The prevalence and incidence of SCC can vary significantly depending on the population being investigated and the individual risk factors involved, making it challenging to provide precise figures. The combination of radiation and surgical interventions has demonstrated efficacy in the management of the majority of cases of SCC. Nevertheless, a considerable fraction of high-risk SCC, constituting 5-10% of instances, poses a formidable therapeutic challenge [24]. The present treatments available for metastatic skin cancer are still inadequate, particularly in the elderly population, highlighting the necessity for more efficient and methodical therapeutic alternatives. The risk of morbidity and mortality from SCC is an unrecognized public health problem. According to statistical reports, SCC has been identified as a significant health burden, resulting in numerous fatalities, particularly in European countries. The scientific community is presently engaged in developing precise anti-cancer therapeutics through the analysis of genomic data, with a focus on identifying a viable remedy for both melanoma and non-melanoma skin cancers [25], [26]. A meta-analysis conducted on the oncogenome of SCC has indicated that while the alterations associated with each tumor may vary, a significant number of SCC cases exhibit shared dysregulated molecular pathway networks resulting from these alterations. The acquisition from these studies can prove to be advantageous in comprehending the cellular and molecular mechanisms involved in the progression of SCC [27].

The etiology of SCC remains uncertain; however, it is hypothesized to be associated with protracted exposure to ultraviolet radiation emanating from the sun or alternative sources, such as tanning devices. Additional risk factors comprise compromised immune function, exposure to specific chemical agents, and medical history of skin disorders [28]. Upon its onset, SCC has the potential to proliferate and metastasize to distant anatomical sites. During its initial phases,

the condition may manifest as a diminutive, scaly lesion on the epidermis that exhibits a reddish or otherwise altered hue. As the condition advances, it has the possibility to develop into a protruding, papilloma-like lesion or an ulcerated wound that exudes blood. During the later stages, the malignancy has a tendency to metastasize to adjacent lymph nodes as well as other bodily structures, thereby rendering therapeutic interventions more challenging [29]. The therapeutic approach for SCC is contingent upon the cancer's stage and location and may involve surgical intervention to excise the malignant tissue, radiation, or chemotherapy. The primary therapeutic modality for SCC is surgical intervention, which entails the excision of malignant tissue. The aforementioned procedure can be accomplished through a local excision, which involves the removal of solely the cancerous tissue, or an expanded excision, which entails the removal of a portion of healthy tissue surrounding the cancer [30]. Radiation therapy employs ionizing radiation, such as X-rays, to induce lethal damage to malignant cells. This therapeutic intervention is capable of being applied either as a monotherapy or in conjunction with a surgical procedure. The efficacy of the treatment of SCC can be considerable; however, it is not without chances for adverse effects, including skin inflammation and fatigue [31]. Chemotherapy involves the use of medications to eliminate cancer cells, and its efficacy is typically inferior to that of these alternative modalities. The administration of chemotherapy may result in adverse reactions, including but not limited to emesis, alopecia, and nausea [32].

Managing SCC can pose challenges, particularly in cases where metastasis has taken place, and the efficacy of existing treatments may not always meet expectations. Researchers are currently endeavoring to devise novel and efficacious interventions for SCC. Also, the challenge of identifying SCC in its initial phases may lead to treatment delays, thereby augmenting the intricacy of managing cancer's progress. Also, it has been observed that SCC may exhibit a suboptimal response to specific therapeutic interventions such as radiation and

chemotherapy, leading to unfavorable outcomes or adverse reactions in some individuals, so the timely diagnosis of SCC is crucial for effective treatment [32], [33].

Researchers are considering using AI-based diagnostic tools, such as ML algorithms, to solve the difficulties of early SCC detection and improve patient outcomes [34]. These technologies have shown promise in improving the accuracy and effectiveness of SCC diagnosis by using medical data and images to find patterns linked to the condition and lowering the number of unwanted biopsies, which can be painful and expensive. They can also assist in monitoring the course of lesions over time, allowing medical experts to change treatment approaches as needed. The integration of AI in the diagnosis of SCC holds the promise of transforming cancer research and enhancing the general health status of patients [35].

Moreover, AI is currently being applied to the pursuit of novel drugs and therapeutic options for individuals afflicted with SCC. The application of AI possesses the ability to considerably expedite the drug development process, consequently enhancing the probability of prompt and effective therapy for patients. Further, it can aid in the recognition of individuals who are best suited to receive specific treatments, thus allowing for the establishment of personalized therapies that are customized to the specific needs and characteristics of each patient. With the continued breakthroughs in AI, it is likely that further pioneering applications of this technology in the fight against SCC and various forms of cancer will emerge [36].

II. Genetic Predisposition to Squamous Cell Carcinoma

A higher probability of developing SCC can be attributed to genetic predispositions. Such predispositions refer to an innate susceptibility to a disease that increases the risk of its occurrence. Individuals who have a family history of SCC or carry certain genetic mutations, like TP53, PIK3CA, CDKN2A, or KRAS genes, may have a genetic predisposition to developing this condition [37].

The PIK3CA (phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha) gene is responsible for coding the PIK3CA protein, which plays a significant role in regulating cell growth and division through the PI3K signaling pathway. When activated, this protein triggers the production of AKT/Protein kinase B, which, in turn, activates other proteins that promote cell growth and division. However, mutations in the PIK3CA gene can cause the protein to become overactive, leading to uncontrolled cell growth and the development of SCC [38].

The TP53 (tumor protein p53) gene produces a protein that assists in regulating cell growth and division. In the event of a mutation in this gene, the regulatory mechanism may be disrupted, potentially resulting in the onset of SCC. Mutations in genes linked to TP53, such as TP63 and TP73, can also contribute to SCC formation [39].

The CDKN2A, which is a cyclin-dependent kinase inhibitor 2A, is of paramount importance in the maintenance of appropriate cellular growth and division control. However, mutations in the CDKN2A gene can cause the CDKIs to become less effective, resulting in uncontrolled cell growth and cancer. Mutations in other genes related to the CDKN2A pathway, such as CDK4 and CDK6, can also contribute to the development of cancer, including SCC [40].

The KRAS (Kirsten rat sarcoma viral oncogene homolog) gene codes for the KRAS protein. If it undergoes a mutation, it can produce an abnormal version of the protein that can cause cells to grow and divide uncontrollably. This uncontrolled growth can lead to the formation of tumors that may eventually develop into SCC. Mutations in the KRAS gene have also been linked to a more aggressive form of SCC that is challenging to treat as the abnormal protein can interfere with other signaling pathways within the cell, leading to further dysregulation of cell growth and division. While the KRAS gene plays a significant role in SCC's development, there may be other genes and pathways involved in its development yet to be discovered [41].

Although having a genetic inclination towards SCC doesn't guarantee its occurrence, other elements like UV radiation may also affect the chances. Apart from certain genetic mutations or a family history of SCC, additional factors contribute to the likelihood of the disease [42]. Therefore, it is crucial for individuals to have a discussion with their healthcare provider to understand their individual risk factors and take preventive measures to reduce their risk of developing SCC. By being proactive and taking steps to protect their skin from harmful UV rays, individuals can help reduce their risk of developing this type of skin cancer.

III. Role of dysregulated genes for identification of SCC biomarkers

The identification of biomarkers for SCC is of the highest importance, and dysregulated genes may serve as viable candidates for this purpose. The aberrant or mutated genes have the potential to trigger anomalous cellular proliferation. By acknowledging these genes and their corresponding pathways, it is feasible to develop biomarkers that can facilitate the identification and monitoring of SCC. In conjunction with other diagnostic modalities, such as biopsies and imaging, biomarkers can provide a comprehensive assessment of the patient's medical status, which can aid in the selection of appropriate therapeutic interventions [43]. The genes in question can provide crucial insights into the mechanisms underlying the development of SCC, thereby facilitating the identification of targeted therapies for this condition [44], [45].

The use of computational tools by researchers has led to the precise recognition of dysregulated genes in cancer. By analyzing publicly accessible gene expression data, unusual expression levels can be identified, potentially leading to the discovery of biomarkers and therapeutic targets [46]. RNA-seq, microarray analysis, and gene ontology analysis are among the computational tools available to researchers for this purpose. Due to the unique benefits and drawbacks of each tool, it is essential to choose the most appropriate one for a particular research query [47].

Advanced computational methodologies are being used to scrutinize diverse data repositories, encompassing genomic, transcriptomic, and proteomic data, with the aim of detecting plausible genes linked to SCC. The approach entails the acquisition and pre-processing of data to eliminate inaccuracies and incongruities, followed by the application of statistical and ML algorithms, such as association analysis, regression analysis, and clustering, to detect patterns and trends. Subsequently, functional analysis techniques such as gene ontology, pathway analysis, and network analysis are utilized to comprehend the biological mechanisms underlying these gene associations [48]. It is imperative to validate the outcomes utilizing autonomous datasets prior to ascertaining the function of the identified genes in SCC via experimental methodologies such as gene knockdown or overexpression experiments. This is necessary to facilitate the development of novel diagnostic and therapeutic strategies for SCC.

IV. A comprehensive overview of Machine Learning and its types

In recent times, there have been notable transformations in the field of cancer research. Researchers have employed diverse techniques, such as preventive screening, to identify cancer prior to the manifestation of symptoms, and have come up with novel approaches for predicting treatment efficacy. The advent of cutting-edge technologies has led to the acquisition of tremendous amounts of data that are accessible to the healthcare research community [49]. Nevertheless, accurately predicting disease outcomes remains a challenging task for physicians. In response to this challenge, the application of ML techniques has gained widespread popularity among medical researchers due to their ability to efficiently discern patterns and correlations within intricate datasets, thereby enabling the prediction of future outcomes for various forms of cancer [50].

ML is a robust tool that facilitates the improvement of system performance on particular tasks without the need for explicit programming to execute those tasks. ML is propelling numerous innovative advancements in various domains, including but not limited to image and speech

recognition, natural language processing, and robotics, by utilizing extensive data. It also has real-world uses in areas like fraud analysis, client profiling, and preventative maintenance. ML, in its many forms, provides researchers and organizations with a formidable tool for taking on difficult challenges [51].

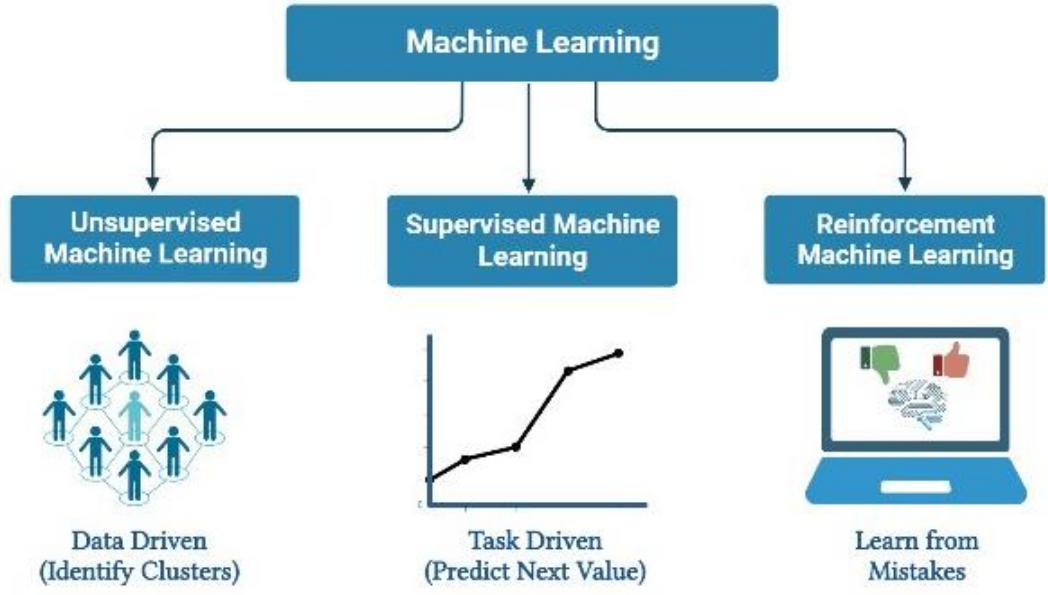


Figure 2.1: The Trifecta of Machine Learning: Supervised ML, Unsupervised ML, and Reinforcement ML.

ML is commonly classified into three fundamental categories (Figure 2.1), which are determined by the specific techniques and methodologies utilized for the learning process. The aforementioned categories comprise:

1. **Unsupervised Machine Learning**

Unsupervised ML pertains to a set of algorithms that facilitate the autonomous identification of patterns and structures, thereby uncovering underlying insights from unannotated data (Figure 2.2). Clustering algorithms, such as k-means and hierarchical clustering, enable the categorization of data points with comparable characteristics, thus enabling the detection of distinct clusters within the data [52].

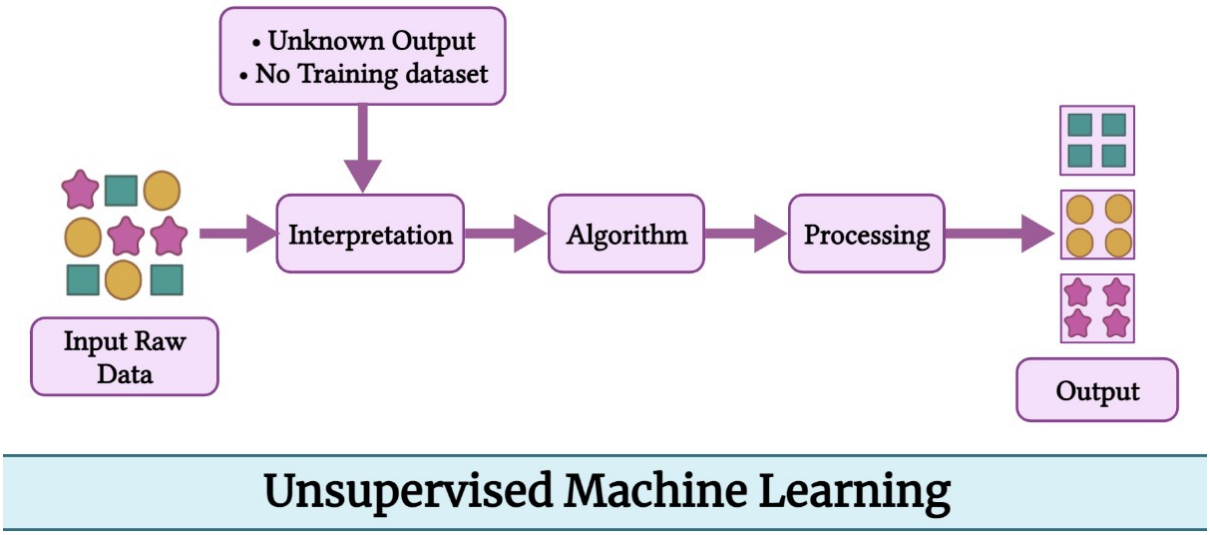


Figure 2.2: Unsupervised ML workflow

Dimensionality reduction methods, such as PCA and t-SNE, are employed to reduce the number of dimensions in a given dataset while preserving its essential characteristics [53]. This procedure facilitates the representation of information and the elimination of superfluous characteristics. Anomaly detection algorithms are employed for the purpose of identifying rare or atypical patterns within data. This renders them particularly valuable for detecting fraudulent activities and improving the security of networks [54]. Association rule learning algorithms, namely Apriori and FP-growth, are employed to uncover associations and dependencies among discrete items in a provided dataset [55].

Unsupervised learning is a widely employed technique in diverse domains, such as customer segmentation, recommendation systems, image, and text analysis, etc. This facilitates the acquisition of valuable insights by data scientists, enabling them to make informed decisions predicated on the underlying structures of the data.

2. Reinforcement Machine Learning

Reinforcement learning is a distinct category of ML that instructs an agent to optimize a reward signal while making decisions within a designated environment. The process of achieving this involves a trial-and-error approach, where the agent is provided with feedback in the form of positive rewards or negative punishments. The provision of feedback is of utmost importance in facilitating the agent's acquisition of knowledge and adjustment of its policy to enhance decision-making capabilities in subsequent instances without the need for explicit directives (Figure 2.3). This methodology is effective in intricate or unfamiliar settings where the identification of the most advantageous course of action necessitates the agent's exploration and experimentation. Reinforcement learning comprises three fundamental constituents, namely the agent, environment, and action space [56].

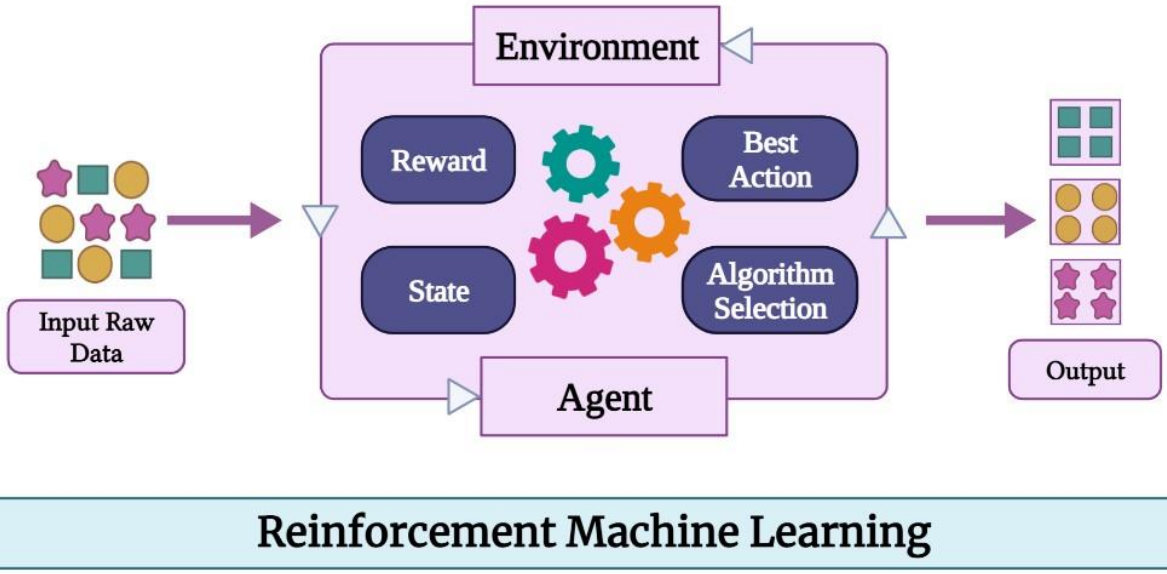


Figure 2.3: Reinforcement ML workflow

Reinforcement learning algorithms encounter the challenge of delayed, noisy, or sparse feedback signals, which poses a difficulty in devising an effective policy. Approaches such as exploration and exploitation, temporal difference learning, and Monte Carlo methods are

employed to assist agents in mitigating this issue. These techniques facilitate the enhancement of agents' learning efficacy [57], [58].

Q-learning is a commonly employed approach in reinforcement learning algorithms for determining the optimal strategy. This method entails the agent maintaining a record of anticipated rewards for each feasible action in all environmental states. The approximations are revised based on the incentives obtained by the agent subsequent to executing manoeuvres within the given context. Over time, these approximations approach the accurate anticipated reward values, facilitating the agent in obtaining the most advantageous approach [59].

Reinforcement learning has proven to be beneficial in various fields, such as recommendation systems, natural language processing, games, and robot control. One example of this phenomenon is the application of reinforcement learning in the development of robotic agents capable of performing complex tasks. The field of reinforcement learning is a significant and evolving area of research within the realm of AI. It has the capacity to tackle complex real-world problems and has demonstrated encouraging results in diverse applications. An exemplary instance of this phenomenon is the advancement of computer game agents that possess the capacity to match or surpass human performance in games such as chess.

3. Supervised Machine learning

Supervised ML is a widely used methodology in which an algorithm is trained on a predetermined dataset that contains annotated outputs for each instance (Figure 2.4). The objective of this procedure is to generate a model that has the capability of predicting results for novel instances that bear a resemblance to the data used for training. This form of learning has diverse practical applications, like natural language processing, image recognition, credit scoring, etc. The application of supervised learning has become a prominent methodology for

making informed decisions and estimations by exploiting data [60], [61]. Various forms of supervised learning are delineated and elaborated upon in the subsequent discourse.

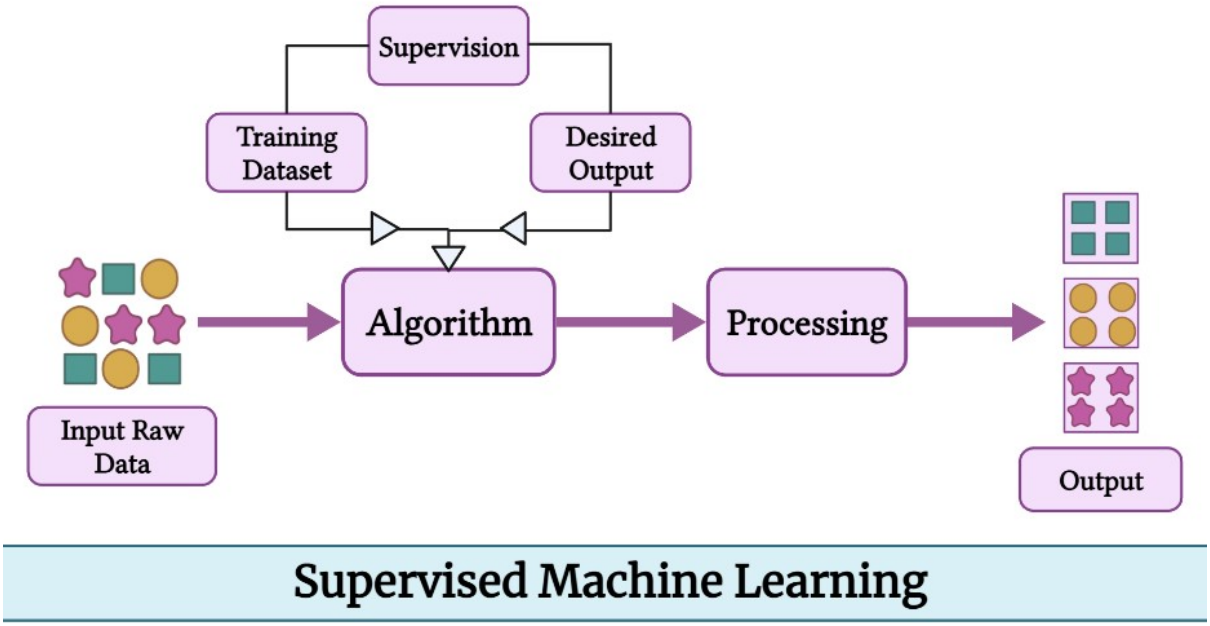


Figure 2.4: Supervised ML workflow

a) Regression

Regression is a commonly used algorithm in the field of supervised learning. Its primary objective is to predict a continuous output variable based on one or more input variables. Regression analysis is utilized in diverse domains, including the estimation of a house's value based on its location, size, and other pertinent variables. In order to reduce the discrepancy between anticipated and observed values within the training dataset, these algorithms identify the optimal line of best fit. The resultant model generates forecasts for unobserved data by applying this equation. There are several regression algorithms available, including linear regression, logistic regression, and non-linear regression. Each algorithm exhibits its own distinct advantages and disadvantages, and the selection of the optimal algorithm is contingent upon the specific problem being addressed [62].

- Linear regression is a statistical technique involving the use of empirical data to construct an equation of linearity that represents the relationship between one or more predictor variables and a response variable. Typically, the dependent variable exhibits continuity, while the independent variables may assume either continuous or categorical forms. The aim is to ascertain the most appropriate line that represents their association [63].
- Logistic regression is a fitting method that is appropriate for situations where the dependent variable has only two potential outcomes, namely 0 or 1. Through the application of this regression methodology, it is possible to approximate the probability of an occurrence, such as evaluating the likelihood of an individual developing a condition or disease based on particular risk elements [64].
- Non-linear regression is a modeling technique employed when a linear equation fails to accurately represent the relationship between a dependent variable and a number of independent variables. This method uses non-linear functions, such as trigonometric or polynomial functions, to accommodate more intricate data patterns. The primary objective of non-linear regression is to ascertain the function that most accurately characterizes the given data points [65].

b) Classification

Classification algorithms are used in supervised learning to predict output variables from input variables. The objective of classification is to classify data into distinct categories, for instance, ascertaining whether an email meets the criteria for being spam or whether a customer intends to renew their subscription. The procedure entails the identification of decision boundaries that demarcate distinct classes within the training dataset. Subsequently, the model makes use of these boundaries to generate predictions on novel data. The commonly used supervised classifiers in ML are k-nearest neighbors, decision trees, and support vector machines [66].

- KNN, an abbreviation for k-nearest neighbors, is an ML methodology that works in both classification and regression tasks. The fundamental principle underlying the KNN algorithm is to determine the k nearest neighboring data points to a specific point in a dataset and subsequently utilize these points to generate predictions concerning the given point. In classification tasks, KNN draws on the labels of the k-closest data points to anticipate the label of an unknown data point. On the other hand, in regression tasks, KNN relies on the values of these data points to figure out the value of a new data point [67].
- The decision tree algorithm is frequently used in tasks involving supervised learning for the purposes of classification and regression both. The development of a tree structure allows for the representation of tests on characteristics as internal nodes, the representation of results as branches, and the indication of the class label as leaf nodes. In order to generate precise projections for novel data, this model acquires knowledge from antecedent training data [68].
- Support Vector Machines (SVMs) represent a type of supervised learning algorithm that is capable of performing classification or regression tasks on data. The primary aim of the analysis is to identify an optimal decision boundary, commonly referred to as a hyperplane, that effectively segregates data into distinct categories. SVMs exhibit a high level of efficacy in handling datasets with a large number of features in high-dimensional spaces using the kernel trick. This allows SVMs to identify a linear boundary for decisions in the transformed space that can effectively segregate the classes in the original data [69].
- The XGBoost algorithm, which is a highly potent form of supervised learning, is commonly employed for the purposes of classification and regression. This belongs to the class of gradient-boosting algorithms that combine several weak models, such

as decision trees, to generate a dependable predictive model. XGBoost is distinguished by its widespread adoption owing to its outstanding performance and capacity to effectively manage complex datasets with a high degree of precision [70]. The XGBoost algorithm is based on the principle of iteratively constructing a series of decision trees to form an ensemble. Gradient boosting is a technique in which each subsequent tree is formulated to correct the mistakes of the preceding trees. The process of optimizing the ensemble entails the computation of gradients or partial derivatives of a loss function specified by the user in relation to the predicted values of the ensemble. The subsequent trees are developed with the aim of mitigating this gradient, resulting in a gradual decrease in the aggregate error of the model (Figure 2.5) [71].

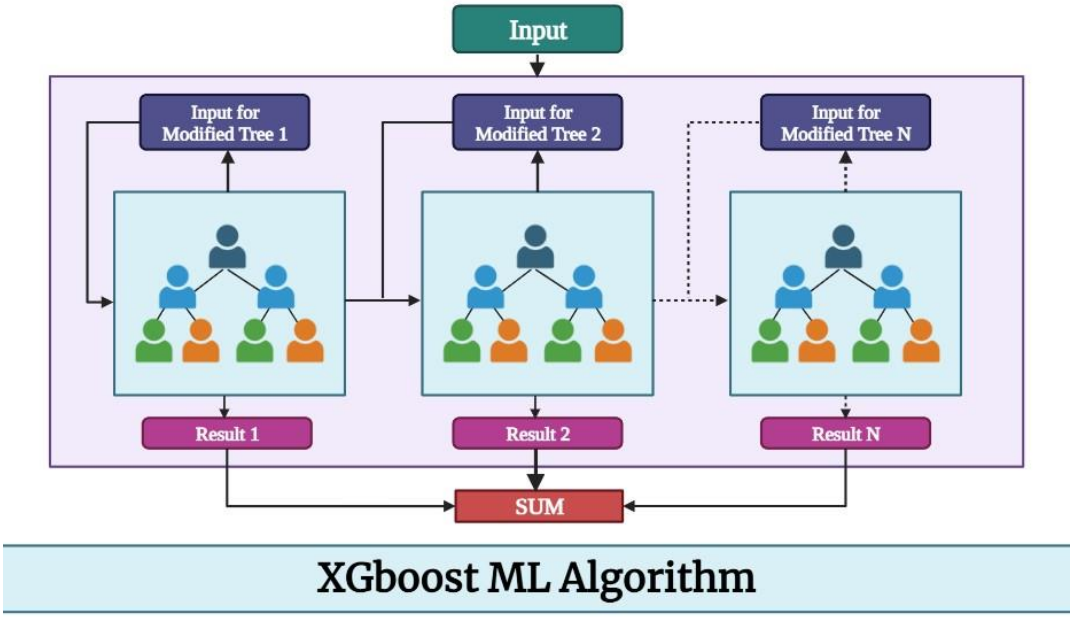


Figure 2.5: XGBoost ML Algorithm

XGBoost works on a range of crucial characteristics to enhance its efficacy and prevent overfitting. The optimization function incorporates L1 and L2 regularization components, which impose penalties on complex models and promote parsimony. This algorithm integrates gamma as a hyperparameter for regulating the minimum

loss reduction which is required to partition a leaf in a tree, leading to the elimination of extraneous branches.

The capability of XGBoost to effectively manage missing values is a significant benefit, as it can acquire knowledge on the optimal direction to allocate them during the tree-building process by applying accessible data without the need for manual accusation. Also, XGBoost offers valuable insights into the significance of features by quantifying the extent to which each feature contributes to minimizing the loss function across all trees. This feature is particularly useful for feature selection and comprehending the interrelationships between features and the target variable [72].

V. XGBoost ML classifier in the prediction of SCC biomarkers

The XGBoost algorithm is widely recognized as a significant ML technique that has proven to be advantageous in the realm of cancer investigation, specifically in the domains of therapy and diagnostics. The capacity to proficiently handle intricate and multi-dimensional datasets has established it as an essential resource for both scholars and healthcare practitioners. XGBoost can analyze complex biological data, enabling the detection of substantial patterns that enhance the precision of predictions. The contributions of XGBoost to this field have also facilitated progress in the development of treatments tailored to individual patients [73]. The importance of XGBoost in the biomedical domain (Figure 2.6) is underscored by the subsequent factors:

Figure 2.6: Applications of XGBoost ML Algorithm

1. **Feature selection and extraction:**

The XGBoost algorithm exhibits proficiency in analyzing large datasets that encompass numerous features, rendering it a fitting choice for cancer research. XGBoost can perform feature selection and extraction, thereby streamlining the research process by identifying crucial factors. By prioritizing the most informative variables, users can enhance their ability to make more accurate predictions [74].

2. **Early detection and diagnosis:**

Early detection considerably enhances the probability of successful cancer treatment. The application of XGBoost algorithms enables medical practitioners to assess patient data, encompassing genomic profiles, medical imaging, and clinical records, for the purpose of

detecting patterns and ascertaining the likelihood of cancer. This procedure expedites the timely detection of the condition, enabling expeditious execution of suitable therapies [75].

3. **Risk stratification and prognosis:**

XGBoost has the capability to perform risk stratification of patients by analyzing a multitude of factors, such as tumor characteristics, biomarkers, and patient demographics. This analysis enables the prediction of the likelihood of disease progression, recurrence, and patient survival. The acquisition of this data is paramount in the development of tailored therapeutic regimens and the judicious allocation of medical provisions [76].

4. **Drug response prediction:**

The XGBoost algorithm possesses the capability of using molecular and genetic data analysis in order to suggest the response of a patient toward cancer therapies. Incorporating discrete features such as genomic modifications and genetic expression patterns into XGBoost algorithms is pivotal in discerning individuals who may derive therapeutic benefits or experience unfavorable reactions to particular pharmacological interventions. The implementation of a beneficial approach can effectively optimize treatment options while simultaneously mitigating the occurrence of unfavorable side effects [77].

5. **Feature importance and interpretability:**

Through the elucidation of salient attributes within the model, XGBoost facilitates researchers' comprehension of the pivotal factors that underlie prediction. As a result, the aforementioned transparency can facilitate the selection of potential biomarkers and enhance our understanding of the fundamental mechanisms underlying cancer progression [78].

6. Ensemble learning and model performance:

XGBoost effectively harnesses the potential of ensemble learning to augment the precision and strength of the ultimate model by amalgamating numerous inferior predictive models. The implementation of this methodology amplifies the efficacy of cancer prognostication models, rendering them more reliable and appropriate for medical applications [79].

7. Real-time monitoring and decision support:

By examining information like vital signs and laboratory results, XGBoost algorithms can help medical practitioners monitor and make decisions about patients in real-time. With these features, clinicians can get alerts about possible cancer-related complications or shifts in a patient's condition, allowing them to act quickly. Consequently, the XGBoost models exhibit promising potential as valuable channels within healthcare environments [80].

VI. Machine Learning Model Evaluation Metrics

The metrics used to assess the efficacy of ML models are commonly referred to as ML model evaluation metrics. These metrics make it possible to evaluate various models and select the most suitable one for a particular task, assisting in the quantitative evaluation of a model's accuracy and efficacy. Commonly used measures for evaluating ML include accuracy, precision, recall, F1 score, and AUC-ROC (Area under the receiver operating characteristic curve). Through the quantification of diverse aspects of a model's performance, these metrics facilitate the assessment of its capacity to generate precise predictions, prevent erroneous positive and negative outcomes, and manage data that is unevenly distributed [81].

Evaluation metrics are an important part of ML because they let us compare different models in a fair way and figure out which one fits a certain problem the best. Also, this method permits us to identify areas in need of improvement and to adjust the models' parameters in order to

achieve superior outcomes. To measure the efficacy of an ML model, there are a variety of widely employed evaluation matrices available [82].

1. Accuracy

Accuracy is a commonly used metric for assessing the performance of ML models. This metric is computed by determining the proportion of the model's predictions that were accurate out of the total number of predictions made. It is frequently used as a fundamental evaluation method for classification models, for instance, predicting the classification of wildlife in an image. The mathematical expression for calculating this metric is:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)}$$

To calculate the accuracy of a model, the total number of accurate predictions is divided by the total number of predictions [83], [84]. Accuracy can be expressed in terms of percentage or a decimal value between 0 and 1, where 1 is perfect accuracy, and 0 is no accuracy. However, when evaluating model performance on imbalanced datasets, accuracy might not be the most meaningful metric to use. In these kinds of situations, precision and recall might provide more useful information.

2. Precision

The precision of an ML model can be evaluated by the ratio of true positive predictions to total positive predictions. Accurate predictions mean that the model has correctly found the presence of a positive class, while false positives happen when the model predicts a positive class in an image with a negative class. The concept of precision is frequently applied in scenarios where the occurrence of false positives is more detrimental than that of false negatives. This is particularly relevant in medical diagnoses, where the prediction of a disease in an otherwise healthy individual can result in severe implications. In such situations, it is imperative to

prioritize the avoidance of false positives to mitigate the risk of unnecessary medical interventions and potential harm to patients. The precision formula can be expressed as:

$$Precision = \frac{TP}{TP + FP}$$

Commonly, this metric is represented as a percentage or decimal value between 0 and 1, with 1 representing flawless precision and 0 representing no precision [83], [84]. The assessment of an ML model's performance solely based on precision may not offer a comprehensive overview. Therefore, it is frequently accompanied by other metrics, such as recall and the F1 score.

3. Recall

Recall is used to measure the performance of a classification model in ML. The objective of this metric is to evaluate the performance of a model's predictions with respect to positive instances. To be more precise, the model's true positive predictions are evaluated against a total of potential positive instances. In this case, a real positive prediction would mean the correct identification of a positive class, like an image of a cat. In contrast, if a model makes a prediction that an image does not contain a cat despite the presence of an actual positive class, it would be deemed as a false negative prediction.

In specific contexts, such as fraud detection, recall is commonly used to prioritize the avoidance of false negatives over false positives. When it comes to finances, a false negative can lead to significant losses, while a false positive might cause less harm overall. The recall of an ML model can be calculated by dividing the predicted number of true positives by the actual number of positive cases [83], [85].

$$Recall = \frac{TP}{TP + FN}$$

If the model accurately anticipated 75 positive cases out of a total of 100, its recall metric would be 75%. The metric of recall is often used in conjunction with other measures, such as precision and the F1 score, in order to obtain a more all-encompassing assessment of the effectiveness of the model.

4. F1 Score

The F1 score, which is the harmonic mean of precision and recall, offers a comprehensive evaluation of a model's efficacy. This calculation entails the utilization of the harmonic mean of precision and recall, which confers greater significance to lower values due to the typical inverse correlation between precision and recall.

The F1 score is considered valuable due to its ability to provide a straightforward and easily understandable metric that encapsulates the overall performance of a model. The value of a model becomes evident when there is a requirement for equal consideration of precision and recall in the evaluation process [85], [86]. In the field of medical diagnosis, it is imperative to ensure both accuracy, which prevents false positives, and completeness, which prevents false negatives; for this reason, the F1 score has emerged as a reliable performance metric. The F1 score is computed through the subsequent formula:

$$F1\ Score = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

5. AUC-ROC curve

The AUC-ROC curve is a commonly used visualization tool in the ML domain to evaluate the performance of a binary classifier. The acronym AUC-ROC denotes the "area under the receiver operating characteristic curve," a graphical representation of the relationship between true positive rate (TPR) and false positive rate (FPR). The computation of the True Positive Rate (TPR) involves the division of the number of positive instances that are correctly predicted

by the total number of positive instances that actually exist. On the other hand, False Positive Rate (FPR) is determined by dividing the number of positive instances that are predicted incorrectly by the total number of negative instances that actually exist.

In the event that a classifier has the ability to effectively distinguish between positive and negative classes, its AUC-ROC curve will manifest as a step function with a score of 1. This means that the true positive rate will be one, and the false positive rate will be zero. In contrast, when a classifier fails to distinguish between positive and negative classes, its AUC-ROC curve will manifest as a diagonal line with a value of 0.5, denoting that the true positive rate will be equivalent to the false positive rate [87].

The AUC-ROC curve provides a single-number summary that is helpful in assessing the effectiveness of a classifier. It is especially useful when comparing the performance of various classifiers, enabling the selection of the optimal solution for a given problem. For example, if two ML classifiers have respective AUC-ROC values of 0.75 and 0.85, the classifier with the higher value would be considered preferable. It is a common statistic used by experts in the field of ML when evaluating and choosing among several classifiers [88].

Applying evaluation metrics permits us to measure the efficacy of incorporating ML into cancer research. The utility of ML algorithms is being investigated for the identification of biomarkers for various cancer forms and the prediction of effective therapies against them (Figure 2.7). With the use of these algorithms, novel biomarkers for SCC can be found by analyzing large datasets to find patterns and relationships. This improves the personalization of treatment for SCC patients by allowing for more accurate forecasts of therapeutic outcomes.

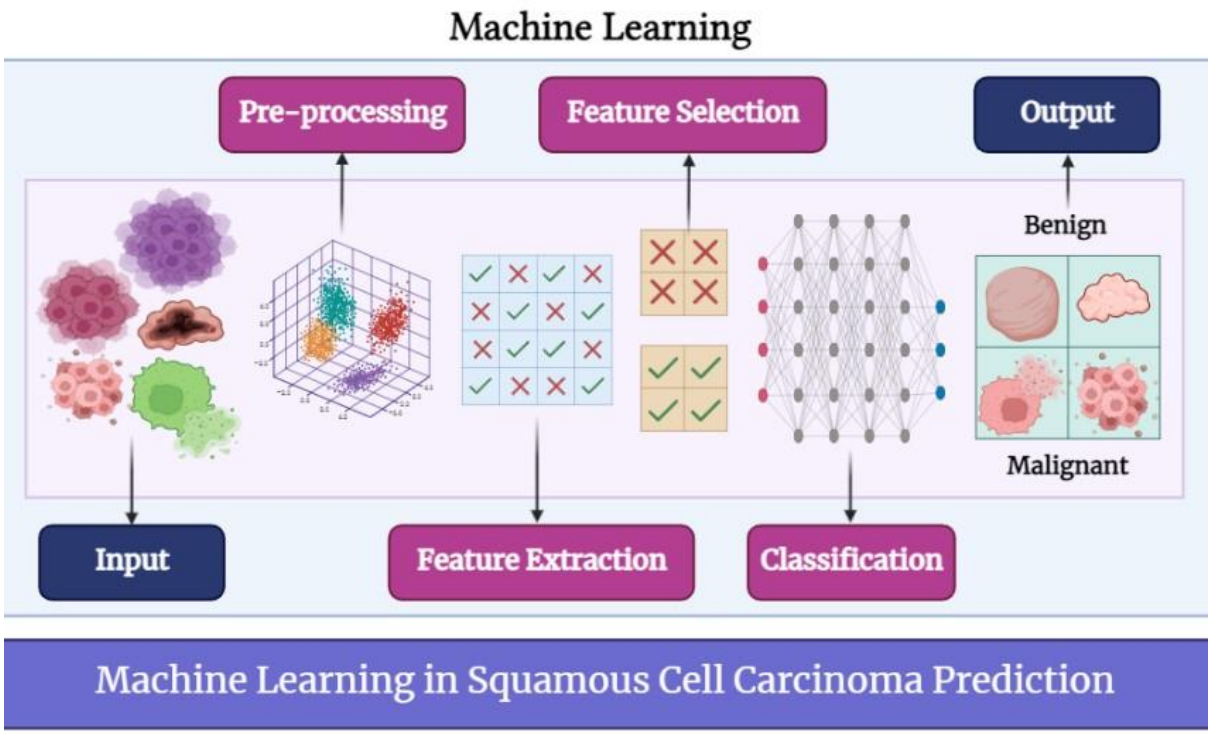


Figure 2.7: ML in SCC Prediction: Enhancing Accuracy through Machine Learning

VII. XAI for SCC Biomarker Identification: Enhancing Interpretability in ML Models

XAI is a set of techniques and methodologies used to enhance the interpretability and comprehension of ML models by humans. The primary objective of XAI is to provide insight into how a model arrives at a particular decision or prediction, thereby enhancing confidence and transparency in AI systems. XAI offers the opportunity to bridge the gap between black-box ML models and human comprehension by allowing users to understand a model's output factors and the reasoning behind them. SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-Agnostic Explanations) are well-known XAI techniques [89], [90].

1. SHAP (SHapley Additive exPlanations)

The SHapley Additive exPlanations (SHAP) method is a widely used technique in the XAI domain. Its primary purpose is to provide an interpretation of the predictions generated by ML

models. It is based on the idea of Shapley values from cooperative game theory and offers a framework independent of any particular model for comprehending the significance and contributions of features. The goal of SHAP is to quantify each feature's influence on the outcome and attribute it to the final prediction in order to explain the output of a black-box model. SHAP provides a fair distribution of importance across all features and their interactions by producing Shapley values, which indicate the average marginal contribution of each feature across distinct feature subsets [91].

SHAP offers the benefit of furnishing both local and global interpretations, which are valuable in comprehending the process of decision-making. The provision of local explanations facilitates comprehension of the influence of distinct features on the prediction of a given instance, whereas global explanations afford a comprehensive understanding of the significance of features across the entirety of the dataset. SHAP is versatile in its applicability to a range of model types, such as tree-based models, neural networks, and ensemble models. Also, it offers consistent and reliable explanations, irrespective of the underlying model architecture. This facilitates the establishment of generalizability and fosters confidence in the interpretive procedure [92].

2. LIME (Local Interpretable Model-Agnostic Explanations)

LIME is a well-known XAI technique used to provide explanations for ML model predictions. The concept behind LIME is to create an interpretable model that replicates the behavior of the original model close to a specific instance by modifying input features and observing the resulting prediction changes. LIME generates a locally accurate and interpretable model by fitting simple models such as linear regression or decision trees to perturbed instances and their respective predictions. LIME's model-agnostic nature allows it to be applied to different types of ML models, making it a versatile tool that can be used in various applications across

domains. Its flexibility makes it easy to use without requiring knowledge of internal architecture [93], [94].

Biomarker identification is essential for SCC diagnosis, prognosis, and treatment decisions. Researchers and medical professionals can comprehend the factors underpinning SCC development and progression with the aid of XAI techniques. By selecting an explainable ML model, such as XGBoost, researchers are able to evaluate the importance of various features in predicting SCC occurrence. Rule extraction algorithms and local interpretability methods such as SHAP facilitate the understanding of individual predictions and the identification of specific biomarkers implicated in SCC diagnosis, prognosis, and treatment decision-making. The interpretability provided by XAI fosters confidence in the predictions of the ML model and facilitates its implementation in clinical practice [1], [95].

VIII. Exploring Drug Repurposing for Innovative Therapeutic Solutions

Drug repurposing, also known as drug repositioning or drug reprofiling, is the process of identifying new therapeutic uses for existing drugs originally developed for different indications. Instead of commencing the process of drug development from scratch, researchers investigate the potential of existing drugs to treat other diseases or conditions. This strategy is gaining popularity in the pharmaceutical industry due to its ability to accelerate the development process, reduce costs, and provide innovative treatment options. To initiate the process of drug repurposing, researchers typically employ data mining, systematic screening, or serendipitous observations to identify promising drugs for a new therapeutic indication. Large databases of clinical, pharmacological, and genomic data are analyzed to identify pharmaceuticals that may be effective against specific targets or pathways pertinent to the new indication.

In contrast to traditional drug discovery and development, which can take years and cost billions of dollars, drug repurposing can significantly reduce the time and cost associated with drug development. Repurposing an existing drug has the added benefit of shortened development timelines, as many aspects, such as formulation, pharmacokinetics, and toxicity, are already known. In addition, the safety profiles of repurposed pharmaceuticals are well-documented because they have already been administered to humans for their original indications. Since safety data is already available, regulatory approval can be expedited, allowing the focus to shift to establishing efficacy for the new indication. Also, drug repurposing enables the discovery of new applications for drugs that have failed in their original indications or have limited market potential [96].

Metformin [97], a pharmaceutical agent predominantly employed in the management of diabetes, exemplifies the concept of drug repurposing, as it is currently under investigation for its prospective anti-cancer attributes. Thalidomide [98], originally utilized as a sedative and antiemetic for expectant mothers, has been repurposed for the treatment of leprosy and multiple myeloma, a form of cancer, owing to its anti-inflammatory characteristics. In both instances, the pharmaceutical agents were found to exhibit additional advantages beyond their initially intended therapeutic application [99], [100]. The preceding examples illustrate the possibility of repurposing existing pharmaceuticals for new therapeutic purposes, which can result in substantial clinical benefits. Repurposing medications enables scientists to utilize their prior safety precautions knowledge. This expedites the availability of new medications, potentially reducing patient and healthcare system costs [101]. With drug repurposing comes the possibility of treating a wide range of diseases and conditions, which could result in novel and efficient healing options for people all over the world.

IX. Molecular Docking and Dynamic Simulations

Using molecular dynamics (MD) is a valuable tool in drug repurposing as it allows researchers to investigate how molecules interact with target proteins. Molecular docking [102], a computational technique within MD, can predict and analyze binding interactions between a drug and a target protein by simulating the docking process. This technique helps researchers prioritize potential drug candidates by estimating their binding affinities and identifying those with the most favorable interactions. However, to obtain a more comprehensive understanding of the underlying mechanisms and dynamic behavior, researchers use molecular dynamics simulations that model the movements and interactions of individual atoms within a system over time using physical laws and mathematical equations [103].

Researchers can use molecular dynamics simulations [104] to observe the dynamic behavior, structural fluctuations, and conformational changes of both the protein and ligand. These simulations capture the interactions and movement of atoms, providing detailed information about the kinetics, thermodynamics, and energetics of the system. Molecular dynamics simulations are especially useful for drug repurposing because they can help researchers understand potential off-target effects and elucidate the mechanisms of action of repurposed drugs. By observing ligand-protein complex behavior, researchers can investigate how drugs interact with target proteins, how binding sites may undergo conformational changes, and how drug presence affects overall protein stability and function [105], [106].

Researchers can obtain a complete comprehension of repurposing candidates by merging molecular docking and molecular dynamics simulations (Figure 2.8). Molecular docking aids in the detection of possible drug candidates by analyzing their binding affinities. Meanwhile, molecular dynamics simulations offer a more profound understanding of the dynamic behavior, structural modifications, and mechanisms of action. This strategy enables researchers to investigate and assess the therapeutic capabilities of current drugs in a more comprehensive

manner, leading to the identification of new therapeutic applications through drug repurposing [107], [108].

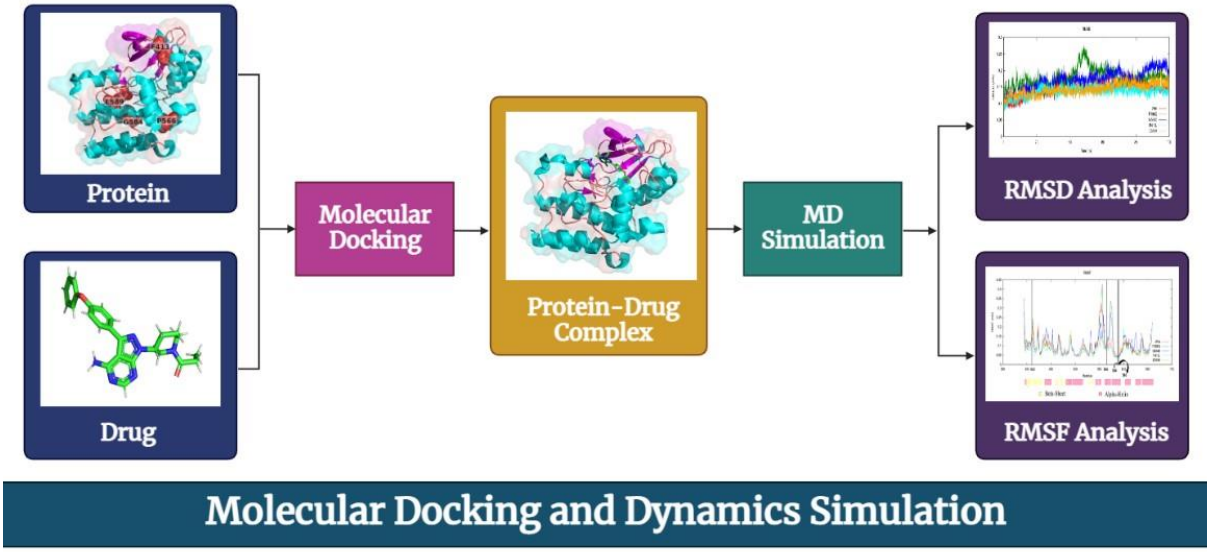


Figure 2.8: Molecular Docking and Dynamic Simulations: Investigating Molecular Interactions and Behavior

X. Unraveling Molecular Interactions: Contact Analysis in Molecular Dynamic Simulations

In molecular dynamics (MD) simulations, contact analysis is an important tool for examining the interactions between various molecular entities such as proteins, ligands, and protein-protein complexes. This method is useful for identifying specific residues or regions that come in contact during the simulation, which provides insights into potential binding sites, protein-protein interfaces, or critical interactions [109]. PyContact and CONAN are two commonly used tools for performing contact analysis in MD simulations.

1. PyContact

PyContact is a software package created using Python that simplifies contact analysis by computing different structural and dynamic features from MD trajectories. It has several metrics that can be used to evaluate contacts, such as hydrogen bonds, residue-residue

distances, hydrophobic interactions, and salt bridges. PyContact utilizes an adaptable and easy-to-use framework for the analysis and visualization of contacts, allowing researchers to determine significant interactions and obtain mechanistic knowledge from MD simulations. Its user-friendly interface has made it prevalent in the scientific community as it enables researchers to perform contact analysis quickly [110].

2. CONAN (CONtact ANalysis)

CONAN, a command-line utility, is a tool utilized in molecular dynamic simulations for contact analysis. Its primary focus includes quantifying protein-protein contacts and interactions by computing various properties such as contact frequency, lifetime, and strength. The software also provides numerous customization options, including contact definitions and analytical choices to cater to the diverse research requirements. Researchers find CONAN helpful in identifying critical residues involved in protein-protein interfaces and interactions [111].

The use of contact analysis can assist in comprehending the behavior and interactions of biomolecular systems. By analyzing contacts formed during MD simulations, significant information about structural and functional aspects can be obtained. Researchers can identify crucial residues, explain binding mechanisms, and gain a better understanding of molecular processes through contact analysis. This knowledge can be pivotal in drug discovery, protein engineering, and comprehending biological functions at a molecular level [112].

XI. Drug Gene Budger

Drug Gene Budger (DGB) is a web and mobile-based application that assists in the prioritization of drugs and small molecule compounds based on their ability to influence the expression of a target gene. It employs datasets such as LINCS L1000 [113] and CMap [114], which characterize the transcriptomic response of human cell lines to numerous small

molecules, such as FDA-approved drugs and preclinical compounds. DGB incorporates and analyzes these datasets, as well as drug-induced gene expression signatures curated from sources such as Gene Expression Omnibus (GEO) database [115] via crowdsourcing initiatives such as CREEDS [116]. DGB seeks to identify small molecules that significantly affect the expression of a specific target gene by employing these exhaustive collections of drug-induced transcriptomic signatures. A user-friendly interface of DGB allows users to select their target gene of interest and interact with the ranked list of small molecules generated as query results. It is a valuable tool for researchers and scientists engaged in drug discovery and personalized medicine research, as it enables the prioritization of drugs and small molecules based on their capacity to modulate the expression of a specific target gene [117].

XII. Conclusion

The present literature review was centered on the identification of prospective biomarkers and therapeutic interventions for SCC through an exploration of genes that have been found to be dysregulated. This was achieved by looking at the involvement of these genes in the development of SCC and highlighting the significance of using ML approaches for biomarker discovery. In order to give ML models interpretability, the use of XAI techniques was investigated to improve the accuracy of SCC biomarker prediction. The review also examined drug repurposing methods, molecular docking, and simulation techniques as plausible approaches in cancer research. These methods have the potential to identify novel biomarkers and treatment options, ultimately leading to enhanced patient outcomes.

Chapter 3

Application of Explainable Artificial Intelligence in the Identification of Squamous Cell Carcinoma Biomarkers

---

Chapter 3. Application of Explainable Artificial Intelligence in the Identification of Squamous Cell Carcinoma Biomarkers

Abstract

NMSCs are the fifth most common type of cancer worldwide, affecting both men and women. Each year, more than a million new occurrences of NMSC are estimated, with SCC representing approximately 20% of all skin malignancies. The purpose of this study was to find potential diagnostic biomarkers for SCC by application of XAI on XGBoost ML models trained on binary classification datasets comprising the expression data of 40 SCC, 38 AK, and 46 normal healthy skin samples. After successfully incorporating SHAP values into the ML models, 23 significant genes were identified and were found to be associated with the progression of SCC. These identified genes may serve as diagnostic and prognostic biomarkers in patients with SCC.

I. Introduction

Skin cancers are commonly divided into two categories: firstly, NMSC, which includes SCC and Basal cell carcinoma (BCC), and secondly, melanoma skin cancers [118]. NMSC is the world's fifth most prevalent form of cancer, affecting both men and women. In the United States, over 1.8 million new cases of NMSCs are reported each year, with cutaneous SCC being the most frequent kind of skin cancer. [119], [120]. African Americans and Asian Indians have a higher incidence of SCC, and also, it ranks the second most prevalent among Hispanics and Chinese/Japanese Asians. [121]. SCC has been recognized as a kind of cancer that originates in keratinocytes. The skin ailment AK, also known as Carcinoma-in-Situ, has been associated with the emergence of SCC in numerous studies. However, a considerable number of high-risk SCC cases, approximately 5-10 percent of all instances, are exceedingly difficult to diagnose and treat, necessitating the use of radiation or surgery in the majority of cases. It is less likely that therapies associated with such high-risk metastatic skin cancer will be effective, particularly in an elderly population [122] in critical need of a promising yet systematic diagnosis and treatment for SCC [123], [124]. Microarray data is growing in volume, and the information it gives on the genes responsible for a disease phenotype is being used more and more for variant categorization and analysis, as well as other applications. Microarrays are a relatively recent method that involves the placement of hundreds of DNA probes that are matched to target genes on a tiny chip that can then be used to analyze gene expression in samples. One of the primary applications of this approach was to compare cancer and normal tissues, as well as distinct cancer subtypes and individuals with varying prognoses, among other things [125], [126]. When it came to identifying microarray samples, the widely used ML technique of support vector machines (SVMs) [127], artificial neural network [128], logistic regression, naïve Bayes, etc., worked admirably. In a large number of studies, metabolomic data is used to gain insight into the metabolites that define each organism's state and the

dynamics of those metabolites under various settings. The ‘omics’ domain is a critical component of systems biology. Due to its emphasis on small molecules and interactions, it has gained widespread adoption in a variety of fields recently, such as biomarker discovery and identification, development of drugs, customized health care, etc. [129]. Some pioneer studies on omics data have made normalization tools like NOREVA [130], [131], [132] and ANPELA, an integrated workflow for Label-free quantification (LFQ) [133] of data. These tools have made significant contributions to numerous facets of scientific investigations.

Technological advances like Next Generation Sequencing (NGS), Genome-wide association studies (GWAS), and computational methods have expanded the scope of precision medicine and diagnostics by enabling the cost-effective analysis and integration of clinical data to examine tumor genomes, transcriptomes, and so on [134], [135]. Additionally, single-cell sequencing enabled the identification of key cancer driver genes, paving the way for personalized cancer management [136]. Most of the research are currently going on to find what causes SCC to create biomarkers that may be used to develop more precise methods of diagnosis and treatment [137]. Differential expression analysis is a widely used technique for determining how a gene is controlled and whether or not it is associated with a particular condition. It is a statistical technique that uses normalized read count data to determine quantifiable differences in the level of expression between the experimental and control group. Statistics are used to establish whether an observed variation in read counts for a particular gene is statistically substantial, that is, more than what would be predicted from natural random fluctuation. In addition to edgeR and DESeq (based on Negative Binomial (NB) distributions) and baySeq and EBSeq (Bayesian techniques based on an NB model), there are numerous other methods for differential expression analysis [138], [139]. Computational strategies have been applied in bioinformatics research for nearly three decades to help in the study of molecular processes and the development of novel medical interventions for a variety of disorders [140].

Deep learning algorithms have witnessed a spike in popularity in recent years in fields such as omics data analysis, sequence data analysis, biomedical imaging, and signal processing, where they have proven extraordinary performance [141], [142].

Using ML algorithms, previous studies have identified crucial biomarkers in the search for genes with greater predictive value for SCC, thereby facilitating the identification of biomolecules with higher predictive value [143]. Researchers are currently utilizing AI-based ML techniques to investigate the genetic variability of cancers, which can be used to enhance the accuracy of cancer diagnosis, the development of potent biomarkers, and the success rate of cancer therapies [144]. Regarding robots, AI refers to their capacity to imitate human behavior, a feature that proves especially advantageous in processing vast quantities of information. ML is a crucial application of AI that enables computer systems to acquire knowledge from their own unique encounters without requiring explicit programming [145], [146]. ML models can be conceptualized as a modeling technique that entails a buildup of knowledge and the improvement of performance. The purpose of these models is to facilitate the recognition of advantageous components and their interrelationships [147]. In recent years, AI has made significant progress, transitioning from a primarily theoretical concept to a practical, application-focused field. The utilization of AI across various domains is presently linked with elevated prospects, particularly in the realm of cancer research. ML has already been employed to investigate survival rates and prognostic models for pancreatic cancer, advanced nasopharyngeal carcinoma, breast cancer, and several other types of cancer [144], [148]. AI algorithms, specifically those utilizing ML, have demonstrated efficacy in producing accurate outcomes and predictions. However, these algorithms suffer from a lack of transparency, which impedes understanding of their fundamental operational processes. This opacity presents a significant challenge, as relying on a system that cannot provide self-explanation poses considerable risks when making critical decisions. The concept of XAI

proposes a fundamental change in the approach to AI, with the aim of achieving greater transparency and comprehensibility to address this challenge. This study aims to develop a set of strategies that lead to the improvement of more comprehensible models while simultaneously maintaining a high level of performance.

1. What is Explainable AI (XAI)?

XAI is a field of research that seeks to make the outcomes of AI systems more human-comprehensible. Recently, academia and practitioners have rekindled interest in the concept of XAI. The term ‘explainable’ or ‘explainability’ refers to a model’s ability to rationalize its output. Additionally, explainability ensures the model’s compliance with accuracy and refers to the complete and precise representation of a model’s output. Explainability can be classified as either local (for a single instance of a decision) or global (for comprehending the model’s decision-making mechanism). The growing significance of explainability underscores the critical nature of tools that assist humans in comprehending the behavior of black-box models [149]. SHAP (SHapely Additive Explanation), LIME (Local Interpretable Model-agnostic Explanation), ELI5, AIX360, and Skaters are just a few of the XAI frameworks available, with SHAP and LIME being the most widely used and interoperable with any deep learning or ML model.

SHAP stands for **SHapley A**dditive ex**P**lanations and is an open-source package that determines if an ML model is trustworthy or not. SHAP evolved from Lloyd Shapley’s 1951 presentation of Shapley values as a solution paradigm for cooperative game theory. SHAP employs game theory in an understandable manner to create a link between optimal credit apportionment and local explanations by using standard Shapley values for a typical model explanation. In brief, SHAP is a wonderful state-of-the-art ML explainer that helps in reverse-engineering any prediction algorithm’s results. It is generally used for a complex model, like in the case of gradient boosting, deep neural networks, etc., to better understand the decisions

that have been made by the model to check its correctness and faithfulness. From here, SHAP values come into existence; the calculation of these values means a lot to explain and interpret the outcomes of the model. SHAP values originated from Shapley values of game theory, which is a concept that employs 'game,' which represents the outcome of the prediction model, and 'players' represent the features of the model. Shapley values play a significant role in quantifying the performance of the players in the betterment of the game equivalently, SHAP values quantify the performance of each feature that contributes to the decision-making of the prediction model locally. SHAP identifies two estimation approaches, KernelSHAP and TreeSHAP, for Shapley values, where KernelSHAP is for local surrogate models to explain black-box ML model predictions and TreeSHAP for complex models based on trees [150], [151].

2. SHAP values explained

SHAP is a cooperative game theory-based approach for determining Shapley values, and its main objective is to compute each attribute's contribution to the forecast of an incident 'x' to account for it. Shapley values provide guidance on how to distribute the prediction evenly across the attributes. The Shapley value explanation is portrayed as a strategy for attributing additive features, a linear model, which is an innovation of SHAP. '$\emptyset s$' can be calculated with the help of a linear cooperative model, where 'x' represents a vector for all the feature values which are present in the model. According to SHAP, the following is the explanation:

$$g(z') = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i z'_i$$

Where 'g' denotes the model of explanation; the coalition vector of game theory is denoted by $z' = \{0,1\}^{\boldsymbol{M}}$; the coalition size is denoted by 'M,' and the feature attribution for a feature 'i' is denoted by $\emptyset_i \in \mathbb{R}$. Efficiency, Symmetry, Dummy, and Additivity are all satisfied by

Shapley values. SHAP computes Shapley values; hence it meets both. Discrepancies between SHAP and Shapley characteristics are found in the various SHAP studies. SHAP identifies local accuracy, missingness, and consistency which are the three desirable characteristics.

Property 1: Local accuracy

$$\hat{f}(x) = g(x') = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i x'_i$$

If we put $\emptyset_0 = E_X\left(\hat{f}(x)\right)$ and all x'_i are set to 1, then this will become the Shapley efficiency property using the coalition vector.

$$\hat{f}(x) = \emptyset_0 + \sum_{i=1}^{M} \emptyset_i x'_i = E_X\left(\hat{f}(X)\right) + \sum_{i=1}^{M} \emptyset_i$$

Property 2: Missingness

$$x'_i = 0 \Rightarrow \emptyset_i = 0$$

According to missingness, a feature that is not present is assigned a value of zero. Take note that x'_i denotes coalitions, with a value of *'0'* denoting the lack of a feature value. In cooperative game theory terminology, all feature values x'_i of the instance to be analyzed must be *'1'*. The existence of a *'0'* implies the fact that the instance's feature value is absent. In contrast to "regular" Shapley values, this trait does not appear on their list of characteristics. In the words of Lundberg, it is a "small book-keeping feature". Because it is multiplied by $x'_i = 0$, a missing feature might theoretically have any Shapley value it desired without impairing the local accuracy property of the feature set. The Missingness property ensures that features that are not present receive a Shapley value of zero. When it comes to practical application, this is significant only for continuous traits.

Property 3: Consistency

If $\hat{f}_x(z')$ denotes $\hat{f}(h_x(z'))$ and $z'_{\setminus i}$ denotes that $z'_i = 0$. For any binary model *f* and *f′* that meet the following condition:

$$\hat{f}'_x(z') - \hat{f}'_x(z'_{\setminus i}) \geq \hat{f}_x(z') - \hat{f}_x(z'_{\setminus i})$$

If $z' \in \{0,1\}^M$ is true for all inputs, then,

$$\emptyset_i(\hat{f}', x) \geq \emptyset_i(\hat{f}, x)$$

In terms of the consistency property, if a model is adjusted in such a way that the relative contribution of a feature value rises or remains constant (independent of the contribution of other features), the Shapley value similarly increases or stays constant. Consistency contributes to the Shapley characteristics of Linearity, Dummy, and Symmetry, which are all derived from it [150], [152].

The purpose of this work was to use *an ML-based approach that uses the XAI method* for elucidation of the molecular mechanism by which AK progresses to SCC and to identify important genes associated with SCC, which may give novel diagnostic options for SCC management. Differentially expressed genes from SCC samples were compared to healthy skin samples and AK samples to elucidate the molecular biological processes behind SCC.

II. Materials and methods

1. Data retrieval

The normalized, calibrated, and pre-processed array data for AK and SCC were obtained from the GEO database available at NCBI. Three datasets were identified using the search terms SCC and AK: GSE45216, GSE98744, and GSE108008. The GSE45216 collection has 30 SCC and 10 AK samples, while the GSE98774 collection contains 18 AK and 36 normal healthy

skin samples. There are 10 SCC, 10 AK, and 10 normal healthy skin samples in the GSE108008 dataset.

2. Data Preprocessing

Given that the dataset was constructed by integrating three GEO datasets, its quality must be verified. RMA (Robust MultiArray Average) normalization method was applied for microarray summarization and quantile normalization of the datasets. We have performed log2-transformation and quantile normalization on the expression data to draw boxplots and expression density plots for both normalized and non-normalized data [153]. Finally, we identified around 10,000 common genes that were present in all three datasets; consequently, we merged the expression profiles for these 10,000 common genes. PCA provides the visualization of variables' correlations and the identification of clusters of comparable data. The resulting dataset had 10,000 genes classified into three types (Healthy, Actinic Keratosis, and Squamous Cell Carcinoma). The dataset was subsequently separated into three binary classification problems, namely Healthy vs AK, Healthy vs SCC, and SCC vs AK, with each case treated separately. Each dataset was subjected to PCA using the scikit-learn package of python to assess whether or not the sample groups separated based on the variance of gene expression in two major components, hence determining the dataset's quality.

3. Machine learning on the datasets

The datasets were randomly apportioned in an 80:20 ratio into training and testing sets. ML techniques like SVMs, KNNs, deep learning, etc., gained popularity recently in disciplines such as omics data analysis, sequence data analysis, biomedical imaging, and signal processing [140], [141], [142], so we chose to conduct ML on our datasets. The training sets were used to train three XGBoost models for the classification of three datasets. The XGBoost algorithm (Extreme Gradient Boosting) is an ML technique based on decision trees that optimizes

performance via a process known as boosting. Since its inception, it has consistently outperformed the majority of other ML methods, including logistic regression, the random forest model, and conventional decision trees. XGBoost frameworks are available for a variety of programming languages, most notably Python, and it interacts well with the popular scikit-learn ML framework used by Python data scientists. After the application of the XGBoost ML classifier using Scikit-learn library on the datasets, the testing sets were used to evaluate the models' performances. Models were evaluated in terms of the confusion matrix and accuracy of the model calculated using the test set.

4. Explainable AI (XAI) on the trained ML models

The XAI analysis on the trained XGBoost models was performed using the Python SHAP (SHapley Additive exPlanations) package. The XAI analysis probes into the process of decision-making by the ML model and assists in identifying the features that contribute significantly to the model's prediction confidence. Thus, XAI analysis will aid in identifying relevant genes from which trained models may identify/classify the phenotype/condition, such as Healthy, AK, or SCC. A local summary plot was built to illustrate SHAP values, where values indicate the feature's contribution to decision confidence. Additionally, a SHAP summary plot was created to show the global feature relevance derived from the training data. We chose the top 14 genes with the highest average SHAP value as features and utilized them for training new XGboost models. The performance of newly trained models was compared to that of XGBoost models trained on 10,000 genes to validate the significance of the selected 14 genes. To check the robustness of the model, we have evaluated the model performance on an independent test set retrieved from GEO, accession no. GSE32628. The data for independent testing was classified into three datasets (Healthy vs AK, Healthy vs SCC and SCC vs AK dataset). Preprocessing of raw data was done using the RMA normalization method, as discussed earlier, and the XGBoost ML classifier was applied to find the accuracy of each class.

Further, a statistical analysis was conducted on the candidate genes derived through the application of SHAP values on ML models to find their significance. The GEO2R tool was used to find the statistical significance of identified key genes. Genes with a P-value (False discovery rate) <0.05 were considered statistically significant, and on their basis, we have checked the expression of each gene for SCC progression. Additionally, a critical examination of the literature was conducted to ascertain the significance of the genes with the highest average SHAP value in the progression of AK and SCC from healthy skin cells.

III. Results

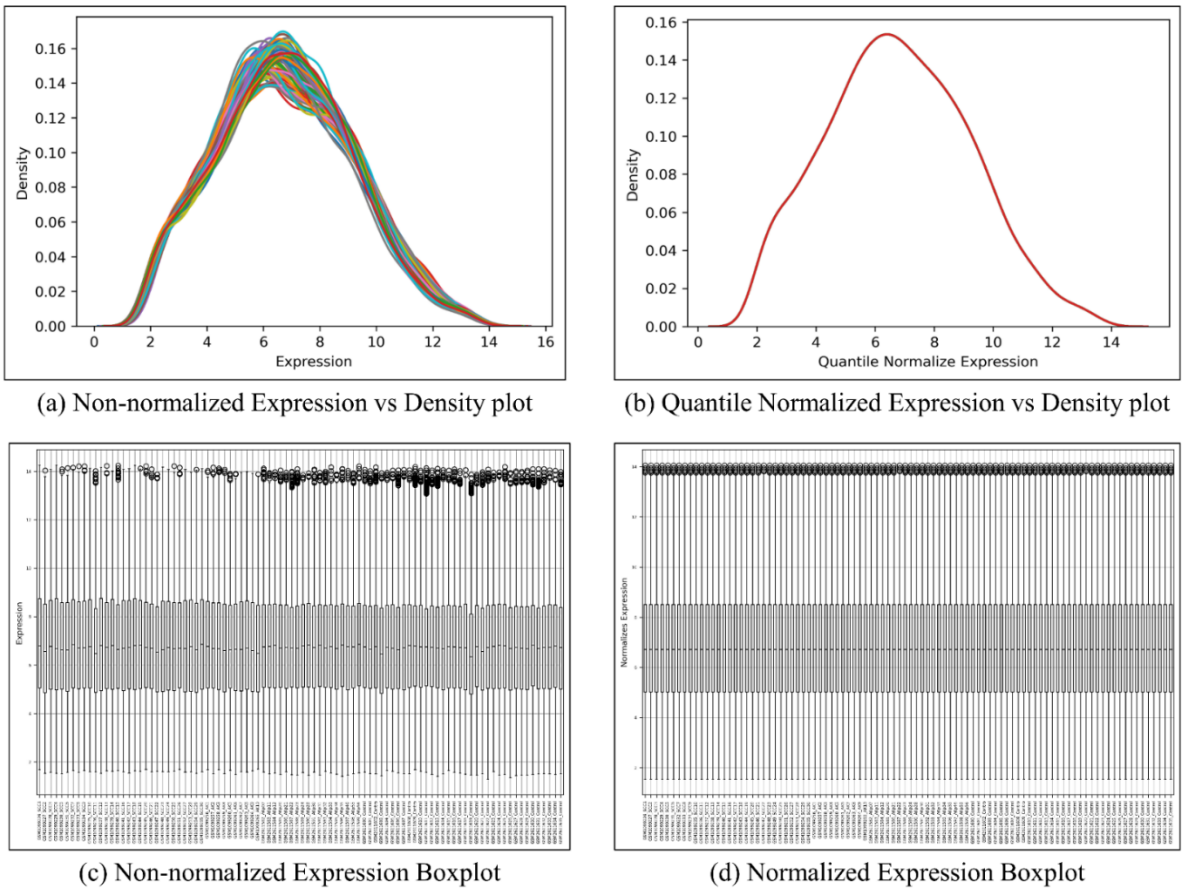
The normalized, calibrated, and pre-processed array data for SCC were obtained from the GEO database available on NCBI, the description given in (Table 3.1). All these datasets were retrieved using the search terms AK and SCC. These datasets are then merged based on common gene symbols. Ten thousand common genes with their expression values were employed as features, and three classes of phenotypes were defined, namely AK, SCC, and Control.

Table 3.1: Microarray data description with their GEO accession number, number of samples in each series, sample type, sample size and the platform.

GEO Accession Number	No. of Samples	Sample Type	Sample Size	Platform
GSE45216	40	Actinic Keratosis	10	GPL570
		Squamous Cell Carcinoma	30	
GSE98774	54	Actinic Keratosis	18	GPL570
		Healthy Skin	36	
GSE108008	30	Healthy Skin	10	GPL16686
		Actinic Keratosis	10	
		Squamous Cell Carcinoma	10	

1. Data Preprocessing

The datasets were normalized using RMA (Robust MultiArray Average) normalization. To remove biases from the expression data, we used log2 transformation and quantile normalization to create the normalized expression boxplot and density plot (Figure 3.1). Box plots show the expression distribution in each array, while the Density plots show the expression density distribution in each array's color channel.



(a) Non-normalized Expression vs Density plot

(b) Quantile Normalized Expression vs Density plot

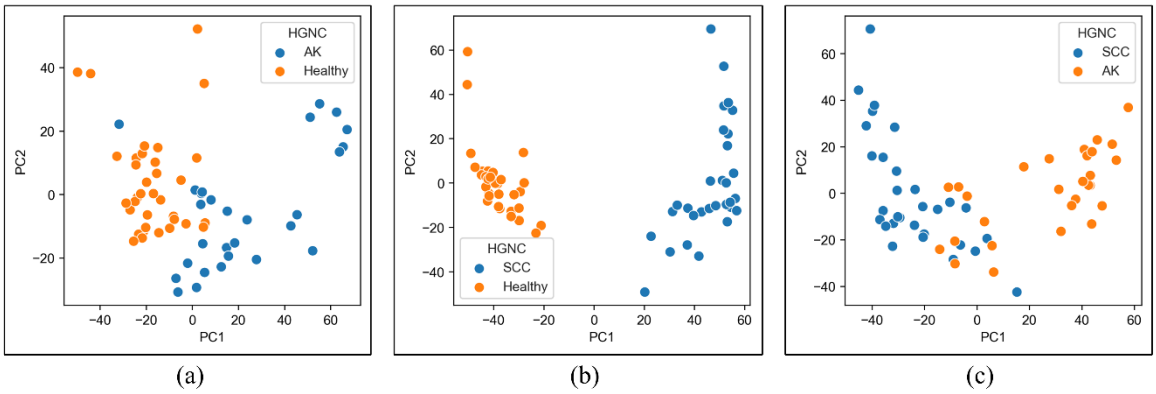
(c) Non-normalized Expression Boxplot

(d) Normalized Expression Boxplot

Figure 3.1: An overview of RMA Normalization. Density plots (a and b) show the expression density distribution in each array's color channel, while the Box plots (c and d) show the expression distribution in each array before and after doing RMA normalization.

A PCA analysis was conducted on the datasets categorized into three binary classification problems, namely Healthy vs AK, Healthy vs SCC, and SCC vs AK, by using a scikit-learn package of python. PCA was done to verify the quality of data and to ensure that our data is

well grouped on the basis of variance among the features, scatter plots were made out of the information collected from PCA analysis, as shown in (Figure 3.2). PCA was applied to integrate highly correlated variables into a smaller collection of variables that account for the majority of the variance in the data. Here, the results of PCs scatter plots describe the classes that are well grouped, and ML can be applied for the classification of the data.



(a) (b) (c)

Figure 3.2: Principal Component Analysis plots for (a) Healthy vs AK dataset, (b) Healthy vs SCC dataset, (c) SCC vs AK dataset. Segregation was observed for both modes between Healthy and AK, Healthy and SCC, and finally, SCC and AK individuals.

2. Machine learning on datasets

For ML purposes, the datasets were prepared and randomly divided in an 80:20 ratio into training and testing sets. In our study, the XGBoost ML algorithm was implemented using the Scikit-learn library. The training sets were used to train three XGBoost models for classification on three different datasets that were separated into three binary classification problems, namely Healthy vs AK, Healthy vs SCC, and SCC vs AK, with each case treated separately. The models' performance was then assessed using the testing sets. Models were evaluated in terms of the confusion matrix and the accuracy of the model generated using the test set. We used accuracy as a predicted performance indicator to assess the performance of an ML classifier. Accuracy is a parameter that can be used to evaluate various classifiers; it

refers to the model's percentage of correct predictions. For binary classification, accuracy can also be stated in terms of positives and negatives. as described by the following equation:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

For our datasets, the XGBoost ML classifier has given a predicting accuracy of 100% for the Healthy vs AK dataset, 92.86% for the Healthy vs SCC dataset, and 91.67% for SCC vs AK, showing that our ML prediction model has performed well in distinguishing the features namely healthy, AK and SCC. The number of instances in each dataset with their computed accuracy by our ML algorithm is compiled in (Table 3.2).

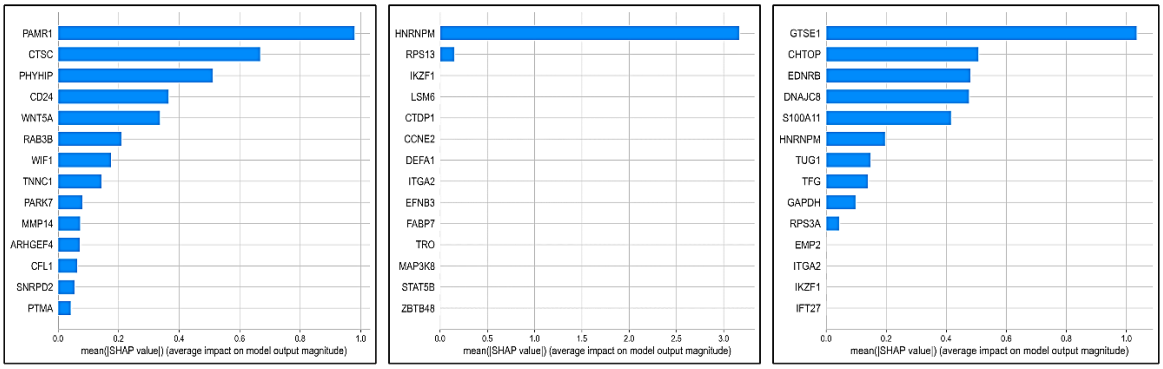
Table 3.2: Performance evaluation of XGBoost ML classifier for each dataset in terms of Accuracy percentage.

Datasets	Number of Instances	Accuracy %
Healthy vs AK	46 Healthy	100
	38 AK	
Healthy vs SCC	46 Healthy	92.86
	40 SCC	
SCC vs AK	40 SCC	91.67
	38 AK	

3. Explainable AI on the Trained ML models

XAI analysis was performed on the trained XGBoost models using the Python SHAP (Shapley Additive exPlanations) package. The XAI analysis delves into the process of decision-making by the model and assists in identifying features that contribute greatly to the model's prediction confidence. Thus, XAI analysis assisted in discovering relevant genes from which trained models were able to identify/classify the phenotype/condition, such as Healthy, AK, or SCC. A local summary plot was created to display SHAP values, which represent the contribution of a feature to decision confidence.

By passing an array of SHAP values to the bar plot function, a global feature importance plot was created, where the global significance for every gene is defined as its average/mean absolute value over all of the given samples. This feature importance plot helped us in revealing which genes are the most significant in descending order. The top genes contribute more to the ML model's prediction than the bottom genes and so have high predictive power. The bar plots, as shown in (Figure 3.3), depict the genes of utmost importance placed on the top and the genes of least significance at the bottom. *PAMR1* (of Healthy vs AK); *HNRNPM* (of Healthy vs SCC) and *GTSE1* (of SCC vs AK) are the genes of high predictive value and are the most significant in our ML prediction model.



(a) (b) (c)

Figure 3.3: SHAP Barplot depicting the genes of highest relevance on top for (a) Healthy vs AK dataset, (b) Healthy vs SCC dataset, (c) SCC vs AK dataset.

From the given bar plots, genes of highest importance were driven by implementing SHAP values on the trained models, showing *PAMR1, CTSC, PHYHIP, CD24, WNT5A, RAB3B, WIF1, TNNC1, PARK7, MMP14, ARHGEF4, and CFL1* to be the most significant genes in Healthy vs AK dataset; *HNRNPM* and *RPS13* in Healthy vs SCC dataset and *GTSE1, CHTOP, EDNRB, DNAJC8, S100A11, HNRNPM, TUG1, TFG, GAPDH* and *RPS3A* in SCC vs AK dataset.

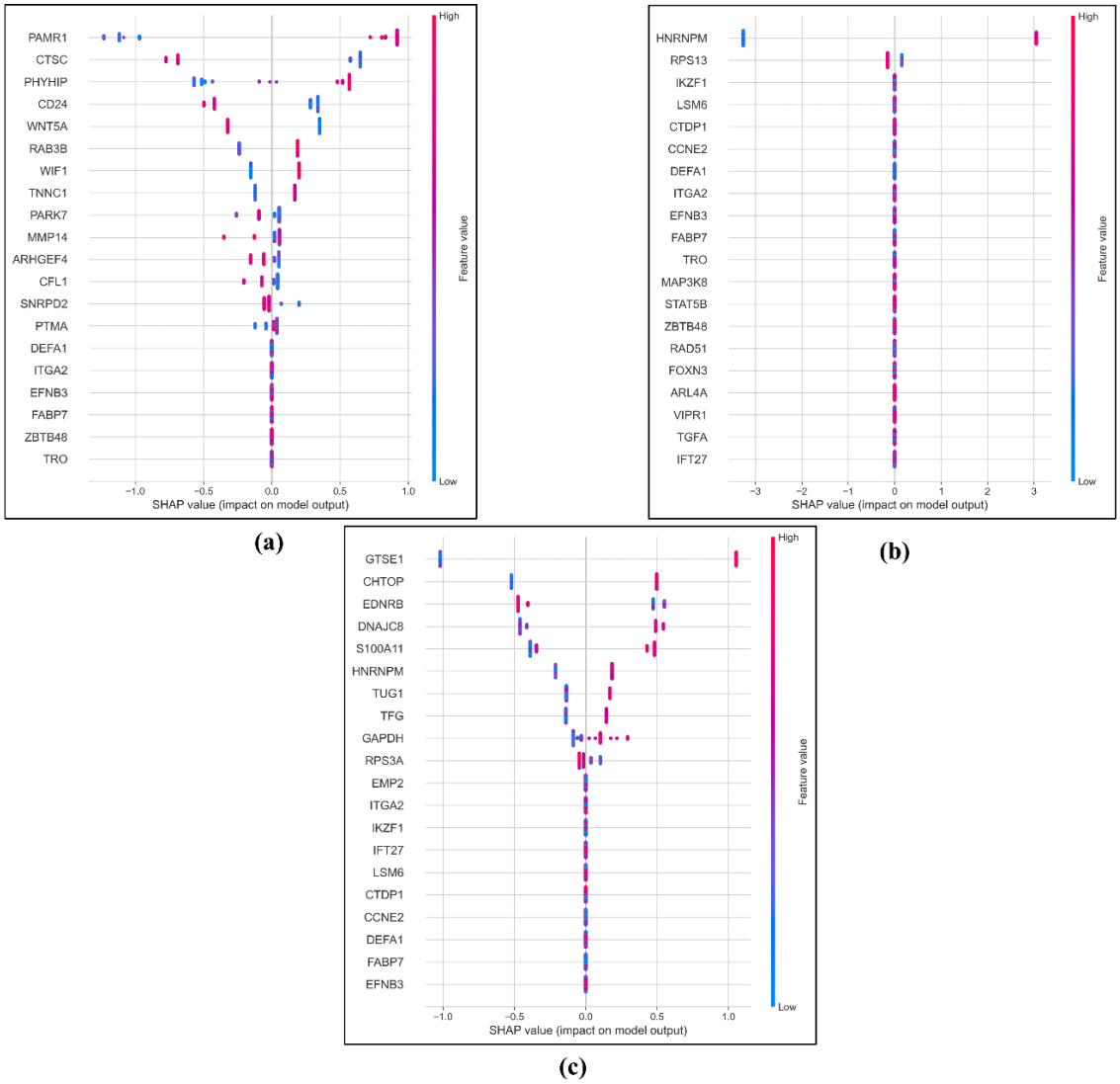


Figure 3.4: SHAP Summary plot depicting the most important genes and their impact in (a) Healthy vs AK dataset, (b) Healthy vs SCC dataset, (c) SCC vs AK dataset.

Additionally, the SHAP summary plots were made and used to visualize the predictors' positive and negative associations with the target gene (Figure 3.4). These SHAP summary plots illustrate the following points: Genes are ranked descendingly according to their feature importance; the horizontal location indicates whether the effect of a gene is related to greater or reduced prediction, indicating its impact on the model output; the color indicates whether the effect of a particular gene is significant (in red) or minimal (in blue) for that observation; a high level of '*PAMR1*' has a strong positive impact on the quality rating, indicating

the correlatedness of that particular gene. The "high" is shown by the red color, while the "positive" influence is indicated by the X-axis. Similarly, we would state that the "*CTSC*" is inversely connected to the target variable. From the following SHAP summary plots, we have inferred that *'PAMR1*', '*HNRNPM*', and '*GTSE1*' are the most significant genes in all the datasets and have a high and positive impact on models' predictions. While '*CTSC'*, '*RPS13*', and '*EDNRB*' are negatively correlated with models' predictions.

4. Evaluation of XAI output

To quantify the ML output, we applied XAI on the dataset and found the most important genes that are involved in the progression of SCC, the list given below in (Table 3.3).

Table 3.3: List of significant genes in each dataset after applying the SHAP values on the XGBoost ML classifier.

Datasets	Significant Genes
Healthy vs AK	*PAMR1, CTSC, PHYHIP, CD24, WNT5A, RAB3B, WIF1, TNNC1, PARK7, MMP14, ARHGEF4, CFL1*
Healthy vs SCC	*HNRNPM, RPS13*
SCC vs AK	*GTSE1, CHTOP, EDNRB, DNAJC8, S100A11, HNRNPM, TUG1, TFG, GAPDH, RPS3A*

To further check the authenticity of the results, we have applied the ML classifier XGBoost on the top 14 genes of each dataset again. The accuracy of each dataset (Table 3.4) helps us to check the performance of our predicted genes and to show their effect on our ML model. The confusion matrix of each dataset, namely Healthy vs AK, Healthy vs SCC and SCC vs AK is shown in (Figure 3.5).

Table 3.4: Comparison of the accuracy before and after the calculation of SHAP values on the XGBoost ML classifier for a 10000 gene set as well as 14 gene set.

Datasets	Accuracy % on 10000 gene set (Before calculating SHAP values)	Accuracy % on 14 gene set (After calculating SHAP values)
Healthy vs AK	100	100
Healthy vs SCC	92.86	92.86
SCC vs AK	91.67	91.67

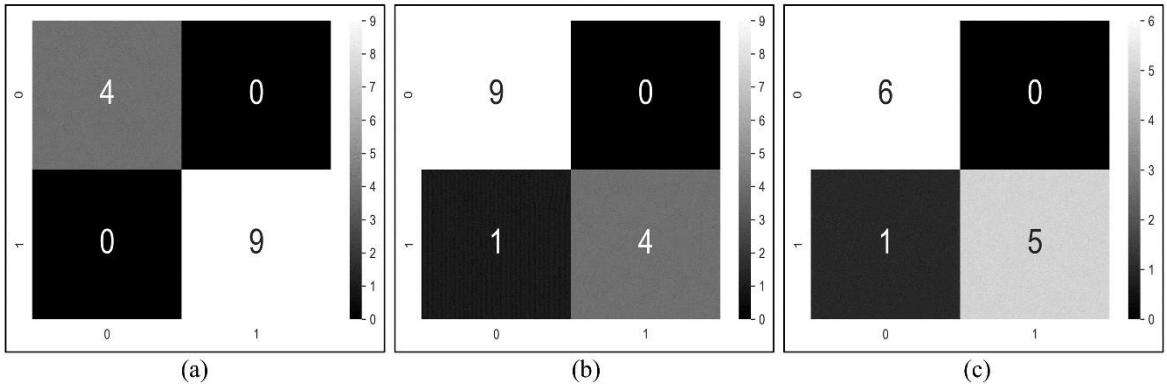


Figure 3.5: Confusion matrix for (a) Healthy vs AK dataset, (b) Healthy vs SCC dataset, (c) SCC vs AK dataset of 14 genes.

To check the robustness of the model, we have done independent testing on a GEO dataset, namely, GSE32628, and classified this dataset into Healthy vs AK, Healthy vs SCC, and SCC vs AK datasets. We have done preprocessing of the raw data as previously explained in the materials and methods section and applied XGBoost ML classifier to find the accuracy for each classification. Accuracy for Healthy vs AK dataset was found to be 96.30 percent; for Healthy vs SCC dataset, accuracy was 85.71 percent, and for SCC vs AK dataset, accuracy was 86.21 percent, as in (Table 3.5).

Table 3.5: Comparison of accuracy for the Independent test set classified into Healthy vs AK, Healthy vs SCC and SCC vs AK datasets.

Datasets	Accuracy with independent test set
Healthy vs AK	96.30%
Healthy vs SCC	85.71%
SCC vs AK	86.21%

5. Statistical analysis of identified genes

We employed the GEO2R computational tool to characterize the relevance of important genes that were differentially expressed during SCC development. P-values < 0.05 were considered statistically significant for the identified genes. *PAMR1, PHYHIP, RAB3B, WIF1, TNNC1, HNRNPM, GTSE1, CHTOP, DNAJC8, S100A11, TUG1, TFG,* and *GAPDH* were found to be down-regulated while *CTSC, CD24, WNT5A, PARK7, MMP14, ARHGEF4, CFL1, RPS13, EDNRB,* and *RPS3A* were found to be up-regulated in SCC progression (Table 3.6).

Table 3.6: Statistical analysis results for each identified genes in the datasets.

Genes	P-value	logFC
Dataset: Healthy vs AK Dataset		
PAMR1	6.83E-19	-2.33005
CTSC	3.35E-11	1.042324
PHYHIP	6.51E-28	-2.37733
CD24	8.25E-17	1.790993
WNT5A	3.09E-14	2.491315
RAB3B	3.13E-19	-1.33558
WIF1	1.76E-27	-4.25276
TNNC1	1.32E-19	-1.87799
PARK7	2.71E-06	0.342546
MMP14	2.08E-05	-0.55263

ARHGEF4	0.005396	0.444668
CFL1	4.67E-08	0.367648
Dataset: Healthy vs SCC		
HNRNPM	3.61E-21	-0.78581
RPS13	5.91E-15	0.529767
Dataset: SCC vs AK		
GTSE1	2.91E-21	-1.57927
CHTOP	7.87E-13	-0.55512
EDNRB	3.85E-18	2.066272
DNAJC8	5.62E-13	-0.54132
S100A11	4.49E-21	-1.1461
TUG1	1.66E-11	-0.67914
TFG	2.50E-12	-0.72343
GAPDH	4.98E-05	-0.42799
RPS3A	0.03709	0.100498

IV. Discussion

SCCs are typically composed of epidermal keratinocytes and exhibit varying degrees of keratosis [154]. In 2015, SCC was estimated to affect up to 2.2 million people [155]. While SCC has a generally favorable prognosis, when invasion and distant metastases occur, the five-year survival rate drops to 34% [156], [157]. Until recently, biomarker research on SCC has been insufficient, and the discovery of biomarkers indicative of progression from AK to SCC is crucial, as they may aid in the management, diagnosis, and treatment of SCC. In this study, 40 SCC samples, 38 AK samples, and 46 normal healthy skin samples from datasets GSE45216, GSE98774 and GSE108008 were divided into 3 binary classifications, namely Healthy vs AK, Healthy vs SCC and SCC vs AK and were utilized for bioinformatics analysis to elucidate the molecular process behind SCC progression. Finally, 23 genes were identified by applying SHAP values on the ML model to quantify the ML predictions. The datasets

comprising of top 14 genes in each group were again subjected to ML to compare the results before and after the application of SHAP values to make the conclusion, and we have found that accuracy in both cases was the same, suggesting that the genes that were identified using SHAP values are equally effective for making predictions and are highly valuable in providing the insights about the data used. We have effectively identified the most important genes that are associated with the progression of SCC and may act as promising biomarkers for the prediction and diagnosis of SCC.

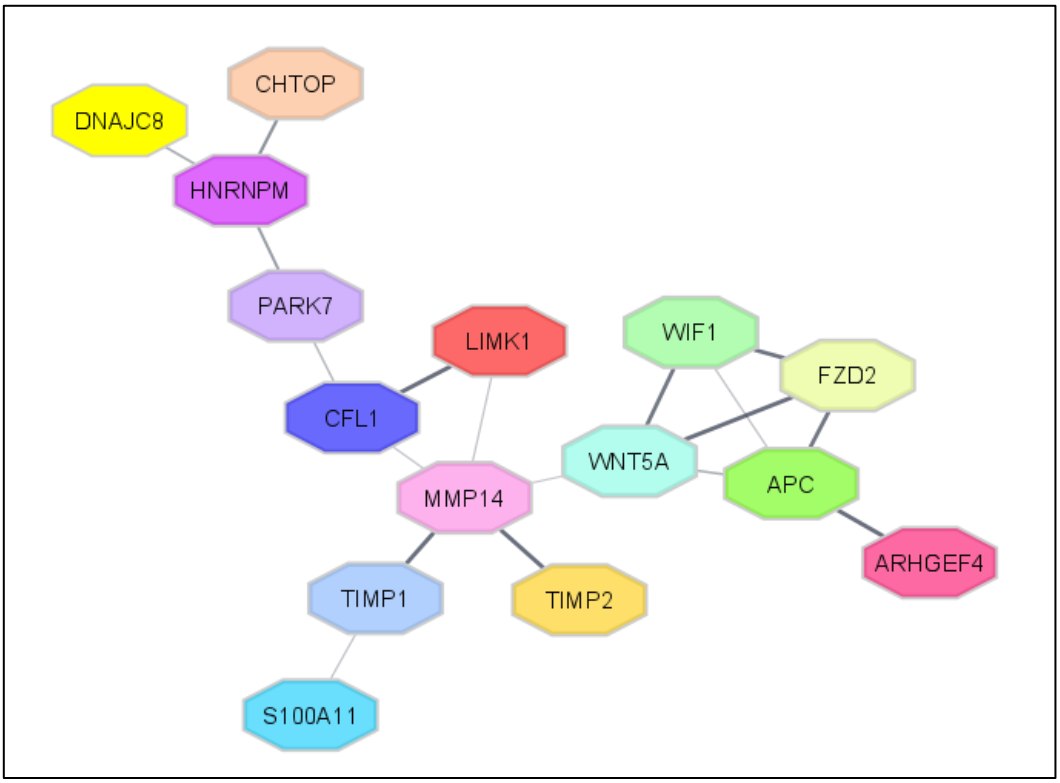


Figure 3.6: A STRING network made from the genes that were retrieved to be of the highest relevance using SHAP values. Here the edge thickness represents confidence in the connection.

A STRING [158] network was generated and visualized using Cytoscape [159], as shown in (Figure 3.6), using those 23 genes which we have retrieved by applying SHAP values that suggest the role of these genes in causing SCC. By studying this network, we have found that *MMP14* and *CFL1* are involved in cellular component disassembly while *ARHGEF4*, *WIF1*,

WNT5A, MMP14, S100A11, CFL1, PARK7, and *CHTOP* have an involvement in the positive regulation of cellular processes. Further, we have found that *WNT5A* is implicated in signaling pathways that govern stem cell pluripotency in Gastric cancer, breast cancer, basal cell carcinoma, and hepatocellular cancer; *WNT5A* and *CFL1* are involved in axon guidance; *ARHGEF4* and *CFL1* are involved in actin cytoskeleton regulation; *WIF1* and *WNAT5A* are involved in WNT signaling pathway, further details are given in functional and pathways enrichment analysis section.

1. Function and Pathway enrichment analysis on the identified key Genes

To perform the Gene Ontology (GO) enrichment analysis as well as pathway enrichment analysis on the key genes, we have used STRING [158]. Based on the identified genes, STRING automatically enriched pathways or functional subsystems using hypergeometric testing with a P-value of <0.05, set as the threshold. Identified key genes were functionally enriched in six GO_BP (Biological Process) terms, one in GO_MF (Molecular Function) terms, and thirteen GO_CC (Cellular Component) terms. Also, these genes were pathway enriched in nine KEGG pathway terms, five REACTOME terms, and twelve WikiPathway terms. Disassembly of cellular components, Positive regulation of the cellular process, Extracellular matrix disassembly, Regulation of the catabolic process of proteins, negatively regulating proteolysis of membrane protein ectodomains, and Negative regulation of the activity of metallopeptidase was the most highly enriched GO_BP terms. Protein binding was the most significantly enriched GO_MF term, while the Anchoring junction, Ruffle, Cell leading edge, Extracellular matrix, Cell junction, Ruffle membrane, Intracellular organelle lumen, Extracellular region, Focal adhesion, Endocytic vesicle membrane with clathrin coating, Cell-cell junction, and Lamellipodium were the most significantly enriched GO_CC terms, refer to the (Table 3.7).

Table 3.7: Significant GO terms with their P-value for STRING network

Term ID	Term Description	P-value	Matching proteins in the network
Biological Process			
GO:0022411	Disassembly of cellular components	0.0039	*TIMP1, APC, TIMP2, MMP14, FZD2, CFL1*
GO:0048522	Positive regulation of cellular process	0.0304	*TIMP1, APC, TIMP2, WNT5A, S100A11, WIF1, MMP14, ARHGEF4, FZD2, LIMK1, CHTOP, PARK7, CFL1*
GO:0022617	Extracellular matrix disassembly	0.0478	*TIMP1, TIMP2, MMP14*
GO:0042176	Regulation of the catabolic process of proteins	0.0478	*TIMP1, APC, TIMP2, WNT5A, PARK7*
GO:0051045	Negatively regulating proteolysis of membrane protein ectodomains	0.0478	*TIMP1, TIMP2*
GO:1905049	Negative regulation of the activity of metallopeptidase	0.0478	*TIMP1, TIMP2*
Molecular Function			
GO:0005515	Protein binding	0.0193	*TIMP1, APC, TIMP2, DNAJC8, WNT5A, S100A11, WIF1, MMP14, ARHGEF4, FZD2, HNRNPM, LIMK1, PARK7, CFL1*
Cellular Component			
GO:0070161	Anchoring junction	0.0019	*APC, S100A11, MMP14, FZD2, LIMK1, PARK7, CFL1*
GO:0001726	Ruffle	0.0079	*APC, S100A11, ARHGEF4, CFL1*
GO:0031252	Cell leading edge	0.0079	*APC, S100A11, ARHGEF4, LIMK1, CFL1*
GO:0031012	Extracellular matrix	0.0148	*TIMP1, TIMP2, WNT5A, MMP14, HNRNPM*
GO:0030054	Cell junction	0.0164	*APC, WNT5A, S100A11, MMP14, FZD2, LIMK1, PARK7, CFL1*
GO:0032587	Ruffle membrane	0.0164	*APC, ARHGEF4, CFL1*
GO:0070013	Intracellular organelle lumen	0.0219	*TIMP1, APC, TIMP2, DNAJC8, WNT5A, S100A11, MMP14, HNRNPM, LIMK1, CHTOP, PARK7, CFL1*

GO:0005576	Extracellular region	0.0338	*TIMP1, TIMP2, DNAJC8, WNT5A, S100A11, WIF1, MMP14, HNRNPM, PARK7, CFL1*
GO:0005925	Focal adhesion	0.0338	*MMP14, FZD2, LIMK1, CFL1*
GO:0005912	Adherens junction	0.0373	*APC, S100A11, PARK7*
GO:0030669	Endocytic vesicle membrane with clathrin coating	0.0407	*WNT5A, FZD2*
GO:0005911	Cell-cell junction	0.0465	*APC, S100A11, PARK7, CFL1*
GO:0030027	Lamellipodium	0.0469	*APC, LIMK1, CFL1*

KEGG pathway analysis indicated that the key genes are enriched in the WNT signaling pathway, Regulation of actin cytoskeleton, Basal cell carcinoma, Axon guidance, Hippo signaling pathway, signaling pathways regulating pluripotency of stem cells, Breast cancer, Hepatocellular carcinoma, and Gastric. Moreover, REACTOME pathway analysis showed the enrichment of key genes in the Activation of Matrix Metalloproteinases, Signaling mediated by TCF in response to WNT, WNT ligand antagonists exerting a negative regulation effect on TCF-dependent signaling, Internalization of FZD2, FZD5, and ROR2 mediated by WNT5A and finally RHO GTPases activating ROCKs while the WikiPathways showed the enrichment of genes in lncRNA in canonical WNT signaling and colorectal cancer, ncRNAs implicated in hepatocellular carcinoma WNT signaling, Matrix metalloproteinases, WNT signaling, Regulation of actin cytoskeleton, WNT signaling pathway and pluripotency, Embryonic stem cell pluripotency pathways, WNT/beta-catenin signaling pathway in leukemia, Breast cancer pathway, and Extracellular vesicle-mediated signaling in recipient cells, WNT signaling in kidney disease and G13 signaling pathway, refer (Table 3.8).

Table 3.8: Significant pathway terms with their P-value for the STRING network.

Term ID	Term Description	P-value	Matching proteins in the network
KEGG Pathways			
hsa04310	WNT signaling pathway	0.0017	*APC, WNT5A, WIF1, FZD2*
hsa04810	Regulation of actin cytoskeleton	0.0026	*APC, ARHGEF4, LIMK1, CFL1*
hsa05217	Basal cell carcinoma	0.0026	*APC, WNT5A, FZD2*
hsa04360	Axon guidance	0.0137	*WNT5A, LIMK1, CFL1*
hsa04390	Hippo signaling pathway	0.0137	*APC, WNT5A, FZD2*
hsa04550	Signaling mechanisms that control stem cell pluripotency	0.0137	*APC, WNT5A, FZD2*
hsa05224	Breast cancer	0.0137	*APC, WNT5A, FZD2*
hsa05225	Hepatocellular carcinoma	0.0137	*APC, WNT5A, FZD2*
hsa05226	Gastric cancer	0.0137	*APC, WNT5A, FZD2*
Reactome Pathways			
HSA-1592389	Activation of Matrix Metalloproteinases	0.0056	*TIMP1, TIMP2, MMP14*
HSA-201681	Signaling mediated by TCF in response to WNT	0.0152	*APC, WNT5A, WIF1, FZD2*
HSA-3772470	WNT ligand antagonists exerting a negative regulation effect on TCF-dependent signaling	0.0414	*WNT5A, WIF1*
HSA-5140745	Internalization of FZD2, FZD5, and ROR2 mediated by WNT5A	0.0414	*WNT5A, FZD2*
HSA-5627117	RHO GTPases Activate ROCKs	0.0414	*LIMK1, CFL1*
WikiPathways			
WP4258	lncRNA in canonical WNT signaling and colorectal cancer	0.00036	*APC, WNT5A, WIF1, FZD2*
WP4336	ncRNAs implicated in hepatocellular carcinoma WNT signaling	0.00036	*APC, WNT5A, WIF1, FZD2*
WP129	Matrix metalloproteinases	0.00037	*TIMP1, TIMP2, MMP14*

WP428	WNT signaling	0.00037	*APC, WNT5A, WIF1, FZD2*
WP51	Regulation of actin cytoskeleton	0.00062	*APC, ARHGEF4, LIMK1, CFL1*
WP399	WNT signaling pathway and pluripotency	0.0071	*APC, WNT5A, FZD2*
WP3931	Embryonic stem cell pluripotency pathways	0.0091	*APC, WNT5A, FZD2*
WP3658	WNT/beta-catenin signaling pathway in leukemia	0.0154	*APC, WIF1*
WP4262	Breast cancer pathway	0.0154	*APC, WNT5A, FZD2*
WP2870	Extracellular vesicle-mediated signaling in recipient cells	0.0165	*APC, WNT5A*
WP4150	WNT signaling in kidney disease	0.0214	*WNT5A, FZD2*
WP524	G13 signaling pathway	0.0219	*LIMK1, CFL1*

2. Biological Significance of the identified key Genes

We have done a literature survey on the identified key genes to find their biological relevance. It was found in many studies that *ARHGEF4* enhances tumor cell motility and invasiveness [160], it communicates with APC through its armadillo repeat domain located at the NH2-terminus of APC-binding region which helps to improve *ARHGEF4's* GEF activity against Rac1 and Cdc42, consequently influencing actin cytoskeleton reorganization, cell shape, adhesion, and migration [161], [162], [163], so the inhibiting the activity of *ARHGEF4* can act as a novel molecular therapeutic marker.

The PAMR1 gene encodes a regeneration-associated muscle protease (RAMP) [164]. It is principally generated in tissues of normal skeletal muscle and brain, and its expression is downregulated in Duchenne muscular dystrophy muscles and type 2 diabetes [164], [165], [166], [167]. *PAMR1* expression is diminished in SCC tissues [168], and it is often deleted

totally in breast cancer samples (20.8–58.3%) [165], as a result of promoter hypermethylation, and hence has been deemed a tumor suppressor gene [169].

Another gene named *GTSE1* has a basic molecular function to adhere to and inhibit the tumor suppressor protein p53's propensity to suppress cancer [170]. Additionally, multiple investigations revealed that overexpression of *GTSE1* was commonly observed in a spectrum of different cancers [171], [172].

WIF1 is a WNT/-catenin pathway downstream gene that acts as an antagonist and a negative regulator of WNT signaling [173]. *WIF1* silencing may be a pre-cancerous epigenetic event that may help tumors grow and spread [174]. Many human tumors have shown that aberrant WNT signaling contributes greatly to cancer genesis. WNT signaling has been shown to reduce apoptosis and make cancerous cells more likely to spread in head and neck SCC [175].

WNT5A is a non-canonical WNT ligand that regulates cell polarity, convergent extension, and epithelial-mesenchymal interaction during embryonic morphogenesis [176]. When *WNT5A* is turned on or blocked, it can be both oncogenic and tumor-suppressive [177]. *WNT5A* has been demonstrated to inhibit thyroid and colorectal cancer cell proliferation, migration, and invasion [177], [178], [179], but an elevated expression of *WNT5A* is associated with belligerence in other types of malignancies, such as melanoma and gastric cancer [180], [181].

MMP14 is a membrane-bound extracellular proteinase. It has been demonstrated that *MMP14* is critical for cancer cells to infiltrate and spread [182]. *MMP14* expression has been related to a poor prognosis in several forms of cancer [183], and downregulation of MMP-14 expression has been shown to decelerate cell growth and metastasis in esophageal SCC. Considering these facts, *MMP14* may be a promising target for esophageal SCC handling, and additional research may be required to elucidate this [184].

CFL1 is a ubiquitous small protein that promotes actin filament abscission and depolymerization. Cytokinesis, cell motility, and morphogenesis are all dependent on *CFL1*. Numerous studies have established a link between *CFL1* and cancer cell migration and invasion, two key hallmarks of malignant tumor cells in a range of solid tumor tissues [185], [186], [187]. According to a study on vulvar SCC, aberrant *CFL1* expression can influence vulvar carcinogenesis and development. *CFL1* gene silencing significantly inhibited the development of vulvar SCC cells, showing that *CFL1* may be exploited as a target for SCC therapy [188].

PARK7 is a protein that is expressed ubiquitously in the vast majority of mammalian tissues [189]. This protein plays a key role in a number of critical physiological processes, including cell proliferation, differentiation, regulation of the transcriptional process, protection from oxidative stress, maintenance of mitochondrial function, inflammation, and metabolic regulation [190], [191], [192], [193], [194]. Silencing *PARK7* decreased oral SCC cell growth and invasion, indicating that *PARK7* may operate as an oncogene during the oral SCC carcinogenesis process [195].

Heterogeneous nuclear ribonucleoprotein M (*HNRNPM*) is a key component of the spliceosome complex. It inhibits pre-mRNA splicing by inhibiting splice site recognition. Interestingly, it has been demonstrated that *HNRNPM* affects the alternative splicing of several cancer-related genes, including FGFR2 and the CD44 cell surface protein [196]. Multiple studies have also reported that *HNRNPM* facilitates the transition of triple-negative breast cancer cells from CD44v to CD44s, which is needed for the transition from epithelial to mesenchymal cells and for cancer to spread [196], [197]. Additionally, in people with breast cancer, increased *HNRNPM* expression is associated with remote metastases, a poor prognosis, and an upsurge in CD44s [198]. This indicates that *HNRNPM* plays a pivotal part in the genesis

of cancer. The signals that control the *HNRNPM*-mediated splicing process in SCC are still unknown, but they are being looked into.

In this study, we examined the applicability of XAI in discovering biomarkers associated with SCC. In the first phase, we created an ML-based XGBoost classification model; in the second step, we used XAI techniques to demonstrate that the model output was interpretable by establishing a relationship between the model output and the relevant genes. SHAP barplot and SHAP summary plots were used to explain the outcomes of the XGBoost ML classifiers predictions for datasets revealing the genes of the highest relevance in causing SCC from AK. *PAMR1, GTSE1, ARHGEF4, WIF1, WNT5A, MMP14, S100A11, CFL1, PARK7, HNRNPM, DNAJC8, and CHTOP* were the top genes that contribute to the model's accuracy and are highly related to cancer development. Also, we have found from this study that the accuracy was the same for the XGBoost ML classifier before and after the application of SHAP values indicating the fact that explaining ML models can be accomplished without jeopardizing the performance of the model. The genes that we have found from this study may serve as candidate targets in SCC management. This study highly recommends the use of XAI in biomarkers discovery for predictive and prognostic purposes in the biomedical field. While we attempted to use the unique approach of XAI to high-dimensional transcriptomics data in order to extract useful information in this study, identifying biomarkers using transcriptome data may have certain inherent biases, these can be overcome by using additional omics data types such as proteomics, metabolomics, and so on, which are not provided in this study but might be an extended part of this work.

V. Conclusion

SCC is the most prominent type of skin cancer, and its prevalence has increased in recent years. Although AK is a precursor lesion for SCC and accounts for around sixty-five percent of SCCs, the genetic defect underpinning SCC development is unknown. In this study, we tried the

applicability of XAI on transcriptomics data to identify candidate genes, namely *PAMR1, GTSE1, ARHGEF4, WIF1, WNT5A, MMP14, S100A11, CFL1, PARK7, HNRNPM, DNAJC8,* and *CHTOP* that may be highly associated with the occurrence and progression of SCC from AK. However, because these conclusions are based on bioinformatics research, they may require confirmation through wet-lab experiments. This study supports the use of XAI on ML models to quantify and thoroughly assess the prediction results, especially in the field of biomedicine, for the discovery of biomarkers relevant to predictive and prognostic purposes.

Chapter 4

Rare deleterious mutations in Bruton's Tyrosine Kinase as biomarkers for Ibrutinib-based therapy: an in-silico insight

---

Chapter 4. Rare deleterious mutations in Bruton's Tyrosine Kinase as biomarkers for Ibrutinib-based therapy: an *in-silico* insight

Abstract

Squamous Cell Carcinoma (SCC) is the second most common type of skin cancer caused by malignant keratinocytes. Multiple studies have shown that protein mutations have a significant impact on the development and progression of cancer, including SCC. We attempted to decode the effect of single amino acid mutations in Bruton's tyrosine kinase (BTK) protein in this study. Molecular dynamic (MD) simulations were performed on selected deleterious mutations of the BTK protein, revealing that the variants adversely affect the protein, indicating that they may contribute to the prognosis of SCC by making the protein unstable. Then, we investigated the interaction between the protein and its mutants with Ibrutinib, a drug designed to treat SCC. Even though the mutations have deleterious effects on protein structure, they bind to Ibrutinib similarly to their wild type counterpart. This study demonstrates that the effect of detected missense mutations is unfavorable and can result in function loss, which is severe for SCC, but that Ibrutinib-based therapy can still be effective on them, and the mutations can be used as biomarkers for Ibrutinib-based treatment. Seven different computational techniques were used to compute the effect of SAVs in accordance with the experimental requirements of this study. To understand the differences in protein and mutant dynamics, MD simulation and trajectory analysis, including RMSD, RMSF, PCA, and contact analysis, were performed. The free binding energy and its decomposition for each protein-drug complex were determined using docking, MM-GBSA, MM-PBSA, and interaction analysis (wild and mutants).

I. Introduction

Squamous Cell Carcinoma (SCC) is the second most highly prevalent skin tumor that develops when keratinocytes alter and turn cancerous. Invasive SCC continues to be influenced by everyday UV exposure to the skin. Global disease rates are on the rise due to aging populations and other demographic shifts [199]. SCC is significant because it occurs twice as frequently as skin cancer in European Caucasians and up to ten times as frequently in fair-skinned Australians, where the incidence is even greater [200]. The fact that SCC strikes men more frequently than women suggests that female immunity may play some role in protecting against the disease, as evidenced by recent studies [201]. People with light skin and light eyes are more likely to develop squamous cell carcinoma beyond the age of 50. It often develops in sites that have been exposed to the sun in the past. Those with a history of extensive exposure to UV, whether through previous medical procedures or the sun, are at a higher risk [202]. Immunosuppressed patients also have a high incidence of squamous cell carcinoma, which can progress into aggressive subtypes [203]. Small squamous cell carcinoma lesions can be removed and are not lethal, but depending on their location, they might cause severe morbidity [204]. Most head and neck squamous cell malignancies necessitate extensive surgery, which, even in the best of hands, can result in poor symptom relief. In addition, the expense of treating these tumors increases each year, posing a critical need to explore low-cost, effective, and efficient treatment options for SCC management [205], [206].

The significance of amino acid variations as hereditary risk factors for human disease, especially cancer, has been recognized for decades [207]. Protein expression and function, its subcellular localization, folding and integrity, and protein-protein interactions (PPI) can all be impacted by mutations. The genesis and progression of cancer are significantly influenced by protein mutations. The variety of mutation's effects on molecular function determines their unique role [208]. Oncogenes and tumor suppressors are examples of proteins where mutations

can activate or inactivate protein activity. After realizing the critical role mutations play in cancer cause and progression, the scientific community began conducting systematic, widespread screening of tumor tissues for mutations. Every year, hundreds of cancer-associated protein mutations are revealed as a result of numerous re-sequencing initiatives, and thousands more are very certainly still to come [209].

In this study, we are trying to decipher the effect of mutations that are found in Bruton's tyrosine kinase (BTK) protein using computational methods. BTK belongs to the Tec tyrosine kinase family and is a nonreceptor cytoplasmic tyrosine kinase. The Tec kinase family, which includes BTK, TEC, ITK, BMX, and RLK, is the second-largest family of cytoplasmic tyrosine kinases [210]. BTK is known primarily for its role in B-lymphocyte development and BCR-mediated signaling [211]. The tumor microenvironment is a multifaceted and intricate system of cells and their precursors, and all of these cells contribute to the emergence of cancer [212]. Acknowledging these discoveries, scientists are considering BTK as a therapeutic target for solid tumors [213]. Further, we tried to find the impact of the mutation on Ibrutinib binding affinity. Ibrutinib, which is also known by its drug code name PCI-32765, is an innovative medication that inhibits BTK in an irreversible manner and was envisioned as a possible therapy for a number of cancers originating in the B-cell lineage; hence we chose this drug for our MD simulation study [214].

Several computational pipelines can be used to investigate genes and variants associated with SCC. These pipelines not only help identify prospective genes and pathogenic mutations but also shed light on aberrations in the target gene product's structure and molecular mechanisms. To examine the structural basis of alterations, reveal underlying molecular mechanisms, and ascertain their pharmacological influence, we tried to make use of MD simulations and other

in silico techniques to explore the influence of mutations namely F413L, P566Q, G584E and E589K on the BTK protein and its binding with the Ibrutinib drug for SCC treatment.

II. Material and Methods

1. Data Retrieval

Mutations across the BTK gene that are specifically observed in SCC were retrieved through the COSMIC (Catalogue of Somatic Mutations in Cancer) database (Url: https://cancer.sanger.ac.uk/cosmic). The COSMIC database is the most comprehensive source of information on somatic mutations that are associated with human cancers. This database contains information on over 2500 distinct cancer types. The database gathers data from two primary sources. As a starting point, the literature is mined for alterations in well-established cancer genes, and the Cancer Gene Census is used to identify the genes that are manually curated [215]. Second, the Cancer Genome Project collects whole genome sequencing data from cancer samples for placement in the database [216]. A comprehensive literature survey was conducted to find the BTK-targeting drug used in SCC treatment. DrugBank (Url: https://go.drugbank.com/) was also searched to validate the role of the identified drug in SCC treatment [217].

2. Mutation Analysis

A total of seven computational tools, namely the Mutation Assessor, SIFT, PON-P2, SNPs&GO, PROVEAN, PolyPhen-2, and MutPred2, were used to assess each of the mutations associated with the BTK gene previously retrieved from the COSMIC database for its deleteriousness. Mutations that were found to be harmful to all seven tools were subjected to the simulation study. All the selected tools can help in evaluating a huge number of mutations by making use of several computational algorithms and ML classifiers, such as Support Vector Machine, Random Forest, Decision Tree, and Neural Network (NN), etc. In addition to putting

pathogenicity into categories, these tools give additional data about the functional influence of mutations [218]. **Mutation Assessor** (Url: http://mutationassessor.org/r3/) uses a multiple sequence alignment (MSA) to describe the functional effects of a missense variation; the MSA is divided into sub-alignments that take into account the conservation scores and the functional specificity of each mutation. Combining a conservation score and a specificity score yields a functional impact score of the input mutation. Protein function is projected to be unaffected by 'neutral' or 'low' variants, whereas it is projected to get affected by 'medium' or 'high' variants [209][219]. **SIFT** (Url: https://sift.bii.a-star.edu.sg/) Web Server is a novel sequence homology-based *in silico* approach that examines the influence of coding mutations at a particular location on the phenotypic effect of proteins [220]. **PON-P2** (Url: http://structure.bmc.lu.se/PON-P2/) is a computational meta-predictor of mutation tolerance. It categorizes amino acid changes into three classes: harmful, neutral, and unknown tolerance. PON-P2 is based on a Random Forest classifier that has been developed and found to be effective on benchmark datasets to predict harmful and pathogenic mutations by employing evolutionary sequence conservation features, amino acid characteristics, GO annotations, and functional annotations, if available [221]. **SNPs&GO** (Url: https://snps.biofold.org/snps-and-go/) is a suitable ML-based tool that visualizes the link between single amino acid variants (SAVs) and a given condition using functional protein annotation. To distinguish normal SAVs from disease-causing ones, this computational tool employs a binary SVM classifier. The functional score provided by SNPs&GO is an empirical estimate of how likely it is that a protein has a harmful SAV based on the related GO terms [222]. **PROVEAN (**Protein Variation Effect Analyzer) (Url: https://www.jcvi.org/research/provean) is another sequence-based computational tool that predicts the potentially damaging effects of mutations in sequences of protein. A non-synonymous SNP (nsSNP) in an MSA triggers a mutation in the symmetry of closely related protein sequences, which is the basis of the prediction output.

PROVEAN uses a delta alignment score calculated from both the native and mutant protein sequences to compare homologous sequences. A threshold of -2.5 is set to identify the deleterious protein variants, so an nsSNP with a score equal to or less than this threshold value will be considered a detrimental mutation [223]. **PolyPhen**-2 (Url: http://genetics.bwh.harvard.edu/pph2/) is a program and a web server that uses structural and evolutionary similarities to foretell how changes in amino acids would affect the structural stability and proper functioning of human proteins. It annotates functional SNPs, links coding SNPs to transcripts, and structurally characterizes proteins to build conservation profiles. All these features are then used to assess the missense mutation's potential pathogenicity [224]. **MutPred2** (Url: http://mutpred.mutdb.org/) is an ML technique based on neural networks that use genetic and molecular information to provide a probabilistic assessment of the pathogenicity of amino acid variations. MutPred2 presently replicates a variety of physical and functional features of proteins, including their secondary structure, transmembrane organization and different signaling pathways, their catalytic efficiency, macromolecular interaction, post-translational alterations, metal binding, and allostery [225].

The amino acid changes that were considered to be the most harmful by all these seven tools used were chosen for molecular dynamics study to check the structural integrity and functionality of the BTK protein with the observed mutations.

3. Molecular Dynamics Study

Molecular dynamics simulation is utilized to comprehend the influence of structural alterations in a mutant protein relative to its natural assembly (wild type) and their interactions with a drug. We retrieved the three-dimensional structure of BTK1 co-crystalized with ligand (PDB id: 5P9J) resolved at 1.08 A° from RCSB PDB (Url: https://www.rcsb.org/) for simulation investigations after removing the ligand [226]. The Swiss PDB Viewer (SPDBV) was used to induce mutations in the wild type structure of BTK protein. SPDBV is a tool for visualizing

protein structures on multiple platforms. In addition to structure modeling, the application allows for the computation and viewing of protein electrostatic potentials [227]. Energy minimization of structures was done through GROMACS version 21 [228], and by using the CHARMM-GUI's (Url: https://www.charmm-gui.org/) solution builder program, five structures were prepared for MD simulation. This simulation study employed a TIP3 rectangular water box function with 10 A° amid box edges. By means of the all-atom additive force field CHARMM36 (C36) [229], protein force field function and the generic protein drug complex force field CHARMM (Chemistry at Harvard Macromolecular Mechanics), the topologies and coordinates for each system were generated [230]. Counter sodium and chlorine ions were added to each system to make them neutral. We used GROMACS version 21 for the simulation purpose. To decrease steric repulsions, each system has undergone fifty thousand steps of sharpest descent minimization of energy. The NVT equilibration was executed for five hundred picoseconds to maintain a consistent temperature in the system, and a brief orientation limitation NPT was also executed for five hundred picoseconds to keep a constant pressure in the system by relaxing them while retaining the protein in place. All systems were subjected to a hundred nanosecond (ns) simulation under no constraints. GROMACS utilities were applied to analyze the trajectory file. The root mean square deviation (RMSD) was computed with the help of the gmx_rmsd file, while the root mean square fluctuation (RMSF) was computed with the help of the gmx_rmsf file.

4. Principle Component Analysis on wild type and mutant BTK proteins

We considered using PCA to quantify the variance between the trajectories of the wild and mutant proteins. Calculating and diagonalizing the covariance matrix for Carbon-alpha (Cα) atoms is an essential part of PCA. This helps uncover the accumulated modes of protein structural variations [231]. For the objective of identifying fluctuations in wild type and mutant proteins from their respective last 25 ns of the simulation trajectories, we conduct PCA on

molecular dynamics trajectory files using the scikit-learn Python package [232]. The PCs, or eigenvectors, were extracted from the wild and mutant BTK protein structures and displayed as cluster groups.

5. CONAN Analysis for wild type and mutant BTK proteins

CONAN was also used to determine how each amino acid residue in the wild and mutant BTK proteins might interact with other amino acid residues and to create contact maps for them. CONAN operates within the GROMACS molecular dynamic engine's mdmat too [233]. It was designed to investigate the statistical and dynamic properties of contacts in order to collect data on how contacts between atoms change over time in molecular dynamics simulations (MD). The software can also read a set of input files that define any variable against which it can determine whether a connection between two contacts was made or broken. CONAN assigned a contact map to each atom in the simulated structure, and the average of these maps was compared for both the wild and mutant BTK proteins.

6. Docking and contact analysis

Ibrutinib is an irreversible, ATP-competitive kinase inhibitor that works by replacing ATP in the substrate binding site of mutant BTK proteins, rendering them inactive and leading to tumor regression. We docked the wild and mutant BTK proteins with Ibrutinib to establish the mutant protein's relative affinity for Ibrutinib binding. Water molecules were excluded from the final simulated structure of each system, i.e., the 100th ns structure. Ibrutinib's three-dimensional structure was obtained from PubChem and then docked via AutoDockTools using AutoDock 4.2 [234]. The grid box was generated by assigning the coordinates of the CA atoms in C481 to its center and spacing the resulting points outward by 60 on each axis. Then the complexes were simulated using the above-mentioned protocol for 100 ns.

The MM-PBSA (Molecular Mechanics Poisson-Boltzmann Surface Area) is a reliable and efficient free energy simulation tool that has been widely used to mimic molecular affinity, including protein-ligand binding interactions [235]. When used in conjunction with molecular dynamics (MD) simulations, MM-PBSA can incorporate conformational and entropic variables into the binding energy. This method has also been utilized to provide a thorough understanding of biomolecular interactions by decomposing the total binding energy into many sections. In this study, we provide the results of MM-PBSA analyses conducted with 1000 snapshots of the MD-trajectory acquired at 25 ps intervals between 76 and 100 ns. We have also tested the wild type and mutant protein-drug complexes with the MM-GBSA technique. MM-GBSA (Molecular mechanics with generalized Born and surface-area solvation) is one of the most widely used *in silico* approaches for estimating protein–ligand binding energies, finding key residues in PPIs, and assessing macromolecular stability [236]. Utilizing the gmx_MMPBSA package [237], the free binding energy of BTK-Ibrutinib complexes was calculated by the interaction entropy method [238].

PyContact was used to perform contact analysis on residues surrounding wild type and mutant protein-drug complexes during the final 25 ns of the trajectory. PyContact is a user-friendly, highly configurable, and intuitive application with a GUI designed to analyze biomolecular interaction in molecular dynamics trajectories. PyContact is intended to assist this effort by facilitating the recognition of significant non-covalent interactions in an understandable way, offering rapid data analysis and visual representation of data without the need for extra programming, by providing entire in-program personalization and comprehensive options for cutting-edge users [239].

III. Results

1. Retrieval of Variations and Drug associated with BTK protein.

Mutations across the BTK gene that are specifically identified in SCC were retrieved using the COSMIC (Catalogue of Somatic Mutations in Cancer) database. A total of 47 distinct SCC mutations were identified using advanced filters, including whole genome screening and target screening as screen types; skin as the SCC cancer tissue location; missense mutations and their pathogenic impact. A comprehensive literature search was conducted to identify the BTK-targeting drug used in SCC treatment, and we have found Ibrutinib for our investigation. DrugBank was also examined to validate the role of Ibrutinib (DrugBank Accession No. DB09053) in BTK inhibition. Ibrutinib (PCI-32765) is categorized as a targeted covalent drug, and it is a propitious anti-cancer treatment for skin cancers [213], [214], [217], [240].

2. Mutational Analysis

A total of seven computational methods, namely Mutation Assessor, SIFT, PON-P2, SNPs&GO, PROVEAN, PolyPhen-2, and MutPred2, were utilized to evaluate the deleteriousness of each variant associated with the BTK gene. We have found F413L, P566Q, G584E, and E589K amino acid variants with the most anticipated adverse impacts, based on all the *in-silico* analysis techniques, as detailed in Table 4.1. Mutation Assessor gives each amino acid in a protein a score between -5.2 and 6.5 based on its functional impact. A score greater than 3.5 is deemed deleterious to the protein's proper functioning [209]. SIFT calculates the likelihood that an amino acid at a specific site will be tolerated, assuming that the most common amino acid is tolerated. If this normalized value falls below a threshold of 0.05, it is believed that the substitution will be damaging [220]. PON-P2 employs a cut-off value of 0.5; for an amino acid variant to be classified as pathogenic, its probability score must be higher than this cut-off value [221]. SNPs&GO computes the prediction's reliability index, where a

score of 0 indicates an unreliable prediction, and a score of 10 indicates the most reliable prediction [222]. For each supporting sequence in PROVEAN, a 'delta alignment score' is worked out. The final PROVEAN score is the average of cluster-level values. The protein variation is likely to have a 'deleterious' effect if the PROVEAN score is -2.5 or less. If the PROVEAN score is more than this cutoff, the effect of the variant is judged to be 'neutral' [223]. PolyPhen-2 looks at the polyphen probability scores to figure out what will happen. Scores between 0 and 0.15 are thought to be harmless. Scores between 0.15 and 1.0 show a variant that could be harmful, while scores between 0.85 and 1.0 are more likely to be harmful in the long run [224]. MutPred2 provides a general score as its final output, which is the likelihood that the amino acid mutation is pathogenic. Each neural network's score in MutPred2 has been averaged to produce this result. A score cut-off of 0.50, if regarded as a probability, would infer pathogenicity [225]. F413L, P566Q, G584E, and E589K scores passed each computational tool's cut-off value, so we selected them for molecular dynamics analysis to find their impact on protein's structural stability.

Table 4.1: Mutations that were determined to be detrimental by all seven tools.

Mutations	Mutation Assessor	SIFT	Pon-P2	SNPs&GO	PROVEAN	PolyPhen-2	MutPred-2
E589K	4.225	0	0.958	9	-3.944	0.999	0.946
G584E	5.075	0	0.842	9	-7.908	1	0.871
P566Q	5.055	0	0.859	8	-7.936	1	0.838
F413L	3.665	0	0.779	9	-5.616	1	0.905

3. Molecular Dynamic Simulation Analysis

Molecular dynamics simulation is utilized to comprehend the influence of structural alterations in a mutant protein relative to its natural structure. We selected the three-dimensional crystallized structure of BTK1 for simulation investigations (PDB id: 5P9J) [226].

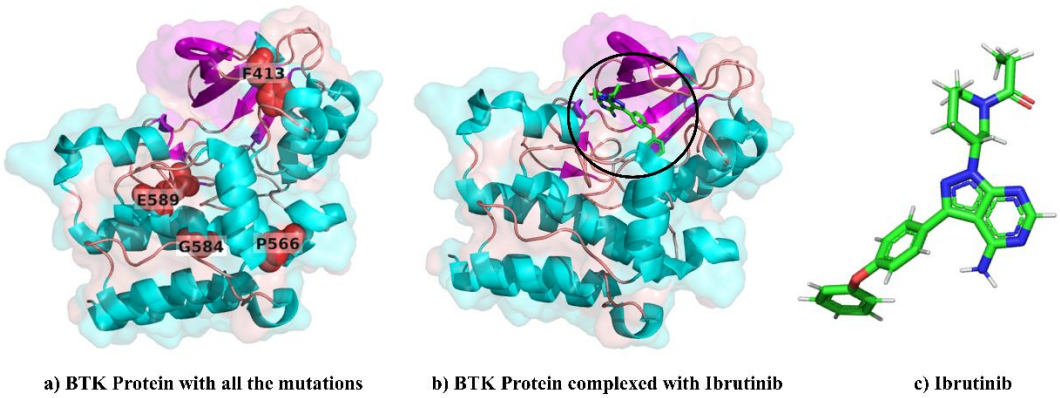


Figure 4.1: Three-dimensional structures of (a) BTK protein with all the mutations depicted by red spheres b) BTK protein complexed with Ibrutinib, highlighted by a black circle (c) Ibrutinib.

The Swiss PDB Viewer (SPDBV) was used to induce mutations in the wild type structure of BTK protein (Figure 4.1), and energy minimization was done by GROMACS version 21. Using the CHARMM GUI input generator, distinct systems for each mutant and native BTK protein were constructed [230]. These wild and mutant systems were solvated and then neutralized using the Monte Carlo Ion placement method. Equilibration input was made with the help of the NVT Ensemble class. GROMACS version 202 [241] and for each system (4 mutants and the wild type BTK proteins), the simulation was run for 100 ns at an ambient temperature of 303.15K and a pressure of 1 bar. All MD trajectories in this study used the single chain "A" of the BTK protein and its mutant structures (F413L, P566Q, G584E, and E589K). There are many domains in BTK protein, starting with the N-terminal PH-domain (Pleckstrin Homology Domain) and continuing through the TEC homology domain, the SH3 and SH2 (SRC homology domains), and finally, the C-terminal BTK-KD domain (BTK-Kinase Domain) [210]. We have taken into consideration BTK-KD (residues 382-659), since not only is there are drug binding sites at position C481 but also the pathogenic mutations that have been found via the use of computational methods present in this kinase domain.

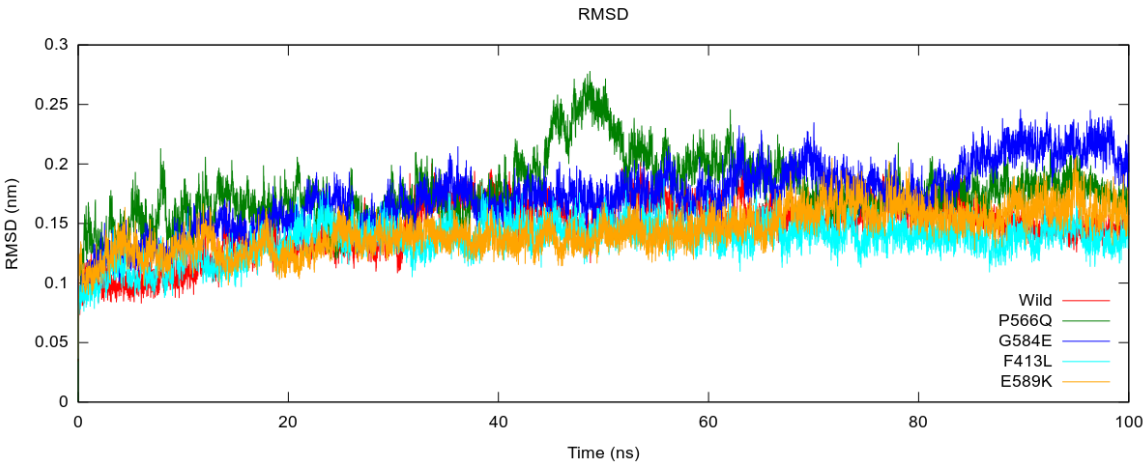


Figure 4.2: RMSD plot of all the four mutated and wild type BTK proteins showing a high degree of variability in mutated BTK proteins as compared to wild type BTK protein. P566Q mutation showing the highest unstability indicated by "Green" color.

The convergence of simulated proteins was determined by comparing the RMSD of average simulated postures derived from all the frames formed during simulation to the initial structure as a result of the time trajectory in MD simulations. Initially, RMSD data for BTK protein (C-alpha), were retrieved from each 100 ns simulation trajectory with respect to the starting pose. All mutants exhibited RMSD variability than the wild system, with P566Q (represented by green color) appearing to be the most unstable (Figure 4.2).

To get a more detailed overview of the RMSD analysis, we evaluate the local regions where the mutations are located since the overall instability may dominate the global RMSD. There is the presence of an ATP binding site and an intrinsically disordered region in this beta-sheet conformation, starting from 402-421. Intrinsically disordered areas are distinguished by a lack of secondary and organized tertiary structure; they interact extensively with macromolecules and may undergo structural modifications upon binding to interacting molecules [242]. The amino acid residue Phe-413 is an important and conserved phenylalanine located at the beginning of the beta-strand. It sits directly on top of the triphosphate group of ATP, forming a crucial hydrogen bond with the oxygen atom of the beta-phosphate. This interaction is

essential for the proper binding and subsequent release of ADP. Substituting the bulky side chain of phenylalanine with a leucine residue at position 413 (F413L mutation) would significantly limit the flexibility of the loop structure required to accommodate ATP, ultimately hindering its binding and subsequent release of ADP. Additionally, this substitution would not be optimal for forming the necessary hydrogen bond with ATP, leading to an altered conformation of the BTK protein. Consequently, this conformational change could interfere with the correct alignment of ATP for catalysis, rendering it incompatible with the enzyme's function. Based on this observation, we can observe that the F413L (represented by cyan color) mutation in this segment is involved in increased rigidity of the disordered region of the protein, which might affect ATP binding with the BTK protein (Figure 4.3).

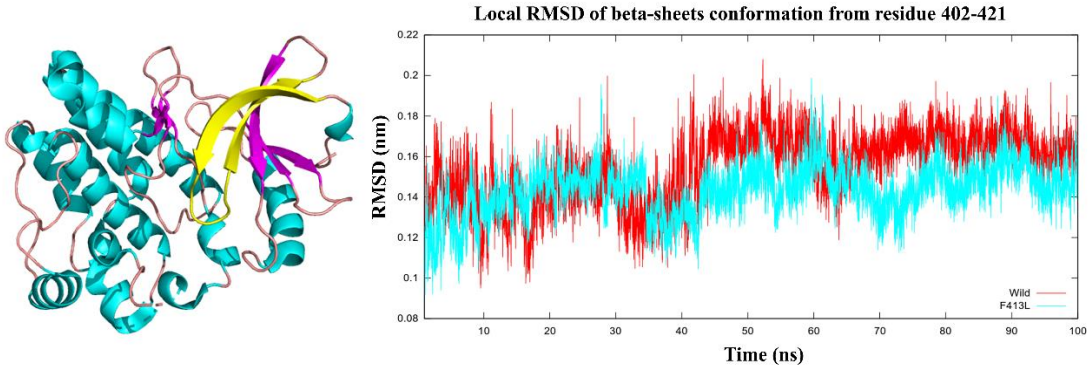


Figure 4.3: Local RMSD plot of Beta-sheets conformation from residue 402-421 highlighted by yellow color in left panel. A rigidity can be seen after 44ns in the mutant protein as compared to wild type showing the F413L mutation's impact on protein.

P566Q variation lies in the helix region starting from residues 560–572, so we have conducted a local RMSD analysis on this segment also. From this RMSD analysis, we have found that the P566Q variation is shown to affect the helix formation; higher RMSD can be due to more flexibility in this helix region. This might also affect the conformation of the active site due to the existence of buried residues that typically form hydrophobic cores to preserve the integrity of the protein structure. Compared to exposed residues that are not associated with the active site, buried residues are very sensitive to mutation [243], and hence the P566Q (represented by

green color) mutation might be causing a very high degree of distortion and flexibility in this BTK protein (Figure 4.4).

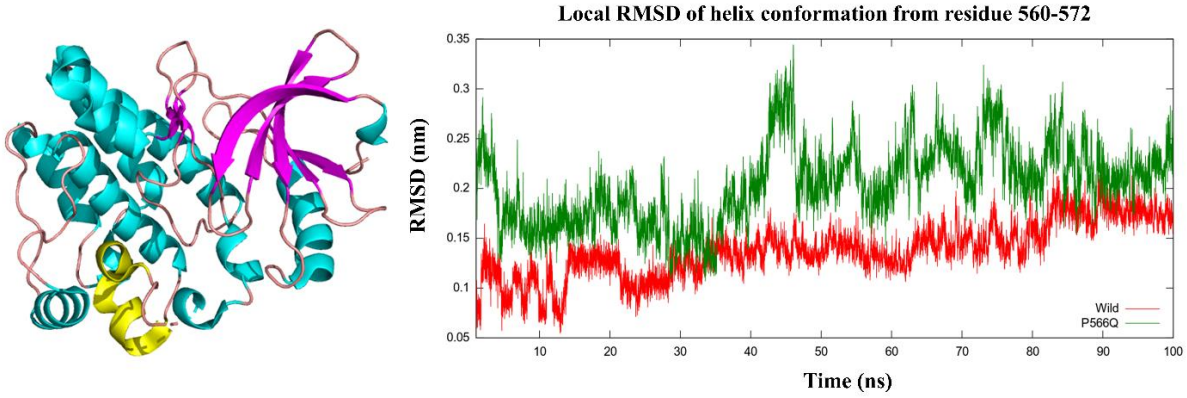


Figure 4.4: Local RMSD plot for P566Q mutant protein located in helix conformation starting from residue 560-572 highlighted by yellow color in left panel. A high rise in RMSD peak can be seen at various positions but a significant rise can be seen after 44ns.

We have observed the presence of buried residues and increased hydropathy in the helix conformation starting from residues 575–592 in wild type BTK proteins, indicating the existence of hydrophobic buried residues in this particular segment. Because of the hydrophobic effect, protein folds are more robust as hydrophobic amino acids are hidden deep inside the protein and are protected from water [244]. The local RMSD analysis on mutations G584E (shown in blue) and E589K (shown in orange) reveals a distortion in the structure of the BTK protein (Figure 4.5). This helix conformation containing our concerned mutations seems to lose their rigidity, indicating damage to the hydrophobic effect and hence disruption of the protein fold, leaving the residues more flexible and unstable; this distortion was seen more for the G584E mutant protein.

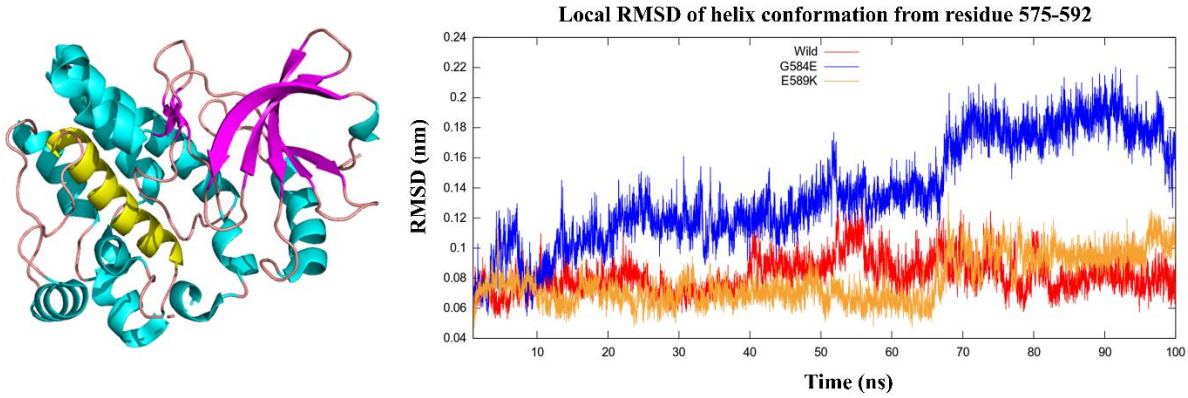


Figure 4.5: Local RMSD plot for G584E and E589K mutant proteins located in helix conformation starting from residue 575-592 highlighted by yellow color in left panel. A high rise in RMSD peak can be seen after 65ns in the mutant protein as compared to wild type BTK protein.

The stimulus of the mutation on the dynamic behavior of the residues was determined by computing RMSF values for both the wild and mutant BTK proteins. The RMSF values were estimated during the final 50 ns of the simulation's trajectory since the wild structure stagnates after a time period. Region 552–559 in the P566Q mutation showed the highest flexibility, as the RMSF peak had a rise of 0.42 nm, while E589K showed the highest rigidity for the same region. Region 479–483 contains the active site at C481 of the BTK protein. This region containing the binding site showed a high level of rigidity, along with another region starting from residues 582–592, for all the mutants, out of which the F413L mutation showed the highest rigidity for both of these sites, so we can conclude that these mutations may hamper the drug binding due to structural instability. The G584E mutation seems to be highly unstable as almost all the residues show variation in flexibility in contrast with the wild type, with the highest flexibility in the region starting from 572–576. The RMSF plots confirmed that the P566Q mutant protein demonstrated considerable residual shifts relative to the wild type protein, suggesting a role in structural damage due to protein flexibility, as evident in the latter half of the RMSD plots. Drug-binding sites are located in several of these dynamic areas, which means they may affect drug-binding interactions (Figure 4.6).

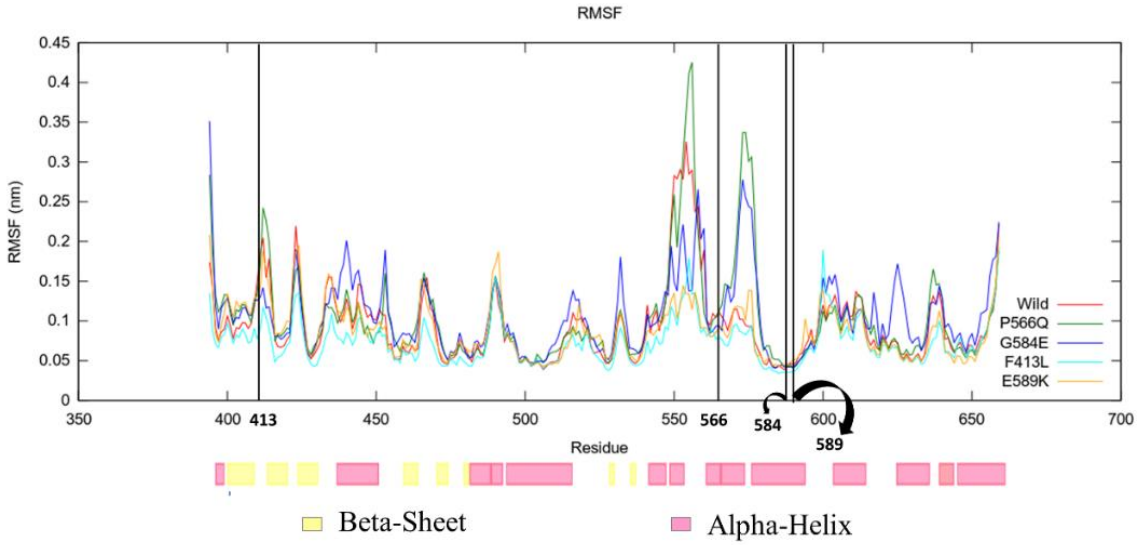


Figure 4.6: RMSF plot of all the four mutated and wild type BTK proteins showing a high degree of variability in mutated BTK proteins with respect to wild type BTK protein. P566Q mutation showing the highest unstability indicated by "Green" color.

4. PCA Analysis on MD Trajectories

PCA tries to find the cumulative patterns of fluctuations in the protein structure by computing and diagonalizing the covariance matrix for Carbon-alpha (C_α) atoms to determine the cumulative modes of fluctuations in the structure of proteins. These produced orthogonal vectors, also known as eigenvectors, have been given the name 'principal components' since they have the biggest eigenvalues (PCs). So, for the purpose of finding fluctuations in native and mutant proteins, we run principal component analysis (PCA) on molecular dynamics' trajectory files using the scikit-learn python library to extract the kinase domain interactions and the atomic gap in wild type and mutant BTK proteins from their corresponding to last 25 ns MD simulation trajectories. The principal components (PCs) or eigenvectors were retrieved and presented as groups from the corresponding MD simulation trajectories for the wild and mutant BTK protein structures. As a result of PCA analysis, we've determined that mutant proteins exhibit large dynamic movements and evident fluctuations in terms of atomic vibrations for the last 25 ns simulation period. PCA plots for extracted PC1 and PC2 indicate

the mutations in the groups throughout the observed simulation interval via a color shift from yellow to purple, representing the significant periodic bounces amid the different conformational postures of the wild type and mutant BTK protein structures. Therefore, it is clear that the mutations are altering the BTK protein, as demonstrated by PCA plots (Figure 4.7).

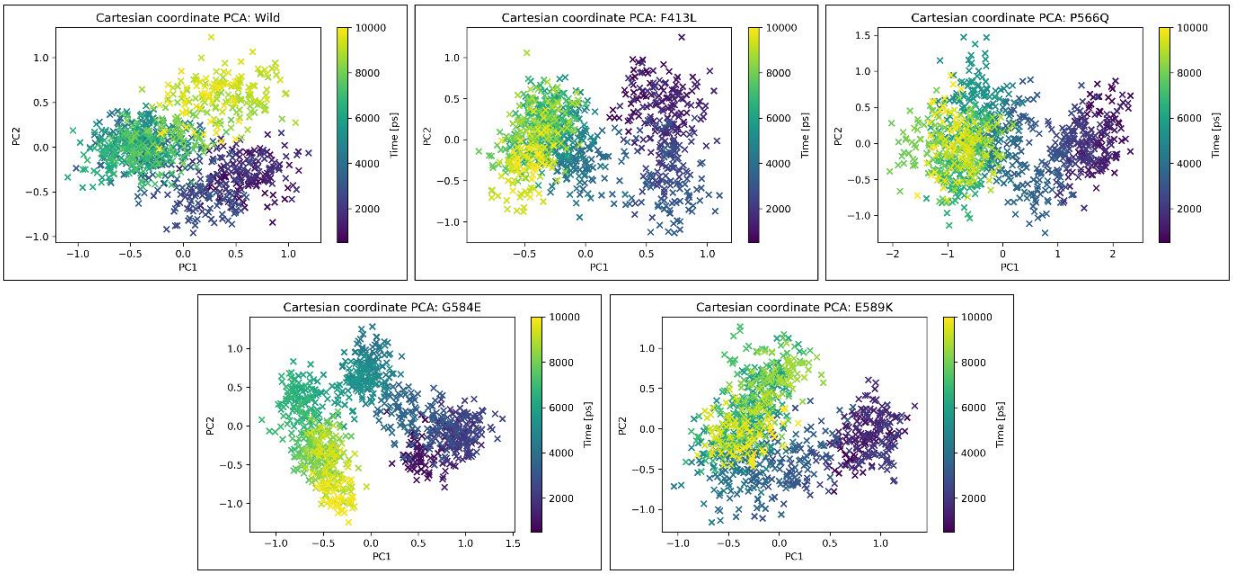


Figure 4.7: Principal Component Analysis for both wild and mutant BTK proteins exhibiting large dynamic movements and evident fluctuations in terms of atomic vibrations as a consequence of 100 ns MD simulation.

5. Contact map analysis through CONAN.

We tried to find out the possible interactions of each residue with other residues in the wild and mutant BTK proteins. For this purpose, we conducted a contact map analysis through CONAN (CONtact Analysis). It works with the GROMACS molecular dynamic engine's mdmat tool. In this study, CONAN was used to generate contact maps for each atom in the simulated structure, and the average of these was then compared for both wild type and mutant BTK proteins. Contact maps capture the secondary and tertiary structural characteristics that describe the molecule. Alpha-helices are shown by a deepening of the matrix diagonal. Parallel and antiparallel beta-strands, on the other hand, are shown by thin contact segments that are

either orthogonal to or parallel to the matrix diagonal. Interestingly, we have found significant changes in contact maps of mutant proteins with reference to wild BTK protein for the last 25ns of MD trajectory, depicting the contact maps in 2D format for a visual interpretation (Figure 4.8).

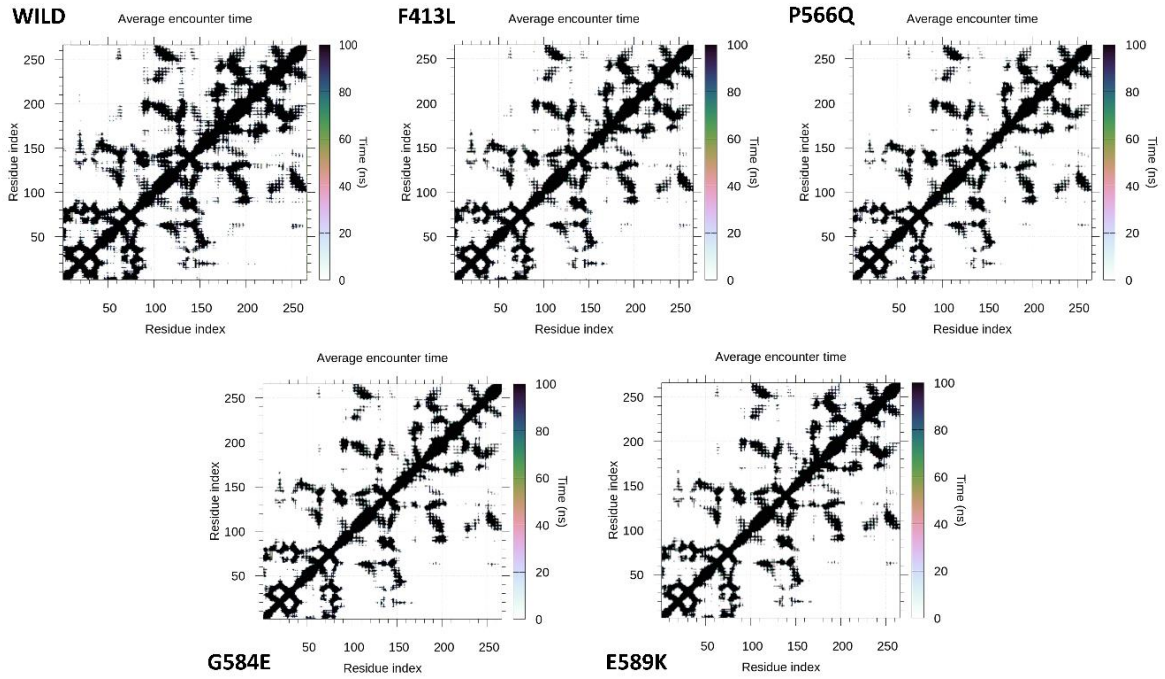


Figure 4.8: Contact maps generated by CONAN for both wild and mutant BTK proteins exhibiting lightning of backbone represented by the diagonal and missingness throughout the MD trajectory for mutant proteins as compared to the wild type BTK protein.

By observing the images closely, we can see that the F413L mutant protein's residues have various weak contacts depicted by the thinning of the backbone (represented by the diagonal) in contrast with the wild type BTK protein, concluding the weakening of contacts among the alpha-helices of this protein. Parallel and anti-parallel beta-strands depicted by thin lines and light spots in CONAN contact maps showed the loss of contacts at various places in contrast with the wild protein. We can also observe lightning and the missingness of some points throughout the entire MD trajectory, which concludes the breaking of contacts in the mutant structure. Throughout the MD trajectory, we can see the same kind of thinning of the backbone,

showing the weakening of contact among residues in alpha-helices and a similar kind of missingness and lighting of various spots at different locations showing the breaking of contacts in residues of beta-strands.

6. Docking and MM-PBSA and MM-GBSA binding energies analysis

To investigate the influence of amino acid substitution on Ibrutinib binding, we docked the mutant and wild BTK protein structures with the FDA-approved drug. Table 4.2, presented the outcomes of the docking analysis containing Binding energy (kcal/mol) and predicted Inhibition Constant (nM). The estimated free energy is determined by combining the torsion and intermolecular free energies. For all the complexes, the calculated binding free energy for Ibrutinib ranges from 9.16 to 9.55 kcal/mol. According to the binding energy data, Ibrutinib seems to be stable in each of the complexes.

Table 4.2: Binding energy of mutated and wild system when docked with Ibrutinib.

System	Binding energy (kcal/mol)	Predicted Inhibition Constant (nM)
Wild	-9.37	135.22
E589K	-9.48	111.82
G584E	-9.43	122.28
P566Q	-9.55	99.59
F413L	-9.16	191.52

The MM-PBSA study has been widely used to simulate molecular affinity, like protein-ligand binding interactions, as an active and reliable free energy simulation tool. When paired with MD simulations, MM-PBSA has the inherent potential to include conformational variations and entropic components in the binding energy. Decomposing the overall binding energy into several components, this approach has also been used to provide a comprehensive understanding of biomolecular interactions [236], [245]. We provided here the findings of

MM-PBSA analyses that were carried out utilizing 1000 snapshots collected from the MD-trajectory at 25 ps intervals between 76 and 100 ns of MD trajectory of wild and mutant complexes with Ibrutinib. We also calculated the -TΔS for both MM-PBSA and MM-GBSA analysis using the interaction entropy method and compiled the enthalpy change and entropy change in Table 11, using the equation given below.

$$\Delta \mathrm{Gbind} = \Delta \mathrm{H} - \mathrm{T}\Delta S$$

Where 'ΔH' (enthalpy change) is calculated by MM-PBSA/MM-GBSA, while '-TΔS' is calculated by interaction entropy.

From the details given in Table 4.3, we can see that F413L (-25.68 ± 3.0 kcal/mol) and G584E (-25.03 ± 3.0 kcal/mol) mutant protein-drug complex has the lowest total free binding energy (ΔG_{bind}) indicating good binding between protein and drug while E589K (-15.78 ± 4.08 kcal/mol) mutant complex has the highest total binding energy followed by wild type (-17.28 ± 4.29 kcal/mol) and P566Q (-22.33 ± 4.33 kcal/mol) mutant protein-drug complexes indicating the poor binding affinity. The van der Waals (ΔE_{vdW}), electrostatic interaction (ΔE_{elec}), solvent-accessible surface area energy (ΔE_{SASA}), enthalpy change (ΔH), and entropy change (-TΔS) were the most significant components of the entire free binding energy of each wild and mutant protein complex.

Table 4.3: MM-PBSA and MM-GBSA analysis results for free binding energy of Ibrutinib with wild type and mutant BTK proteins.

MM-PBSA					
Energy Component	**Wild**	**F413L**	**P566Q**	**G584E**	**E589K**
ΔE_{vdW}	-61.20 ± 0.61	-56.71 ± 0.65	-57.06 ± 1.09	-53.74 ± 0.82	-60.34 ± 1.05
ΔE_{elec}	-23.27 ± 1.16	-31.22 ± 1.62	-30.36 ± 3.14	-29.72 ± 2.00	-24.87 ± 1.97

ΔPB	62.30 ± 0.71	61.43 ± 0.31	62.33 ± 0.56	54.64 ± 0.73	66.22 ± 0.80
ΔE_{SASA}	-5.44 ± 0.02	-5.10 ± 0.02	-5.12 ± 0.00	-4.88 ± 0.03	-5.40 ± 0.02
ΔH	-27.61 ± 1.77	-31.59 ± 1.98	-30.21 ± 3.50	-33.70 ± 2.44	-24.38 ± 2.54
-TΔS	10.33 ± 0.11	5.91 ± 0.09	7.87 ± 0.19	8.67 ± 0.24	8.6 ± 0.11
ΔG	-17.28 ± 4.29	-25.68 ± 3.0	-22.33 ± 4.33	-25.03 ± 3.0	-15.78 ± 4.08
MM-GBSA					
ΔE_{vdW}	-61.20 ± 0.61	-56.71 ± 0.65	-57.06 ± 1.09	-53.74 ± 0.82	-60.34 ± 1.05
ΔE_{elec}	-23.27 ± 1.16	-31.22 ± 1.62	-30.36 ± 3.14	-29.72 ± 2.00	-24.87 ± 1.97
ΔGB	51.48 ± 0.16	55.08 ± 0.59	55.13 ± 0.97	50.79 ± 0.76	52.21 ± 0.99
ΔE_{SURF}	-8.06 ± 0.01	-7.54 ± 0.03	-7.65 ± 0.01	-7.19 ± 0.01	-7.83 ± 0.01
ΔH	-41.05 ± 1.63	-40.39 ± 2.04	-39.94 ± 3.59	-39.87 ± 2.45	-40.82 ± 2.61
-TΔS	10.33 ± 0.11	5.91 ± 0.09	7.87 ± 0.19	8.67 ± 0.24	8.6 ± 0.11
ΔG	-30.72 ± 2.96	-34.48 ± 2.49	-32.07 ± 2.65	-31.19 ± 2.79	-32.22 ± 2.81

We have also conducted MM-GBSA tests on the wild and mutant protein-drug complexes. MM-GBSA is one of the most widely used techniques for estimating protein–ligand binding energies, identifying key residues in protein–protein interactions, and investigating macromolecular stability [246]. The free binding energy, ΔG_{bind} calculated through MM-GBSA was found to be highest (-30.72 ± 2.96 kcal/mol) for the wild type BTK-Ibrutinib complex, showing poor binding affinity between the BTK protein and the Ibrutinib drug, while this was found to be lower for the rest of the mutant protein-drug complexes, showing good binding affinity. For the F413L mutant protein-drug complex, the free binding energy was found to be the highest (-34.48 ± 2.49 kcal/mol) followed by E589K (-32.22 ± 2.81 kcal/mol), P566Q (-32.07 ± 2.65 kcal/mol) and G584E (-31.19 ± 2.79 kcal/mol) mutant protein-drug complexes, showing a significant increase in binding affinity between the BTK protein and the Ibrutinib drug. The van der Waals energy (ΔE_{vdW}), electrostatic energy (ΔE_{elec}), solvation energy (ΔE_{SURF}), enthalpy change (ΔH), and entropy change (-TΔS) were the contributing factors in identifying the binding nature of protein-drug complexes, and their respective values

are given in Table 4.3. We also found the average decomposition value for each residue in both wild type and mutant protein-drug complexes, details are given in (Figure 4.9).

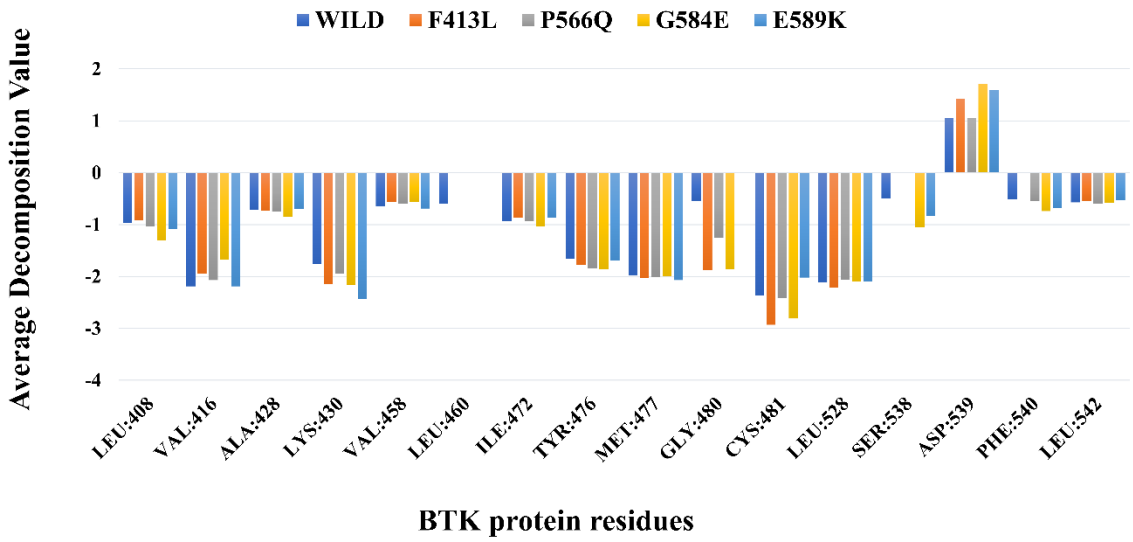


Figure 4.9: Average decomposition values for each residue in both wild type and mutant BTK-Ibrutinib complexes.

7. Contact analysis for wild and mutant BTK-Ibrutinib complexes.

Noncovalent interactions have substantial significance in terms of molecule binding and identification. Biomolecular recognition relies considerably on nonbonded interactions, including hydrogen bonding, ionic, and hydrophobic [247]. The MD trajectories of wild and mutant BTK-Ibrutinib complexes were analyzed using GROMACS and PyContact to identify hydrogen bonding interactions. All hydrogen bond interactions were set to have a range of bond distances from 1.5 to 2.5, and the cut-off angle for hydrogen bonds was set to 120°. The percentage of hydrogen bonding for wild and mutant protein complexes' MD trajectories at an interval of 100 ps is shown in (Figure 4.10). In the wild type BTK-Ibrutinib complex, the CYS481 residue showed the maximum hydrogen bond occupancy of 100 percent, whereas the remaining residues followed the order MET477=GLU475>THR474. For the F413L mutant protein-drug complex, GLU475 showed a maximum hydrogen bond occupancy of 100 percent,

and the order of remaining residues follows the order CYS481=MET477>THR474>LYS430 for hydrogen bond occupancy. We have found the following sequence of residues CYS481=GLU475>MET477>THR474 for the mutant P566Q protein-drug complex, where CYS481 and GLU475 occupied the highest percentage for hydrogen bond occupancy. The occupancy for hydrogen bonds was found to be in the order GLU475>MET477>CYS481>THR474>ASN484 for the mutant G584E protein-drug complex. For the mutant E589K protein-drug complex, the order of the hydrogen bonds was found to be GLU475>MET477>CYS481>LYS430>THR474, with GLU475 having the most hydrogen bonds (99.50%) occupied.

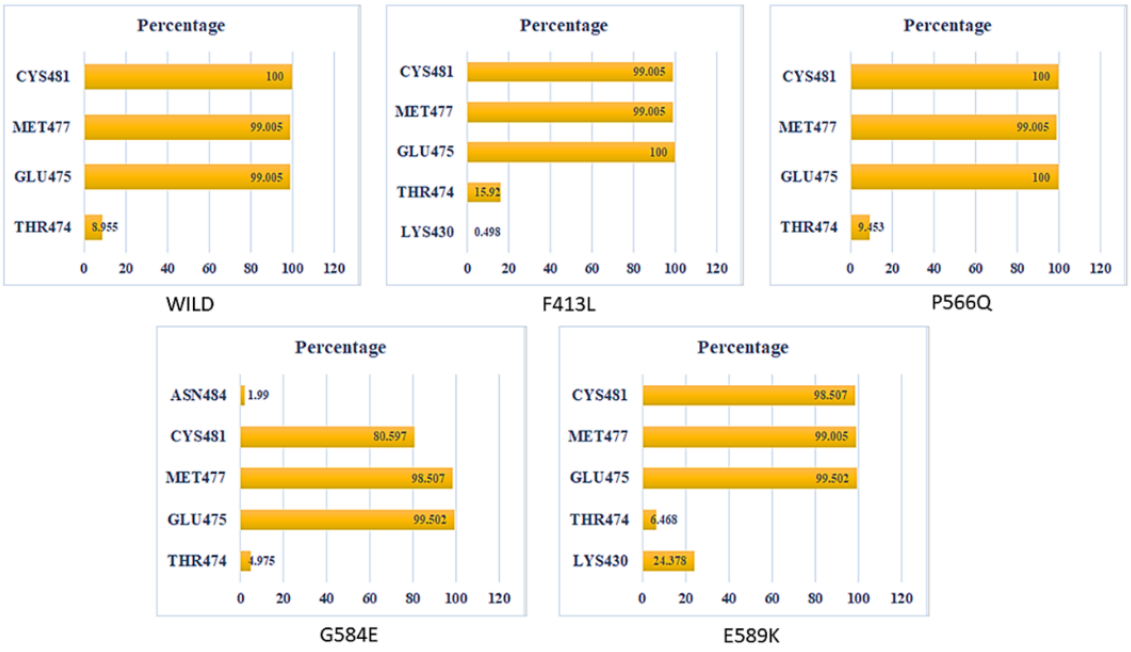


Figure 4.10: PyContact analysis graphs for wild and mutant BTK-Ibrutinib protein complexes' MD trajectories for potential hydrogen bond occupancy.

IV. Discussion

Integrating data from multiple genetic variation databases could provide a thorough understanding of the genes associated with SCC. We accessed the COSMIC database to get

mutational data on the BTK gene, which is responsible for SCC. This mutation analysis helps in the identification of potentially cancer-causing genomic variants. In this study, a total of 47 mutations of the BTK gene that are unique to Squamous Cell Carcinoma (SCC) were extracted from the COSMIC database using advanced filters that considered whole genome screening as well as target screening; skin as the SCC tumor tissue site; missense mutations and their impact as pathogenic. BTK is a cytoplasmic nonreceptor tyrosine kinase that belongs to the Tec family of tyrosine kinases. The Tec family of kinases includes BTK, TEC, ITK, BMX, and RLK. In the cytoplasmic tyrosine kinase superfamily, this family is the second largest. BCR-mediated signaling and the formation of B-lymphocytes may be its most well-known function. BTK is found in all types of hematopoietic cells, not just B cells. Because of this, BTK was found to play a vital role in the tumor microenvironment, which is a complex network of cells and their precursors. This network includes pericytes and smooth muscle cells, fibroblasts with different phenotypes, neutrophils, T- and B-cell lymphocytes, natural killer (NK) cells, antigen-presenting cells (APCs), etc., to name some. Each of these cell types contributes to the progression of cancer. Numerous preclinical and clinical cancer studies have investigated Ibrutinib (PCI-32765), a small-molecule pharmacological inhibitor of BTK. An example of a targeted therapeutic strategy in cancer would be the use of pharmacological inhibitors to stop the BTK gene from functioning. The BTK gene and its targeted drug Ibrutinib are the least-studied gene-drug combination for SCC, according to the survey of the relevant literature; hence, additional research is required. Computational genetics is focussed on learning more about certain aspects of cancer biology. Pathogenic and damaging genetic mutations have been isolated with the aid of a combination of strong computational methods. In this study, 47 nsSNPs for the BTK protein were retrieved from the COSMIC database, and after running each of these mutations through a total of seven different mutation impact prediction methods, four different nsSNPs, specifically, F413L, P566Q, G584E, and E589K, were found to be

detrimental, malignant, pathogenic, and deleterious, correspondingly. Also, there are some studies where these mutations have a role in carcinoma progression. A study based on tumor molecular profiling in precision oncology identified F413L and E589K as cancer biomarkers. This study analyzed tumors and matched normal sequence data from over 10,000 patients with advanced cancer and found clinically relevant somatic mutations, novel noncoding alterations, and mutational signatures shared by common and rare tumor types [248]. Additionally, a study on colorectal carcinogenesis analyzed eight pairs of colorectal adenomas and carcinomas using whole-exome sequencing and found that P566Q has an associated role in the carcinogenesis of colorectal cancer [249]. In another study, P566Q and G584E have been found in tumorigenesis of basal cell carcinoma (BCC) by conducting the genetic profiling of 293 BCC samples for the identification of cancer driver genes [250]. Based on the above studies, we chose these mutations and simulated them in the BTK protein structure in order to analyze their interaction and the effect of mutations on the protein structure. This work used molecular simulations to identify the underlying molecular mechanism by which non-synonymous variants may cause damage to protein structure. In these 100 ns molecular dynamic simulation studies, the mutations, namely F413L, P566Q, G584E, and E589K, of the BTK protein exhibited markedly altered molecular properties in contrast with the wild BTK protein, notably RMSD and RMSF values at the residue level. PCA plots for both wild and mutant proteins made it abundantly evident that the mutations are distorting the BTK protein. CONAN analysis was also done to find significant changes in contact maps of mutant proteins as compared to wild type for the entire MD trajectory, and we have found that the contact maps of mutants of the BTK protein, namely F413L, P566Q, G584E, and E589K, showed visible dissimilarity with the contact map of the wild type protein, which also supported the fact that the protein is actually distorted upon these mutations. When docked with Ibrutinib, all the systems showed similar free energy suggesting no effect in binding with the FDA-approved drug. The MM-PBSA method was

developed and proven to be an effective tool for assessing protein-drug interactions resulting from simulation trajectory information. So, the last 25 ns of the trajectories of the complexes were analyzed by means of MM-PBSA, revealing that the G584E mutant protein-drug complex has the lowest free binding energy, accompanied by F413L and P566Q mutants, whereas wild type and E589K mutants have the highest total free binding energy, indicating that these mutations cause structural instability but are not affecting the binding of Ibrutinib to the BTK protein. This observation was also validated by the MM-GBSA study, which demonstrated that the wild type BTK-Ibrutinib complex has the lowest free binding energy, accompanied by the G584E, P566Q, F413L, and E589K mutant protein-drug complexes. Since additional mutations have demonstrated modest differences in their free binding energy, the preceding observation can be supported.

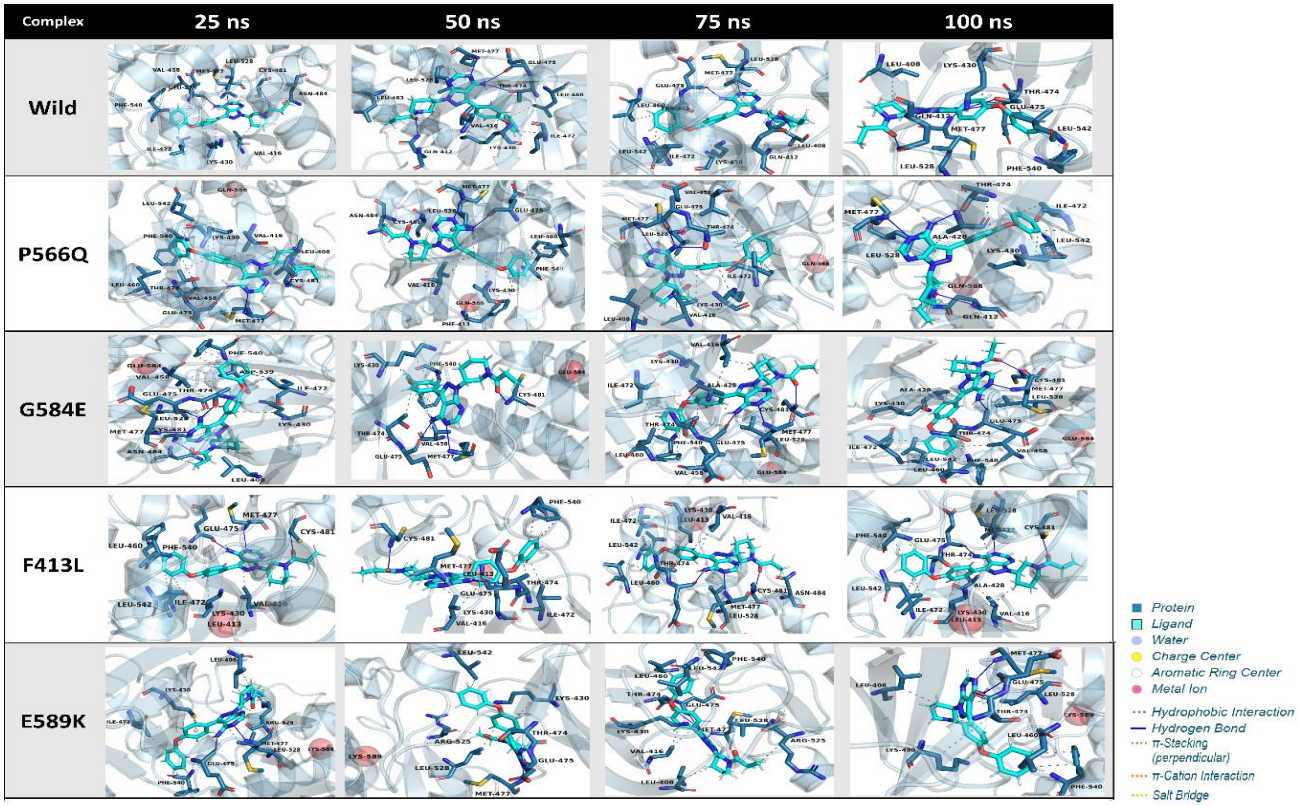


Figure 4.11: Interaction and proximity of residues around Ibrutinib in wild type and mutant BTK-Ibrutinib complexes throughout different time frames.

Table 4.4: Analysis of Ibrutinib's interactions with BTK protein residues throughout different time frames.

Systems	Time Frames	Interactions of BTK-protein residues with the Ibrutinib drug		
		Hydrophobic Interactions	**Hydrogen Bonds**	**pi-cation interactions**
Wild	25ns	VAL416, LYS430, VAL458, ILE472, ASN484, LEU528, PHE540	GLU475, MET477, CYS481	-
	50ns	VAL416, LYS430, LEU460, ILE472, LEU483, LEU528	GLN412, THR474, GLU475, MET477	-
	75ns	LEU408, LYS430, LEU460, ILE472, THR474, LEU528, LEU542	GLN412, GLU475, MET477	-
	100ns	LEU408, LYS430, LEU528, PHE540, LEU542	GLN412, THR474, GLU475, MET477	-
F413L	25ns	VAL416, LYS430, LEU460, ILE472, PHE540, LEU542	GLU475, MET477, CYS481	-
	50ns	VAL416, LYS430, ILE472, THR474, PHE540	THR474, GLU475, MET477, CYS481	-
	75ns	VAL416, LYS430, LEU460, ILE472, THR474, ASN484, LEU528, LEU542	GLU475, MET477, CYS481	-
	100ns	VAL416, ALA428, LYS430, ILE472, THR474, LEU528, PHE540, LEU542	GLU475, MET477, CYS481	-
P566Q	25ns	LEU408, VAL416, LYS430, VAL458, LEU460, THR474, PHE540, LEU542	GLU475, MET477, CYS481	-
	50ns	PHE413, VAL416, LYS430, LEU460, ASN484, LEU528, PHE540	GLU475, MET477, CYS481	-
	75ns	LEU408, VAL416, LYS430, VAL458, ILE472, LEU528,	THR474, GLU475, MET477	-
	100ns	ALA428, LYS430, ILE472, THR474, LEU528, LEU542	GLN412, THR474, MET477	-
G584E	25ns	LEU408, LYS430, VAL458, ILE472, THR474, ASN484, LEU528, PHE540	THR474, GLU475, MET477, CYS481, ASP539	-

	50ns	LYS430, VAL458, THR474, PHE540	GLU475, MET477, CYS481	-
	75ns	VAL416, ALA428, VAL458, LEU460, ILE472, THR474, LEU528, PHE540	LYS430, GLU475, MET477, CYS481	-
	100ns	ALA428, LYS430, VAL458, LEU460, ILE472, THR474, LEU528, PHE540, LEU542	GLU475, MET477, CYS481	-
E589K	25ns	LEU408, ILE472, ARG525, LEU528, PHE540	GLU475, MET477	LYS430
	50ns	LYS430, THR474, ARG525, LEU528, LEU542	GLU475, MET477	-
	75ns	LEU408, VAL416, LYS430, LEU460, ARG525, LEU528, PHE540, LEU542	THR474, GLU475, MET477	-
	100ns	LEU408, LYS430, LEU460, THR474, LEU528, PHE540	GLU475, MET477	-

By PyContact analysis, we have found that in the wild type BTK-Ibrutinib complex, CYS481 residue showed the maximum hydrogen bond occupancy. In F413L, G584E, and E589K mutant protein-drug complexes, GLU475 showed the maximum hydrogen bond occupancy, while P566Q showed the maximum hydrogen occupancy for both CYS481 and GLU475. PLIP (Protein Ligand Interaction Profiler) analysis for 4-time frames, namely 25 ns, 50 ns, 75 ns, and 100 ns, showed the hydrophobic interactions, hydrogen bonding, and pi-cation interactions in wild and mutant proteins complexed with Ibrutinib (Table 4.4) (Figure 4.11). We have observed that pi-cation interaction is found in a single E589K mutant complex's residue, namely LYS430, at a 25 ns time frame. Being the strongest non-covalent interaction, this pi-cation interaction is actually giving stability to the E589K mutant protein complex, providing another proof that the mutations are distorting the structure, but the protein and drug binding are still unaffected, and hence we can use Ibrutinib in populations where we found these mutations of BTK protein.

V. Conclusions

In this study, we have tried to decipher the effect of mutations that are found in BTK protein and may have a role in SCC progression using computational methods. BTK belongs to the Tec tyrosine kinase family and is a nonreceptor cytoplasmic tyrosine kinase. The Tec kinase family, which includes BTK, TEC, ITK, BMX, and RLK, is the second-largest family of cytoplasmic tyrosine kinases [251]. BTK protein contains a number of domains, beginning with the N-terminal PH-domain (Pleckstrin Homology Domain) and progressing through the TEC homology domain, SH3 and SH2 (SRC homology domains), and lastly, the C-terminal BTK-KD domain (BTK-Kinase Domain). Phosphatidylinositol lipids like PIP3 can be linked to the PH domain, bringing proteins close to the cell membrane. The TH domain comprises a zinc-finger motif that is required for the protein's optimal function and integrity. SH domains communicate with other proteins and bind to proline-rich areas and phosphorylated tyrosinase. LYN or SYK may phosphorylate the catalytic kinase domain's Y551 site, resulting in autophosphorylation of the SH3 domain's Y233 site [210]. It was observed from many studies that mutations in the BTK protein may have a role in SCC disorder, and hence various studies are going on to validate its connection. Ibrutinib is a drug that permanently suppresses BTK. Ibrutinib was initially proposed as a potential treatment for a range of B-cell lineage-derived diseases [214]. Treatment resistance in most cancers may be induced by factors including genomic heterogeneity, intratumoral genetic polymorphism, and field cancerization. These mutations can influence protein expression and subcellular localization, as well as protein folding and integrity, protein function, and protein-protein interactions (PPI). This study was aimed at discovering the possible deleterious role of mutations in the BTK protein. We selected 4 mutants of the BTK protein and observed by conducting simulation studies that these variants are actually distorting the proteins, which we confirmed through MD simulation analysis. Mutants differed greatly from the native protein, suggesting that they contribute to SCC

prognosis by rendering protein unstable. Even the mutant proteins' MD trajectories were distinct from those of normal proteins, and they bound to Ibrutinib in the same manner. According to the analysis of the protein-ligand interaction, the mutations had no effect on how the drug bound to the proteins. MM-PBSA and MM-GBSA both supported this. Even mutant proteins bind to the drug more effectively than wild-type proteins. Consequently, this study reveals that the effect of detected missense mutations is unfavorable and can lead to function loss, which is severe for SCC, but that Ibrutinib-based therapy can still be effective on them, and the mutations can be utilized as biomarkers for Ibrutinib-based treatment. This study suggests that the drug Ibrutinib, which targets BTK and is used to treat SCC, may be unaffected by these mutations. Nonetheless, this finding must be validated by exhaustive clinical investigations.

Chapter 5

Exploring Dysregulated Genes for Novel Targeted Therapies in Squamous Cell Carcinoma

---

Chapter 5. Exploring Dysregulated Genes for Novel Targeted Therapies in Squamous Cell Carcinoma

Abstract

The aim of the present investigation was to identify genes that are dysregulated in SCC and explore potential therapeutic interventions against them. In this study, a comparative analysis was performed on the gene expression profiles retrieved from the GEO database, containing samples taken from individuals diagnosed with SCC, AK, and healthy skin samples to identify any significant differences in gene expression between these groups. As a result of the analysis, several genes were found to be differentially expressed. The study conducted functional and pathway enrichment analysis to determine the involvement of these dysregulated genes in biological processes related to SCC progression. The results showed that these dysregulated genes play a crucial role in cellular disassembly, regulation of protein catabolism, extracellular matrix disassembly, etc. The analysis of pathways has highlighted the significance of WNT signaling, regulation of actin cytoskeleton, etc., in the SCC development. DGB was used to investigate the potential efficacy of specific chemical perturbations, including drugs like Doxorubicin, Dasatinib, and Tretinoin, in restoring abnormal gene expression profiles of these dysregulated genes. The results of this study offer new possibilities for targeted therapeutic interventions in customized therapies by revealing the molecular mechanisms that drive SCC and their effects on the surrounding tumor environment.

I. Introduction

NMSCs, which include SCC and BCC, and melanocytic skin cancers, are the two main classifications for skin cancers [252]. NMSC is a neoplastic disease that rates fifth in worldwide incidences and affects both genders equally. In the United States, over 1.8 million new cases of NMSCs are identified annually, with cutaneous SCC being the most prevalent form of skin cancer. African Americans and Asian Indians have been observed to be more susceptible to developing SCC. In addition, this cancer has the second-highest incidence rate among Hispanic and Chinese/Japanese Asian populations [253]. SCC has been recognized as a cancer of the epidermal keratinocytes. AK, a dermatological condition, has been linked in numerous studies to the development of SCC. A significant proportion of high-risk SCC cases, comprising roughly 5-10% of all occurrences, present a formidable diagnostic and therapeutic challenge, necessitating the application of radiation or surgical interventions. The likelihood of successful therapeutic interventions for metastatic skin cancer, which entails high risk, appears to be diminished, especially in the elderly population, which has an urgent need for a systematic diagnosis and treatment for SCC [254].

Taking advantage of cutting-edge technologies alongside the unique pathophysiological characteristics of tumors may facilitate the discovery of novel therapeutic agents aimed at enhancing the overall survival of individuals afflicted with SCC. Nonetheless, the endeavor of translating these pharmacological agents into the realm of clinical practice is a formidable task, given the low efficacy observed in clinical trials, as well as the excessive expenses and long-lasting timelines that can span over a decade [255]. The strategy of drug repositioning or repurposing presents a viable alternative by discerning novel applications for pre-existing drugs beyond their initial medical indication. The strategic approach of drug repositioning is deemed advantageous due to the pre-existing establishment of pharmacodynamic, pharmacokinetic, and toxicity profiles of these drugs [101]. The plausibility of utilizing publicly available

databases housing data on gene expression and biological pathways in the context of cancer is within reach. Besides being useful to their therapeutic applications, biomarkers may also serve as valuable tools for the diagnosis or prognosis of individuals afflicted with cancer [99], [256].

One such publicly available comprehensive data repository is the Library of Integrated Network-based Cellular Signatures (LINCS) [113]. This repository contains a vast collection of approximately two million distinct files that provide exhaustive insights into the gene expression and metadata of cell lines that have been subjected to chemical perturbation at specific doses and durations. Through the application of computational techniques such as systems biology, bioinformatics, ML, and network biology, it is feasible to establish a relationship between the gene expression profile of cancer and the signature induced by drugs or perturbing agents. The relationship between the two variables at hand can serve as a viable approach to discern novel indications for pre-existing drugs and reutilize them for intricate ailments, including cancer. An additional repository, known as CREEDS [116], exhibits tested gene expression profile annotations that demonstrate associations among drugs, genes, and diseases. The signatures of the data are thoroughly evaluated with great attention to detail in terms of their distinctiveness and quality. Packed circles are utilized to cluster analogous signatures, thereby presenting interrelationships in a graphical format. Moreover, CREEDS offers interactive heatmaps that display hierarchically grouped matrices for all signatures.

Transcriptomics-guided drug repositioning has garnered significant attention, as indicated by the multitude of studies conducted to date [257]. This methodology exhibits particular potential for the treatment of cancer, given the restricted understanding of cancer and the requisite drug categories [108]. Prior research has focused on identifying dysregulated genes and enriched pathways in diverse cancer types, yet no investigations have delved into SCC utilizing gene expression data. The current investigation has identified a group of genes that are responsible

for tumorigenesis and the progression of SCC. These genes have exhibited a negative correlation with antineoplastic signatures that were obtained from the DGB application. The DGB tool is a web-based application designed to prioritize small molecules that are anticipated to have an impact on the expression of target genes [258].

II. Materials and methods

1. Data retrieval and Pre-processing

The data for this study was taken from the previous study, details provided in Chapter 2. Three datasets from the GEO database [115] of NCBI, namely, GSE45216, GSE98744, and GSE108008 were used for this study. These GEO series contained the normalized, calibrated, and pre-processed array data for healthy, AK, and SCC skin samples. The GSE45216 collection contains 30 SCC and 10 AK samples, whereas the GSE98774 collection includes 18 AK and 36 samples of healthy skin. The GSE108008 dataset contains 10 SCC, 10 AK, and 10 healthy skin samples. These datasets were divided into three binary classification problems: Healthy vs. AK, Healthy vs. SCC, and SCC vs. AK, with each case being treated separately for ML application. These above-mentioned binary datasets were then subjected to pre-processing using the RMA-normalization method for microarray summarization and quantile normalization [259].

2. Identification of key dysregulated genes and their statistical analysis

Through the implementation of ML techniques and the SHAP method of XAI, an attempt was made to identify the key dysregulated genes that play a crucial role in the advancement of SCC. The application of the GEO2R computational tool facilitated the characterization of the significance of pivotal dysregulated genes that exhibited differential expression patterns in the course of SCC progression. The statistical significance of these identified genes was

established by verifying that their p-values were below the threshold of 0.05. Furthermore, the regulatory pattern of all these dysregulated genes was ascertained based on their logFC values.

3. Function and Pathway enrichment analysis on the identified key Genes

To perform the Gene Ontology (GO) enrichment analysis as well as pathway enrichment analysis on the key dysregulated genes, we have used STRING [260], [261]. Based on these identified dysregulated genes, STRING automatically enriched pathways or functional subsystems using hypergeometric testing with a P-value of <0.05, set as the threshold. Functional and pathway enrichment was done to find the role of all identified dysregulated genes in the tumour microenvironment.

4. Identification of Chemical perturbations for each dysregulated gene

Recent high-throughput genome-wide expression-based drug screening has resulted in the generation of extensive collections of drug-induced transcriptomic signatures. The LINCS L1000 dataset [113] documents the transcriptomic responsiveness of human cell lines to over 20,000 small molecules, including FDA-approved drugs and preclinical compounds. Crowdsourcing efforts, like CREEDS [116], have been utilized to curate numerous drug-induced gene expression signatures from GEO, with the aim of prioritizing small molecules that significantly modulate single genes. In this study, we utilized the DGB [258], a web-based and mobile application, to identify chemical perturbations against each identified gene. DGB utilizes various datasets to prioritize drugs and small molecule compounds that have the potential to significantly impact the expression of a target gene. The DGB platform provides a user-friendly interface that enables the selection of a target gene and facilitates interaction with the resulting list of small molecules, which are ranked based on the query results.

III. Results

1. Identification of key dysregulated genes and their statistical analysis

The identification of significant genes in the provided datasets was accomplished by employing ML techniques, specifically the XGBoost algorithm [77] and SHAP tool [92], [94] of XAI. The ML algorithm conducted an analysis of the datasets and identified pertinent features that differentiate between various conditions. The study conducted a comparative analysis of gene expression between healthy vs AK skin samples and identified the following genes as significantly differentially expressed: PAMR1, CTSC, PHYHIP, CD24, WNT5A, RAB3B, WIF1, TNNC1, PARK7, MMP14, ARHGEF4, and CFL1. The comparison between healthy vs SCC skin samples revealed the statistical significance of two genes, namely HNRNPM and RPS13. Finally, the comparative analysis of SCC vs AK skin samples revealed the identification of several significant dysregulated genes, namely GTSE1, CHTOP, EDNRB, DNAJC8, S100A11, HNRNPM, TUG1, TFG, GAPDH, and RPS3A [1]. These results contribute to our comprehension of the molecular mechanisms underlying the genetic differences between these conditions, refer to Table 5.1.

Table 5.1: List of identified dysregulated genes identified using ML and SHAP

Datasets	Significant Genes
Healthy vs AK	*PAMR1, CTSC, PHYHIP, CD24, WNT5A, RAB3B, WIF1, TNNC1, PARK7, MMP14, ARHGEF4, CFL1*
Healthy vs SCC	*HNRNPM, RPS13*
SCC vs AK	*GTSE1, CHTOP, EDNRB, DNAJC8, S100A11, HNRNPM, TUG1, TFG, GAPDH, RPS3A*

We employed the GEO2R computational tool to characterize the relevance of important dysregulated genes that were differentially expressed during SCC development. P-values < 0.05 were considered statistically significant for the identified genes. LogFC values were

also determined to find the expression of all the significant genes. According to LogFC values, *PAMR1, PHYHIP, RAB3B, WIF1, TNNC1, HNRNPM, GTSE1, CHTOP, DNAJC8, S100A11, TUG1, TFG,* and *GAPDH* were found to be down-regulated while *CTSC, CD24, WNT5A, PARK7, MMP14, ARHGEF4, CFL1, RPS13, EDNRB,* and *RPS3A* were found to be up-regulated in SCC progression, details given in Table 5.2.

Table 5.2: Statistical analysis results for each identified dysregulated genes.

Genes	P-value	logFC
Dataset: Healthy vs AK Dataset		
PAMR1	6.83E-19	-2.33005
CTSC	3.35E-11	1.042324
PHYHIP	6.51E-28	-2.37733
CD24	8.25E-17	1.790993
WNT5A	3.09E-14	2.491315
RAB3B	3.13E-19	-1.33558
WIF1	1.76E-27	-4.25276
TNNC1	1.32E-19	-1.87799
PARK7	2.71E-06	0.342546
MMP14	2.08E-05	-0.55263
ARHGEF4	0.005396	0.444668
CFL1	4.67E-08	0.367648
Dataset: Healthy vs SCC		
HNRNPM	3.61E-21	-0.78581
RPS13	5.91E-15	0.529767
Dataset: SCC vs AK		
GTSE1	2.91E-21	-1.57927
CHTOP	7.87E-13	-0.55512
EDNRB	3.85E-18	2.066272
DNAJC8	5.62E-13	-0.54132
S100A11	4.49E-21	-1.1461
TUG1	1.66E-11	-0.67914
TFG	2.50E-12	-0.72343

GAPDH	4.98E-05	-0.42799
RPS3A	0.03709	0.100498

2. **Function and Pathway enrichment analysis on the identified key dysregulated Genes to find their role in SCC.**

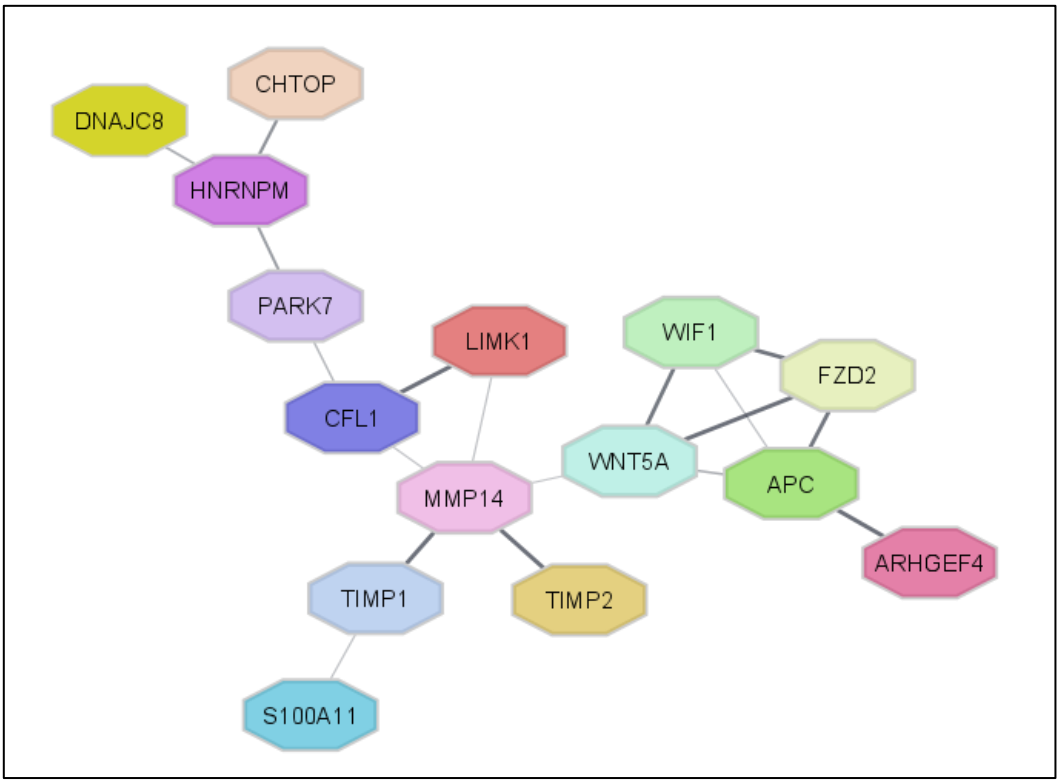


Figure 5.1: STRING network made from the identified dysregulated genes that were retrieved to be of the highest relevance using SHAP values.

The Gene Ontology (GO) enrichment analysis as well as pathway enrichment analysis was performed on the key dysregulated genes using STRING (Figure 5.1) [260], [261]. Identified key dysregulated genes were functionally enriched in six GO_BP (Biological Process) terms, one in GO_MF (Molecular Function) terms, and thirteen GO_CC (Cellular Component) terms. Disassembly of cellular components, Positive regulation of the cellular process, Extracellular matrix disassembly, Regulation of the catabolic process of proteins, negatively regulating proteolysis of membrane protein ectodomains, and Negative regulation of the activity of metallopeptidase was the most highly enriched GO_BP terms. Protein binding was the most

significantly enriched GO_MF term, while the Anchoring junction, Ruffle, Cell leading edge, Extracellular matrix, Cell junction, Ruffle membrane, Intracellular organelle lumen, Extracellular region, Focal adhesion, Endocytic vesicle membrane with clathrin coating, Cell-cell junction, and Lamellipodium were the most significantly enriched GO_CC terms, details in Table 5.3.

Table 5.3: GO terms with their P-value from STRING network.

Term ID	Term Description	P-value	Matching proteins in the network
Biological Process			
GO:0022411	Disassembly of cellular components	0.0039	*TIMP1, APC, TIMP2, MMP14, FZD2, CFL1*
GO:0048522	Positive regulation of cellular process	0.0304	*TIMP1, APC, TIMP2, WNT5A, S100A11, WIF1, MMP14, ARHGEF4, FZD2, LIMK1, CHTOP, PARK7, CFL1*
GO:0022617	Extracellular matrix disassembly	0.0478	*TIMP1, TIMP2, MMP14*
GO:0042176	Regulation of the catabolic process of proteins	0.0478	*TIMP1, APC, TIMP2, WNT5A, PARK7*
GO:0051045	Negatively regulating proteolysis of membrane protein ectodomains	0.0478	*TIMP1, TIMP2*
GO:1905049	Negative regulation of the activity of metallopeptidase	0.0478	*TIMP1, TIMP2*
Molecular Function			
GO:0005515	Protein binding	0.0193	*TIMP1, APC, TIMP2, DNAJC8, WNT5A, S100A11, WIF1, MMP14, ARHGEF4, FZD2, HNRNPM, LIMK1, PARK7, CFL1*
Cellular Component			
GO:0070161	Anchoring junction	0.0019	*APC, S100A11, MMP14, FZD2, LIMK1, PARK7, CFL1*

GO:0001726	Ruffle	0.0079	*APC, S100A11, ARHGEF4, CFL1*
GO:0031252	Cell leading edge	0.0079	*APC, S100A11, ARHGEF4, LIMK1, CFL1*
GO:0031012	Extracellular matrix	0.0148	*TIMP1, TIMP2, WNT5A, MMP14, HNRNPM*
GO:0030054	Cell junction	0.0164	*APC, WNT5A, S100A11, MMP14, FZD2, LIMK1, PARK7, CFL1*
GO:0032587	Ruffle membrane	0.0164	*APC, ARHGEF4, CFL1*
GO:0070013	Intracellular organelle lumen	0.0219	*TIMP1, APC, TIMP2, DNAJC8, WNT5A, S100A11, MMP14, HNRNPM, LIMK1, CHTOP, PARK7, CFL1*
GO:0005576	Extracellular region	0.0338	*TIMP1, TIMP2, DNAJC8, WNT5A, S100A11, WIF1, MMP14, HNRNPM, PARK7, CFL1*
GO:0005925	Focal adhesion	0.0338	*MMP14, FZD2, LIMK1, CFL1*
GO:0005912	Adherens junction	0.0373	*APC, S100A11, PARK7*
GO:0030669	Endocytic vesicle membrane with clathrin coating	0.0407	*WNT5A, FZD2*
GO:0005911	Cell-cell junction	0.0465	*APC, S100A11, PARK7, CFL1*
GO:0030027	Lamellipodium	0.0469	*APC, LIMK1, CFL1*

The identified genes were found to be pathway enriched in nine KEGG pathway [262] terms, five REACTOME [263] terms, and twelve WikiPathway [264] terms, details provided in Table 5.4. KEGG pathway analysis indicated that the key genes are enriched in the WNT signaling pathway, Regulation of actin cytoskeleton, Basal cell carcinoma, Axon guidance, Hippo signaling pathway, signaling pathways regulating pluripotency of stem cells, Breast cancer, Hepatocellular carcinoma, and Gastric carcinoma. Moreover, REACTOME pathway analysis showed the enrichment of key genes in the Activation of Matrix Metalloproteinases, Signaling mediated by TCF in response to WNT, WNT ligand antagonists exerting a negative regulation effect on TCF-dependent signaling, Internalization of FZD2, FZD5, and ROR2 mediated by WNT5A and finally RHO GTPases activating ROCKs while the

WikiPathways showed the enrichment of genes in lncRNA in canonical WNT signaling and colorectal cancer, ncRNAs implicated in hepatocellular carcinoma WNT signaling, Matrix metalloproteinases, WNT signaling, Regulation of actin cytoskeleton, WNT signaling pathway and pluripotency, Embryonic stem cell pluripotency pathways, WNT/beta-catenin signaling pathway in leukemia, Breast cancer pathway, and Extracellular vesicle-mediated signaling in recipient cells, WNT signaling in kidney disease and G13 signaling pathway.

Table 5.4: Pathway terms with their P-value for the STRING network.

Term ID	Term Description	P-value	Matching proteins in the network
KEGG Pathways			
hsa04310	WNT signaling pathway	0.0017	*APC, WNT5A, WIF1, FZD2*
hsa04810	Regulation of actin cytoskeleton	0.0026	*APC, ARHGEF4, LIMK1, CFL1*
hsa05217	Basal cell carcinoma	0.0026	*APC, WNT5A, FZD2*
hsa04360	Axon guidance	0.0137	*WNT5A, LIMK1, CFL1*
hsa04390	Hippo signaling pathway	0.0137	*APC, WNT5A, FZD2*
hsa04550	Signaling mechanisms that control stem cell pluripotency	0.0137	*APC, WNT5A, FZD2*
hsa05224	Breast cancer	0.0137	*APC, WNT5A, FZD2*
hsa05225	Hepatocellular carcinoma	0.0137	*APC, WNT5A, FZD2*
hsa05226	Gastric cancer	0.0137	*APC, WNT5A, FZD2*
Reactome Pathways			
HSA-1592389	Activation of Matrix Metalloproteinases	0.0056	*TIMP1, TIMP2, MMP14*
HSA-201681	Signaling mediated by TCF in response to WNT	0.0152	*APC, WNT5A, WIF1, FZD2*
HSA-3772470	WNT ligand antagonists exerting a negative regulation effect on TCF-dependent signaling	0.0414	*WNT5A, WIF1*

HSA-5140745	Internalization of FZD2, FZD5, and ROR2 mediated by WNT5A	0.0414	*WNT5A, FZD2*
HSA-5627117	RHO GTPases Activate ROCKs	0.0414	*LIMK1, CFL1*
WikiPathways			
WP4258	lncRNA in canonical WNT signaling and colorectal cancer	0.00036	*APC, WNT5A, WIF1, FZD2*
WP4336	ncRNAs implicated in hepatocellular carcinoma WNT signaling	0.00036	*APC, WNT5A, WIF1, FZD2*
WP129	Matrix metalloproteinases	0.00037	*TIMP1, TIMP2, MMP14*
WP428	WNT signaling	0.00037	*APC, WNT5A, WIF1, FZD2*
WP51	Regulation of actin cytoskeleton	0.00062	*APC, ARHGEF4, LIMK1, CFL1*
WP399	WNT signaling pathway and pluripotency	0.0071	*APC, WNT5A, FZD2*
WP3931	Embryonic stem cell pluripotency pathways	0.0091	*APC, WNT5A, FZD2*
WP3658	WNT/beta-catenin signaling pathway in leukemia	0.0154	*APC, WIF1*
WP4262	Breast cancer pathway	0.0154	*APC, WNT5A, FZD2*
WP2870	Extracellular vesicle-mediated signaling in recipient cells	0.0165	*APC, WNT5A*
WP4150	WNT signaling in kidney disease	0.0214	*WNT5A, FZD2*
WP524	G13 signaling pathway	0.0219	*LIMK1, CFL1*

The results of the functional and pathway enrichment analysis provided robust evidence for the implication of the identified dysregulated genes in the advancement of SCC and their impact on the microenvironment of the tumor.

3. Identification of Chemical perturbations for each identified gene

Upon performing a statistical analysis of the identified genes, we determined their significance and observed their respective expression patterns. The DGB, an online software tool, was employed to investigate prospective therapeutic agents for SCC. This application facilitates the assessment of small molecules that are anticipated to modulate the transcription of specific genes. The LIMMA approach [265] confers statistical significance by means of p-value, q-value, fold change, and specificity. The DGB conducted a comprehensive analysis of transcriptomic responses in diverse human cell lines to systematically profile over 20,000 small molecules, encompassing preclinical compounds and FDA-approved drugs.

DGB analysis on identified dysregulated genes presents the details pertaining to the patterns of gene expression and the associated drugs or small molecules capable of reversing these expression patterns. The results in Table 5.5, have been obtained from the LINCS L1000 dataset, a comprehensive gene expression dataset that encompasses data on the impact of numerous drugs and small molecules on gene expression profiles. The initial dataset, namely Healthy vs AK, PAMR1 exhibited a decrease in expression levels when comparing healthy cells to those with AK. According to the DGB analysis, it has been determined that the administration of Doxorubicin at a concentration of 10.2 µM for a duration of 24 hours can successfully restore the anomalous expression profile. This analysis revealed that Doxorubicin had a considerable effect, as evidenced by a low p-value of 6.40e-8 and a q-value of 3.82e-7. The observed log2 fold change denotes a significant upregulation of gene expression with a value of 1.432, and the drug demonstrated a remarkable degree of specificity, as evidenced by a value of 1.16e-4. Another up-regulated gene CTSC expression was found to be effectively

reversed through treatment with Dasatinib at a concentration of 10.0 µM for a duration of 24 hours. The drug exhibited a noteworthy impact, as indicated by the statistical significance of the low p-value (9.60e-6) and q-value (1.63e-4). The observed log2 fold change value of -1.348 denotes a significant reduction in the expression of the gene. It is noteworthy that Dasatinib exhibited a high degree of specificity, with a value of 2.74e-4. Another down-regulated gene, PHYHIP, was investigated in this dataset. The administration of Tretinoin at a concentration of 3.33333 µM for a duration of 6 hours has been demonstrated to be efficacious in reinstating the expression pattern of this target gene. The drug demonstrated a noteworthy effect, as shown by its low p-value of 1.11e-4 and q-value of 1.86e-3. Despite the moderate change in expression indicated by the log2 fold change of 0.398, the drug exhibited a reasonable level of specificity, as evidenced by the p-value of 4.53e-4.

The dataset Healthy vs SCC reports that the application of Cediranib treatment (10.0 µM, 6 hours) can reverse the downregulation of HNRNPM. The drug showed significant effectiveness, as evidenced by the statistically significant low p-value (5.21e-4) and q-value (3.96e-2). The observed log2 fold change of 0.678 indicated a moderate upregulation in gene expression. Additionally, the drug demonstrated satisfactory specificity, as evidenced by a statistically significant value of 4.72e-3.

In the SCC vs AK dataset, the expression of a down-regulated gene CHTOP was found to be moderately up-regulated by administration of Dasatinib at a concentration of 0.4 µM for a duration of 24 hours. The drug exhibited a notable effect, as demonstrated by the statistically significant low p-value (2.55e-3) and q-value (7.88e-3). The data indicate a moderate increase in gene expression as suggested by the log2 fold change value of 0.275. Additionally, the drug demonstrated a statistically significant level of specificity with a value of 1.96e-4. Similarly, all the genes listed in the table were subjected to analysis using DGB in order to identify drugs that target them and assess their efficacy in reversing the observed expression pattern.

Table 5.5: List of drugs/small molecules obtained from L1000 dataset.

Genes	Expression	Drugs/ Small molecules	LINCS sig_id	Cell Line	LINCS pert_id	Time	Dose	p - value	q - value	log2 Fold Change	Specificity
Dataset: Healthy vs AK Dataset											
PAMR1	Down-regulated	Doxorubicin	CRCGN004_HA1E_24H:BRD-A76941896-003-04-6:10.2UM	HA1E	BRD-A76941896	24 h	10.2 µM	6.40e-8	3.82e-7	1.432	1.16e-4
CTSC	Up-regulated	Dasatinib	LJP001_BT20_24H:BRD-K49328571-001-06-9:10	BT20	BRD-K49328571	24 h	10.0 µM	9.60e-6	1.63e-4	-1.348	2.74e-4
PHYHIP	Down-regulated	Tretinoin	AML001_HL60_6H:BRD-K71879491:3.33333	HL60	BRD-K71879491	6 h	3.33333 µM	1.11e-4	1.86e-3	0.398	4.53e-4
CD24	Up-regulated	Resveratrol	CPC010_A375_6H:BRD-K80738081-001-32-8:10	A375	BRD-K80738081	6 h	10.0 µM	6.80e-3	3.98e-2	2.107	4.37e-4
WNT5A	Up-regulated	Doxorubicin	CPC003_PC3_24H:BRD-K92093830-003-05-0:10	PC3	BRD-K92093830	24 h	10.0 µM	8.92e-4	2.76e-3	-1.125	1.78e-4
RAB3B	Down-regulated	Tretinoin	CPC006_HCC515_24H:BRD-K64634304-001-01-5:40	HCC515	BRD-K64634304	24 h	40.0 µM	1.80e-3	1.22e-2	0.343	3.18e-4
WIF1	Down-regulated	Imatinib	LJP001_MCF10A_24H:BRD-K92723993-066-12-8:2	MCF10A	BRD-K92723993	24 h	2.0 µM	2.28e-3	2.13e-2	1.264	5.23e-4

TNNC1	Down-regulated	Dasatinib	LJP001_MDAMB231_24H:BRD-K49328571-001-06-9:0.4	MDAMB231	BRD-K49328571	24 h	0.4 µM	6.65e-3	1.77e-2	0.630	1.96e-4
PARK7	Up-regulated	Bortezomib	CPC006_HT29_24H:BRD-K88510285-001-01-2:0.04	HT29	BRD-K88510285	24 h	0.04 µM	2.32e-6	8.18e-6	-0.784	1.46e-4
MMP14	Down-regulated	Imatinib	LJP001_BT20_6H:BRD-K92723993-066-12-8:0.4	BT20	BRD-K92723993	6 h	0.4 µM	6.48e-5	4.01e-3	0.878	1.06e-3
ARHGEF4	Up-regulated	Vemurafenib	CPC006_HCC515_24H:BRD-K56343971-001-02-3:10	HCC515	BRD-K56343971	24 h	10.0 µM	1.46e-3	2.48e-2	-0.575	1.06e-3
CFL1	Up-regulated	Tretinoin	AML001_HL60_24H:BRD-K71879491:0.37037	HL60	BRD-K71879491	24 h	0.37037 µM	2.82e-3	2.61e-2	-0.264	4.85e-4
Dataset: Healthy vs SCC											
HNRNPM	Down-regulated	Cediranib	CPC014_VCAP_6H:BRD-K86930074-001-01-9:10	VCAP	BRD-K86930074	6 h	10.0 µM	5.21e-4	3.96e-2	0.678	4.72e-3
RPS13	Up-regulated	Cytarabine	CPC011_VCAP_24H:BRD-K33106058-001-12-7:10	VCAP	BRD-K33106058	24 h	10.0 µM	6.05e-3	2.99e-2	-0.135	3.61e-4
Dataset: SCC vs AK											
GTSE1	Down-regulated	Doxorubicin	CPC015_ASC_24H:BRD-K92093830-003-05-0:10	ASC	BRD-K92093830	24 h	10.0 µM	3.31e-5	2.82e-4	0.837	2.04e-4

CHTOP	Down-regulated	Dasatinib	LJP001_MDAMB231_24H:BRD-K49328571-001-06-9:0.4	MDAMB231	BRD-K49328571	24 h	0.4 µM	2.55e-3	7.88e-3	0.275	1.96e-4
EDNRB	Up-regulated	Estradiol	ERG005_VCAP_6H:BRD-K18910433-001-14-3:5	VCAP	BRD-K18910433	6 h	5.0 µM	1.11e-2	4.61e-2	-0.525	3.59e-4
DNAJC8	Down-regulated	Doxorubicin	CRCGN004_HA1E_24H:BRD-A76941896-003-04-6:10.2UM	HA1E	BRD-A76941896	24 h	10.2 µM	1.89e-8	1.34e-7	0.803	1.16e-4
S100A11	Down-regulated	Dasatinib	LJP001_MDAMB231_6H:BRD-K49328571-001-06-9:0.08	MDAMB231	BRD-K49328571	6 h	0.08 µM	5.23e-4	1.75e-2	0.895	1.55e-3
TUG1	Down-regulated	Vemurafenib	CPC006_A375_24H:BRD-K56343971-001-02-3:10	A375	BRD-K56343971	24 h	10.0 µM	1.01e-10	8.16e-10	1.230	1.31e-4
TFG	Down-regulated	Bortezomib	CPC006_MCF7_24H:BRD-K88510285-001-01-2:0.04	MCF7	BRD-K88510285	24 h	0.04 µM	3.36e-13	8.60e-12	1.032	1.59e-4
GAPDH	Down-regulated	Doxorubicin	CPC004_HT29_6H:BRD-A52530684-001-01-1:10	HT29	BRD-A52530684	6 h	10.0 µM	2.36e-4	1.67e-3	0.693	2.39e-4
RPS3A	Up-regulated	Estradiol	CPC014_VCAP_24H:BRD-K86930074-001-01-9:10	VCAP	BRD-K86930074	24 h	10.0 µM	1.71e-3	1.25e-2	-0.232	3.93e-4

Crowdsourced data from the Connectivity Map's resource CREEDS was utilized to identify drugs targeting dysregulated genes through the use of DGB. This analysis was done to discover prospective drugs or small molecules that could counteract the aberrant gene expression profiles linked with SCC with the aim of investigating novel therapeutic avenues, and the results are compiled in Table 5.6. In the Healthy vs AK dataset, a number of genes were being dysregulated, and corresponding drugs were found to effectively reverse their abnormal expression patterns. The gene PAMR1 exhibited down-regulation, and the administration of Doxorubicin resulted in a notable effect in reinstating its expression. The findings of this analysis are in line with prior analysis conducted on the LINCS L1000 dataset, which serves to reinforce the effectiveness of Doxorubicin in its ability to target PAMR1. Similarly, the observation held true for all dysregulated genes with LINCS L1000 dataset in the remaining datasets also, namely Healthy vs. SCC and SCC vs. AK.

These findings obtained by the DGB underscore the capacity of certain drugs to restore the expression profiles of aberrantly regulated genes linked to SCC. The findings also offer significant perspectives on possible therapeutic remedies for SCC and advocate for the investigation of personalized treatment approaches that focus on the regulation of genes that are not functioning properly.

Table 5.6: List of drugs/small molecules obtained from CREEDS.

Genes	Expression	Drugs/Small molecules	CREEDS ID	GEO ID	Drugbank ID	PubChem ID	p - value	q - value	log2 Fold Change	Specificity
Dataset: Healthy vs AK Dataset										
PAMR1	Down-regulated	Doxorubicin	drug:3263	GSE6930	DB00997	31703	4.84e-8	1.07e-7	0.295	1.69e-4
CTSC	Up-regulated	Dasatinib	drug:3306	GSE59357	DB01254	3062316	8.62e-6	3.94e-5	-0.427	2.27e-4
PHYHIP	Down-regulated	Tretinoin	drug:2828	GSE23702	DB00755	444795	1.07e-4	2.23e-4	0.447	1.08e-4
CD24	Up-regulated	Resveratrol	drug:3500	GSE25412	DB02709	445154	2.38e-6	3.18e-5	1.090	2.96e-4
WNT5A	Up-regulated	Doxorubicin	drug:3263	GSE6930	DB00997	31703	7.29e-17	4.39e-15	-2.126	1.78e-4
RAB3B	Down-regulated	Tretinoin	drug:2828	GSE23702	DB00755	444795	5.73e-7	2.74e-6	0.650	1.08e-4
WIF1	Down-regulated	Imatinib	drug:2764	GSE24493	DB00619	5291	6.08e-4	1.79e-3	0.140	1.54e-4
TNNC1	Down-regulated	Dasatinib	drug:3306	GSE59357	DB01254	3062316	5.17e-13	1.28e-10	1.736	1.27e-4
PARK7	Up-regulated	Bortezomib	drug:2686	GSE30931	DB00188	387447	6.12e-4	1.09e-2	-0.284	1.02e-3
MMP14	Down-regulated	Imatinib	drug:2764	GSE24493	DB00619	5291	2.16e-6	2.46e-5	0.445	1.54e-4
ARHGEF4	Up-regulated	Vemurafenib	drug:2495	GSE42872	DB08881	42611257	1.68e-4	6.32e-4	-0.334	1.62e-4
CFL1	Up-regulated	Tretinoin	drug:3233	GSE23702	DB00755	444795	3.47e-4	5.92e-4	-0.137	1.31e-4

Dataset: Healthy vs SCC										
HNRNPM	Down-regulated	Cediranib	drug:2642	GSE32569	DB04849	9933475	9.14e-6	2.00e-5	0.030	1.07e-4
RPS13	Up-regulated	Cytarabine	drug:3422	GSE6930	DB00987	6253	3.00e-3	1.25e-2	-0.173	5.84e-4
Dataset: SCC vs AK										
GTSE1	Down-regulated	Doxorubicin	drug:3265	GSE6930	DB00997	31703	2.39e-2	4.26e-2	0.290	4.46e-4
CHTOP	Down-regulated	Dasatinib	drug:3306	GSE59357	DB01254	3062316	9.70e-7	5.97e-6	0.490	1.27e-4
EDNRB	Up-regulated	Estradiol	drug:3203	GSE12446	DB00783	5757	1.89e-4	2.43e-3	-1.518	4.87e-4
DNAJC8	Down-regulated	Doxorubicin	drug:2720	GSE12972	DB00997	31703	2.85e-17	6.29e-17	0.204	2.03e-4
S100A11	Down-regulated	Dasatinib	drug:3306	GSE59357	DB01254	3062316	4.82e-8	4.75e-7	0.535	1.27e-4
TUG1	Down-regulated	Vemurafenib	drug:2564	GSE37441	DB08881	42611257	1.59e-5	1.17e-4	0.469	2.89e-4
TFG	Down-regulated	Bortezomib	drug:2686	GSE30931	DB00188	387447	1.59e-5	7.02e-4	0.455	1.02e-3
GAPDH	Down-regulated	Doxorubicin	drug:2720	GSE12972	DB00997	31703	5.66e-10	8.90e-10	0.091	2.03e-4
RPS3A	Up-regulated	Cediranib	drug:2642	GSE32569	DB04849	9933475	2.85e-5	5.66e-5	-0.024	1.20e-4

By utilizing DGB analysis on dysregulated genes, valuable insights were gained on potential drug and small molecule reversals of these patterns. The LINCS L1000 dataset and CREEDS were crucial in obtaining these results, which revealed successful restoration of anomalous expression profiles by certain drugs like Doxorubicin and Dasatinib, as well as the reinstatement of target gene expression by Tretinoin and Cediranib. These drugs were statistically significant in p-values and q-values, demonstrating varying degrees of specificity, suggesting their ability to target specific genes. These findings suggest the potential of drugs to counteract dysregulated gene expression, and DGB analysis can provide further insight into identifying therapeutic interventions for other dysregulated genes.

IV. Discussion

The study was conducted to identify primary dysregulated genes linked to SCC and explore the potential drugs targeting them. The study applied the ML approach, specifically the XGBoost algorithm and the SHAP tool of XAI, to identify significant genes in the given datasets, namely healthy vs AK, healthy vs SCC and SCC vs AK. A set of dysregulated genes were identified through this analysis, which showed a significant association with SCC and AK. The study identified several upregulated genes in SCC, including CTSC, CD24, WNT5A, PARK7, MMP14, ARHGEF4, CFL1, RPS13, EDNRB, and RPS3A and also found the genes that were downregulated, including PAMR1, PHYHIP, TNNC1, HNRNPM, GTSE1, CHTOP, DNAJC8, S100A11, TUG1, TFG, and GAPDH. These dysregulated genes have been found to play a crucial role in the molecular mechanisms that drive the progression of SCC, as indicated by the findings.

Gene ontology (GO) enrichment [266] analysis and pathway enrichment [267] analysis were conducted to obtain insights into the functional roles of these dysregulated genes. The analysis of gene ontology indicated that these dysregulated genes exhibited enrichment in diverse

cellular components, molecular functions, and biological processes. The study also conducted pathway enrichment analysis to identify significant pathways that may be involved in the development of SCC. The findings from these analyses suggest that these genes may play a crucial role in SCC development. The DGB tool was utilized to identify chemical perturbations for dysregulated genes that have the potential to reverse the abnormal expression patterns of these dysregulated genes and discovered that Doxorubicin, Dasatinib, Tretinoin, etc., were effective in restoring the expression of PAMR1, CTSC, PHYHIP, etc., respectively.

The findings of this investigation enhance our comprehension of the molecular pathways involved in the advancement of SCC and emphasize the possible therapeutic approaches for SCC. These findings provide valuable information for further research and the development of personalized treatment approaches for SCC. Additional experimental validation and clinical investigations are required to verify the efficacy of these potential therapeutic interventions and their implementation in clinical settings.

V. Conclusion

A deeper understanding of the molecular mechanisms driving SCC growth and the therapeutic treatments that could target these mechanisms has been made possible by the systematic investigation of dysregulated genes in SCC. The utilization of ML algorithms and XAI to identify significant dysregulated genes holds promise for their potential application in the healthcare industry, specifically for advanced diagnostic and therapeutic purposes. The results of the investigation provide valuable insights into the understanding of SCC development and have the potential to facilitate the establishment of precise treatment approaches. The analysis of dysregulated genes through functional and pathway enrichment has identified significant biological processes and pathways implicated in SCC, providing direction for future investigations into potential diagnostic markers or therapeutic targets. The detection of

chemical perturbations that can reverse anomalous gene expression patterns presents novel prospects for focused treatments, underscoring the promise of precision medicine. This research results possess significant implications in the healthcare sector, as they can serve as biomarkers for the diagnosis, prognosis, and monitoring of SCC. Furthermore, they provide opportunities for the development of personalized treatment approaches aimed at enhancing patient outcomes and quality of life.

Chapter 6

Summary and Future Prospects

Chapter 6. Summary and Future Prospects

AI and ML have brought about significant change in various sectors. These state-of-the-art technologies have altered the traditional approach to problem-solving and decision-making. AI is responsible for creating systems that can mimic human intelligence, thereby facilitating complex tasks like speech recognition, image classification, and data analysis. Meanwhile, ML has become a subset of AI that concentrates mainly on algorithms that improve machines' workflow without any explicit programming via data reinforcement. AI and ML have proven to be beneficial in many industries. In the healthcare sector, AI-based systems can analyze extensive medical data to assist in disease diagnosis, recommend personalized treatments, and forecast patient outcomes. ML algorithms can detect fraudulent transactions, optimize investment strategies, and automate customer service in the finance industry. Additionally, virtual assistants, recommendation systems, and tailored advertisements powered by AI have revolutionized how we engage with technology and access information, improving user experiences [268].

As AI and ML models become increasingly intricate and widespread, questions regarding their transparency and accountability arise. The opaque nature of AI systems often renders their decision-making processes incomprehensible to humans, impeding acceptance, restricting their implementation in crucial areas, and raising ethical apprehensions. In recent years, the development of XAI has become increasingly important to help address concerns and challenges related to artificial intelligence. XAI aims to create AI systems that can provide clear explanations for their decisions and actions, making it easier for humans to understand and trust them. This enhanced transparency and accountability can help build greater confidence in AI technology and its potential applications across a range of fields. By employing XAI techniques, AI and ML can be improved in various ways. One benefit is that

XAI aids in establishing user and stakeholder confidence and approval. When people comprehend the decision-making process of AI models, they are more inclined to trust and depend on the technology. This is especially critical in fields such as healthcare, where precise diagnoses and treatment suggestions are essential. Second, XAI plays a crucial role in promoting regulatory compliance and ethical considerations within AI systems [269]. Due to the potential impact on people's lives, it is essential to ensure fairness, avoid bias, and adhere to legal and ethical guidelines. Explainability can aid in identifying and mitigating potential biases or discriminatory patterns by providing insights into the decision-making process. This allows for corrective actions, if necessary, promoting fair and ethical practices in AI systems. XAI encourages the joint efforts of humans and AI [270]. With XAI's ability to enhance the comprehensibility of AI systems, it enables humans to team up with AI models, taking advantage of their strengths while reducing their weaknesses. Experts can authenticate and approve the judgments made by AI systems, resulting in more resilient and dependable results [271].

In recent years, there has been a growing interest in the potential of XAI to identify biomarkers associated with various conditions, so for the first study, we have used XAI to identify biomarkers related to Squamous Cell Carcinoma (SCC). A two-phase methodology was used for the development of a classification model based on the predictive performance of the ML algorithm XGBoost, followed by the application of XAI techniques to establish interpretability by linking model outputs to relevant genes. The XGBoost model was trained using a dataset comprised of genes linked to SCC to create the classification model. To ensure interpretability, the XAI techniques were implemented, which include the use of SHAP barplots and summary plots. These visualization tools allowed for a better understanding of the model's predictions by showing the significance of individual gene contributions in relation to SCC classification results and hence, establishing a link between genes associated with SCC and the output of the

model. After analyzing the data, it was discovered that there were certain genes that played a major role in the XGBoost model's accuracy, and these genes could have implications for cancer development. Adding SHAP values for interpretability did not affect the model's accuracy, indicating the usefulness of XAI methods. Additionally, the identified genes could potentially be targeted for managing SCC, making them important for predictive and prognostic reasons in the biomedical field.

The potential for XAI looks very positive as the industry progresses and gains more experience. A significant focus is on improving the ways of interpreting XAI through different techniques. Experts are currently experimenting with unique mechanisms to offer more transparent and straightforward explanations for AI model decisions. As feature importance analysis, rule extraction algorithms, and visualization techniques advance, it'll lead to more sophisticated explanations [55]. An interactive interface or immersive environment could be the future to facilitate the exploration and comprehension of AI decision-making processes at a deeper level. One potential area for future development in XAI is the incorporation of specialized knowledge and expertise. By merging AI models with expert knowledge from specific domains like finance or healthcare, the explanations generated by AI systems can become more interpretable [272]. This blend of AI and domain-specific knowledge can lead to more comprehensive and accurate explanations, enabling individuals to make informed decisions based on AI recommendations. Consequently, cooperative endeavors between AI researchers and industry experts could result in more effective and insightful explanations. The field of XAI must also take into account the ethical implications of AI and ML [268]. XAI will be instrumental in detecting and lessening biases, guaranteeing equity, and curbing discrimination in AI-based decision-making. Transparent explanations provided by XAI can reveal possible biases and reveal the decision-making procedure, leading to corrective measures. Future advancements will concentrate on creating XAI approaches that actively tackle ethical concerns and

encourage the ethical and impartial use of AI technologies. In the future, XAI development will give significant attention to human-centric design. The objective is to put the needs and preferences of human users first. By doing so, XAI explanations become more intuitive and meaningful, and the techniques used to generate these explanations become actionable. To ensure trust and transparency in AI reasoning, XAI methods will be designed to communicate effectively with non-experts. This approach empowers individuals to make informed decisions based on AI recommendations, enhancing acceptance and adoption of AI technologies. To ensure that AI systems can function optimally in dynamic environments and make real-time decisions, future XAI developments will focus on providing instantaneous and dynamic explanations. Real-time XAI will play a vital role in maintaining user trust and confidence in AI systems by enabling them to understand the reasoning behind AI decisions as they occur. In high-stakes scenarios like autonomous vehicles or critical healthcare interventions, real-time explanations can enhance safety and foster more efficient human-AI collaborations [52].

The purpose of the second objective was two-fold. It aimed to assess how somatic non-synonymous mutations influence the BTK protein and its effect on FDA-approved therapies for skin cancer. Molecular dynamics simulations were employed to examine the impact of individual amino acid mutations on the stability of the BTK protein. According to the results, these mutations could potentially make the protein unstable and affect the prognosis of SCC. Subsequently, a study was undertaken to examine how the BTK protein and its mutants interacted with Ibrutinib, and it was found that the mutants had similar binding to Ibrutinib as the wild-type protein, suggesting that Ibrutinib could be an effective therapy for treating SCC mutations. To expand on the given directions, more research can be conducted on the effectiveness of Ibrutinib-based therapy through clinical trials and preclinical studies for SCC patients. This initiative can aid in determining the actual influence of BTK protein mutations on SCC's prognosis and confirm Ibrutinib's potential as a directed treatment for such

mutations. Moreover, the results could act as a catalyst in the creation of new medications or treatments focused on treating BTK protein mutations in SCC, hoping to enhance patient outcomes and diversify skin cancer treatment options [2], [210].

The primary focus of the third objective was to pinpoint chemical perturbations against biomarkers or dysregulated genes that are associated with SCC development. By thoroughly scrutinizing the gene expression profiles of individuals afflicted with SCC, healthy counterparts, and AK patients, we managed to identify a number of dysregulated genes that significantly contribute to SCC progression. These genes are closely linked to crucial biological processes and pathways known for their importance in the evolution and advancement of SCC. We used the DGB tool, which is tailor-made for studying potential therapeutic interventions, to improve our research results. Our analysis uncovered the impressive efficacy of certain medications, including Doxorubicin, Dasatinib, and Tretinoin, among others, in correcting abnormal expression patterns of these identified dysregulated genes linked with SCC [273]–[275]. This discovery provides hope for precise, personalized treatment and opens up possibilities for innovative drug development and repurposing strategies in cancer research.
The outlook for this study is very positive as it has highlighted the dysregulated genes and their associated biological processes in SCC. This discovery has presented new possibilities for targeted therapies where researchers can focus on specific genes and related pathways to develop novel drugs that can precisely target the molecular abnormalities underlying SCC. Repurposing existing drugs such as Doxorubicin, Dasatinib, and Tretinoin, as evidenced by their effectiveness, offers an economical and time-saving way to develop treatments for SCC. Tailoring treatments to address the specific dysregulated genes of each SCC patient, based on their genetic profile and gene expression patterns, has the potential to significantly improve therapeutic outcomes. This personalized approach can boost the efficacy of interventions while

reducing unwanted side effects, offering hope for more effective treatment options for SCC patients.

Chapter 7

References

References

[1] J. Meena and Y. Hasija, "Application of explainable artificial intelligence in the identification of Squamous Cell Carcinoma biomarkers," *Comput Biol Med*, vol. 146, Jul. 2022, doi: 10.1016/J.COMPBIOMED.2022.105505.

[2] J. Meena and Y. Hasija, "Rare deleterious mutations in Bruton's tyrosine kinase as biomarkers for ibrutinib-based therapy: an in silico insight," *J Mol Model*, vol. 29, no. 4, p. 120, Apr. 2023, doi: 10.1007/S00894-023-05515-6/FIGURES/11.

[3] V. Gligorijević and N. Pržulj, "Methods for biological data integration: perspectives and challenges," *J R Soc Interface*, vol. 12, no. 112, Nov. 2015, doi: 10.1098/RSIF.2015.0571.

[4] V. Gligorijević and N. Pržulj, "Computational Methods for Integration of Biological Data," pp. 137–178, 2016, doi: 10.1007/978-3-319-39349-0_8.

[5] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, "Multi-omics Data Integration, Interpretation, and Its Application," *Bioinform Biol Insights*, vol. 14, 2020, doi: 10.1177/1177932219899051.

[6] D. Minh, H. X. Wang, Y. F. Li, and T. N. Nguyen, "Explainable artificial intelligence: a comprehensive review," *Artificial Intelligence Review 2021 55:5*, vol. 55, no. 5, pp. 3503–3568, Nov. 2021, doi: 10.1007/S10462-021-10088-Y.

[7] H. W. Loh, C. P. Ooi, S. Seoni, P. D. Barua, F. Molinari, and U. R. Acharya, "Application of explainable artificial intelligence for healthcare: A systematic review of the last decade (2011–2022)," *Comput Methods Programs Biomed*, vol. 226, p. 107161, Nov. 2022, doi: 10.1016/J.CMPB.2022.107161.

[8] J. Vamathevan *et al.*, "Applications of machine learning in drug discovery and development," *Nature Reviews Drug Discovery 2019 18:6*, vol. 18, no. 6, pp. 463–477, Apr. 2019, doi: 10.1038/s41573-019-0024-5.

[9] S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu, and M. J. Ahsan, “Machine Learning in Drug Discovery: A Review,” *Artif Intell Rev*, vol. 55, no. 3, p. 1947, Mar. 2022, doi: 10.1007/S10462-021-10058-4.

[10] Y. Masoudi-Sobhanzadeh, Y. Omidi, M. Amanlou, and A. Masoudi-Nejad, “Drug databases and their contributions to drug repurposing,” *Genomics*, vol. 112, no. 2, pp. 1087–1095, Mar. 2020, doi: 10.1016/J.YGENO.2019.06.021.

[11] X. Xia, “Bioinformatics and Drug Discovery,” *Curr Top Med Chem*, vol. 17, no. 15, p. 1709, Apr. 2017, doi: 10.2174/1568026617666161116143440.

[12] D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia, and R. K. Tekade, “Artificial intelligence in drug discovery and development,” *Drug Discov Today*, vol. 26, no. 1, p. 80, Jan. 2021, doi: 10.1016/J.DRUDIS.2020.10.010.

[13] D. Cirillo and A. Valencia, “Big data analytics for personalized medicine,” *Curr Opin Biotechnol*, vol. 58, pp. 161–167, Aug. 2019, doi: 10.1016/J.COPBIO.2019.03.004.

[14] G. S. Krishnan, A. Joshi, and V. Kaushik, “Bioinformatics in Personalized Medicine,” *Adv Bioinformatics*, pp. 303–315, Jan. 2021, doi: 10.1007/978-981-33-6191-1_15/COVER.

[15] E. D. Esplin, L. Oei, and M. P. Snyder, “Personalized sequencing and the future of medicine: discovery, diagnosis and defeat of disease,” *Pharmacogenomics*, vol. 15, no. 14, p. 1771, Nov. 2014, doi: 10.2217/PGS.14.117.

[16] S. Al-Amrani, Z. Al-Jabri, A. Al-Zaabi, J. Alshekaili, and M. Al-Khabori, “Proteomics: Concepts and applications in human medicine,” *World J Biol Chem*, vol. 12, no. 5, p. 57, Sep. 2021, doi: 10.4331/WJBC.V12.I5.57.

[17] Y. Hasin, M. Seldin, and A. Lusis, “Multi-omics approaches to disease,” *Genome Biology 2017 18:1*, vol. 18, no. 1, pp. 1–15, May 2017, doi: 10.1186/S13059-017-1215-1.

[18] J. Montaner *et al.*, “Multilevel omics for the discovery of biomarkers and therapeutic targets for stroke,” *Nature Reviews Neurology 2020 16:5*, vol. 16, no. 5, pp. 247–264, Apr. 2020, doi: 10.1038/s41582-020-0350-6.

[19] C. M. Micheel *et al.*, “Omics-Based Clinical Discovery: Science, Technology, and Applications,” Mar. 2012, Accessed: Jun. 14, 2023. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK202165/

[20] M. Zambrano-Román, J. R. Padilla-Gutiérrez, Y. Valle, J. F. Muñoz-Valle, and E. Valdés-Alvarado, “Non-Melanoma Skin Cancer: A Genetic Update and Future Perspectives,” *Cancers (Basel)*, vol. 14, no. 10, May 2022, doi: 10.3390/CANCERS14102371.

[21] M. Ciążyńska *et al.*, “The incidence and clinical analysis of non-melanoma skin cancer,” *Sci Rep*, vol. 11, no. 1, p. 4337, Dec. 2021, doi: 10.1038/S41598-021-83502-8.

[22] A. Sánchez-Danés and C. Blanpain, “Deciphering the cells of origin of squamous cell carcinomas,” *Nature Reviews Cancer 2018 18:9*, vol. 18, no. 9, pp. 549–561, May 2018, doi: 10.1038/s41568-018-0024-5.

[23] D. E. Johnson, B. Burtness, C. R. Leemans, V. W. Y. Lui, J. E. Bauman, and J. R. Grandis, “Head and neck squamous cell carcinoma,” *Nature Reviews Disease Primers 2020 6:1*, vol. 6, no. 1, pp. 1–22, Nov. 2020, doi: 10.1038/s41572-020-00224-3.

[24] L. Fania *et al.*, “Cutaneous Squamous Cell Carcinoma: From Pathophysiology to Novel Therapeutic Approaches,” *Biomedicines*, vol. 9, no. 2, pp. 1–33, Feb. 2021, doi: 10.3390/BIOMEDICINES9020171.

[25] H. Li, Y. Zhang, M. Xu, and D. Yang, “Current trends of targeted therapy for oral squamous cell carcinoma,” *Journal of Cancer Research and Clinical Oncology 2022 148:9*, vol. 148, no. 9, pp. 2169–2186, May 2022, doi: 10.1007/S00432-022-04028-8.

[26] A. S. Halim and N. Ramasenderan, “High-risk cutaneous squamous cell carcinoma (CSCC): Challenges and emerging therapies,” *Asian J Surg*, vol. 46, no. 1, pp. 47–51, Jan. 2023, doi: 10.1016/J.ASJSUR.2022.04.079.

[27] K. Y. Sarin *et al.*, “Genome-wide meta-analysis identifies eight new susceptibility loci for cutaneous squamous cell carcinoma,” *Nature Communications 2020 11:1*, vol. 11, no. 1, pp. 1–8, Feb. 2020, doi: 10.1038/s41467-020-14594-5.

[28] A. C. Green and C. M. Olsen, “Cutaneous squamous cell carcinoma: an epidemiological review,” *Br J Dermatol*, vol. 177, no. 2, pp. 373–381, Aug. 2017, doi: 10.1111/BJD.15324.

[29] I. Farooq and A. Bugshan, “Oral squamous cell carcinoma: metastasis, potentially associated malignant disorders, etiology and recent advancements in diagnosis,” *F1000Res*, vol. 9, 2020, doi: 10.12688/F1000RESEARCH.22941.1.

[30] D. T. Debela *et al.*, “New approaches and procedures for cancer treatment: Current perspectives,” *SAGE Open Med*, vol. 9, 2021, doi: 10.1177/20503121211034366.

[31] A. Likhacheva *et al.*, “Definitive and Postoperative Radiation Therapy for Basal and Squamous Cell Cancers of the Skin: Executive Summary of an American Society for Radiation Oncology Clinical Practice Guideline,” *Pract Radiat Oncol*, vol. 10, no. 1, pp. 8–20, Jan. 2020, doi: 10.1016/J.PRRO.2019.10.014.

[32] R. C. Deconti, “Chemotherapy of squamous cell carcinoma of the skin,” *Semin Oncol*, vol. 39, no. 2, pp. 145–149, Apr. 2012, doi: 10.1053/J.SEMINONCOL.2012.01.002.

[33] C. Shreve, C. Shropshire, and D. G. Cotter, “Metastatic Squamous Cell Carcinoma: A Cautionary Tale,” *Cureus*, vol. 12, no. 10, Oct. 2020, doi: 10.7759/CUREUS.10879.

[34] O. Elemento, C. Leslie, J. Lundin, and G. Tourassi, “Artificial intelligence in cancer research, diagnosis and therapy,” *Nature Reviews Cancer 2021 21:12*, vol. 21, no. 12, pp. 747–752, Sep. 2021, doi: 10.1038/s41568-021-00399-1.

[35] Y. Kumar, A. Koul, R. Singla, and M. F. Ijaz, “Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda,” *J Ambient Intell Humaniz Comput*, vol. 14, no. 7, p. 8459, 2023, doi: 10.1007/S12652-021-03612-Z.

[36] N. Savage, “Tapping into the drug discovery potential of AI,” *Biopharma Dealmakers*, May 2021, doi: 10.1038/D43747-021-00045-7.

[37] G. P. Dotto and A. K. Rustgi, “Squamous cell cancers: a unified perspective on biology and genetics,” *Cancer Cell*, vol. 29, no. 5, p. 622, May 2016, doi: 10.1016/J.CCELL.2016.04.004.

[38] F. F. Munari *et al.*, “PIK3CA mutations are frequent in esophageal squamous cell carcinoma associated with chagasic megaesophagus and are associated with a worse patient outcome,” *Infect Agent Cancer*, vol. 13, no. 1, Dec. 2018, doi: 10.1186/S13027-018-0216-3.

[39] T. de Bakker *et al.*, “Restoring p53 Function in Head and Neck Squamous Cell Carcinoma to Improve Treatments,” *Front Oncol*, vol. 11, Jan. 2021, doi: 10.3389/FONC.2021.799993.

[40] S. S. Padhi *et al.*, “Role of CDKN2A/p16 expression in the prognostication of oral squamous cell carcinoma,” *Oral Oncol*, vol. 73, pp. 27–35, Oct. 2017, doi: 10.1016/J.ORALONCOLOGY.2017.07.030.

[41] F. Acker *et al.*, “KRAS Mutations in Squamous Cell Carcinomas of the Lung,” *Front Oncol*, vol. 11, Dec. 2021, doi: 10.3389/FONC.2021.788084.

[42] P. M. Rodust, E. Stockfleth, C. Ulrich, M. Leverkus, and J. Eberle, “UV-induced squamous cell carcinoma--a role for antiapoptotic signaling pathways,” *Br J Dermatol*, vol. 161 Suppl 3, no. SUPPL. 3, pp. 107–115, Nov. 2009, doi: 10.1111/J.1365-2133.2009.09458.X.

[43] M. Zhang *et al.*, “Identification and validation of potential novel biomarkers for oral squamous cell carcinoma,” *Bioengineered*, vol. 12, no. 1, p. 8845, 2021, doi: 10.1080/21655979.2021.1987089.

[44] V.-M. Voiculescu *et al.*, “Squamous Cell Carcinoma: Biomarkers and Potential Therapeutic Targets,” *Human Skin Cancers - Pathways, Mechanisms, Targets and Treatments*, Dec. 2017, doi: 10.5772/INTECHOPEN.70767.

[45] J. Pillai, T. Chincholkar, R. Dixit, and M. Pandey, “A systematic review of proteomic biomarkers in oral squamous cell cancer,” *World Journal of Surgical Oncology 2021 19:1*, vol. 19, no. 1, pp. 1–28, Oct. 2021, doi: 10.1186/S12957-021-02423-Y.

[46] P. Kaur, A. Singh, and I. Chana, “Computational Techniques and Tools for Omics Data Analysis: State-of-the-Art, Challenges, and Future Directions,” *Archives of Computational Methods in Engineering 2021 28:7*, vol. 28, no. 7, pp. 4595–4631, Feb. 2021, doi: 10.1007/S11831-021-09547-0.

[47] R. Thirunavukarasu, G. P. D. C, G. R, M. Gopikrishnan, and V. Palanisamy, “Towards computational solutions for precision medicine based big data healthcare system using deep learning models: A review,” *Comput Biol Med*, vol. 149, p. 106020, Oct. 2022, doi: 10.1016/J.COMPBIOMED.2022.106020.

[48] P. Kaur, A. Singh, and I. Chana, “Computational Techniques and Tools for Omics Data Analysis: State-of-the-Art, Challenges, and Future Directions,” *Archives of Computational Methods in Engineering 2021 28:7*, vol. 28, no. 7, pp. 4595–4631, Feb. 2021, doi: 10.1007/S11831-021-09547-0.

[49] C. Pucci, C. Martinelli, and G. Ciofani, “Innovative approaches for cancer treatment: current perspectives and new challenges,” *Ecancermedicalscience*, vol. 13, Sep. 2019, doi: 10.3332/ECANCER.2019.961.

[50] K. Kourou, K. P. Exarchos, C. Papaloukas, P. Sakaloglou, T. Exarchos, and D. I. Fotiadis, “Applied machine learning in cancer research: A systematic review for patient diagnosis, classification and prognosis,” *Comput Struct Biotechnol J*, vol. 19, p. 5546, Jan. 2021, doi: 10.1016/J.CSBJ.2021.10.006.

[51] K. K. Ramachandran, A. Apsara Saleth Mary, S. Hawladar, D. Asokk, B. Bhaskar, and J. R. Pitroda, “Machine learning and role of artificial intelligence in optimizing work performance and employee behavior,” *Mater Today Proc*, vol. 51, pp. 2327–2331, Jan. 2022, doi: 10.1016/J.MATPR.2021.11.544.

[52] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Comput Sci*, vol. 2, no. 3, pp. 1–21, May 2021, doi: 10.1007/S42979-021-00592-X/FIGURES/11.

[53] G. Armstrong *et al.*, “Applications and Comparison of Dimensionality Reduction Methods for Microbiome Data,” *Frontiers in Bioinformatics*, vol. 2, p. 821861, Feb. 2022, doi: 10.3389/FBINF.2022.821861.

[54] D. Samariya and A. Thakkar, “A Comprehensive Survey of Anomaly Detection Algorithms,” *Annals of Data Science*, vol. 10, no. 3, pp. 829–850, Jun. 2021, doi: 10.1007/S40745-021-00362-9/TABLES/12.

[55] M. Shawkat, M. Badawi, S. El-ghamrawy, R. Arnous, and A. El-desoky, “An optimized FP-growth algorithm for discovery of association rules,” *Journal of Supercomputing*, vol. 78, no. 4, pp. 5479–5506, Mar. 2022, doi: 10.1007/S11227-021-04066-Y/FIGURES/9.

[56] A. Coronato, M. Naeem, G. De Pietro, and G. Paragliola, “Reinforcement learning for intelligent healthcare applications: A survey,” *Artif Intell Med*, vol. 109, Sep. 2020, doi: 10.1016/J.ARTMED.2020.101964.

[57] X. Xu, L. Zuo, and Z. Huang, “Reinforcement learning algorithms with function approximation: Recent advances and applications,” *Inf Sci (N Y)*, vol. 261, pp. 1–31, Mar. 2014, doi: 10.1016/J.INS.2013.08.037.

[58] H. Zhang and T. Yu, “Taxonomy of reinforcement learning algorithms,” *Deep Reinforcement Learning: Fundamentals, Research and Applications*, pp. 125–133, Jan. 2020, doi: 10.1007/978-981-15-4095-0_3/COVER.

[59] J. Clifton and E. Laber, “Q-Learning: Theory and Applications,” *https://doi.org/10.1146/annurev-statistics-031219-041220*, vol. 7, pp. 279–301, Mar. 2020, doi: 10.1146/ANNUREV-STATISTICS-031219-041220.

[60] V. K. Verma and S. Verma, “Machine learning applications in healthcare sector: An overview,” *Mater Today Proc*, vol. 57, pp. 2144–2147, Jan. 2022, doi: 10.1016/J.MATPR.2021.12.101.

[61] S. Roy, T. Meena, and S. J. Lim, “Demystifying Supervised Learning in Healthcare 4.0: A New Reality of Transforming Diagnostic Medicine,” *Diagnostics (Basel)*, vol. 12, no. 10, Oct. 2022, doi: 10.3390/DIAGNOSTICS12102549.

[62] H. Belyadi and A. Haghighat, “Supervised learning,” *Machine Learning Guide for Oil and Gas Using Python*, pp. 169–295, Jan. 2021, doi: 10.1016/B978-0-12-821929-4.00004-4.

[63] A. Schneider, G. Hommel, and M. Blettner, “Linear Regression Analysis: Part 14 of a Series on Evaluation of Scientific Publications,” *Dtsch Arztebl Int*, vol. 107, no. 44, p. 776, Nov. 2010, doi: 10.3238/ARZTEBL.2010.0776.

[64] S. Sperandei, “Understanding logistic regression analysis,” *Biochem Med (Zagreb)*, vol. 24, no. 1, p. 12, 2014, doi: 10.11613/BM.2014.003.

[65] S. Tateishi, H. Matsui, and S. Konishi, “Nonlinear regression modeling via the lasso-type regularization,” *J Stat Plan Inference*, vol. 140, no. 5, pp. 1125–1134, May 2010, doi: 10.1016/J.JSPI.2009.10.015.

[66] I. Issah, O. Appiah, P. Appiahene, and F. Inusah, “A systematic review of the literature on machine learning application of determining the attributes influencing academic performance,” *Decision Analytics Journal*, vol. 7, p. 100204, Jun. 2023, doi: 10.1016/J.DAJOUR.2023.100204.

[67] S. Uddin, I. Haque, H. Lu, M. A. Moni, and E. Gide, “Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction,” *Scientific Reports 2022 12:1*, vol. 12, no. 1, pp. 1–11, Apr. 2022, doi: 10.1038/s41598-022-10358-x.

[68] S. B. Kotsiantis, “Decision trees: A recent overview,” *Artif Intell Rev*, vol. 39, no. 4, pp. 261–283, Apr. 2013, doi: 10.1007/S10462-011-9272-4/METRICS.

[69] A. Shmilovici, “Support Vector Machines,” *Data Mining and Knowledge Discovery Handbook*, pp. 257–276, 2005, doi: 10.1007/0-387-25465-X_12.

[70] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artif Intell Rev*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/S10462-020-09896-5/TABLES/12.

[71] B. Noh *et al.*, “XGBoost based machine learning approach to predict the risk of fall in older adults using gait outcomes,” *Scientific Reports 2021 11:1*, vol. 11, no. 1, pp. 1–9, Jun. 2021, doi: 10.1038/s41598-021-91797-w.

[72] J. Montomoli *et al.*, “Machine learning using the extreme gradient boosting (XGBoost) algorithm predicts 5-day delta of SOFA score at ICU admission in COVID-19 patients,” *Journal of Intensive Medicine*, vol. 1, no. 2, pp. 110–116, Oct. 2021, doi: 10.1016/J.JOINTM.2021.09.002.

[73] B. Ma, F. Meng, G. Yan, H. Yan, B. Chai, and F. Song, “Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data,” *Comput Biol Med*, vol. 121, p. 103761, Jun. 2020, doi: 10.1016/J.COMPBIOMED.2020.103761.

[74] X. Y. Liew, N. Hameed, and J. Clos, “An investigation of XGBoost-based algorithm for breast cancer classification,” *Machine Learning with Applications*, vol. 6, p. 100154, Dec. 2021, doi: 10.1016/J.MLWA.2021.100154.

[75] A. Bhandari, B. K. Tripathy, K. Jawad, S. Bhatia, M. K. I. Rahmani, and A. Mashat, “Cancer Detection and Prediction Using Genetic Algorithms,” *Comput Intell Neurosci*, vol. 2022, 2022, doi: 10.1155/2022/1871841.

[76] B. Noh *et al.*, “XGBoost based machine learning approach to predict the risk of fall in older adults using gait outcomes,” *Scientific Reports 2021 11:1*, vol. 11, no. 1, pp. 1–9, Jun. 2021, doi: 10.1038/s41598-021-91797-w.

[77] X. Deng, M. Li, S. Deng, and L. Wang, “Hybrid gene selection approach using XGBoost and multi-objective genetic algorithm for cancer classification,” *Med Biol Eng Comput*, vol. 60, no. 3, pp. 663–681, Mar. 2022, doi: 10.1007/S11517-021-02476-X/TABLES/19.

[78] S. Chen *et al.*, “A Novel XGBoost Method to Infer the Primary Lesion of 20 Solid Tumor Types From Gene Expression Data,” *Front Genet*, vol. 12, Feb. 2021, doi: 10.3389/FGENE.2021.632761.

[79] B. Ma, F. Meng, G. Yan, H. Yan, B. Chai, and F. Song, “Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data,” *Comput Biol Med*, vol. 121, p. 103761, Jun. 2020, doi: 10.1016/J.COMPBIOMED.2020.103761.

[80] D. J. Park, M. W. Park, H. Lee, Y. J. Kim, Y. Kim, and Y. H. Park, “Development of machine learning model for diagnostic disease prediction based on laboratory tests,” *Scientific Reports 2021 11:1*, vol. 11, no. 1, pp. 1–11, Apr. 2021, doi: 10.1038/s41598-021-87171-5.

[81] S. A. Hicks *et al.*, “On evaluation metrics for medical applications of artificial intelligence,” *Scientific Reports 2022 12:1*, vol. 12, no. 1, pp. 1–9, Apr. 2022, doi: 10.1038/s41598-022-09954-8.

[82] M. Steurer, R. J. Hill, and N. Pfeifer, “Metrics for evaluating the performance of machine learning based automated valuation models,” *https://doi.org/10.1080/09599916.2020.1858937*, vol. 38, no. 2, pp. 99–129, 2021, doi: 10.1080/09599916.2020.1858937.

[83] Y. C. Goh, X. Q. Cai, W. Theseira, G. Ko, and K. A. Khor, “Evaluating human versus machine learning performance in classifying research abstracts,” *Scientometrics*, vol. 125, no. 2, pp. 1197–1212, Nov. 2020, doi: 10.1007/S11192-020-03614-2/TABLES/5.

[84] F. Deng, J. Huang, X. Yuan, C. Cheng, and L. Zhang, “Performance and efficiency of machine learning algorithms for analyzing rectangular biomedical data,” *Laboratory Investigation 2021 101:4*, vol. 101, no. 4, pp. 430–441, Feb. 2021, doi: 10.1038/s41374-020-00525-x.

[85] C. Goutte and E. Gaussier, “A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation,” *Lecture Notes in Computer Science*, vol. 3408, pp. 345–359, 2005, doi: 10.1007/978-3-540-31865-1_25/COVER.

[86] D. J. Hand, P. Christen, and N. Kirielle, “F*: an interpretable transformation of the F-measure,” *Mach Learn*, vol. 110, no. 3, pp. 451–456, Mar. 2021, doi: 10.1007/S10994-021-05964-1/FIGURES/2.

[87] P. A. Jaskowiak, I. G. Costa, and R. J. G. B. Campello, “The area under the ROC curve as a measure of clustering quality,” *Data Min Knowl Discov*, vol. 36, no. 3, pp. 1219–1245, May 2022, doi: 10.1007/S10618-022-00829-0/FIGURES/5.

[88] J. Muschelli, “ROC and AUC with a Binary Predictor: a Potentially Misleading Metric,” *J Classif*, vol. 37, no. 3, pp. 696–708, Oct. 2020, doi: 10.1007/S00357-019-09345-1/FIGURES/4.

[89] A. Barredo Arrieta *et al.*, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/J.INFFUS.2019.12.012.

[90] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, and W. Samek, “Explainable AI Methods - A Brief Overview,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 13200 LNAI, pp. 13–38, 2022, doi: 10.1007/978-3-031-04083-2_2/FIGURES/3.

[91] H. A. H. Al-Najjar, B. Pradhan, G. Beydoun, R. Sarkar, H. J. Park, and A. Alamri, “A novel method using explainable artificial intelligence (XAI)-based Shapley Additive Explanations for spatial landslide prediction using Time-Series SAR dataset,” *Gondwana Research*, Aug. 2022, doi: 10.1016/J.GR.2022.08.004.

[92] A. Quan Ngo, L. Quy, N. Id, V. Quan, and T. Id, “Developing interpretable machine learning-Shapley additive explanations model for unconfined compressive strength of cohesive soils stabilized with geopolymer,” *PLoS One*, vol. 18, no. 6, p. e0286950, Jun. 2023, doi: 10.1371/JOURNAL.PONE.0286950.

[93] A. Barredo Arrieta *et al.*, “Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/J.INFFUS.2019.12.012.

[94] A. Gramegna and P. Giudici, “SHAP and LIME: An Evaluation of Discriminative Power in Credit Risk,” *Front Artif Intell*, vol. 4, p. 752558, Sep. 2021, doi: 10.3389/FRAI.2021.752558.

[95] X. Zhang, I. Jonassen, and A. Goksøyr, “Machine Learning Approaches for Biomarker Discovery Using Gene Expression Data,” *Bioinformatics*, pp. 53–64, Mar. 2021, doi: 10.36255/EXONPUBLICATIONS.BIOINFORMATICS.2021.CH4.

[96] A. Talevi, “Drug repositioning: Current approaches and their implications in the precision medicine era,” *Expert Rev Precis Med Drug Dev*, vol. 3, no. 1, pp. 49–61, Jan. 2018, doi: 10.1080/23808993.2018.1424535.

[97] W. M. C. Top, A. Kooy, and C. D. A. Stehouwer, “Metformin: A Narrative Review of Its Potential Benefits for Cardiovascular Disease, Cancer and Dementia,” *Pharmaceuticals (Basel)*, vol. 15, no. 3, Mar. 2022, doi: 10.3390/PH15030312.

[98] J. H. Kim and A. R. Scialli, “Thalidomide: the tragedy of birth defects and the effective treatment of disease,” *Toxicol Sci*, vol. 122, no. 1, pp. 1–6, Jul. 2011, doi: 10.1093/TOXSCI/KFR088.

[99] S. Pushpakom *et al.*, “Drug repurposing: progress, challenges and recommendations,” *Nature Reviews Drug Discovery 2018 18:1*, vol. 18, no. 1, pp. 41–58, Oct. 2018, doi: 10.1038/nrd.2018.168.

[100] A. Talevi and C. L. Bellera, “Challenges and opportunities with drug repurposing: finding strategies to find alternative uses of therapeutics,” *https://doi.org/10.1080/17460441.2020.1704729*, vol. 15, no. 4, pp. 397–401, Apr. 2019, doi: 10.1080/17460441.2020.1704729.

[101] M. Allarakhia, “Open-source approaches for the repurposing of existing or failed candidate drugs: Learning from and applying the lessons across diseases,” *Drug Des Devel Ther*, vol. 7, pp. 753–766, Aug. 2013, doi: 10.2147/DDDT.S46289.

[102] G. M. Morris and M. Lim-Wilby, “Molecular docking,” *Methods Mol Biol*, vol. 443, pp. 365–382, 2008, doi: 10.1007/978-1-59745-177-2_19.

[103] M. Macchiagodena, M. Pagliai, and P. Procacci, “Identification of potential binders of the main protease 3CLpro of the COVID-19 via structure-based ligand design and molecular modeling,” *Chem Phys Lett*, vol. 750, Jul. 2020, doi: 10.1016/J.CPLETT.2020.137489.

[104] A. Hospital, J. R. Goñi, M. Orozco, and J. L. Gelpí, “Molecular dynamics simulations: advances and applications,” *Adv Appl Bioinform Chem*, vol. 8, no. 1, p. 37, 2015, doi: 10.2147/AABC.S70333.

[105] V. Salmaso and S. Moro, “Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview,” *Front Pharmacol*, vol. 9, no. AUG, Aug. 2018, doi: 10.3389/FPHAR.2018.00923/FULL.

[106] A. A. Al-Karmalawy *et al.*, “Molecular Docking and Dynamics Simulation Revealed the Potential Inhibitory Activity of ACEIs Against SARS-CoV-2 Targeting the hACE2 Receptor,” *Front Chem*, vol. 9, p. 661230, May 2021, doi: 10.3389/FCHEM.2021.661230/BIBTEX.

[107] T. T. Ashburn and K. B. Thor, “Drug repositioning: Identifying and developing new uses for existing drugs,” *Nat. Rev. Drug Discov.*, vol. 3, no. 8, pp. 673–683, 2004, doi: 10.1038/nrd1468.

[108] P. Wu *et al.*, “Integrating gene expression and clinical data to identify drug repurposing candidates for hyperlipidemia and hypertension,” *Nature Communications 2022 13:1*, vol. 13, no. 1, pp. 1–12, Jan. 2022, doi: 10.1038/s41467-021-27751-1.

[109] A. Bruzzese, J. A. R. Dalton, and J. Giraldo, “Statistics for the analysis of molecular dynamics simulations: providing P values for agonist-dependent GPCR activation,”

Scientific Reports 2020 10:1, vol. 10, no. 1, pp. 1–9, Nov. 2020, doi: 10.1038/s41598-020-77072-4.

[110] M. Scheurer *et al.*, "PyContact: Rapid, Customizable, and Visual Analysis of Noncovalent Interactions in MD Simulations," *Biophys J*, vol. 114, no. 3, pp. 577–583, Feb. 2018, doi: 10.1016/J.BPJ.2017.12.003.

[111] D. Mercadante, F. Gräter, and C. Daday, "CONAN: A Tool to Decode Dynamical Information from Molecular Interaction Maps," *Biophys J*, vol. 114, no. 6, pp. 1267–1273, Mar. 2018, doi: 10.1016/J.BPJ.2018.01.033.

[112] M. Kayikci, A. J. Venkatakrishnan, J. Scott-Brown, C. N. J. Ravarani, T. Flock, and M. M. Babu, "Visualization and analysis of non-covalent contacts using the Protein Contacts Atlas," *Nature Structural & Molecular Biology 2018 25:2*, vol. 25, no. 2, pp. 185–194, Jan. 2018, doi: 10.1038/s41594-017-0019-z.

[113] A. Subramanian *et al.*, "A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles," *Cell*, vol. 171, no. 6, pp. 1437-1452.e17, Nov. 2017, doi: 10.1016/j.cell.2017.10.049.

[114] J. Lamb *et al.*, "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease," *Science*, vol. 313, no. 5795, pp. 1929–1935, Sep. 2006, doi: 10.1126/SCIENCE.1132939.

[115] E. Clough and T. Barrett, "The Gene Expression Omnibus database," *Methods in Molecular Biology*, vol. 1418, pp. 93–110, 2016, doi: 10.1007/978-1-4939-3578-9_5/COVER.

[116] Z. Wang *et al.*, "Extraction and analysis of signatures from the Gene Expression Omnibus by the crowd," *Nature Communications 2016 7:1*, vol. 7, no. 1, pp. 1–11, Sep. 2016, doi: 10.1038/ncomms12846.

[117] Z. Wang, E. He, K. Sani, K. M. Jagodnik, M. C. Silverstein, and A. Ma'Ayan, "Drug Gene Budger (DGB): an application for ranking drugs to modulate a specific gene based on transcriptomic signatures," *Bioinformatics*, vol. 35, no. 7, pp. 1247–1248, Apr. 2019, doi: 10.1093/BIOINFORMATICS/BTY763.

[118] M. Cives *et al.*, "Non-Melanoma Skin Cancers: Biological and Clinical Features," *Int J Mol Sci*, vol. 21, no. 15, pp. 1–24, Aug. 2020, doi: 10.3390/IJMS21155394.

[119] C. Garbe and U. Leiter, "Epidemiology of melanoma and nonmelanoma skin cancer--the role of sunlight," *Adv Exp Med Biol*, vol. 624, pp. 89–103, 2008, doi: 10.1007/978-0-387-77574-6_8.

[120] "Our New Approach to a Challenging Skin Cancer Statistic - The Skin Cancer Foundation." https://www.skincancer.org/blog/our-new-approach-to-a-challenging-skin-cancer-statistic/ (accessed Feb. 03, 2022).

[121] H. M. Gloster and K. Neal, "Skin cancer in skin of color," *J Am Acad Dermatol*, vol. 55, no. 5, pp. 741–760, Nov. 2006, doi: 10.1016/J.JAAD.2005.08.063.

[122] "GENETICS OF HUMAN AGE RELATED DISORDERS - PubMed." https://pubmed.ncbi.nlm.nih.gov/26856084/ (accessed Mar. 24, 2022).

[123] A. S. Weinberg, C. A. Ogle, and E. K. Shim, "Metastatic cutaneous squamous cell carcinoma: an update," *Dermatol Surg*, vol. 33, no. 8, pp. 885–899, Aug. 2007, doi: 10.1111/J.1524-4725.2007.33190.X.

[124] V. R. Yanofsky, S. E. Mercer, and R. G. Phelps, "Histopathological Variants of Cutaneous Squamous Cell Carcinoma: A Review," *J Skin Cancer*, vol. 2011, pp. 1–13, 2011, doi: 10.1155/2011/210813.

[125] A. Gluba, J. Rysz, and T. Pietrucha, "Microarray technology in the study of genetic determinants of cardiovascular diseases," *Cent Eur J Med*, vol. 4, no. 1, pp. 1–10, Mar. 2009, doi: 10.2478/S11536-009-0012-Y/PDF.

[126] S. Mohr, G. D. Leikauf, G. Keith, and B. H. Rihn, “Microarrays as cancer keys: An array of possibilities,” *Journal of Clinical Oncology*, vol. 20, no. 14, pp. 3165–3175, Jul. 2002, doi: 10.1200/JCO.2002.12.073.

[127] Q. X. Yang *et al.*, “Identification of the gene signature reflecting schizophrenia’s etiology by constructing artificial intelligence-based method of enhanced reproducibility,” *CNS Neurosci Ther*, vol. 25, no. 9, p. 1054, 2019, doi: 10.1111/CNS.13196.

[128] S. Cui, Q. Wu, J. West, and J. Bai, “Machine learning-based microarray analyses indicate low-expression genes might collectively influence PAH disease,” *PLoS Comput Biol*, vol. 15, no. 8, p. e1007264, 2019, doi: 10.1371/JOURNAL.PCBI.1007264.

[129] A. Cambiaghi, M. Ferrario, and M. Masseroli, “Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration,” *Brief Bioinform*, vol. 18, no. 3, pp. 498–510, May 2017, doi: 10.1093/BIB/BBW031.

[130] J. Fu *et al.*, “Optimization of metabolomic data processing using NOREVA,” *Nature Protocols 2021 17:1*, vol. 17, no. 1, pp. 129–151, Dec. 2021, doi: 10.1038/s41596-021-00636-9.

[131] B. Li *et al.*, “NOREVA: normalization and evaluation of MS-based metabolomics data,” *Nucleic Acids Res*, vol. 45, no. W1, pp. W162–W170, Jul. 2017, doi: 10.1093/NAR/GKX449.

[132] Q. Yang *et al.*, “NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data,” *Nucleic Acids Res*, vol. 48, no. W1, pp. W436–W448, Jul. 2020, doi: 10.1093/NAR/GKAA258.

[133] J. Tang *et al.*, “ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies,” *Brief Bioinform*, vol. 21, no. 2, pp. 621–636, Mar. 2020, doi: 10.1093/BIB/BBY127.

[134] S. Thakran *et al.*, *Genetic landscape of common epilepsies: Advancing towards precision in treatment*, vol. 21, no. 20. 2020. doi: 10.3390/ijms21207784.

[135] R. Nayak and Y. Hasija, “A hitchhiker’s guide to single-cell transcriptomics and data analysis pipelines,” *Genomics*, vol. 113, no. 2, pp. 606–619, Mar. 2021, doi: 10.1016/J.YGENO.2021.01.007.

[136] P. Suwinski, C. K. Ong, M. H. T. Ling, Y. M. Poh, A. M. Khan, and H. S. Ong, “Advancing personalized medicine through the application of whole exome sequencing and big data analytics,” *Front Genet*, vol. 10, no. FEB, p. 49, 2019, doi: 10.3389/FGENE.2019.00049/BIBTEX.

[137] P. Economopoulou, R. de Bree, I. Kotsantis, and A. Psyrri, “Diagnostic tumor markers in head and neck squamous cell carcinoma (HNSCC) in the clinical setting,” *Front Oncol*, vol. 9, no. AUG, p. 827, 2019, doi: 10.3389/FONC.2019.00827/BIBTEX.

[138] C. Soneson and M. Delorenzi, “A comparison of methods for differential expression analysis of RNA-seq data,” *BMC Bioinformatics*, vol. 14, no. 1, pp. 1–18, Mar. 2013, doi: 10.1186/1471-2105-14-91/TABLES/3.

[139] M. Crow, N. Lim, S. Ballouz, P. Pavlidis, and J. Gillis, “Predictability of human differential gene expression,” *Proc Natl Acad Sci U S A*, vol. 116, no. 13, pp. 6491–6500, Mar. 2019, doi: 10.1073/PNAS.1802973116/-/DCSUPPLEMENTAL.

[140] A. S. Rifaioglu, H. Atas, M. J. Martin, R. Cetin-Atalay, V. Atalay, and T. Doğan, “Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases,” *Brief Bioinform*, vol. 20, no. 5, p. 1878, Sep. 2019, doi: 10.1093/BIB/BBY061.

[141] J. Hong *et al.*, “Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning,” *Brief Bioinform*, vol. 21, no. 4, p. 1437, Jul. 2020, doi: 10.1093/BIB/BBZ081.

[142] J. Hong *et al.*, “Convolutional neural network-based annotation of bacterial type IV secretion system effectors with enhanced accuracy and reduced false discovery,” *Brief Bioinform*, vol. 21, no. 5, pp. 1825–1836, Sep. 2020, doi: 10.1093/BIB/BBZ120.

[143] M. X. Li *et al.*, “Using a machine learning approach to identify key prognostic molecules for esophageal squamous cell carcinoma,” *BMC Cancer*, vol. 21, no. 1, pp. 1–11, Dec. 2021, doi: 10.1186/S12885-021-08647-1/FIGURES/3.

[144] Z. Dlamini, F. Z. Francies, R. Hull, and R. Marima, “Artificial intelligence (AI) and big data in cancer and precision oncology,” *Comput Struct Biotechnol J*, vol. 18, pp. 2300–2311, Jan. 2020, doi: 10.1016/J.CSBJ.2020.08.019.

[145] T. Davenport, A. Guha, D. Grewal, and T. Bressgott, “How artificial intelligence will change the future of marketing,” *J Acad Mark Sci*, vol. 48, no. 1, pp. 24–42, Jan. 2020, doi: 10.1007/S11747-019-00696-0/FIGURES/2.

[146] Y. Dong, J. Hou, N. Zhang, and M. Zhang, “Research on How Human Intelligence, Consciousness, and Cognitive Computing Affect the Development of Artificial Intelligence,” *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/1680845.

[147] A. Bohr and K. Memarzadeh, “The rise of artificial intelligence in healthcare applications,” *Artificial Intelligence in Healthcare*, p. 25, 2020, doi: 10.1016/B978-0-12-818438-7.00002-2.

[148] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Computer Science 2021 2:3*, vol. 2, no. 3, pp. 1–21, Mar. 2021, doi: 10.1007/S42979-021-00592-X.

[149] A. Adadi and M. Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” *IEEE Access*, vol. 6, pp. 52138–52160, Sep. 2018, doi: 10.1109/ACCESS.2018.2870052.

[150] “9.6 SHAP (SHapley Additive exPlanations) | Interpretable Machine Learning.” https://christophm.github.io/interpretable-ml-book/shap.html (accessed Feb. 03, 2022).

[151] S. M. Lundberg, P. G. Allen, and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions”.

[152] S. M. Lundberg, G. Erion, and S.-I. Lee, “Consistent Individualized Feature Attribution for Tree Ensembles,” *undefined*, 2018.

[153] R. A. Irizarry *et al.*, “Exploration, normalization, and summaries of high density oligonucleotide array probe level data,” *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003, doi: 10.1093/BIOSTATISTICS/4.2.249.

[154] R. Marks, “Squamous cell carcinoma,” *The Lancet*, vol. 347, no. 9003, pp. 735–738, Mar. 1996, doi: 10.1016/S0140-6736(96)90081-1.

[155] A. Stratigos *et al.*, “Diagnosis and treatment of invasive squamous cell carcinoma of the skin: European consensus-based interdisciplinary guideline,” *Eur J Cancer*, vol. 51, no. 14, pp. 1989–2007, Sep. 2015, doi: 10.1016/J.EJCA.2015.06.110.

[156] H. W. Rogers, M. A. Weinstock, S. R. Feldman, and B. M. Coldiron, “Incidence Estimate of Nonmelanoma Skin Cancer (Keratinocyte Carcinomas) in the US Population, 2012,” *JAMA Dermatol*, vol. 151, no. 10, pp. 1081–1086, Oct. 2015, doi: 10.1001/JAMADERMATOL.2015.1187.

[157] N. Eisemann *et al.*, “Non-Melanoma Skin Cancer Incidence and Impact of Skin Cancer Screening on Incidence,” *Journal of Investigative Dermatology*, vol. 134, no. 1, pp. 43–50, Jan. 2014, doi: 10.1038/JID.2013.304.

[158] D. Szklarczyk *et al.*, “STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets,” *Nucleic Acids Res*, vol. 47, no. Database issue, p. D607, Jan. 2019, doi: 10.1093/NAR/GKY1131.

[159] P. Shannon *et al.*, “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks,” *Genome Res*, vol. 13, no. 11, p. 2498, Nov. 2003, doi: 10.1101/GR.1239303.

[160] K. Taniuchi, M. Furihata, S. Naganuma, and T. Saibara, “ARHGEF4 predicts poor prognosis and promotes cell invasion by influencing ERK1/2 and GSK-3α/β signaling in pancreatic cancer,” *Int J Oncol*, vol. 53, no. 5, pp. 2224–2240, Nov. 2018, doi: 10.3892/IJO.2018.4549/HTML.

[161] Y. Kawasaki, R. Sato, and T. Akiyama, “Mutated APC and Asef are involved in the migration of colorectal tumour cells,” *Nature Cell Biology 2003 5:3*, vol. 5, no. 3, pp. 211–215, Feb. 2003, doi: 10.1038/ncb937.

[162] T. Akiyama and Y. Kawasaki, “Wnt signaling and the actin cytoskeleton,” *Oncogene*, vol. 25, no. 57, pp. 7538–7544, Dec. 2006, doi: 10.1038/SJ.ONC.1210063.

[163] N. Mitin, L. Betts, M. E. Yohe, C. J. Der, J. Sondek, and K. L. Rossman, “Release of autoinhibition of ASEF by APC leads to CDC42 activation and tumor suppression,” *Nature Structural & Molecular Biology 2007 14:9*, vol. 14, no. 9, pp. 814–823, Aug. 2007, doi: 10.1038/nsmb1290.

[164] Y. Nakayama *et al.*, “Cloning of cDNA Encoding a Regeneration-Associated Muscle Protease Whose Expression Is Attenuated in Cell Lines Derived from Duchenne Muscular Dystrophy Patients,” *Am J Pathol*, vol. 164, no. 5, p. 1773, 2004, doi: 10.1016/S0002-9440(10)63735-2.

[165] L. Hawthorn, J. Luce, L. Stein, and J. Rothschild, “Integration of transcript expression, copy number and LOH analysis of infiltrating ductal carcinoma of the breast,” *BMC Cancer*, vol. 10, Aug. 2010, doi: 10.1186/1471-2407-10-460.

[166] P. R. Sharma *et al.*, “An Islet-Targeted Genome-Wide Association Scan Identifies Novel Genes Implicated in Cytokine-Mediated Islet Stress in Type 2 Diabetes,” *Endocrinology*, vol. 156, no. 9, pp. 3147–3156, Sep. 2015, doi: 10.1210/EN.2015-1203.

[167] H. B. Wang *et al.*, “Identification of differentially expressed genes and preliminary validations in cardiac pathological remodeling induced by transverse aortic constriction,” *Int J Mol Med*, vol. 44, no. 4, pp. 1447–1461, 2019, doi: 10.3892/IJMM.2019.4291.

[168] W. Wei *et al.*, “Identification of biomarker for cutaneous squamous cell carcinoma using microarray data analysis,” *J Cancer*, vol. 9, no. 2, pp. 400–406, 2018, doi: 10.7150/JCA.21381.

[169] P. H. Y. Lo, C. Tanikawa, T. Katagiri, Y. Nakamura, and K. Matsuda, “Identification of novel epigenetically inactivated gene PAMR1 in breast carcinoma,” *Oncol Rep*, vol. 33, no. 1, pp. 267–273, Jan. 2015, doi: 10.3892/OR.2014.3581/HTML.

[170] F. Lin *et al.*, “GTSE1 is involved in breast cancer progression in p53 mutation-dependent manner,” *Journal of Experimental and Clinical Cancer Research*, vol. 38, no. 1, Apr. 2019, doi: 10.1186/S13046-019-1157-4.

[171] Y. Zheng *et al.*, “GTSE1, CDC20, PCNA, and MCM6 Synergistically Affect Regulations in Cell Cycle and Indicate Poor Prognosis in Liver Cancer,” *Analytical Cellular Pathology*, vol. 2019, 2019, doi: 10.1155/2019/1038069.

[172] A. Liu *et al.*, “Overexpression of G2 and S phase-expressed-1 contributes to cell proliferation, migration, and invasion via regulating p53/FoxM1/CCNB1 pathway and

predicts poor prognosis in bladder cancer," *Int J Biol Macromol*, vol. 123, pp. 322–334, Feb. 2019, doi: 10.1016/J.IJBIOMAC.2018.11.032.

[173] N. Reguart *et al.*, "Cloning and characterization of the promoter of human Wnt inhibitory factor-1," *Biochem Biophys Res Commun*, vol. 323, no. 1, pp. 229–234, Oct. 2004, doi: 10.1016/J.BBRC.2004.08.075.

[174] C. Wissman *et al.*, "WIFI, a component of the Wnt pathway, is down-regulated in prostate, breast, lung, and bladder cancer," *Journal of Pathology*, vol. 201, no. 2, pp. 204–212, Oct. 2003, doi: 10.1002/PATH.1449.

[175] F. Yang, Q. Zeng, G. Yu, S. Li, and C. Y. Wang, "Wnt/β-catenin signaling inhibits death receptor-mediated apoptosis and promotes invasive growth of HNSCC," *Cell Signal*, vol. 18, no. 5, pp. 679–687, May 2006, doi: 10.1016/J.CELLSIG.2005.06.015.

[176] I. Oishi *et al.*, "The receptor tyrosine kinase Ror2 is involved in non-canonical Wnt5a/JNK signaling pathway," *Genes to Cells*, vol. 8, no. 7, pp. 645–654, Jul. 2003, doi: 10.1046/J.1365-2443.2003.00662.X.

[177] S. L. McDonald and A. Silver, "The opposing roles of Wnt-5a in cancer," *British Journal of Cancer 2009 101:2*, vol. 101, no. 2, pp. 209–214, Jul. 2009, doi: 10.1038/sj.bjc.6605174.

[178] A. Säfholm, K. Leandersson, J. Dejmek, C. K. Nielsen, B. O. Villoutreix, and T. Andersson, "A Formylated Hexapeptide Ligand Mimics the Ability of Wnt-5a to Impair Migration of Human Breast Epithelial Cells," *Journal of Biological Chemistry*, vol. 281, no. 5, pp. 2740–2749, Feb. 2006, doi: 10.1074/JBC.M508386200.

[179] N. Kremenevskaja, R. Von Wasielewski, A. S. Rao, C. Schöfl, T. Andersson, and G. Brabant, "Wnt-5a has tumor suppressor activity in thyroid carcinoma," *Oncogene 2005 24:13*, vol. 24, no. 13, pp. 2144–2154, Feb. 2005, doi: 10.1038/sj.onc.1208370.

[180] A. T. Weeraratna *et al.*, “Wnt5a signaling directly affects cell motility and invasion of metastatic melanoma,” *Cancer Cell*, vol. 1, no. 3, pp. 279–288, Apr. 2002, doi: 10.1016/S1535-6108(02)00045-4.

[181] M. Kurayoshi *et al.*, “Expression of Wnt-5a Is Correlated with Aggressiveness of Gastric Cancer by Stimulating Cell Migration and Invasion,” *Cancer Res*, vol. 66, no. 21, pp. 10439–10448, Nov. 2006, doi: 10.1158/0008-5472.CAN-06-2359.

[182] C. Valacca, E. Tassone, and P. Mignatti, “TIMP-2 Interaction with MT1-MMP Activates the AKT Pathway and Protects Tumor Cells from Apoptosis,” *PLoS One*, vol. 10, no. 9, p. e0136797, Sep. 2015, doi: 10.1371/JOURNAL.PONE.0136797.

[183] A. M. Knapinska and G. B. Fields, “The Expanding Role of MT1-MMP in Cancer Progression,” *Pharmaceuticals 2019, Vol. 12, Page 77*, vol. 12, no. 2, p. 77, May 2019, doi: 10.3390/PH12020077.

[184] N. Chen, G. Zhang, J. Fu, and Q. Wu, “Matrix metalloproteinase-14 (MMP-14) downregulation inhibits esophageal squamous cell carcinoma cell migration, invasion, and proliferation,” *Thorac Cancer*, vol. 11, no. 11, pp. 3168–3174, Nov. 2020, doi: 10.1111/1759-7714.13636.

[185] W. Wang *et al.*, “The activity status of cofilin is directly related to invasion, intravasation, and metastasis of mammary tumors,” *Journal of Cell Biology*, vol. 173, no. 3, pp. 395–404, May 2006, doi: 10.1083/JCB.200510115.

[186] P. Hotulainen, E. Paunola, M. K. Vartiainen, and P. Lappalainen, “Actin-depolymerizing factor and cofilin-1 play overlapping roles in promoting rapid F-actin depolymerization in mammalian nonmuscle cells,” *Mol Biol Cell*, vol. 16, no. 2, pp. 649–664, Feb. 2005, doi: 10.1091/MBC.E04-07-0555/ASSET/IMAGES/LARGE/ZMK0020529950010.JPEG.

[187] J. Van Rheenen *et al.*, “EGF-induced PIP2 hydrolysis releases and activates cofilin locally in carcinoma cells,” *Journal of Cell Biology*, vol. 179, no. 6, pp. 1247–1259, Dec. 2007, doi: 10.1083/JCB.200706206/VIDEO-1.

[188] Q. Wu, Y. Jiang, S. Cui, Y. Wang, and X. Wu, “The role of cofilin-l in vulvar squamous cell carcinoma: A marker of carcinogenesis, progression and targeted therapy,” *Oncol Rep*, vol. 35, no. 5, pp. 2743–2754, May 2016, doi: 10.3892/OR.2016.4625/HTML.

[189] R. Bandopadhyay *et al.*, “The expression of DJ-1 (PARK7) in normal human CNS and idiopathic Parkinson’s disease,” *Brain*, vol. 127, no. Pt 2, pp. 420–430, Feb. 2004, doi: 10.1093/BRAIN/AWH054.

[190] K. J. Won *et al.*, “DJ-1/park7 protects against neointimal formation via the inhibition of vascular smooth muscle cell growth,” *Cardiovasc Res*, vol. 97, no. 3, pp. 553–561, Mar. 2013, doi: 10.1093/CVR/CVS363.

[191] J. M. Kim *et al.*, “DJ-1 promotes angiogenesis and osteogenesis by activating FGF receptor-1 signaling,” *Nat Commun*, vol. 3, 2012, doi: 10.1038/NCOMMS2313.

[192] C. Martinat *et al.*, “Sensitivity to oxidative stress in DJ-1-deficient dopamine neurons: an ES- derived cell model of primary Parkinsonism,” *PLoS Biol*, vol. 2, no. 11, Nov. 2004, doi: 10.1371/JOURNAL.PBIO.0020327.

[193] T. Hayashi *et al.*, “DJ-1 binds to mitochondrial complex I and maintains its activity,” *Biochem Biophys Res Commun*, vol. 390, no. 3, pp. 667–672, Dec. 2009, doi: 10.1016/J.BBRC.2009.10.025.

[194] J. M. Kim *et al.*, “DJ-1 contributes to adipogenesis and obesity-induced inflammation,” *Sci Rep*, vol. 4, Jun. 2014, doi: 10.1038/SREP04805.

[195] S. Xu *et al.*, “DJ-1 Is Upregulated in Oral Squamous Cell Carcinoma and Promotes Oral Cancer Cell Proliferation and Invasion,” *J Cancer*, vol. 7, no. 8, p. 1020, 2016, doi: 10.7150/JCA.14539.

[196] Y. Xu *et al.*, “Cell type-restricted activity of hnRNPM promotes breast cancer metastasis via regulating alternative splicing,” *Genes Dev*, vol. 28, no. 11, pp. 1191–1203, Jun. 2014, doi: 10.1101/GAD.241968.114.

[197] R. L. Brown *et al.*, “CD44 splice isoform switching in human and mouse epithelium is essential for epithelial-mesenchymal transition and breast cancer progression,” *J Clin Invest*, vol. 121, no. 3, pp. 1064–1074, Mar. 2011, doi: 10.1172/JCI44540.

[198] H. Sun *et al.*, “HnRNPM and CD44s expression affects tumor aggressiveness and predicts poor prognosis in breast cancer with axillary lymph node metastases,” *Genes Chromosomes Cancer*, vol. 56, no. 8, pp. 598–607, Aug. 2017, doi: 10.1002/GCC.22463.

[199] L. Fania *et al.*, “Cutaneous Squamous Cell Carcinoma: From Pathophysiology to Novel Therapeutic Approaches,” *Biomedicines*, vol. 9, no. 2, pp. 1–33, Feb. 2021, doi: 10.3390/BIOMEDICINES9020171.

[200] A. Stratigos, C. Garbe, C. Lebbe, … J. M.-E. journal of, and undefined 2015, “Diagnosis and treatment of invasive squamous cell carcinoma of the skin: European consensus-based interdisciplinary guideline,” *Elsevier*, Accessed: Jul. 29, 2022. [Online]. Available: https://sci-hub.do/https://www.sciencedirect.com/science/article/pii/S0959804915006255

[201] G. Geidel, I. Heidrich, J. Kött, S. W. Schneider, K. Pantel, and C. Gebhardt, “Emerging precision diagnostics in advanced cutaneous squamous cell carcinoma,” *npj Precision Oncology 2022 6:1*, vol. 6, no. 1, pp. 1–8, Mar. 2022, doi: 10.1038/s41698-022-00261-z.

[202] M. Watson, D. M. Holman, and M. Maguire-Eisen, “Ultraviolet Radiation Exposure and Its Impact on Skin Cancer Risk,” *Semin Oncol Nurs*, vol. 32, no. 3, p. 241, Aug. 2016, doi: 10.1016/J.SONCN.2016.05.005.

[203] S. Li *et al.*, “Molecular Subtypes of Oral Squamous Cell Carcinoma Based on Immunosuppression Genes Using a Deep Learning Approach,” *Front Cell Dev Biol*, vol. 9, p. 687245, Aug. 2021, doi: 10.3389/FCELL.2021.687245/FULL.

[204] M. J. Veness, “High-Risk Cutaneous Squamous Cell Carcinoma of the Head and Neck,” *J Biomed Biotechnol*, vol. 2007, 2007, doi: 10.1155/2007/80572.

[205] J. Y. Howell and M. L. Ramsey, “Squamous Cell Skin Cancer,” *StatPearls*, Aug. 2022, Accessed: Nov. 20, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK441939/

[206] R. J. Sanderson, J. A. D. Ironside, and W. I. Wei, “Squamous cell carcinomas of the head and neck,” *BMJ : British Medical Journal*, vol. 325, no. 7368, p. 822, Oct. 2002, doi: 10.1136/BMJ.325.7368.822.

[207] M. M. Pomerantz and M. L. Freedman, “The Genetics of Cancer Risk,” *Cancer J*, vol. 17, no. 6, p. 416, Nov. 2011, doi: 10.1097/PPO.0B013E31823E5387.

[208] G. Zhou, M. Chen, C. J. T. Ju, Z. Wang, J. Y. Jiang, and W. Wang, “Mutation effect estimation on protein-protein interactions using deep contextualized representation learning,” *NAR Genom Bioinform*, vol. 2, no. 2, Jun. 2020, doi: 10.1093/NARGAB/LQAA015.

[209] B. Reva, Y. Antipin, and C. Sander, “Predicting the functional impact of protein mutations: application to cancer genomics,” *Nucleic Acids Res*, vol. 39, no. 17, pp. e118–e118, Sep. 2011, doi: 10.1093/NAR/GKR407.

[210] C. I. E. Smith, T. C. Islam, P. T. Mattsson, A. J. Mohamed, B. F. Nore, and M. Vihinen, “The Tec family of cytoplasmic tyrosine kinases: mammalian Btk, Bmx, Itk,

Tec, Txk and homologs in other species," *Bioessays*, vol. 23, no. 5, pp. 436–446, 2001, doi: 10.1002/BIES.1062.

[211] S. Ponader and J. A. Burger, "Bruton's tyrosine kinase: from X-linked agammaglobulinemia toward targeted therapy for B-cell malignancies," *J Clin Oncol*, vol. 32, no. 17, pp. 1830–1839, Jun. 2014, doi: 10.1200/JCO.2013.53.1046.

[212] T. L. Whiteside, "The tumor microenvironment and its role in promoting tumor growth," *Oncogene*, vol. 27, no. 45, p. 5904, Oct. 2008, doi: 10.1038/ONC.2008.271.

[213] K. Szklener, A. Michalski, K. Żak, M. Piwoński, and S. Mańdziuk, "Ibrutinib in the Treatment of Solid Tumors: Current State of Knowledge and Future Directions," *Cells 2022, Vol. 11, Page 1338*, vol. 11, no. 8, p. 1338, Apr. 2022, doi: 10.3390/CELLS11081338.

[214] M. S. Davids and J. R. Brown, "Ibrutinib: a first in class covalent inhibitor of Bruton's tyrosine kinase," *Future Oncol*, vol. 10, no. 6, p. 957, 2014, doi: 10.2217/FON.14.51.

[215] P. A. Futreal *et al.*, "A CENSUS OF HUMAN CANCER GENES," *Nat Rev Cancer*, vol. 4, no. 3, p. 177, 2004, doi: 10.1038/NRC1299.

[216] S. A. Forbes *et al.*, "COSMIC (the Catalogue of Somatic Mutations in Cancer): a resource to investigate acquired mutations in human cancer," *Nucleic Acids Res*, vol. 38, no. Database issue, Nov. 2010, doi: 10.1093/NAR/GKP995.

[217] A. Berglöf *et al.*, "Targets for Ibrutinib Beyond B Cell Malignancies," *Scand J Immunol*, vol. 82, no. 3, p. 208, Sep. 2015, doi: 10.1111/SJI.12333.

[218] M. S. Hassan, A. A. Shaalan, M. I. Dessouky, A. E. Abdelnaiem, and M. ElHefnawi, "Evaluation of computational techniques for predicting non-synonymous single nucleotide variants pathogenicity," *Genomics*, vol. 111, no. 4, pp. 869–882, Jul. 2019, doi: 10.1016/J.YGENO.2018.05.013.

[219] B. Reva, Y. Antipin, and C. Sander, “Determinants of protein function revealed by combinatorial entropy optimization,” *Genome Biol*, vol. 8, no. 11, Nov. 2007, doi: 10.1186/GB-2007-8-11-R232.

[220] P. C. Ng and S. Henikoff, “Predicting Deleterious Amino Acid Substitutions,” *Genome Res*, vol. 11, no. 5, p. 863, May 2001, doi: 10.1101/GR.176601.

[221] A. Niroula, S. Urolagin, and M. Vihinen, “PON-P2: Prediction Method for Fast and Reliable Identification of Harmful Variants,” *PLoS One*, vol. 10, no. 2, Feb. 2015, doi: 10.1371/JOURNAL.PONE.0117380.

[222] E. Capriotti, P. L. Martelli, P. Fariselli, and R. Casadio, “Blind prediction of deleterious amino acid variations with SNPs&GO,” *Hum Mutat*, vol. 38, no. 9, pp. 1064–1071, Sep. 2017, doi: 10.1002/HUMU.23179.

[223] Y. Choi and A. P. Chan, “PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels,” *Bioinformatics*, vol. 31, no. 16, p. 2745, Aug. 2015, doi: 10.1093/BIOINFORMATICS/BTV195.

[224] I. Adzhubei, D. M. Jordan, and S. R. Sunyaev, “Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2,” *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.]*, vol. 0 7, no. SUPPL.76, p. Unit7.20, 2013, doi: 10.1002/0471142905.HG0720S76.

[225] V. Pejaver *et al.*, “Inferring the molecular and phenotypic impact of amino acid variants with MutPred2,” *Nature Communications 2020 11:1*, vol. 11, no. 1, pp. 1–13, Nov. 2020, doi: 10.1038/s41467-020-19669-x.

[226] A. T. Bender *et al.*, “Ability of Bruton’s Tyrosine Kinase Inhibitors to Sequester Y551 and Prevent Phosphorylation Determines Potency for Inhibition of Fc Receptor but not B-Cell Receptor Signaling,” *Mol Pharmacol*, vol. 91, no. 3, pp. 208–219, Mar. 2017, doi: 10.1124/MOL.116.107037.

[227] N. Guex, M. C. Peitsch, and T. Schwede, “Automated comparative protein structure modeling with SWISS-MODEL and Swiss-PdbViewer: a historical perspective,” *Electrophoresis*, vol. 30 Suppl 1, no. SUPPL. 1, Jun. 2009, doi: 10.1002/ELPS.200900140.

[228] M. J. Abraham *et al.*, “GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers,” *SoftwareX*, vol. 1–2, pp. 19–25, Sep. 2015, doi: 10.1016/J.SOFTX.2015.06.001.

[229] J. Huang *et al.*, “CHARMM36m: An Improved Force Field for Folded and Intrinsically Disordered Proteins,” *Nat Methods*, vol. 14, no. 1, p. 71, Dec. 2017, doi: 10.1038/NMETH.4067.

[230] J. Lee *et al.*, “CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field,” *J Chem Theory Comput*, vol. 12, no. 1, pp. 405–413, Jan. 2016, doi: 10.1021/ACS.JCTC.5B00935/ASSET/IMAGES/LARGE/CT-2015-00935E_0005.JPEG.

[231] C. C. David and D. J. Jacobs, “Principal Component Analysis: A Method for Determining the Essential Dynamics of Proteins,” *Methods Mol Biol*, vol. 1084, p. 193, 2014, doi: 10.1007/978-1-62703-658-0_11.

[232] F. Pedregosa FABIANPEDREGOSA *et al.*, “Scikit-learn: Machine Learning in Python Gaël Varoquaux Bertrand Thirion Vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011, doi: 10.5555/1953048.2078195.

[233] D. Mercadante, F. Gräter, and C. Daday, “CONAN: A Tool to Decode Dynamical Information from Molecular Interaction Maps,” *Biophys J*, vol. 114, no. 6, p. 1267, Mar. 2018, doi: 10.1016/J.BPJ.2018.01.033.

[234] G. M. Morris *et al.*, “User Guide AutoDock Version 4.2 Updated for version 4.2.6 Automated Docking of Flexible Ligands to Flexible Receptors,” 1991, Accessed: Nov. 20, 2022. [Online]. Available: http://autodock.scripps.edu/

[235] C. Wang, D. Greene, L. Xiao, R. Qi, and R. Luo, “Recent developments and applications of the MMPBSA method,” *Front Mol Biosci*, vol. 4, no. JAN, p. 87, Jan. 2018, doi: 10.3389/FMOLB.2017.00087/BIBTEX.

[236] S. Genheden and U. Ryde, “The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities,” *Expert Opin Drug Discov*, vol. 10, no. 5, p. 449, May 2015, doi: 10.1517/17460441.2015.1032936.

[237] M. S. Valdés-Tresanco, M. E. Valdés-Tresanco, P. A. Valiente, and E. Moreno, “Gmx_MMPBSA: A New Tool to Perform End-State Free Energy Calculations with GROMACS,” *J Chem Theory Comput*, vol. 17, no. 10, pp. 6281–6291, Oct. 2021, doi: 10.1021/ACS.JCTC.1C00645/ASSET/IMAGES/LARGE/CT1C00645_0005.JPEG.

[238] L. Duan, X. Liu, and J. Z. H. Zhang, “Interaction entropy: A new paradigm for highly efficient and reliable computation of protein-ligand binding free energy,” *J Am Chem Soc*, vol. 138, no. 17, pp. 5722–5728, May 2016, doi: 10.1021/JACS.6B02682/ASSET/IMAGES/JA-2016-026824_M008.GIF.

[239] M. Scheurer *et al.*, “PyContact: Rapid, Customizable, and Visual Analysis of Noncovalent Interactions in MD Simulations,” *Biophys J*, vol. 114, no. 3, pp. 577–583, Feb. 2018, doi: 10.1016/J.BPJ.2017.12.003.

[240] F.-D. Sun, P.-C. Wang, J. Shang, S.-H. Zou, and X. Du, “Ibrutinib presents antitumor activity in skin cancer and induces autophagy”.

[241] D. van der Spoel, E. Lindahl, B. Hess, G. Groenhof, A. E. Mark, and H. J. C. Berendsen, “GROMACS: Fast, flexible, and free,” *J Comput Chem*, vol. 26, no. 16, pp. 1701–1718, Dec. 2005, doi: 10.1002/JCC.20291.

[242] K. Glover, Y. Mei, and S. C. Sinha, “Identifying Intrinsically Disordered Protein Regions Likely to Undergo Binding-Induced Helical Transitions,” *Biochim Biophys Acta*, vol. 1864, no. 10, p. 1455, Oct. 2016, doi: 10.1016/J.BBAPAP.2016.05.005.

[243] M. Bhasin and R. Varadarajan, “Prediction of Function Determining and Buried Residues Through Analysis of Saturation Mutagenesis Datasets,” *Front Mol Biosci*, vol. 8, Mar. 2021, doi: 10.3389/FMOLB.2021.635425/FULL.

[244] E. van Dijk, A. Hoogeveen, and S. Abeln, “The Hydrophobic Temperature Dependence of Amino Acids Directly Calculated from Protein Structures,” *PLoS Comput Biol*, vol. 11, no. 5, p. e1004277, May 2015, doi: 10.1371/JOURNAL.PCBI.1004277.

[245] E. Wang *et al.*, “End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design,” *Chem Rev*, vol. 119, no. 16, pp. 9478–9508, Aug. 2019, doi: 10.1021/ACS.CHEMREV.9B00055.

[246] F. Godschalk, S. Genheden, P. Söderhjelm, and U. Ryde, “Comparison of MM/GBSA calculations based on explicit and implicit solvent simulations,” *Physical Chemistry Chemical Physics*, vol. 15, no. 20, pp. 7731–7739, May 2013, doi: 10.1039/C3CP00116D.

[247] K. Daze and F. Hof, “Molecular Interaction and Recognition,” *Encyclopedia of Physical Organic Chemistry, 5 Volume Set*, pp. 1–51, Nov. 2016, doi: 10.1002/9781118468586.EPOC3001.

[248] A. Zehir *et al.*, “Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients,” *Nat Med*, vol. 23, no. 6, pp. 703–713, Jun. 2017, doi: 10.1038/NM.4333.

[249] T. M. Kim *et al.*, “Clonal origins and parallel evolution of regionally synchronous colorectal adenoma and carcinoma,” *Oncotarget*, vol. 6, no. 29, pp. 27725–27735, 2015, doi: 10.18632/ONCOTARGET.4834.

[250] X. Bonilla *et al.*, “Genomic analysis identifies new drivers and progression pathways in skin basal cell carcinoma,” *Nature Genetics 2016 48:4*, vol. 48, no. 4, pp. 398–406, Mar. 2016, doi: 10.1038/ng.3525.

[251] H. Mano, “Tec family of protein-tyrosine kinases: an overview of their structure and function,” *Cytokine Growth Factor Rev*, vol. 10, no. 3–4, pp. 267–280, Sep. 1999, doi: 10.1016/S1359-6101(99)00019-2.

[252] M. Cives *et al.*, “Non-Melanoma Skin Cancers: Biological and Clinical Features,” *Int J Mol Sci*, vol. 21, no. 15, pp. 1–24, Aug. 2020, doi: 10.3390/IJMS21155394.

[253] W. Hu, L. Fang, R. Ni, H. Zhang, and G. Pan, “Changing trends in the disease burden of non-melanoma skin cancer globally from 1990 to 2019 and its predicted level in 25 years,” *BMC Cancer*, vol. 22, no. 1, pp. 1–11, Dec. 2022, doi: 10.1186/S12885-022-09940-3/FIGURES/5.

[254] L. Fania *et al.*, “Cutaneous Squamous Cell Carcinoma: From Pathophysiology to Novel Therapeutic Approaches,” *Biomedicines*, vol. 9, no. 2, pp. 1–33, Feb. 2021, doi: 10.3390/BIOMEDICINES9020171.

[255] Y. Gao, L. Lyu, Y. Feng, F. Li, and Y. Hu, “A review of cutting-edge therapies for hepatocellular carcinoma (HCC): Perspectives from patents,” *Int J Med Sci*, vol. 18, no. 14, p. 3066, 2021, doi: 10.7150/IJMS.59930.

[256] J. P. Jourdan, R. Bureau, C. Rochais, and P. Dallemagne, “Drug repositioning: a brief overview,” *J Pharm Pharmacol*, vol. 72, no. 9, p. 1145, Sep. 2020, doi: 10.1111/JPHP.13273.

[257] H. Yang *et al.*, “A network-based approach reveals the dysregulated transcriptional regulation in non-alcoholic fatty liver disease,” *iScience*, vol. 24, no. 11, p. 103222, Nov. 2021, doi: 10.1016/J.ISCI.2021.103222.

[258] Z. Wang, E. He, K. Sani, K. M. Jagodnik, M. C. Silverstein, and A. Ma’Ayan, “Drug Gene Budger (DGB): an application for ranking drugs to modulate a specific gene based on transcriptomic signatures,” *Bioinformatics*, vol. 35, no. 7, pp. 1247–1248, Apr. 2019, doi: 10.1093/BIOINFORMATICS/BTY763.

[259] F. M. Giorgi, A. M. Bolger, M. Lohse, and B. Usadel, “Algorithm-driven Artifacts in median polish summarization of Microarray data,” *BMC Bioinformatics*, vol. 11, no. 1, pp. 1–12, Nov. 2010, doi: 10.1186/1471-2105-11-553/FIGURES/5.

[260] D. Szklarczyk *et al.*, “The STRING database in 2023: protein-protein association networks and functional enrichment analyses for any sequenced genome of interest,” *Nucleic Acids Res*, vol. 51, no. D1, pp. D638–D646, Jan. 2023, doi: 10.1093/NAR/GKAC1000.

[261] D. Szklarczyk *et al.*, “The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets,” *Nucleic Acids Res*, vol. 49, no. D1, pp. D605–D612, Jan. 2021, doi: 10.1093/nar/gkaa1074.

[262] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, and K. Morishima, “KEGG: new perspectives on genomes, pathways, diseases and drugs,” *Nucleic Acids Res*, vol. 45, no. D1, pp. D353–D361, Jan. 2017, doi: 10.1093/NAR/GKW1092.

[263] M. Gillespie *et al.*, “The reactome pathway knowledgebase 2022,” *Nucleic Acids Res*, vol. 50, no. D1, pp. D687–D692, Jan. 2022, doi: 10.1093/NAR/GKAB1028.

[264] M. Martens *et al.*, “WikiPathways: connecting communities,” *Nucleic Acids Res*, vol. 49, no. D1, pp. D613–D621, Jan. 2021, doi: 10.1093/NAR/GKAA1024.

[265] M. E. Ritchie *et al.*, “limma powers differential expression analyses for RNA-sequencing and microarray studies,” *Nucleic Acids Res*, vol. 43, no. 7, p. e47, Jan. 2015, doi: 10.1093/NAR/GKV007.

[266] M. Ashburner *et al.*, “Gene Ontology: tool for the unification of biology,” *Nature Genetics 2000 25:1*, vol. 25, no. 1, pp. 25–29, May 2000, doi: 10.1038/75556.

[267] J. Reimand *et al.*, “Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap,” *Nat Protoc*, vol. 14, no. 2, p. 482, Feb. 2019, doi: 10.1038/S41596-018-0103-9.

[268] A. Sarirete, Z. Balfagih, T. Brahimi, M. D. Lytras, and A. Visvizi, “Artificial intelligence and machine learning research: towards digital transformation at a global scale,” *J Ambient Intell Humaniz Comput*, vol. 13, no. 7, pp. 3319–3321, Jul. 2022, doi: 10.1007/S12652-021-03168-Y/FIGURES/1.

[269] M. Saarela and S. Jauhiainen, “Comparison of feature importance measures as explanations for classification models,” *SN Appl Sci*, vol. 3, no. 2, pp. 1–12, Feb. 2021, doi: 10.1007/S42452-021-04148-9/TABLES/4.

[270] G. Vilone and L. Longo, “Explainable Artificial Intelligence: a Systematic Review,” May 2020, Accessed: Jun. 17, 2023. [Online]. Available: https://arxiv.org/abs/2006.00093v4

[271] M. Chromik and A. Butz, “Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12933 LNCS, pp. 619–640, 2021, doi: 10.1007/978-3-030-85616-8_36/COVER.

[272] E. Cambria, L. Malandri, F. Mercorio, M. Mezzanzanica, and N. Nobani, “A survey on XAI and natural language explanations,” *Inf Process Manag*, vol. 60, no. 1, p. 103111, Jan. 2023, doi: 10.1016/J.IPM.2022.103111.

[273] I. B. S. Sitohang, W. I. Makes, N. Sandora, and J. Suryanegara, “Topical tretinoin for treating photoaging: A systematic review of randomized controlled trials,” *Int J Womens Dermatol*, vol. 8, no. 1, p. e003, Mar. 2022, doi: 10.1097/JW9.0000000000000003.

[274] C. F. Thorn *et al.*, “Doxorubicin pathways: pharmacodynamics and adverse effects,” *Pharmacogenet Genomics*, vol. 21, no. 7, p. 440, 2011, doi: 10.1097/FPC.0B013E32833FFB56.

[275] D. Levêque, G. Becker, K. Bilger, and S. Natarajan-Amé, “Clinical Pharmacokinetics and Pharmacodynamics of Dasatinib,” *Clin Pharmacokinet*, vol. 59, no. 7, pp. 849–856, Jul. 2020, doi: 10.1007/S40262-020-00872-4/METRICS.

Chapter 8

Publications

Publications

PUBLICATIONS FROM THE THESIS

1. **Meena J**, Hasija Y. Application of explainable artificial intelligence in the identification of Squamous Cell Carcinoma biomarkers. Computers in Biology and Medicine. 2022 Jul 1; 146:105505.

2. **Meena J**, Hasija Y. Rare deleterious mutations in Bruton's tyrosine kinase as biomarkers for ibrutinib-based therapy: an in-silico insight. Journal of Molecular Modeling. 2023 Apr;29(4):1-7.

3. **Meena J,** Hasija Y. Recent Advances in Bioinformatics Techniques for Identifying Biomarkers for Early Diagnosis and Targeted Therapy of Squamous Cell Carcinoma (communicated).

CONFERENCES AND PRESENTATIONS

1. **Meena J**, Hasija Y. Assimilating Bioinformatics Knowledge to Bioremediation. In 2019 International Conference on Global Environmental Challenges, Human Health and Sustainable Development - 2019, organized by Jawaharlal Nehru University, Delhi in association with Metropolitan University, USA.

2. **Meena, J**, Chauhan, A, Hasija, Y. (2019). Deciphering the Association of Single Amino Acid Variations with Dermatological Diseases Applying Machine Learning Techniques. In: Luhach, A., Jat, D., Hawari, K., Gao, XZ., Lingras, P. (eds) Advanced Informatics for Computing Research. ICAICR 2019. Communications in Computer and Information Science, vol 1075. Springer, Singapore. https://doi.org/10.1007/978-981-15-0108-1_22.

3. Tanwar N, **Meena J**, Hasija Y. Explicate Toxicity By eXplainable Artificial Intelligence. In 2022 International Conference on Industry 4.0 Technology (I4Tech) 2022 Sep 23 (pp. 1-6). IEEE.

Computers in Biology and Medicine 146 (2022) 105505

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Application of explainable artificial intelligence in the identification of Squamous Cell Carcinoma biomarkers

Jaishree Meena, Yasha Hasija *

Department of Biotechnology, Delhi Technological University, Delhi, India, 110042

ARTICLE INFO

Keywords:
Explainable AI
Machine learning
SHAP values
Principal component analysis
XGBoost machine learning classifier
Squamous cell carcinoma

ABSTRACT

Non-melanoma skin cancers (NMSCs) are the fifth most common type of cancer worldwide, affecting both men and women. Each year, more than a million new occurrences of NMSC are estimated, with Squamous Cell Carcinoma (SCC) representing approximately 20% of all skin malignancies. The purpose of this study was to find potential diagnostic biomarkers for SCC by application of e**X**plainable **A**rtificial **I**ntelligence (XAI) on XGBoost machine learning (ML) models trained on binary classification datasets comprising the expression data of 40 SCC, 38 AK, and 46 normal healthy skin samples. After successfully incorporating SHAP values into the ML models, 23 significant genes were identified and were found to be associated with the progression of SCC. These identified genes may serve as diagnostic and prognostic biomarkers in patients with SCC.

1. Introduction

Skin cancers are commonly divided into two categories: firstly, non-melanoma skin cancers (NMSC), which include Squamous cell carcinoma (SCC) and Basal cell carcinoma (BCC), and secondly, melanoma skin cancers [1]. NMSC is the world's fifth most prevalent form of cancer, affecting both men and women. In the United States, over 1.8 million new cases of NMSCs are reported each year, with cutaneous SCC being the most frequent kind of skin cancer. [2,3]. African Americans and Asian Indians have a higher incidence of SCC and also, it ranks the second most prevalent in Hispanics and Chinese/Japanese Asians. [4]. SCC has been recognized as a kind of cancer that originates in keratinocytes. The skin ailment Actinic Keratosis (AK), also known as Carcinoma-in-Situ, has been associated with the emergence of SCC in numerous studies. A considerable number of high-risk SCC cases, approximately 5–10% of all instances, is exceedingly difficult to diagnose and treat, necessitating the use of radiation or surgery in the majority of cases. It is less likely that therapies associated with such high-risk metastatic skin cancer will be effective, particularly in an elderly population [5], highlighting a critical need of a promising yet systematic diagnosis and treatment for SCC [6,7]. Microarray data is growing in volume, and the information it gives on the genes responsible for a disease phenotype is being used more and more for variant categorization and analysis, as well as other applications. Microarrays are a relatively recent method that involves the placement of hundreds of DNA probes that are matched to target genes on a tiny chip that can then be used to analyse gene expression in samples. One of the primary applications of this approach was to compare cancer and normal tissues, as well as distinct cancer subtypes and individuals with varying prognoses, among other things [8,9] When it came to identifying microarray samples, the widely used machine learning technique of support vector machines (SVMs) [10], artificial neural network [11], logistic regression, naïve bayes etc, worked admirably. In a large number of studies metabolomic data is used to gain insight into the metabolites that define each organism state and the dynamics of those metabolites under various settings. The 'omics' domain is a critical component of systems biology. Due to its emphasis on small molecules and interactions, it has gained widespread adoption in a variety of fields recently, such as biomarker discovery and identification, development of drug and customised health care etc [12]. Some pioneer studies on omics data have made normalization tool like NOREVA [13–15] and ANPELA, an integrated workflow for Label-free quantification (LFQ) [16] of data. These tools have made significant contributions to numerous facets of scientific investigations.

Technological advances like Next Generation Sequencing (NGS), Genome wide association studies (GWAS) and computational methods

Abbreviations: SCC, Squamous Cell Carcinoma; AK, Actinic Keratosis; XAI, Explainable Artificial Intelligence; ML, Machine Learning; RMA, Robust MultiArray Average; PCA, Principal Component Analysis; XGBoost, Extreme Gradient Boosting; SHAP, Shapley Additive Explanations; GO, Gene Ontology.

* Corresponding author.
E-mail addresses: yashahasija@dtu.ac.in, yashahasija06@gmail.com (Y. Hasija).

https://doi.org/10.1016/j.compbiomed.2022.105505
Received 20 February 2022; Received in revised form 3 April 2022; Accepted 5 April 2022
Available online 17 April 2022

Journal of Molecular Modeling (2023) 29:120
https://doi.org/10.1007/s00894-023-05515-6

ORIGINAL PAPER

Rare deleterious mutations in Bruton's tyrosine kinase as biomarkers for ibrutinib-based therapy: an *in silico* insight

Jaishree Meena[1] · **Yasha Hasija**[1]

Received: 21 November 2022 / Accepted: 14 March 2023 / Published online: 29 March 2023

Abstract

Context Squamous cell carcinoma (SCC) is the second most common type of skin cancer caused by malignant keratinocytes. Multiple studies have shown that protein mutations have a significant impact on the development and progression of cancer, including SCC. We attempted to decode the effect of single amino acid mutations in the Bruton's tyrosine kinase (BTK) protein in this study. Molecular dynamic (MD) simulations were performed on selected deleterious mutations of the BTK protein, revealing that the variants adversely affect the protein, indicating that they may contribute to the prognosis of SCC by making the protein unstable. Then, we investigated the interaction between the protein and its mutants with ibrutinib, a drug designed to treat SCC. Even though the mutations have deleterious effects on protein structure, they bind to ibrutinib similarly to their wild type counterpart. This study demonstrates that the effect of detected missense mutations is unfavorable and can result in function loss, which is severe for SCC, but that ibrutinib-based therapy can still be effective on them, and the mutations can be used as biomarkers for Ibrutinib-based treatment.

Methods Seven different computational techniques were used to compute the effect of SAVs in accordance with the experimental requirements of this study. To understand the differences in protein and mutant dynamics, MD simulation and trajectory analysis, including RMSD, RMSF, PCA, and contact analysis, were performed. The free binding energy and its decomposition for each protein-drug complex were determined using docking, MM-GBSA, MM-PBSA, and interaction analysis (wild and mutants).

Keywords Bruton's tyrosine kinase · Ibrutinib · MD simulation · Molecular docking · Contact map analysis · Mutational analysis · Protein-ligand interaction

Introduction

Squamous cell carcinoma (SCC) is the second most highly prevalent skin tumor that develops when keratinocytes alter and turn cancerous. Invasive SCC continues to be influenced by everyday UV exposure to the skin. Global disease rates are on the rise due to aging populations and other demographic shifts [1]. SCC is significant because it occurs twice as frequently as skin cancer in European Caucasians and up to ten times as frequently in fair-skinned Australians, where the incidence is even greater [2]. The fact that SCC strikes men more frequently than women suggests that female immunity may play some role in protecting against the disease, as evidenced by recent studies [3]. People with light skin and light eyes are more likely to develop SCC beyond the age of 50. It often develops in sites that have been exposed to the sun in the past. Those with a history of extensive exposure to UV, whether through previous medical procedures or the sun, are at a higher risk [4]. Immunosuppressed patients also have a high incidence of, which can progress into aggressive subtypes [5]. Small SCC lesions can be removed and are not lethal, but depending on their location, they might cause severe morbidity [6]. Most head and neck squamous cell malignancies necessitate extensive surgery, which, even in the best of hands, can result in poor symptom relief. In addition, the expense of treating these tumors increases each year, posing a critical need to explore low cost, effective, and efficient treatment options for SCC management [7, 8].

✉ Yasha Hasija
yashahasija6oct@gmail.com; yashahasija@dtu.ac.in

[1] Department of Biotechnology, Delhi Technological University, Delhi 110042, India

Jaishree Meena

My life objective is to relentlessly pursue knowledge, strive for perfection with unwavering enthusiasm, and become a lifelong researcher, passionately exploring the intersection of life science and information technology.

jaishreedtu@gmail.com

9717562941

D-356, Street No. 8, D-Block, Bhajanpura, Delhi, India

ORCID ID: 0000-0002-4357-9409

SKILLS

- Bioinformatics
- Machine Learning
- Explainable AI
- Computational Biology
- Molecular Dynamics

LANGUAGES

English and Hindi
Full Professional Proficiency

ACHIEVEMENTS

Commendable Research Award by Delhi Technological University (06/04/2023), Awarded for the paper, titled, "Application of Explainable Artificial Intelligence in the Identification of Squamous Cell Carcinoma Biomarkers", published in Computers in Biology and Medicine (Impact Factor: 6.698)

Qualified Graduate Aptitude Test in Engineering (March 2016)

Overall Citations: 48

h-Index since 2018: 4

WORK EXPERIENCE

Senior Research Fellow (NFST Scheme)
Delhi Technological University, Delhi
04/2022 - Present,
Investigated the impact of somatic non-synonymous mutations on Bruton's Tyrosine Kinase protein and their potential influence on FDA approved therapies for Squamous Cell Carcinoma

Junior Research Fellow (NFST Scheme)
Delhi Technological University, Delhi
08/2019 - 03/2022,
Applied Explainable Artificial Intelligence on Machine learning model for the Identification of Squamous Cell Carcinoma Biomarkers

PUBLICATIONS

J. Meena, Y. Hasija, Rare deleterious mutations in Bruton's tyrosine kinase as biomarkers for ibrutinib-based therapy: an in silico insight. J Mol Model 29, 120 (2023), doi.org/10.1007/s00894-023-05515-6

J. Meena, Y. Hasija, Application of explainable artificial intelligence in the identification of Squamous Cell Carcinoma biomarkers. Computers in Biology and Medicine 146, 105505, 2022

N. Tanwar, J. Meena and Y. Hasija, Explicate Toxicity By eXplainable Artificial Intelligence, 2022, International Conference on Industry 4.0 Technology (I4Tech), Pune, India, 2022, pp. 1-6, doi: 10.1109/I4Tech55392.2022.9952865

J. Meena, A. Chauhan, Y. Hasija, Deciphering the Association of Single Amino Acid Variations with Dermatological Diseases Applying Machine Learning Techniques, International Conference on Advanced Informatics for Computing Research, 227-238, 2019

LS. Meena, J. Meena, Cloning and characterization of a novel PE_PGRS60 protein (Rv3652) of Mycobacterium tuberculosis H37Rv exhibit fibronectinbinding property, Biotechnology and Applied Biochemistry 63 (4), 525-531, 2016

J. Meena, LS. Meena, Scope and perspectives of new TB drugs and vaccines, American Journal of Infectious Diseases 11 (3), 63, 2015

J. Meena, M. Singh, PD. Sahare, LS. Meena, Interaction of nanoparticles in biological systems and their role in therapeutical treatment of tuberculosis and cancer, Journal Luminescene Applications 1, 7-22, 2014

M, Tanwar, J. Meena, LS. Meena, Nanoparticles: scope in drug delivery, Advanced biomaterials and biodevices, 487, 2014

EDUCATION

Ph.D. (Specialization: Bioinformatics)
Delhi Technological University, Delhi
08/2018 - Present

M.Tech. (Specialization: Bioinformatics)
Delhi Technological University, Delhi
07/2016 - 07/2018

B.Tech. (Specialization: Biotechnology)
University School of Biotechnology, Guru Gobind Singh Indraprastha University
06/2010 - 06/2015