MACHINE-LEARNING BASED PREDICTION AND COMPUTATIONAL INVESTIGATIONS ON PROTEIN METHYLTRANSFERASES INVOLVED IN HUMAN MALIGNANCIES

Thesis submitted in fulfillment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY

IN

BIOINFORMATICS

BY

ARVIND KUMAR YADAV

Department of Biotechnology and Bioinformatics

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
WAKNAGHAT, SOLAN, H.P.-173234, INDIA
DECEMBER 2022

DECLARATION BY THE SCHOLAR

I hereby declare that the work reported in the Ph.D. thesis entitled ***"Machine-Learning Based Prediction and Computational Investigations on Protein Methyltransferases Involved in Human Malignancies",*** submitted at **Jaypee University of Information Technology, Waknaghat, India** is an authentic record of my work carried out under the supervision of **Dr. Tiratha Raj Singh** and co-supervision of **Dr. Pradeep Kumar Gupta**. I have not submitted this work elsewhere for any other degree or diploma. I am fully responsible for the contents of my Ph.D. thesis.

Mr. Arvind Kumar Yadav **Date:** 14/12/2022

Enrollment No. 176502

Department of Biotechnology and Bioinformatics
Jaypee University of Information Technology
Waknaghat, Solan, H.P. India-173234

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the Ph.D. thesis entitled "***Machine-Learning Based Prediction and Computational Investigations on Protein Methyltransferases Involved in Human Malignancies***", submitted by **Arvind Kumar Yadav (Enrolment No. 176502)** in fulfilment for the award of degree of **Doctor of Philosophy** in **Bioinformatics** at **Jaypee University of Information Technology, Waknaghat, Solan (HP) India,** is a bonafide record of him original work carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of any other degree or diploma.

(Dr. Tiratha Raj Singh)
Associate Professor
Department of Biotechnology & Bioinformatics
Jaypee University of Information Technology
Waknaghat, Solan, HP-173234, India
Date: 14/12/2022

(Dr. Pradeep Kumar Gupta)
Associate Professor
Department of Computer Science & Engineering
Jaypee University of Information Technology
Waknaghat, Solan, HP-173234, India
Date: 14/12/2022

“In memory of my beloved Grandfather,

who never saw this adventure”

ACKNOWLEDGMENT

"The roots of education are bitter, but the fruit is sweet." ~Aristotle

*With the divine grace of "***Almighty God***", the journey of my Ph.D. came to an end and my heart overflows with appreciation and emotions towards every person who shows immense support, belief, and effort to complete this journey. This part of work is not just an acknowledgment it's gratitude towards each person who assisted me in carving out my dreams into veracity.*

First and foremost, with a feeling of immense respect and gratitude, I bow down to my supervisor ***Dr. Tiratha Raj Singh*** *for believing me and making this doctoral thesis possible. His friendly encouragement, professional suggestions, continuous support, and motivation were indispensable throughout all phases of the work. His generosity helped me blossom through my doctoral journey at JUIT. I also feel a deep sense of gratitude towards my co-supervisor* ***Dr. Pradeep Kumar Gupta*** *for mentoring they provided and constructive support throughout this work. I am also grateful for their constant presence, positive attitude, encouragement, and great suggestions during the research process.*

I am grateful to ***Prof. (Dr.) Sudhir Kumar****, (HOD, Department of Biotechnology and Bioinformatics) for his support, cooperation, and providing desired facilities within the department to carry out experimentation.*

I pay my sincere thanks to JUIT Founder ***Jai Prakash Gaur Ji, Prof. (Dr.) Rajendra Kumar Sharma*** *(Vice-Chancellor),* ***Prof. (Dr.) Ashok Kumar Gupta*** *(Dean Academics & Research) and* ***Maj. Gen. Rakesh Bassi (Retd.)*** *(Registrar) for providing essential amenities, research funding, and infrastructure for pursuing the research work. I also express my profound gratitude to the members of my thesis advisory committee* ***Dr. Rahul Shrivastava, Dr. Raj Kumar,*** *and external member* ***Dr. Neel Kanth*** *for their help, constructive remarks, precious suggestion, and encouragement in this entire work. I would also like to express my sincere thanks to all faculty members, non-teaching, and lab staff of the department for their help and support throughout the research. I feel a deep sense of gratitude towards* ***Dr. Ragini Raj Singh*** *for hir emotional support, guidance, and delicious food throughout the Ph.D. I am also thankful to* ***Mrs. Somlata Sharma, Mrs. Sonika Gupta, and Mr. Baleshwar Shukla*** *for their constant support and encouragement.*

I want to mention special thanks to my most beloved colleagues ***Dr. Ankush Bansal*** *and* ***Rohit Shukla*** *for their unconditional support. I am deeply grateful for their time, suggestions, love, and lots of scientific and non-scientific discussion. This journey wouldn't have been complete without lots of people, although, it is not possible to mention everyone here. In particular, I am deeply grateful to my lab mates and friends in JUIT* ***Dr. Neelam Gupta, Dr. Vikrant Abbot, Nadia, Ayushi Sharma, Monika Chaudhary, Sargeet, Diksha, Raja Adnan Butt, Neelam Sharma, Amit Kumar, Dr. Deepti, Sweta Singh, Rahul,*** *and* ***Mansi Arora*** *for providing me a great and supportive working environment during my Ph.D. journey.*

I am feeling profound happiness and pride in acknowledging the role of some great friends that were no less than the pillars of strength. I am very grateful to have spent time with Totiya building's friends, ***Dr. Swati Sharma, Neha Thakur, Deepak Sharma, Neha Luhakhra, Dr. Sameer, Dr. Kanishka Katoch,*** *and* ***Rohit****. Thank you for all of the good times we have shared during the lockdown and for the variety of self-cooked foods and for all the fun we had together in mountain weather. It's giving me lifelong memories to cherish. Also, I am highly thankful to my dearest friend* ***Neetesh Pandey*** *for his motivation and support. He was always there to share my every high and low moment both in my professional and personal life. I am also thankful to* ***Indra Prakash Jha,*** *who helped me to clear all the ML-related doubts.*

No words can ever be strong enough to explain the gratitude to my family specially my grandmother, my mother ***Smt. Genda Devi****, my father* ***Shri Suresh Chandra Yadav****, my beloved brother* ***Praveen Yadav*** *and my sisters for their belief, unconditional love, and support in all possible ways. I express my profound thanks to my loving wife* ***Sangeeta Yadav*** *for her concern, immense love, understanding, and constant encouragement during the thesis writing. This acknowledgment would not be completed without expressing gratitude to the one and the only loving person my late grandfather* ***Shri Jagdev Prasad Yadav****, might watching me from heaven. What I am today is just because of him. Putting a page in memory of him is very small to describe his love, blessing, and greatness. Their kindness to me was the seed from which many pleasant memories have since grown.*

At last, to all my loved ones, thank you so much for all that you bring into my life. Science would be a very empty pursuit without you.

(Arvind Kumar Yadav)

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ΔΔG	Delta Delta G
BBB	Blood-brain barrier
BLCA	bladder urothelial carcinoma
BRCA	Breast invasive carcinoma
CESC	Cervical squamous cell carcinoma, and endocervical adenocarcinoma
CKSAAP	Composition of k-spaced amino acid pairs
COAD	Colon adenocarcinoma
CYP450	Cytochrome- P450
Cα	Alpha carbon
DFS	Disease-free survival
DLBC	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma
ESCA	Esophageal carcinoma
GO	Gene ontology
H3K4	Histone H3 lysine K4
HIA	Human intestinal absorption
HNSC	Head and Neck squamous cell carcinoma
$Kcal.mol^{-1}$	Kilocalorie per mole
KIRC	Kidney renal clear cell carcinoma
KMT	Lysine methyltransferases
LIHC	Liver hepatocellular carcinoma
MDS	Molecular Dynamics Simulation
mM	Millimolar
nm	Nanometer
nm^2	Square nanometer
ns	Nanosecond

nsSNPs	Non-synonymous SNPs
OS	Overall survival
OV	Ovarian serous cystadenocarcinoma
PAAD	Pancreatic adenocarcinoma
PCA	Principal component analysis
PKMTs	Protein lysine methyltransferases
PMTs	Protein methyltransferases
PPI	Protein-protein interaction
PRMTs	Protein arginine methyltransferases
PTM	Post-translational modification
RB	Retinoblastoma
Rg	Radius of gyration
RMSD	Root-mean-square deviation
RMSF	Root-mean-square fluctuation analysis
ROC curve	Receiver operating characteristics curve
SAH/ AdoHcy	S-adenosyl homocysteine
SAM/ AdoMet	S-adenosyl-L-methionine
SASA	Solvent accessible surface area
SNPs	Single-nucleotide polymorphisms
SVM	Support Vector Machine
TCGA	The cancer genome atlas
TPM	Transcript per million
ug/L	micrograms per liter

TABLE OF CONTENTS

CONTENT **PAGE NO.**

CHAPTER 2: DEVELOPMENT OF MACHINE-LEARNING BASED PREDICTION SERVER FOR METHYLTRANSFERASE

CHAPTER 3: **INTEGRATIVE ANALYSIS OF THE ONCOGENIC ROLE OF SMYD2 IN MULTIPLE HUMAN MALIGNANCIES**

CHAPTER 4: **SEQUENCE AND STRUCTURE LEVEL SNPS ANALYSIS FOR SMYD2**

CHAPTER 5: STRUCTURAL INVESTIGATION AND SIMULATION STUDIES TO DESIGN NOVEL INHIBITORS FOR SMYD2

CHAPTER 6: OVERALL CONCLUSION & FUTURE PROSPECTS

ABSTRACT

This study is focused on the prediction and association of the protein methyltransferases (PMTs) in malignancies. PMTs are the groups of enzymes that help to catalyze the transfer of a methyl group from universal methyl donor S-adenosyl-L-methionine (SAM) to its substrates. This group of enzymes plays a significant role in the epigenetic regulation of gene expression through the methylation of various substrates. PMTs target the lysine or arginine residues for the methylation of their protein substrates. Based on methylation activity it is divided into two major classes such as protein lysine methyltransferases (PKMTs) and protein arginine methyltransferases (PKMTs). Over the years, protein methylation has appeared as an important post-translational modification (PTM) event and is involved in various cellular processes. Dysregulation of methyltransferases is involved in different types of human cancers. However, in light of the well-recognized significance of PMTs, it becomes crucial to have reliable and fast methods for identifying these proteins. In this thesis, a machine-learning-based method was developed for the identification of PMTs. Various sequence-based features were calculated and model training was performed by using several machine-learning algorithms. A ten-fold cross-validation method was applied to train the models. The proposed SVM-based CKSAAP model was identified as the best model for the prediction of PMTs. The best model achieved the highest accuracy of 87.94% with balance Sensitivity (88.8%) and Specificity (87.11%) with MCC of 0.759 and AUROC of 0.945. The best model was implemented in standalone software of PMTPred that will facilitate to predict PMTs. In the recent decade, protein lysine methylation events got more attention from researchers globally. SMYD2 is a protein of the SMYD (SET and MYND domain) family having lysine methyltransferase activity that methylates both histones and non-histones proteins. Numerous tumor suppressor non-histone proteins such as p53, RB1, ERα, and PTEN get methylate through SMYD2 and lead the cancer formation. The emerging evidence supports the association of SMYD2 in the progression of cancers but remains mostly unknown. Therefore further in this study, we computationally analyzed the potential association of SMYD2 in multiple tumors using TCGA data. The results elucidated that a higher expression of SMYD2 was present in tumor tissues as compared to normal tissues in most cancers. A significant association was observed between the SMYD2 gene expressions and the survival of cancer patients. The prognostic analysis showed a strong association of SMYD2 with cancers. We detected 15 missenses, 4 truncating mutations, and 5 others. Gene ontological properties and pathways were found to be significantly linked to the development of cancer. These data-driven results provide a relatively comprehensive insight into the understanding of the association of SMYD2 with cancer patients and its correlation with prognosis. There are many mutations in SMYD2, and some of them are thought to have a significant impact on the enzymatic activity of methyltransferase. Missense mutations are single amino acid point mutations that can have a variety of effects depending on the mutation site and the consequent amino acid substitution. So here, we assessed the nsSNPs in SMYD2 and investigated their structural and functional consequences using a rigorous computational method. Out of the 264 nsSNPs, three nsSNPs (H207D, C209W, and C209R) have the most deleterious impact. According to a molecular dynamics simulation (MDS) study, these mutations have a greater effect on the SMYD2 protein structure and function. Furthermore, SMYD2-specific inhibitors were identified to design targeted therapy strategies. A total of 98071 small natural compounds were taken and virtual screening was performed with SMYD2 as a target protein. Based on the binding energy cut-off of >= -11.7 Kcal.mol^{-1} total of 391 potential compounds were selected for ADMET analysis. Based on ADMET parameters, nine compounds were selected that fit the drug-likeness standard for docking analysis. Finally, three compounds (ZINC03844862, ZINC08490711, and ZINC08764231) were selected as probable inhibitors based on docking score and interaction analysis between protein and ligands. Then 100 ns MDS study was performed and the results revealed that selected compounds with the SMYD2 structure had a stable binding. Finally, these three compounds were identified as a potential lead compounds. The *in vivo* and *in vitro* research for these compounds could be viable leads for developing cancer treatments. Thus, from this thesis work we concluded that the developed prediction method, identified novel SMYD2-specific nsSNPs and ligands would be useful for improved understanding and for aiding better cancer therapeutics.

CHAPTER- 1

INTRODUCTION &

REVIEW OF LITERATURE

1.1 INTRODUCTION

Protein methyltransferases (PMTs) help to catalyze the transfer of a methyl group to their substrate with the help of universal methyl donor S-adenosyl-L-methionine (SAM). This family of enzymes is important for epigenetic regulation because they can methylate a wide range of substrates, including DNA, RNA, protein, and small-molecule secondary metabolites [1]**.** Methylation caused by protein has appeared as a critical post-translational modification (PTM) which regulates a variety of physiological activities for example transcription, pre-mRNA splicing, protein synthesis, protein localization, signal transduction, DNA repair, and apoptosis [2]. According to the activity of methyltransferases, PMTs are distributed into two groups based on their methylation activity: 1**)** protein lysine methyltransferases (PKMTs), which methylate lysine residues and 2) protein arginine methyltransferases (PRMTs) that methylate the arginine residues of their protein substrates.

PKMTs are divided into two categories: those with Suppressor of variegation, Enhancer of Zeste, Trithorax (SET) domain, and those without SET domain [3], [4]. Both are capable to methylate the residue of lysine as mono (me1), di (me2), or tri (me3) methylation on its omega-amine group. PRMT is a methyltransferase that catalyzes arginine methylation. Arginine can be methylated on omega-amino groups as monomethylated (MMA; Rme1), symmetric dimethylarginine (SDMA; Rme2s) or asymmetric dimethylarginine (ADMA; Rme2a) [5]. PKMTs and PRMTs can methylate nonhistone proteins in addition to histones [6]. PKMTs play a key role in both normal physiology and the disease environment [7]**.** Dysregulation of PRMTs is connected to a series of diseases such as cancers, cardiovascular, and pulmonary diseases [8]. PMT inhibitors have been developed as chemical probes and medicinal agents with great success. These tiny chemical inhibitors are crucial in understanding the biological roles and disease progressions of the enzymes they target. According to the reports, several PKMTs are related to human malignancies. There are multiple PKMT families with various key enzymes as targets for the development of anti-cancer therapy [9]. The SET and Myeloid-Nervy-DEAF1 (MYND) domain (SMYD) protein families have gotten a lot of interest recently since of their probable role in cancer progression. The SMYD family has five members (SMYD1, SMYD2, SMYD3, SMYD4, and SMYD5), who are all engaged in biological processes like gene expression and regulation. SMYD2 is extensively expressed in cancer and plays a crucial role in cancer genesis and progression [10].

The SMYD2 and SMYD3 proteins are well characterized. Recently, SMYD2 has also become an important target to cure cancer [11]. In addition, H3-K4 and H3-K36me2 in histone lysine methylations, SMYD2 is believed to modulate the function of tumor suppressors over its methylation of retinoblastoma (RB1) at K860 and p53 at K370 [10], [12], [13]. The monomethylation of these two eminent genomic guardians causes their functions to mediate apoptosis and stop cell cycle progression. The clear role of SMYD2 in tumorigenesis has not been determined, but the ability of SMYD2 to methylate proteins involved in cancer progression provides potential mechanisms for its putative tumorigenic function [14]. Therefore, due to its close relationship with various human cancers, SMYD2 has gained attention because of its feasible target for human cancer therapeutics.

The objective of my thesis is to develop a prediction method for PMTs using machine learning, the oncogenic role of PKMT protein SMYD2, and polymorphism analysis along with novel inhibitor identification of SMYD2 protein from the class of PKMT. The first chapter of my thesis presents a prediction method for PMTs using machine-learning techniques. In chapter two, we have performed integrative investigation to identify the oncogenic role of *SMYD2* in multiple cancers using TCGA data. In chapter three, the structural and functional impact of SMYD2-associated nsSNPs is performed. Finally, in chapter four, we have identified SMYD2-specific inhibitors through *in silico* means.

1.2 Cancer as a public health burden

Cancer has become a main community health problem and the prominent cause of death globally. Due to the increase in incidence and mortality each year, cancer turns out to be a serious health problem worldwide. Cancer is a leading cause of morbidity and mortality in all countries of the world (Figure 1.1). It is measured as the second prominent cause of death after cardiovascular diseases [15]. According to the latest WHO 2020 press release, 19.3 million new patients were diagnosed with cancer (Figure 1.1A) and nearly 10.0 million deaths occurred (Figure 1.1B) due to cancer worldwide, and 1.3 million new cases (Figure 1.1C) and 0.85 million deaths (Figure 1.1D) have occurred in India by 2020. Cancer incidence rates are growing rapidly globally. The burden of cancer is projected to be 28.4 million cases in 2040 worldwide [16]. This is mainly due to the constant aging of the populations in both developing and developed countries. Most of the increase is expected to be concentrated in developing countries such as India due to heavy smoking, poor diet, lack of physical activity, and environmental pollution. In 2020, top cancer that is frequently diagnosed in India is breast cancer (10.6%) followed by lung cancer (11.4%) [16] (Figure 1.1C).

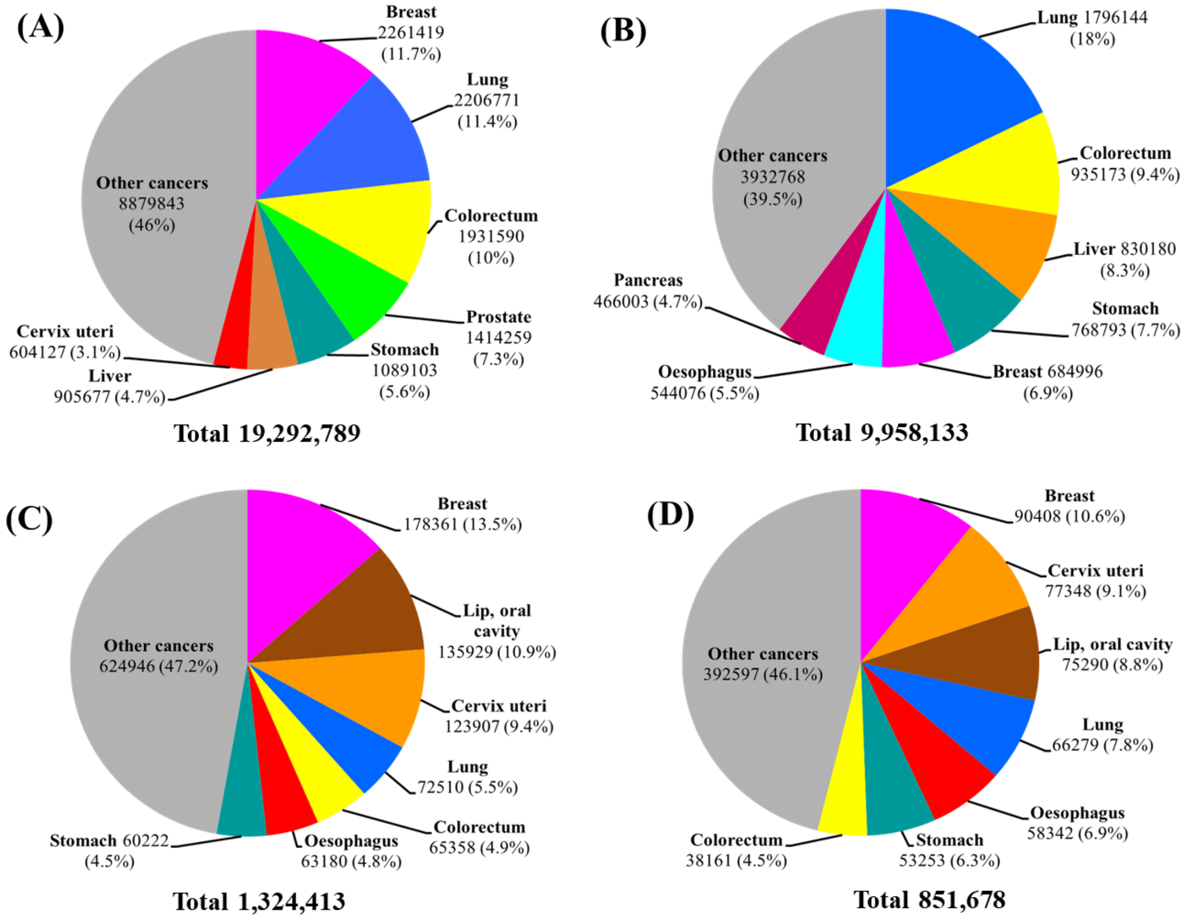


Figure 1.1: Cancer incidence and mortality statistics for the most occurring cancer worldwide and in India by 2020 for both sexes including all ages. (A) The number of new cancer cases worldwide (B) The number of death worldwide (C) A number of new cases in India (D) A number of deaths in India. *Source: Globocan 2020* (https://gco.iarc.fr/).

1.3 Genetic abnormalities in cancer

Cancer is a collection of genetic diseases which is caused by certain changes in the gene that controls the function of a cell. These genetic changes promote cancer. The increased alterations in the parameters of genetics and epigenetics boost oncogenesis which has a positive correlation to the diagnosis of cancer patients. PMTs are the important group of an enzyme that catalyzes the methylation of site-specific lysine or arginine residues on nonhistone or histone proteins. In the chromatin regulation of the gene transcription pathway, site-specific methylation histone protein plays a key element that is inherently changed in cancers [17]. Oncogenic changes such as mutations, and chromosomal translocations of PMTs, or other allied proteins, have an exclusive addiction of cancer cells to the action of PMTs [18]. It is reported that in specific human cancers, both PKMTs and PRMTs are genetically altered [19]. Alterations of these enzymes interrupt the catalytic activity that can lead the pathogenesis in the cell. For example, in any specific PMTs, alterations in the enzymatic activity lead to the transcription of oncogenes by the mechanism of hyperproliferative, tumorigenic phenotype [20]. Multiple examples of genetic alterations are available in PMTs with a specific cancer type. These alterations have been established as a potential cancer driver in some cases [19]. The important structural features that play a key role in the catalysis activity of the PMTs make these enzymes more suitable for the inhibitors identification to combat the cancer progression.

1.4. Protein Methyltransferases

Histone methylation was well-known histone PTM and was first discovered in 2000 [21]. It was widely regarded as a permanent modification before histone demethylase was first revealed in 2004 [19]. PMT-catalyzed histone methylation is the most essential and well-studied PTM because it is associated with an extensive variety of biological processes, containing transcriptional control, maintenance, and formation of heterochromatin, X-chromosome inactivation, DNA repair, and RNA maturation [7]. Many nonhistone proteins have been demonstrated to be targeted by PMTs [23], [24].

PMTs are a group of enzymes that help to enhance one or more methyl groups to add a diversity of substrates, containing protein, DNA, RNA, and small molecules. Due to their enzymatic action on histone molecules and their capability to modify the transcriptional regulation, a subset of proteins that act on the side chains of lysine and arginine are classified as the target for epigenetics. Methyltransferase is one of the most common types of epigenetic enzymes, and it is an active research topic in the industry as well as academia. Recently, several putative epigenetic enzymes have been discovered and the catalytic activity and three-dimensional structures of some of them have been determined [25], [26].

In humans, PMTs are one of them, which help to catalyze the methylation of lysine or arginine residues in proteins. The largest writer enzyme class that operates on protein substrates is found in PMTs [1]. It is generally understood that epigenetic target dysregulation can result in pathogenic alterations that contribute to human disease [27]. To discover and design a new drug, PMTs are of great significance. The role of these enzymes is essential to controlling the regulation of gene, and a large amount of biological and biochemical data has suggested that the activity of these enzymes are associated with cancer progression [28], [29]. Several human PKMTs and PRMTs are closely related to various human cancers. Several studies indicated that in addition to cancer, both the PKMTs and PRMTs are associated with several other human diseases [30], [31]. As a result, PKMTs and PRMTs have become exciting targets for drug discovery, so the identification of PMT inhibitors has been progressively sought in the past era.

1.4.1 PMT mechanism and specificity

To understand the structure of PMTs, the mechanism of these enzymes needs to be understood. All PMTs need two enzyme catalytic substrates, S-5-adenosyl-L-methionine (SAM) and the target methylated residues. PMTs catalyze the transfer of methyl groups from the SAM to its substrate protein. SAM binds to protein and produces S-adenosyl homocysteine (SAH) by transferring the methyl group to the arginine or lysine residue of the protein substrate. The SN2-based reaction mechanism is utilized in the transfer of the methyl group (Figure 1.2A). The PKMT enzyme used a hydrophobic narrow channel in the SET domain to bind with lysine resides and use carbon oxygen-hydrogen bonds to locate the N-terminal α-amino acid [32]. In the process of deprotonation incidence of the PKMT enzyme, water molecules play a significant role [33]. PRMT-substrate enzyme complex does not have a preserved water molecule for the process of deprotonation on the side chain of arginine residue [34].

Lysine can covalently connect with unmodified, one, two, or three methyl groups in four different methylation states. Monomethylated, dimethylated, and unmodified arginine are all possible methylation states. When arginine residues are dimethylated, the geometry can be symmetric or asymmetric (Figure 1.2B). Single or different enzymes can take parts in addition to more methyl groups to the targeted residues of protein [22]. The methylated residues of arginine or lysine do not change the charge of the residues but alter the protein hydrophobicity and bulkiness, therefore they affect the recognition of methylated protein through protein-protein interactions by methyl-lysine/arginine readers. The reader proteins recognize the mark of lysine or arginine methylation through a specific signal generated by it [6]. Based on the methylation activity of targeted residues, PMTs can be mainly divided into two classes i.e.

protein arginine methyltransferases (PRMTs) and protein lysine methyltransferases (PKMTs) [21]. Although PMTs have similar sequence and structural properties to catalytic domains of arginine or lysine residues, their sequence and structure differ significantly. Some PMTs, such as the SMYD protein family, are globular proteins, while others have numerous different domains [4].

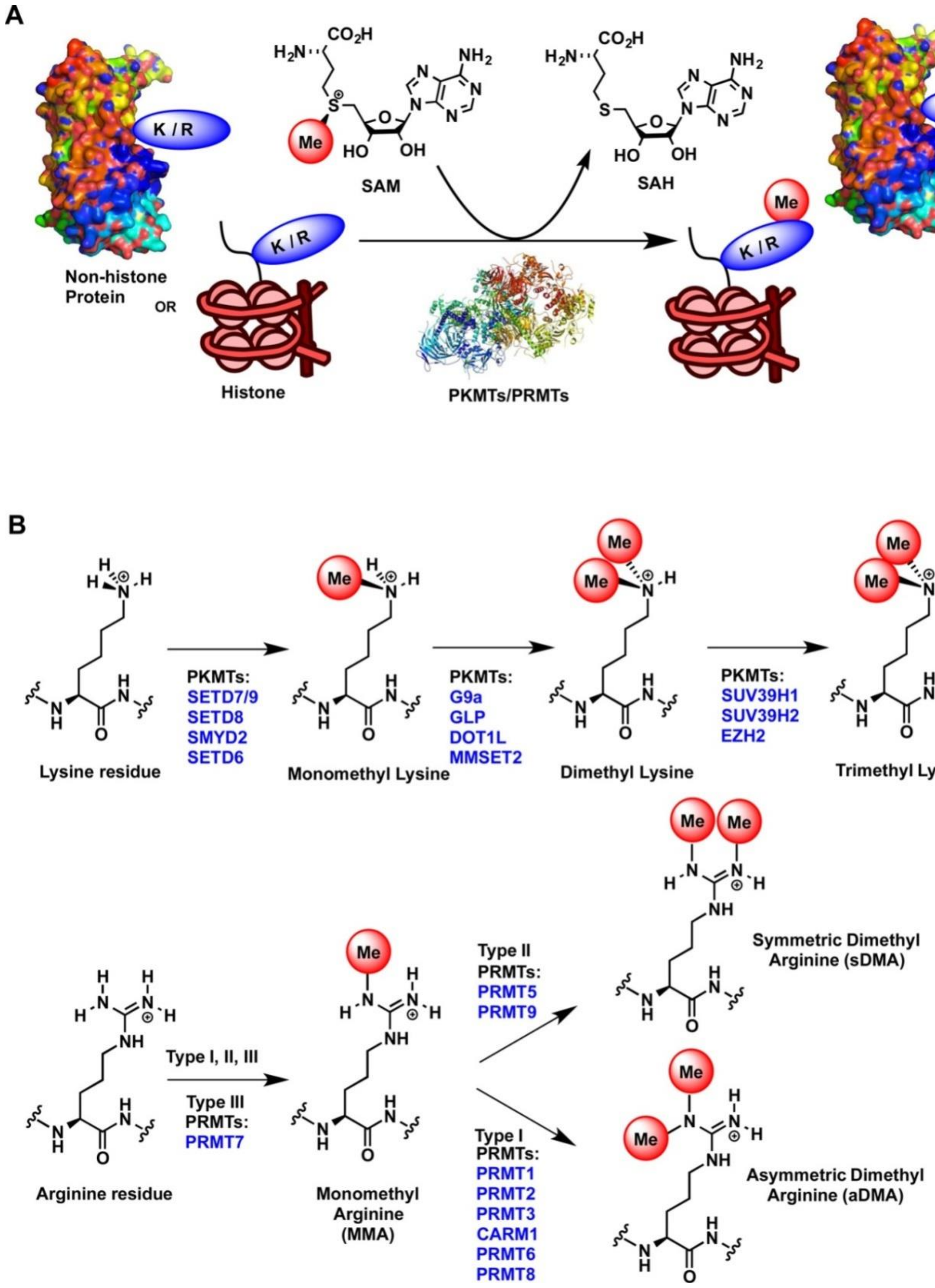


Figure 1.2: (A) Methylation mechanism of nonhistone and histone proteins by PRMTs and PKMTs (B) Different states of methylation in the residues of lysine and arginine. (Adopted from [21]).

1.5 Protein arginine methyltransferases

PRMT enzymes are engaged in an extensive variety of cellular communication processes in the eukaryotes. Along with the acetylation, phosphorylation, and glycosylation, methylation of protein arginine is a crucial and well-studied post-translational modification (PTM) [33],[34]. Arginine establishes five possible hydrogen bonds through their adjacent hydrogen bond acceptors with the help of the guanidinium group. Thus, each methylated arginine avoids forming a possible hydrogen bond, increasing hydrophobicity, and creating steric bulkiness [37]. The methylated arginine residues are very crucial in the recognition of protein and to make changes in its physiological functions [38]. The methylation of PRMT affects various biological processes such as transcriptional regulation, DNA repair, signal transduction [38], viral infection [39], chromatin remodeling [40], and differentiation of neuronal cells [41]. The molecular mechanism of these events confirms that methylated residues of arginine have the capacity to boot or interrupt the convinced interactions such as protein-DNA, protein-RNA, and protein-protein interactions. Of the three domains of PRMT proteins, one is similar to the Rossmann fold called the methyltransferase domain (MTase domain) which contains all residues required for SAM binding. The structure of this domain remains conserved. The second domain is the β-barrel of PRMT enzymes, while the third is the dimerization domain [42]. The substrate peptide binding pocket is found at the edge between the β-barrel and MTase domain [34], [43].

1.5.1 PRMT classes and activities

In humans, there is a total of eleven PRMT isoforms named PRMT1-11 (Figure 1.3) have been discovered and their orthologs are present in *Drosophila melanogaster, Caenorhabditis elegans*, protozoa, yeast, fish, and plants. Based on the methylation activity, PRMTs are categorized into three subcategories such as PRMT I, II, and III [44]. PRMT type I, contains PRMT1, PRMT2, PRMT3, and PRMT4, (also known as CARM1 (coactivator-associated arginine methyltransferase 1)), PRMT6, and PRMT8 that help to catalyze the arginine residues as asymmetric dimethylation and monomethylation. The type II PRMTs cover the PRMT5 and PRMT9 which helps to catalyze the arginine residues via monomethylation and symmetrical dimethylation [45]. In type III PRMT, the only family present is PRMT7 and it catalyzes only the arginine monomethylation. Two putative PRMT genes such as PRMT10 and PRMT11 have also been identified, but they have not shown any methylation activity yet [44], [45]. The diversity of these enzymes demonstrates the biological significance of arginine methylation in both the animal and plant kingdoms. Studies reported that the methylated protein arginines have been associated with carcinogenesis [44], multiple sclerosis [48], viral pathogenesis [49], spinal muscular atrophy [50], cardiovascular disease [51], and lupus [38]. Hence, the interpretation of the mechanism and regulation of these enzymes will demonstrate valuable.

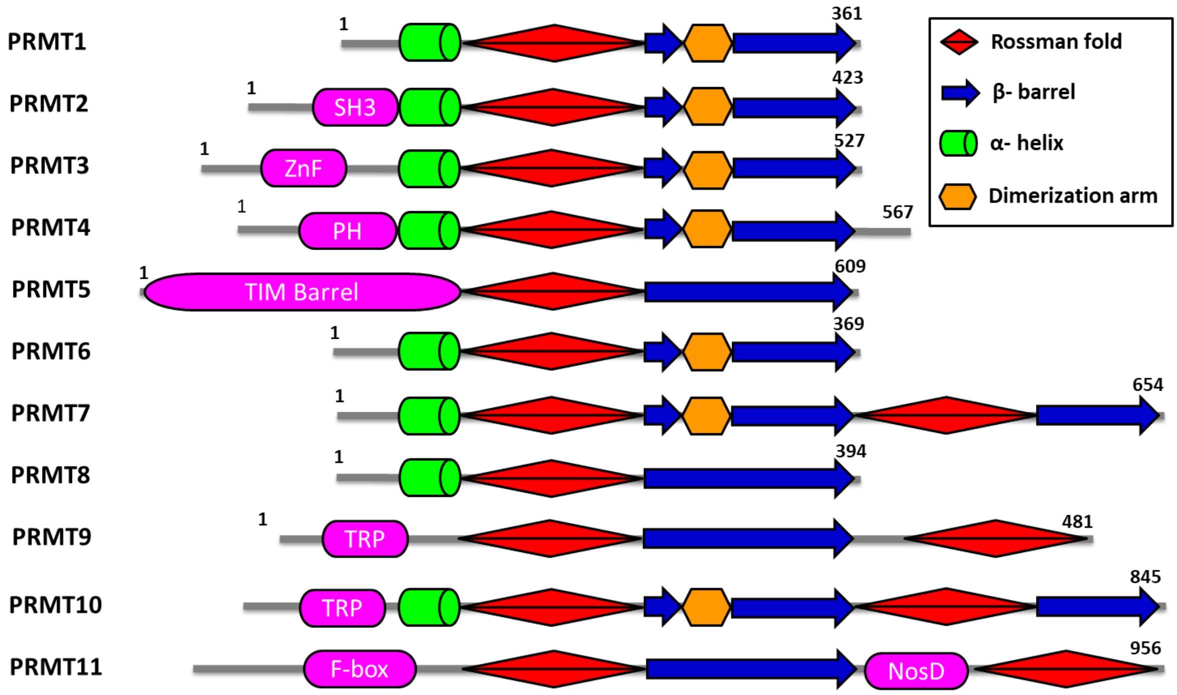


Figure 1.3: The schematic diagram for domains present in human PRMTs.

PRMTs are associated with the transfer of methyl group from AdoMet/SAM to a protein substrate of a positively charged arginine residue in a protein substrate, and AdoHcy/SAH molecule produced in the process. PRMT type I (PRMT1, 3, 4, 6, and 8) help in the catalysis of the production of asymmetric dimethylarginine (ADMA) and monomethyl arginine (MMA) by transferring a methyl group from SAM onto protein arginine residues, while PRMT type II enzymes (PRMT5 and 7) produce MMA and symmetric dimethylarginine (SDMA) (Figure 2). Whether a protein is dimethylated asymmetrically or symmetrically has opposing physiologic implications [44]. All PRMT proteins have 310 amino acids with a conserved core region. Basically, they have N-terminal additions, whereas CARM1 has C-terminal additions [52]. The PRMT monomeric structure has an MTase domain, a unique β barrel, and a dimerization arm. PRMT type I has a homodimeric structure from head to tail. The dimerization arm of one monomeric subunit that extends out to the β-barrel interacts with another Rossman fold subunit [50], [51]. Generally, PRMTs methylate the Glycine and arginine-rich (GAR) motifs with the exemption of CARM1, which helps to methylate glycine-, methionine-rich (PGM), and proline- motifs [54], [55]. PRMT5 can also symmetrically dimethylated both of these motifs [302]. Both nonhistone and histone proteins can be methylated by the PRMTs [28], [35], [38], [56]. Cancer and other disorders have been linked to PRMT and dysregulation and arginine methylation [35], [44].

1.5.2 Biological functions regulated by PRMTs

Gene transcription is regulated by histone proteins through a variety of PTMs, containing lysine methylation, acetylation, phosphorylation, SUMOylation, and ubiquitination [57], [58]. Previous PRMT research has focused on their epigenetic effects. PRMTs help to make methylated arginine on nucleosomes, and the methylated arginine assists as important epigenetic markers [59], [60]. The epigenetic reader proteins are used to recognize the epigenetic modifications that lead to the staffing of repressing or activating the machinery for the transcriptional event. Besides histone proteins, PRMTs are involved in the methylation of several other proteins associated with the transcription process including coactivators, corepressors, and transcription factors (Figure 1.4) [61]. Therefore, PRMTs also take part in specific the transcription regulation process. Several RNA-binding proteins (RBPs) possess RGG/RG-rich protein motifs which have been recognized as typical consensus sequences of PRMTs [62]. Proteomic and theoretical insights analysis of shown that PRMTs methylate various RBPs and these alterations are crucial for translation processes, RNA localization, and mRNA splicing [63]. In addition to the regulation of gene expression, PRMTs are engaged with a collection of physiological processes, for example cell signaling, cell cycle regulation, and DNA damage response [64].

1.5.3 Association of PRMTs in cancer

The PRMT family is one of the important groups of PTM enzymes, and it has a different set of substrates, such as tumor suppressor genes, oncogenes, and histones. Additionally, PRMTs have shown abnormal expression in various human cancers hence these enzymes are considered significant therapeutic targets [62]. In the PRMT family, PRMT1 is one of the crucial enzymes that contribute nearly 90% of the total activity of arginine methylation in mammalian cells [65].

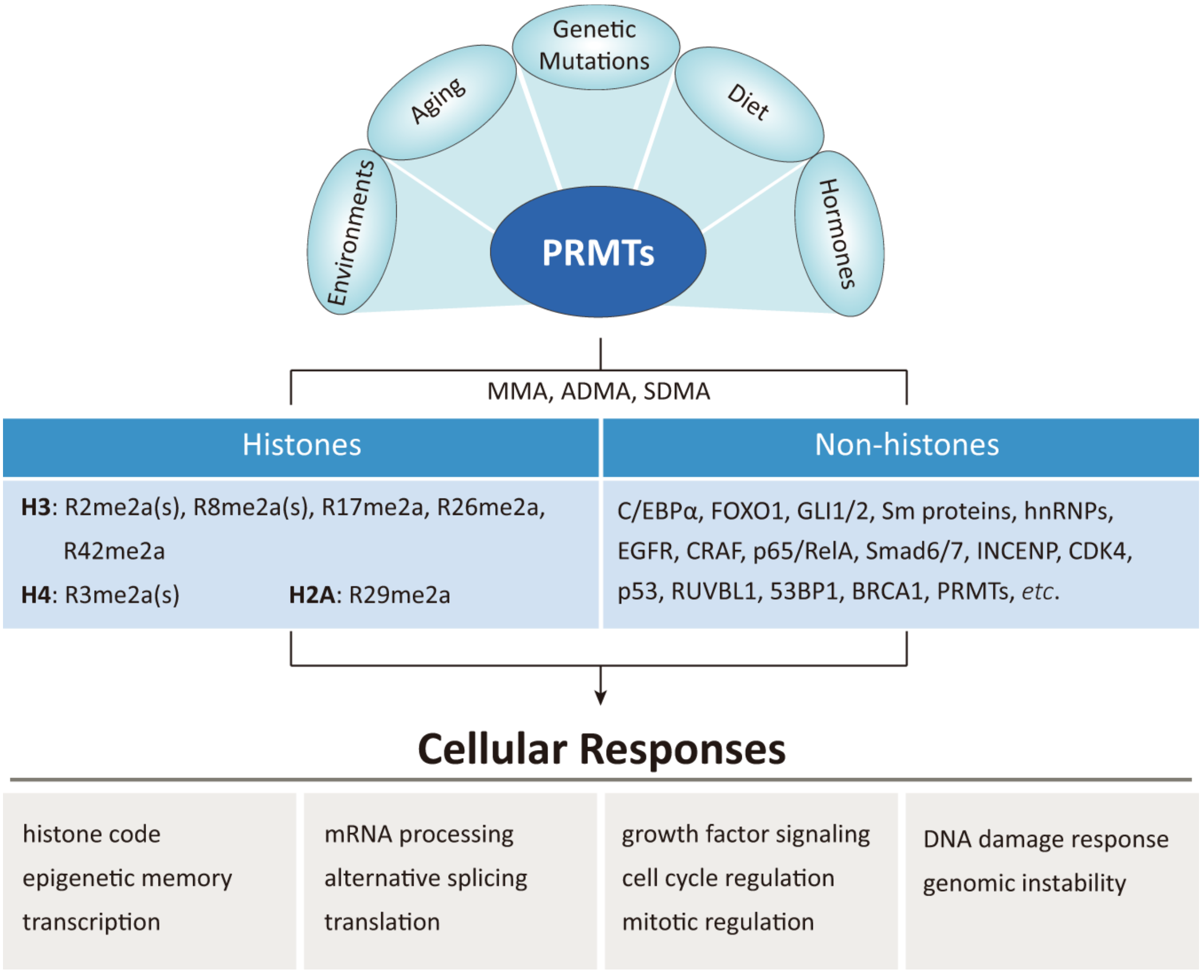


Figure 1.4: Histones and nonhistone proteins are methylated by PRMTs, which play a role in a variety of cellular responses and assist to maintain the cellular homeostasis in various biological systems. The developmental and pathogenic processes, various environmental factors, and genetic mutations are intricate in the regulation of PRMTs expression and activity. (Adopted from [62]).

As shown in Table 1.1, Hwang et al. summarised the dysregulation of PRMT1 expression and its relationship in diverse human carcinomas [62]. The role of PRMT2 in cancer is still up for debate. The oncogenic association of PRMT2 with glioblastoma has been reported in previous studies. Its increased gene expression is associated with tumor grade. Several studies have illustrated that the level of CARM1 is prominent in androgen-resistant prostate cancer [66], [67] in addition to violent breast tumors [68]. CARM1 is overexpressed in colorectal cancer but not in breast or prostate cancer, according to another study [69]. It is a well-known positive ER-regulator that is required for the production of estrogen-response genes by methylating arginine at histone H3 promoters [70]. It has been reported that PRMT5 has shown an important role in carcinogenesis and evolving as the most favorable target for various blood and solid cancers. In several cancer types the dysregulation or higher expression of PRMT5 has been detected [77,78,79,80,81,82,90]. PRMT5 affects the repression of gene expression by direct modification of histone H3 and H4 [73].

Table 1.1: Role of PRMTs and their association in cancer progression.

PRMTs	Expression	Function	Associated cancer	Reference
PRMT1	High	Oncogenic	Breast, Pancreatic, Colorectal, Lung, Hepatocarcinoma, Melanoma, Head and neck cancer, Esophageal squamous-cell carcinoma	[74]–[76]
PRMT2	High	Oncogenic	Breast and glioblastoma cancer	[72], [73]
	Low	Tumor suppressive	Breast cancer	[79]
PRMT3	High	Oncogenic	Pancreatic cancer	[80]
CARM1	High	Oncogenic	Breast, Colorectal, Ovarian, Acute myeloid leukemia cancer	[68], [81]–[85]
	High	Tumor suppressive	Breast,	[86]
	Low	Tumor suppressive	Pancreatic, Hepatocarcinoma	[87]
PRMT5	High	Oncogenic	Lymphoma, Leukemia/lymphoma, Diffuse large B-cell lymphoma, Acute myeloid leukemia, Breast, Lung, Prostate, Gastric, Hepatocarcinoma, Pancreatic, Colorectal, Melanoma, Glioblastoma, and Bladder cancer	[70], [75], [83]–[86], [86]–[96]
PRMT6	High	Oncogenic	Gastric, Endometrial, and Lung cancer	[102]–[104]
	Low	Tumor suppressive	Hepatocarcinoma	[105]
PRMT7	High	Oncogenic	Breast, Lung Non-small cell lung carcinoma, and Renal cell carcinoma	[106], [107]
PRMT9	High	Oncogenic	Hepatocarcinoma	[108]

PRMT5 expression of cancer-specific miRNAs linked to epigenetic control is linked to tumor growth, progression, and metastasis. Through methylation of arginine residues in several tumor suppressors and oncoproteins, PRMT5 plays a role in carcinogenesis [62]. In general, aberrant methylation of histone and non-histone substrates caused by PRMTs' methyltransferase activity results in altered chromatin epigenetic modification, which affects the expression of tumor suppressor genes or oncogenes and promotes cancer (Table 1.1).

1.6 Protein lysine methyltransferases

The protein from PKMTs group contain a conserved SET domain, that have nearly 130 amino acids [109], [110]. Initially, the SET domain was first discovered in three genes of drosophila, such as En(zeste) (the enhancer of zeste), Su(var)3-9 (the suppressor of variegation 3-9), and Trithorax [110]. The human genome contains about 60 PKMTs, which are divided into two groups: SET and without SET domain-containing PKMTs of which DOT1L is one of the important members. The SET domain-containing PKMTs are accounting more than 90% of the total PKMTs. The SET domain folds into a knot-like structure that passes together with the SET domain's two extremely conserved motifs and establishes an active site close to the binding pocket of SAM [111]. The catalytic SET domain is well-defined by certain amino acid patterns (RFINHxCxPN and ELxF/YDY, where x represents any amino acid) and a structure of pseudo-knot. SET-containing protein methyltransferases also have I-SET (Immunoglobulin-SET) and post-SET domains, which contain all catalytic residues. In many cases, these additional domains aid in the formation of the peptide-binding pocket as well as the S-adenosylmethionine (SAM) binding site [112]. SUV, SET1, SET2, EZ, AND RIZ are five major families of SET domain-containing PKMTs that are classified based on SET domain sequence similarity [113], [114]. Because these enzymes also target nonhistone proteins, new classification and nomenclature based on the kind of enzymatic activity and target residue(s) were proposed later [115]. As a result, these enzymes were divided into eight categories, starting with lysine methyltransferases 1 (KMT1) and ending with KMT8. The SET domain can be present in a significant variety of eukaryotic proteins as well as a few bacterial proteins. As a result, is not restricted to PKMTs [116]. PKMT-catalyzed lysine methylation has been recognized as a vital regulator of transcription through gene expression [25], [117]. Based on the methylation position and the nature of the methylation, histone lysine methylation can result in activation or repression of the transcription process. While H3K4, H3K36, and H3K79 methylation are linked to the activation of transcription, di- and trimethylation of H3K9, and trimethylation of H3K27 are linked to the repression of transcription [25], [57], [118].

1.6.1 PKMT classes and activities

Protein lysine methyltransferases help to methylate histones and non-histone proteins, which are divided into eight classes on the basis of the sequence and structure of PKMTs [119]. PKMTs are categorized into two groups, one is SET proteins, having distinct SET domains, and another is seven-strand (7BS) proteins, having a core shape with a seven-strand [120]. Further, the SET-domain-containing proteins clustered into seven families according to the similarity of sequence and domain organizations [121]. These seven families are named SET1, SET2, SUV3/9, SMYD, SUV4-20, RIZ, and EZ. SUV3/9 (G9a [122], GLP [123], SETDB1 [124]), SET1 (NSD1 [125]), SMYD (SMYD2 [10], SMYD3 [126]) and EZ (EZH2 [127]) members methylate both histone and nonhistone substrates. Two protein family such as SET8 and SET7/9 are involved in an extensive variety of protein methylation. Usually, the proteins containing the SET-domain target the flexible histone tails for the methylation of lysine residues. The amino group of lysine can able to receive three methyl groups, mono-, di-, or trimethyl lysine. The different states of methylations have different roles [128]. The majority of closely linked 7BS KMTs, focus on a group of proteins or single for methylation. The functional consequences of methylating 7BS KMT sites are unknown, and the link between biochemistry and biological function is sometimes difficult to recognize [129]. Out of the approximately 100 potential PKMTs encoded by the human genome, More than 60% PKMTs from 100 potential PKMTs are determined through the human genome, are associated with the activity of lysine methyltransferase in the variety of histone and nonhistone proteins [130].

1.6.2 Mechanism of protein lysine methylation

In the course of lysine methylation, enzymes add or remove methyl groups on the specific type of substrates [131], [132] (Figure 1.5). Over 50 PKMTs and 20 protein lysine demethylases (PKDMs) have been discovered so far [119]. Although most PKMTs have the SET domain, lysine N methyltransferase activity has also been shown in the proteins having non-SET domains, for example, methyltransferase-like 21A (METTL21), DOT1L, and METTL10 [133], [134]. The SET-domain PKMTs are anticipated to accelerate a sequential bi-bi kinetic process involving random substrate contact and product release [135]. The methylated lysine residue has increased hydrophobicity and basicity and these properties are helpful in the recognition of other methylated lysine proteins.

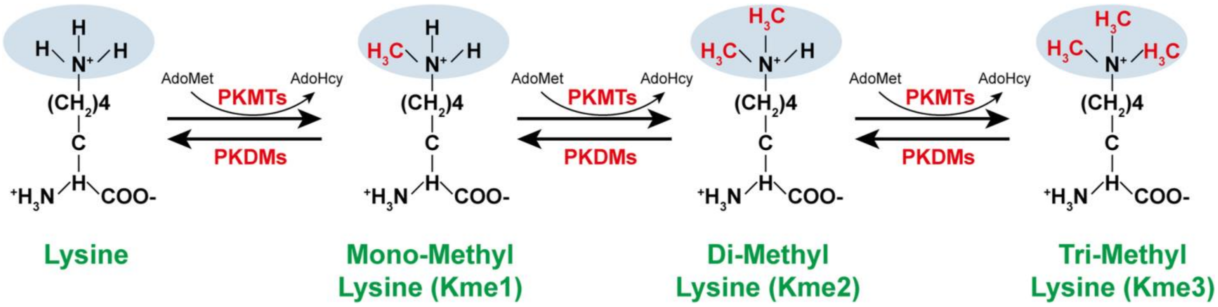


Figure 1.5: The schematic representation of methylation and demethylation of a lysine residue. S-adenosyl-L-methionine (AdoMet) is used by the PKMT as a universal methyl group donor to catalyze mono (Kme1), di (Kme2), and tri (Kme3) methylation on lysine amino acid. This alteration is reversible, and PKDMs can remove methyl groups. (Adopted from [10]).

1.6.3 Lysine demethylases

Until 2004, when Lys-specific demethylase 1 (LSD1, also known as KDM1A, AOF2, and BHC110), was found, lysine methylation was thought to be irreversible. In the human genome, only LSD2 (also known as KDM1B) shows the homology with LSD1. The amino oxidases can only take part in the mono and di lysine dimethylation not in trimethylation because it required lone pair of electrons. LSD1 and LSD2 only monomethyl and dimethyl lysine residues which belong to the first KDM family of flavin-dependent monoamine oxidases [136]. The second type of KDM is Jumonji C (JMJC) domain-containing proteins, which demethylate monomethylated, dimethylated, and trimethylated lysine residues using an oxygenase mechanism [137] [3].

1.6.4 Somatic cancer mutations in PKMTs

A rising number of researches suggest that inappropriate lysine methylation is linked to abnormal PKMT expression and plays a vital role in carcinogenesis. Current whole transcriptome and genome sequencing endeavors have also revealed numerous somatic mutations that are highly susceptible for cancer development [138]–[140]. Such type of mutations is also present in PKMTs, which are common in malignancies. These contain translocations of the chromosome that induce peculiar or mistargeted PKMT expression, as well as nonsense or frameshift mutations that render the protein useless. Missense mutations cause enzyme activity loss, but they can also affect the PKMTs properties due to changes occurring in the specificity of substrate or product and increased enzyme activity, and subsequent phenotypic gain-of-function occur. Mutations such as Loss-of-function contain frameshifts, critical point mutations, deletions, and nonsense mutations, which result in the inactivation of the affected protein. Instead, Gain-of-function mutations alter the function of the distressed protein, potentially contributing to cancer. Changing just one allele is usually adequate for gain-of-function mutations, but loss-of-function mutations usually disturb both alleles. Furthermore, by blocking a mutated cancer gene, they have the potential to assist as a specialized therapeutic strategy in personalized medicine [141], [142]. Somatic mutations present in PKMTs can also alter the interactions among the partner proteins or cause the PKMT to be recruited to particular genomic loci, which ultimately affect the expression of tumor suppressor genes or any specific oncogene. A more active PKMT may produce abnormal histone methylation at specific genes, which can either boost or inhibit oncogene or tumor suppressor gene transcription. The peptide sequence specificity of PKMT can be altered due to the mutations in the SET domain that leads to the methylation of unique targets. The molecular problems of somatic mutations present in PKMTs are generally unclear, and explanation would necessitate extensive biochemical research [141].

In addition to somatic mutations, the selection of germline variants may cause chromatin modulators to play a significant role in the development of disease and response to treatment. A single base variation in a DNA sequence that affects a sizable part of a population (≥1%) is known as a single nucleotide polymorphism (SNP) [143]. Non-synonymous SNPs (altering the amino acid sequence of a protein) are less common in genes known to influence disease [144]. SNPs can be linked to a disease's diagnosis and risk factors, but they can also influence a disease's course without directly causing it [145], [146]. Therefore, it is important to take into account how these polymorphisms in methylation-modifying enzymes affect methylation control, the genomic environment, and possible diseases. Numerous cancer forms have been associated with various histone mutations. Here we have briefly summarized about the computational analysis of SNPs in PKMTs.

1.6.4.1 Single nucleotide polymorphism

The majority of Single Nucleotide Polymorphisms (SNPs) are projected to be non-coding SNPs, according to the first comprehensive analyses of SNPs in the human gene coding region. Both synonymous (quiet, without changing the amino acid in the protein) and non-synonymous SNPs (nsSNPs) are equally prevalent in the coding area. When comparing two human genomes, the average diversity is one difference every 1,200 base pairs, but the typical gene has about four coding SNPs [147], [148]. A significant portion of nsSNPs are projected to influence the protein's structure, which would likely affect its function [149]. These variations, for instance, could have an impact on post-translational modifications, ligand binding, or protein stability or folding [149]. This raises the possibility that some coding SNPs have phenotypes. There aren't many researches linking KMT SNPs to disease or treatment outcomes [150]. Although it is now known that KDMs have a role in illnesses and that SNPs in these enzymes can be associated with disease, it is yet unknown how these SNPs influence the regulation or function of the enzymes.

Choosing which SNP to look at is one of the challenges of investigating SNPs and their potential implications in disease process and/or response to therapy. Models for the action of missense coding SNPs have been developed, providing a basis for understanding the impact of a coding SNP at the molecular level and, consequently, an approach of predicting which coding SNP are potentially involved in disease [149], [151]. This is because nsSNPs with coding repeats are more likely to affect gene function. The ability to find variations that raise the risk factor for a particular disease has risen due to the recent exponential growth in the amount of data collected by new sequencing technology. Due to availability of large number of SNPs, computational methods for the identification of potential candidate SNPs become more helpful.

Computational method was used to identify a SNP in PKMT's protein MLL, two most deleterious nsSNPs (Q1198P and K1203Q) were identified by applying various sequence and structure-based computational methods. Out of these two nsSNP, the mutation at amino acid position Q1198P with id rs1784246 found as significant mutation in acute leukemia caused by MLL gene [152]. Recently, Gautam *et al.* perform the computational analysis for SNPs of EZH2 to determine the structural and functional association with breast cancer susceptibility. Two EZH2 SNPs such as rs41277434 and rs201135441 (A490T) were explored computationally and validated experimentally on the population of in north Indian region of Punjob. The analysis revealed that SNP rs201135441 (A490T) has significant association with breast cancer susceptibility [153]. Such a computational analysis may be useful in locating significant SNPs for further research. The creation of individualised therapy would be much easier with this analysis. However, more research is required to conclusively connect these SNPs and any potential impact on protein function or stability in respect to disease.

However, while a number of recent studies uncovered the alteration of the genes encoding PKMTs, a lot of work is left to do in order to understand the mechanisms by which PKMTs are involved in the onset of diseases and response to treatments. The understanding of the importance and functional impact of SNPs in PKMTs is still in the nascent stages. Our literature survey showed that there is a wide choice of literature on SMYD2 gene associated with cancers through experimental studies but the computational analysis undertaken for an *in silico* investigations on the mutation of nsSNPs in SMYD2 gene are scarce. Improving our understanding of SNPs and mutations in PKMTs will help to assess risk and develop personalized medicine.

1.6.5 Functional role and disease implication of protein lysine methylation

The role of protein methyltransferases and demethylases in epigenetic control via histone methylation has been widely studied. Now, histone methylation is well recognized as a key regulator of chromatin functions, particularly transcriptional control (Figure 1.6). The majority of methylation sites reported in core histones so far have been located in histones H3 and H4. The histone mark present at the site of methylation is assumed to represent a unique type of function. Other histone lysine methylations such as H3K4 methylation is represented the active transcriptional mark, and monomethylation of H3K4 is prevalent at the site for enhancers [154]. In addition to actively transcribed genes, promoter and enhancer regions contain dimethylated histone H3K4 (H3K4me2). Histone H3K4trimethylation (H3K4me3) is a renowned characteristic of actively transcribed genes' promoter regions [155]. H3K27 dimethylation and trimethylation (H3K27me2 me3), which are linked to transcriptional regulation, are common in the target polycomb genes [156]. Monomethylation and dimethylation of H4K20 are essential for DNA damage repair and DNA replication, while trimethylation of H4K20 is associated with decreased heterochromatic regions [157]. Non-histone protein methylation is significantly associated with the modulation of numerous signaling cascades, according to the research. Lysine methylation is linked to the functional control of two important tumor suppressor proteins, such as RB1 and p53 [3]. Lysine methylated non-histone proteins provides five key functions through post-translation modification (PTM): It modulates I phosphorylation; (ii) protein-protein interactions; (iii) substrate protein stability; (iv) substrate subcellular localization; and (v) substrate protein promoter binding affinity (Figure 1.6) [9]. Due to these features, non-histone proteins with methylated lysine are associated with a number of cellular functions.

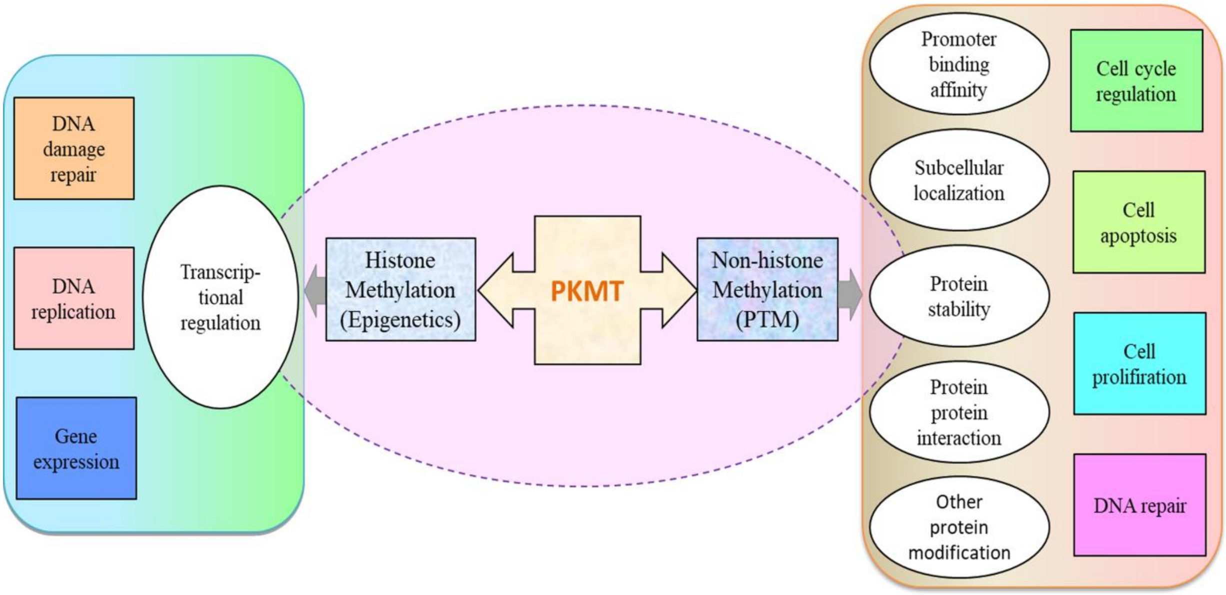

DNA damage repair
DNA replication
Gene expression
Transcrip-tional regulation
Histone Methylation (Epigenetics)
PKMT
Non-histone Methylation (PTM)
Promoter binding affinity
Subcellular localization
Protein stability
Protein protein interaction
Other protein modification
Cell cycle regulation
Cell apoptosis
Cell prolifiration
DNA repair

Figure 1.6: The biological processes regulated by protein lysine methyltransferases in two ways. One is non-histone targets methylation as of the post-translation modification (PTM). The other is the transcriptional regulation as epigenetics through the methylation of histone proteins.

Cancer is widely thought to be a hereditary disease; however epigenetic alterations have just recently been discovered to play an important role in critical stages of malignant transformation and development. Lysine residues methylated with Mono-, di-, and/or tri-methylated on both histone and nonhistone substrates are modified by PKMTs. These chromatin modifiers regulate post-translational changes, protein stability, protein-protein interactions, and non-histone substrate subcompartment cellular location, as well as the transcription of certain downstream target genes [3]. A growing amount of evidence implies that PKMTs are involved in the PTM of histone and non-histone proteins, as well as cancer progression and development [9]. Several forms of cancer commonly have dysregulation of this alteration (Table 1.2).

Table 1.2: Associated cancer types with the protein families of PKMTs.

Protein family	**Protein name**	**Associated cancer**	**References**
SET-domain containing proteins (SETD)	SETD8	Bladder cancer, chronic myelogenous leukemia, hepatocellular carcinoma, prostate cancer, non-small-cell lung carcinoma, small-cell lung carcinoma	[23], [158]
	SETD7	Breast cancer and multiple myeloma	[24], [159]–[164]

	SETD1A	Bladder, breast, colorectal, and lung cancer, hepatocellular carcinoma, and renal cell carcinoma	[165], [166]
SET and MYND domain-containing proteins (SMYD)	SMYD3	Breast, cervical, colorectal, esophageal, gastric, lung, medullary thyroid, pancreatic, and prostate cancer, cholangiocarcinoma, and hepatocellular carcinoma	[133], [167]–[173]
	SMYD2	Breast, bladder, cervical, esophageal, colorectal, head and neck, lymphoma, ovarian, and pancreatic cancer, renal cell carcinoma, and hepatocellular carcinoma	[13], [14], [135], [174]–[179]
Nuclear receptor-binding SET-domain proteins (NSD)	WHSC1L1	Acute myeloid leukemia, NUT midline carcinoma, small-cell lung carcinoma, lymphoma, bladder, and breast cancer	[180]–[182]
	WHSC1	Cholangiocarcinoma, chronic myelogenous leukemia, non-small-cell lung carcinoma, small-cell lung carcinoma, osteosarcoma, renal cell carcinoma, hepatocellular carcinoma, multiple myeloma, breast, prostate, bladder, and esophageal cancer	[183]–[185]
	NSD1	Acute myeloid leukemia, glioblastoma, lung cancer, multiple myeloma	[186], [187]
Polycomb complex	EZH2	Acute myeloid leukemia, cholangiocarcinoma, chronic myelogenous leukemia, glioblastoma, lymphoma, non-small-cell lung carcinoma, small-cell lung carcinoma, T-cell acute lymphoblastic leukemia, osteosarcoma, renal cell carcinoma, bladder, breast, colorectal, and esophageal cancer	[188]–[191]
Euchromatic histone-lysine N-methyl-transferase	EHMT2	Acute myeloid leukemia, cholangiocarcinoma, chronic myelogenous leukemia, esophageal, bladder, and breast cancer	[192]–[194]
MLL family	MLL	Acute myeloid leukemia	[195]
	MLL2	Melanoma, mixed-lineage leukemia, breast, bladder, colorectal, and lung cancer	[196]
	MLL3	Glioblastoma, melanoma, mixed-lineage leukemia, pancreas, stomach, breast, and esophagus cancer	[197]–[199]
Suppressor of Variegation 3-9 Homolog	SUV39H2	Acute lymphoblastic leukemia, bladder, cervical, and esophageal cancer	[200], [201]
DOT1-like histone H3K79 methyltransferase	DOT1L	Mixed-lineage leukemia	[202], [203]

According to The Cancer Genome Atlas (TCGA), PKMTs have been discovered to have frequent expression and genetic alterations in a range of malignancies [204], [205]. Preclinical studies have revealed some of these enzymes' modes of action, opening the path for the improvement of PKMT-specific inhibitors for cancer prevention. Analysis of expression and genetic alterations of PKMTs in multiple types of cancer have been performed on transcriptomics and genomics level. The datasets for these analyses are available on various public databases, but there is no systematic analysis presented on whether these genetic and expression alterations are replicated at the level of protein. Deregulation of PKMTs has been linked to a variety of cancers [128], [206]–[209]. The up-regulated SMYD2 expression was detected in bladder cancer cells [175], and esophageal squamous cell carcinoma [174]. The overexpression of SMYD3 is detected in breast carcinoma and plays a role in tumor proliferation whereas higher expression of G9a is present in hepatocellular carcinoma and plays a crucial role in the invasion of prostate and lung cancer [174], [210], [211], [212]. As a result, it has been demonstrated that methylated lysine can affect the carcinogenic pathways, which suggested the role of PKMTs in cancer progression. Cell proliferation is boosted by SMYD2 methylation of pRb, which is assumed to be owing to E2F transcriptional activity [175]. Similarly, SET7/9's modification of K372 prevents SMYD2 methyltransferase activity from activating the p53 pro-apoptotic function [14]. Consequently, these enzymes are considered potential cancer markers and therapeutic targets for cancer [209], [213]–[217]. Protein methylation dysregulation has been linked to a range of disorders, including cancer, and multiple articles [9], [62], [218], [219] have documented that abnormal states of PMTs and demethylases linked with human cancer, including aberrant expression and somatic mutations. PMTs-specific small molecular inhibitors are also being enthusiastically identified as a potential drug for cancer. Several PMT-specific inhibitors are currently under clinical trials [3].

1.6.6 Small-molecule Inhibition of PKMTs

Given their increasing functions as epigenetic modulators, PKMTs have gotten a lot of attention in the last decade. PKMTs serve as a multifunctional protein with a catalytic domain of methyltransferase and other motifs for interacting with a variety of binding partners [2], [21], [220], [221]. PKMTs-based small-molecule inhibitors can be produced based on the selectivity of methyltransferase activity [220], [222]. Small-molecule inhibitors, like other pharmacological agents, can disrupt PKMTs in a dose-dependent, temporal (precise time), and spatial (defined location) way [220], [222]. In the future, certain PKMT inhibitors could be employed as therapeutic treatments [21]. Many prior endeavors in industry and academia have been done to develop PKMT-specific inhibitors that resulted in hundreds of inhibitors for human PKMTs being available [21]. The quality of PKMT-specific inhibitors may differ from a small number of compounds that are well described *in vitro*, and *in vivo* to a large number of compounds that have only been examined in vitro biochemical assays [220]. It's difficult to choose relevant chemicals for biological investigations because many PKMT inhibitors haven't been thoroughly characterized. There's always the potential of misusing well-studied PKMT inhibitors and misinterpreting their biological outcomes [220].

As a result, it's critical to identify the criteria for high-quality PKMT inhibitors as well as the circumstances in which they should be used. Copeland et al. [223] found small chemical inhibitors affecting 11 PKMTs, with methods including SAM competition, complex disruption, peptide-site binding inhibition, and allosteric inhibition. They created a list of all available PKMT inhibitors, as well as information on their modes of action and clinical trials [223]. EHMT2, EZH2, SETD7, SETD8, SUV420H1/2, MLL, SMYD2, and SMYD3 are among the PKMTs for which small molecule inhibitors are available [21], [224]. Clinical trials with EZH2 inhibitors and Disruptor of Telomeric Silencing 1-like (DOT1L) are previously underway. Five EZH2 targeting drugs are currently in clinical trials, all with a solid tumor indication. In therapeutic trials for hematologic malignancies, one DOT1L targeted inhibitor is now being investigated [224].

1.6.7 The SMYD family of PKMTs

SMYD are a group of five soluble enzymes that methylate histone as well as non-histone substrates [4], [21], [169]. SMYD1–5 are the five members of the SMYD family that have been discovered so far (Figure 1.7). All members of the SMYD family have a conserved SET domain and MYND domain. The SET domain is a conserved catalytic unit that is responsible for the lysine methylation, that is present in approximately all histone methyltransferases (HMT) [225]. MYND domain has a zinc finger motif with proline-rich regions that is responsible for making protein-protein interactions [4], [226]. Three of the five members of the family have full-length crystal structures: SMYD1 [227], SMYD2 [228], and SMYD3 [229]. SET, I-SET, post-SET, and MYND domains are present in the N-terminal lobe, whereas the sequences having similarity with Tetratrico-peptide repeat (TPR) domains are present in the C-terminal region. The C-terminal domain (CTD) only present in the family of

SMYD1–4 but absent in SMYD5, is another distinguishing trait. The most structurally diverse compounds in the family have yet to be solved. SMYD4 is roughly twice as large as the other SMYD molecules and has extra TPR domains at the region of the N-terminus, whereas SMYD5 has a unique sequence on the C-terminal that is not related to the domains of the C-terminal. The intersection of the N- and C-terminal lobes generates a large, deep binding site for protein substrate. The sequence and orientation of the C-terminal lobes are responsible for the molecular structure, which makes the diverse surface topologies for peptide binding sites in SMYD1, 2, and 3 [4]. SMYD2 and SMYD3 are the most studied protein family from the SMYD group. SMYD1 [230] and SMYD3 [179] are responsible for trimethylate whereas SMYD2 is responsible for monomethylating the various residues of lysine on non-histone and histone proteins. SMYD2 was first discovered to methylate H3K36 [179], but later, it was revealed that it alters its specificity to H3K4 when it comes into contact with HSP90 [178]. According to the earlier study, SMYD2 also methylates K266 residue of estrogen receptor alpha [177], K810, and K860 of the retinoblastoma (RB) protein [13], [175], and K370 of p53 [13]. During cell-cycle development and differentiation, the methylation of Lys860 in RB is regulated [13], [175]. In conjunction with SMYD2, the structures of the p53 [231] and ER [232] peptides have been identified. For SMYD4 and SMYD5, structural data is currently unavailable. According to a recent study, methylated histone as well as non-histone substrates through SMYD proteins that are linked with various cellular functions for example cell regulation, including signal transduction, chromatin remodeling, transcription, and control of cell cycle.

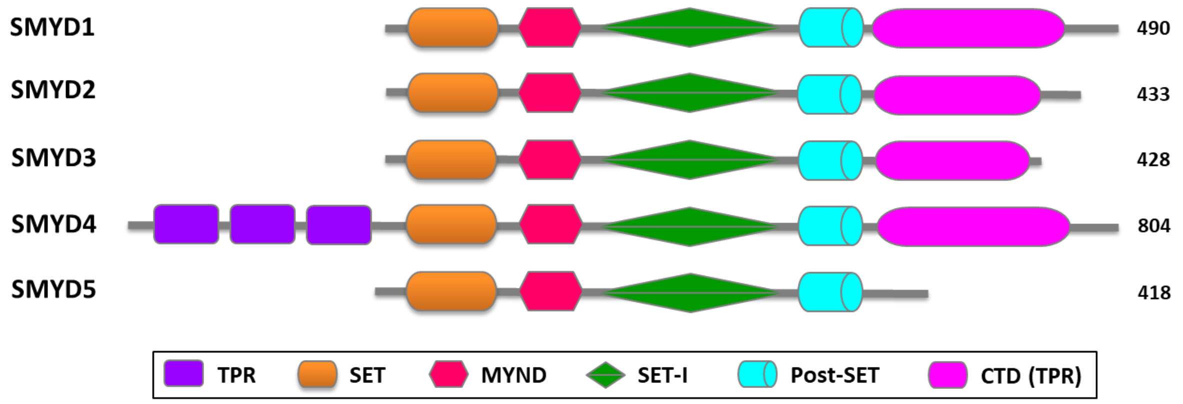


Figure 1.7: Schematic demonstration for a member of the SMYD family. The different colors represented the structural domains present in SMYD1-SMYD5 proteins. The number of amino acids present in each SMYD protein is denoted at the end.

1.6.8 SMYD proteins as a Drug Design Perspective

The research on SMYD proteins is intriguing since they are associated with a variety of cancer-related pathways. As a result, it opens up new possibilities for cancer and cardiovascular treatment. SMYD protein overexpression has been linked to nearly all cancer types [169], [174], [175], [233], [234]. Overexpression of SMYD1 inhibits the transcription of genes required for the production of ion channels in the heart, resulting in heart failure [102]. Overexpression of the SMYD1 gene has also been linked to hypoplastic left heart syndrome (HLHS), a condition typified by an immature left ventricle [235]. In ESCC or p53-related malignancies, SMYD2 is overexpressed, and knocking it down decreases tumor cell growth [174], [175]. SMYD3 is overexpressed in a range of malignancies, with colon, breast, prostate, pancreatic, and lung cancer [126]. SMYD3 overexpression is linked to a bad prognosis, and wiping it out stops tumor growth [133], [236]. As a result, therapeutic intervention with any of the SMYD proteins can be beneficial in cancer treatment. Efforts are now being made to develop SMYD inhibitors. The high-throughput chemical screening method was used to discover the SMYD2-specific competitive inhibitor AZ505 [237]. The lysine access channel in SMYD2 protein may not be the main target for therapeutic pharmacological action. SMYD2 is associated with multiple diverse processes of cellular function and is necessary for the methylation of a range of targets [232]. Complete SMYD2 knockdown could not be a practical choice because unselective inhibition of SMYD2 might create undesirable and potentially lethal side effects. One strategy for developing a therapeutic treatment that can selectively lower SMYD2 function in cancer is to target different binding sites that will affect the partial function of SMYD2. Substrate-binding pockets of SMYD are linked to the CTD orientations based on shape and size, due to that various protein members from SMYD family are considred for drug design. For potential drug design, it is essential to understand the conformational changes occur in SMYD proteins in addition to functional significance of every conformational state. Drug research efforts should not be restricted to the channel for specific lysine access, despite the fact that SMYD structures may be effective for therapeutic intervention. Alternative binding sites and conformations can be used to successfully cut down the malignant function of SMYD2 and SMYD3 proteins without distracting the general function of proteins from SMYD family. As a result, SMYD proteins are crucial for both functional and therapeutic purposes.

1.7 SMYD2 Protein

SMYD2 is a protein from the family of SMYD. All five members of this family (SMYD1–5) have a catalytic conserved SET domain and a zinc-finger MYND motif. SMYD2 is highly expressed in the heart and skeletal muscle [238], [239]. In cardiomyocytes, endogenous SMYD2 is expressed in both the nucleus and the cytoplasm [240]. SMYD2 is involved in the development of the heart and muscles [10], [11]. SMYD2 is involved in cellular proliferation and the development of transcriptional control. By methylating histones or interacting with RNA polymerases, SMYD2 controls gene transcription [178]. Many histone and non-histone proteins can be methylated by SMYD2 [176], [177], [241]. According to crystallographic investigations of SMYD2 structures, the broad substrate specificity of SMYD2 is achieved via a variety of processes, including various peptide-binding modalities and the inherent dynamics of peptide ligands [11], [242]. SMYD2 is highly expressed in variety of cancer types, and its role as an oncogene is being studied [9], [243], [244]. SMYD2 substrates were later identified as p53 and RB protein [11], and it was revealed that when SMYD2 interacts with the HSP90 protein, it methylates H3K4 rather than H3K36 [178].

1.7.1 Discovery and structure of SMYD2

In 2006, Brown and coworkers found Smyd2, a histone methyltransferase, in the 1q32.3 locus [179]. It was determined that a higher mRNA level of Smyd2 is present in the brain, liver, heart kidney, ovary, and thymus using northern blotting. The SMYD2 is located in both the cytoplasm and nucleus using immunohistochemical labeling [179]. The crystal structure of SMYD2 was determined in 2011 by two groups of researchers [178], [187]. SMYD2 has five structurally separate domains (Table 1.3) that come together to create two enormous lobs [178]. The N-terminal region of SMYD2 has a heterogeneous structure comprising helices, strands, and long extended loops, whereas the C-terminal region has a twisted seven-helical bundle [245]. S-sequence is important for the ideal enzymatic activity of SMYD2 but the post-SET domain is absolutely necessary [182]. According to metal studies, SMYD2 needs three bound zinc ions for their catalytic activity and structural integrity [182]. The C-terminal domain, which stabilizes the auto-inhibited SMYD2 conformation and blocks the substrates to enter the catalytic site by broad contact with the domain of methyltransferase, has been reported to regulate the auto-inhibition of SMYD2's methyltransferase activity [228]. The open-closed motion of SMYD2's bilobal shape may alter substrate specificity [188].

Table 1.3: Five different domains present in human SMYD2 protein and their position.

S.N.	Domain name	Position in protein
1	S-sequence	1-46 amino acid
2	MYND	47-96 amino acid
3	SET-I	97-243 amino acid
4	post-SET	244-271 amino acid
5	TPR	272-433 amino acid

1.7.1.1 Catalytic SET domain

Similar to other proteins of the SMYD family, SMYD2 has a "split" type of SET domain [165], [166], [167]. The SET domain is divided into two parts with the help of the MYND domain, one is S-sequence and another is the core SET domain. These two sections form a structure that is evolutionarily conserved in SET proteins, resulting in a fold. The SET domain is required for SMYD2's methyltransferase activity. The enzyme activity is reduced when residues from the cofactor binding site, lysine access channel, or binding sites of substate get mutated. The SET domain is surrounded in structure by the two domains such as insertion SET (SET-I) and post-SET. The binding of cofactors and substrates is also aided by these two domains [166], [172], [178]. In the SET domain, the SET-I is represented by a helix bundle. The post-SET domain consist cysteine-rich region folded downstream of the SET domain with one zinc atom. SMYD2 methyltransferase activity is abolished when the post-SET domain is deleted [241]. SMYD2-mediated methylation is dramatically reduced when a zinc-chelating residue (C264S) inside the post-SET domain is mutated [241]. This suggests that SMYD2 methyltransferase activity requires an intact post-SET domain.

1.7.1.2 Zinc-finger MYND domain

The MYND domain contains of a zinc-finger motif with two zinc atoms (Figure 2A). The MYND domain interacts directly with the SET domain, although it does not take part in the binding of the substrate. This is in line with the discovery that the MYND domain isn't required for SMYD2's histone methylation activity [178]. By binding to a proline-rich region, MYND domains facilitate particular protein interaction [4]. SMYD2's MYND domain interacts with EBP41L3. With a PXLXP motif, EBP41L3 [178] is a tumor suppressor. The EBP41L3 binding site on the MYND domain's surface is thought to be in a shallow groove [166]. Because the groove is entirely exposed, potential binding proteins can easily access it (Figure 2A). This suggests that the MYND domain is predominantly a component of protein-

protein interaction that manages SMYD2 protein with other proteins to control tumor development.

1.7.1.3 TPR-like carboxy-terminal domain

The CTD domain is made up of seven antiparallel helices and has a structure like helix-turn-helix. The structure is comparable to that of TPR, albeit the sequences are not identical. Because TPR motifs drive the building of multi-protein complexes [247], this structural similarity suggests that the CTD may operate as a protein-protein interaction module. According to current research, the CTD is critical for protein interaction [248]. The C-terminal and N-terminal regions of SMYD2 help to bind the sarcomeric protein [240]. The CTD binds a PEG molecule in a SMYD2 structure, implying that the CTD has a second peptide-binding site, potentially allowing the binding of two distinct proteins [232]. Except for SMYD5 [249], the CTD is substantially conserved in the family of SMYD proteins. Between SMYD1 and SMYD3, SMYD2 seems to represent a conformational intermediate [228], [232]. Furthermore, depending on which cofactor analogs attach to the protein, the CTD in SMYD2 can adopt distinct conformations [28]. These findings point to CTD domains' intra- and interdomain flexibility. The CTD in SMYD2 is required to make the binding pocket of the substrate and has been shown to stabilize p53 interactions [172], [178]. CTD, on the other hand, appears to have a substrate-dependent effect on SMYD2 activity. The methylation activity of SMYD2 on p53 proteins was drastically reduced when the CTD was deleted, although histone H3 methylation was unaffected [231]. CTD deletion boosted H3K4 activity but did not affect H3K36 or p53 methylation when peptides were utilized as substrates. However, replacing Tyr374 with alanine led to the methylation of the p53 peptide being lost [231]. Tyr374 is a substrate-binding residue that is found in the CTD (Figure 2B). This implies that the CTD is involved in substrate recognition.

1.7.2 Catalytic activity of SMYD2

AdoMet is used as a methyl donor by the SMYD2 enzyme, which catalyzes lysine methylation. AdoMet is converted to AdoHcy when the methyl group is transferred to a target molecule. SMYD2 utilizes a rapid-equilibrium random Bi-Bi mechanism in catalysis, according to steady-state kinetic studies [241]. H3 or p53 methylation peaks at alkaline pH (pH 9.0–10.0) [241], whereas Hsp90 methylation peaks at pH 7.5–8.0 [249]. The influence of pH on the catalytic efficiency of SMYD2 is mostly dictated by kcat. At different pH levels, both AdoMet and the substrate have identical KMs [241]. This suggests that solvent basicity may facilitate deprotonation of the -amine group in target lysine. SMYD2 activity is affected by ionic strength, the optimal activity is found when ionic strength gets lowered. SMYD2 activity peaks at 32°C and quickly drops at temperatures over 37°C [241]. The catalytic effectiveness of SMYD2 varies greatly depending on the substrate. When it comes to p53 peptides, SMYD2 has a 10-fold stronger activity than histone peptide substrates [231]. SMYD2 had a 3- to 6-fold greater action on p53 protein than on histone H3 or nucleosome substrates [231]. H4 and H2B are the most proficient substrates than H3, by 3 to 5-fold advanced activity [182]. Full-length H3 is a more potent substrate for SMYD2 than H3 peptides [172], [182]. SMYD2 has very little activity with H3K36 peptides, although it does have activity with H3K4 [178]. Substrate binding is linked to differences in SMYD2 activity. According to ITC research, SMYD2 binding to p53 peptides (residues 361–380) has a dissociation constant (Kd) of 20 M, whereas binding to H3K4 peptides (residues 1–20) is around 35-fold weaker [231]. These findings imply that in vitro, SMYD2 preferentially binds and methylates p53. Hsp90 increases the activity of SMYD2 methyltransferase [178]. This activity augmentation occurs on the H3K4 substrate but not on the H3K36 substrate [178]. Hsp90 also boosts SMYD1 and SMYD3's H3K4 methylation activity [133], [230]. The exact mechanism of Hsp90-induced activity augmentation, however, remains uncertain. Some Hsp90 co-chaperones (like Hop) influence the activity of SMYD2 on Hsp90. The presence of Hop induces a substantial decrease in methylation, whereas AHA1 and p50 have little effect on Hsp90 methylation [249]. Hop is bound by Hsp90 via the MEEVD motif on the region of C-terminal. The CTD domain of SMYD2 is thought to bind to this area [232].

1.7.3 Methylation targets of SMYD2

H3K4 and H3K36 have been found to be methylated by SMYD2 [178], [179], implying a role as a transcriptional activator in epigenetic gene regulation [178], [179]. SMYD2 help to methylate a variety of non-histone proteins, indicating that it has broad substrate specificity. Tumor suppressor protein such as p53 [14], RB [11], [145], cytoplasmic Hsp90 [249], [250], ER [177], and recently Poly(ADP-ribose) polymerase-1 (PARP1) [176] are some of the target for non-histone methylation. SMYD2 methylates p53 at lysine 370, reducing the activity of p53-mediated transactivation [251]. During the process of the cell cycle, differentiation, and DNA damage response, SMYD2 modulates RB activity by mono-methylating it at lysine 810 and 860 [11], [145]. The establishment of a complex with the sarcomeric protein titin by

SMYD2-mediated methylation of Hsp90 at lysine 615 affects myofilament structure [238], [250]. Under estrogen-depleted circumstances, ER methylation at lysine 266 by SMYD2 reduces ER chromatin recruitment and inhibits ER target gene activation [177]. In response to oxidative DNA damage, SMYD2 mono-methylates PARP1 at lysine 528 and controls poly-(ADP-ribosyl)-ation activity [176]. Due to observed methylation of tumor suppressor gene and their suppressive effects on cancers, these findings imply that SMYD2 may work as an oncogene.

1.7.4 Role of SMYD2 in cancer

Studying the association of SMYD2 in cancer has gotten a lot of attention recently. The most well-known tumor suppressor is p53, which was the first nonhistone protein substrate of SMYD2 [14]. Tumor suppressor gene p53 gets monomethylate by SMYD2 at Lys370, decreasing p21 and Mdm2 expression and allowing cancer cells to grow more easily. SMYD2 plays a vital role in a range of malignancies, including esophageal squamous cell carcinoma, breast cancer, gastric cancer, and leukemia according to a growing body of studies [9], [233], [243], [252]. Higher expression of SMYD2 was found in cell lines and tissues of breast cancer, knocking down SMYD2 in triple-negative breast cancer (TNBC). The reduced activities of SMYD2 with AZ505 inhibitor expressively reduce the growth of tumors in vivo, according to the study of Li et al. [252]. SMYD2 enhances the survival and cell proliferation of TNBC with the help of activation and methylation of the p65 and STAT3 subunit of NF-κB (Figure 1.8) [252]. Estrogen signaling is involved in differentiation and cell proliferation also it has been connected with lot of human diseases, including cancer for example ovarian and breast cancer [253], [254]. To govern gene stimulation or suppression in response to estrogen stimulation, the ligand-activated transcription factor ER binds multiple coregulators to estrogen response elements. The breast and ovarian tissues have high levels of ER [255]. The regulatory mechanisms that affect ER expression and activity must be discovered in order to understand human illnesses [256]. Zhang et al. discovered that MCF7 breast cancer cells, SMYD2 help to methylate ER at Lysine 266 to inhibit the activation of the ER target gene [177]. The structural analysis suggested that SMYD2 bind with ER in the conformation of U-shaped [232]. In the ER, SMYD2 increases the methylation of lysine 266 in ER by interacting with the chaperones HSP90/p23 molecular [257]. Furthermore, in the breast cancer cells, SMYD2 is used with Phosphatase and Tensin Homolog (PTEN) as a substrate. Lysine 313 in the *in vitro* and *in vivo* experiment, SMYD2 methylates lysine 313 of PTEN that decreasing the tumor suppressor function of PTEN and activate the pathway of phosphatidylinositol 3-kinase (PI3K)-AKT. PTEN takes part in the phosphorylation of Serine 380 that increased the knockdown of SMYD2 in cancer cell, although the phosphorylation of AKT gets reduced. These data show that SMYD2-mediated PTEN methylation at Lysine 313 lowers PTEN phosphorylation at Serine 380, resulting in AKT activation and breast cancer cell proliferation [258]. High-grade serous ovarian cancer is the most frequent type of ovarian cancer in female [259]. According to Kukita et al., higher expression level of SMYD2 was reported in clinical tissues of high-grade serous ovarian cancer (HGSOC) that suppress or inhibit the SMYD2 with LLY-507 and increase the cell death in apoptotic.

Furthermore, LLY-507 indicated a preservative impact with the olaparib inhibitor of PARP in the formation of colony assays. It suggested that LLY-507 can be utilized only or in blend with Olaparib to treat HGSOC patients [260]. Olaparib is a PARP1 inhibitor that was established for HGSOC of BRCA1/2 mutant [261], [262]. PARP1 is considered a crucial target for anticancer therapeutic research. The synergistic consequence of Olaparib and LLY-507 beside HGSOC could be linked to the involvement of SMYD2 in the methylation of PARP1. The studies reveal that higher expression of SMYD2 in the HGSOC and BC patients suggesting that SMYD2 could be employed as a significant biomarker for the diagnosis of such cancer types. Furthermore, SMYD2 inhibition may possibly consider an effective treatment for patients with HGSOC and BC. Though new SMYD2-specific inhibitors need to be developed, and clinical trials must be completed to approve their inhibitory mechanism.

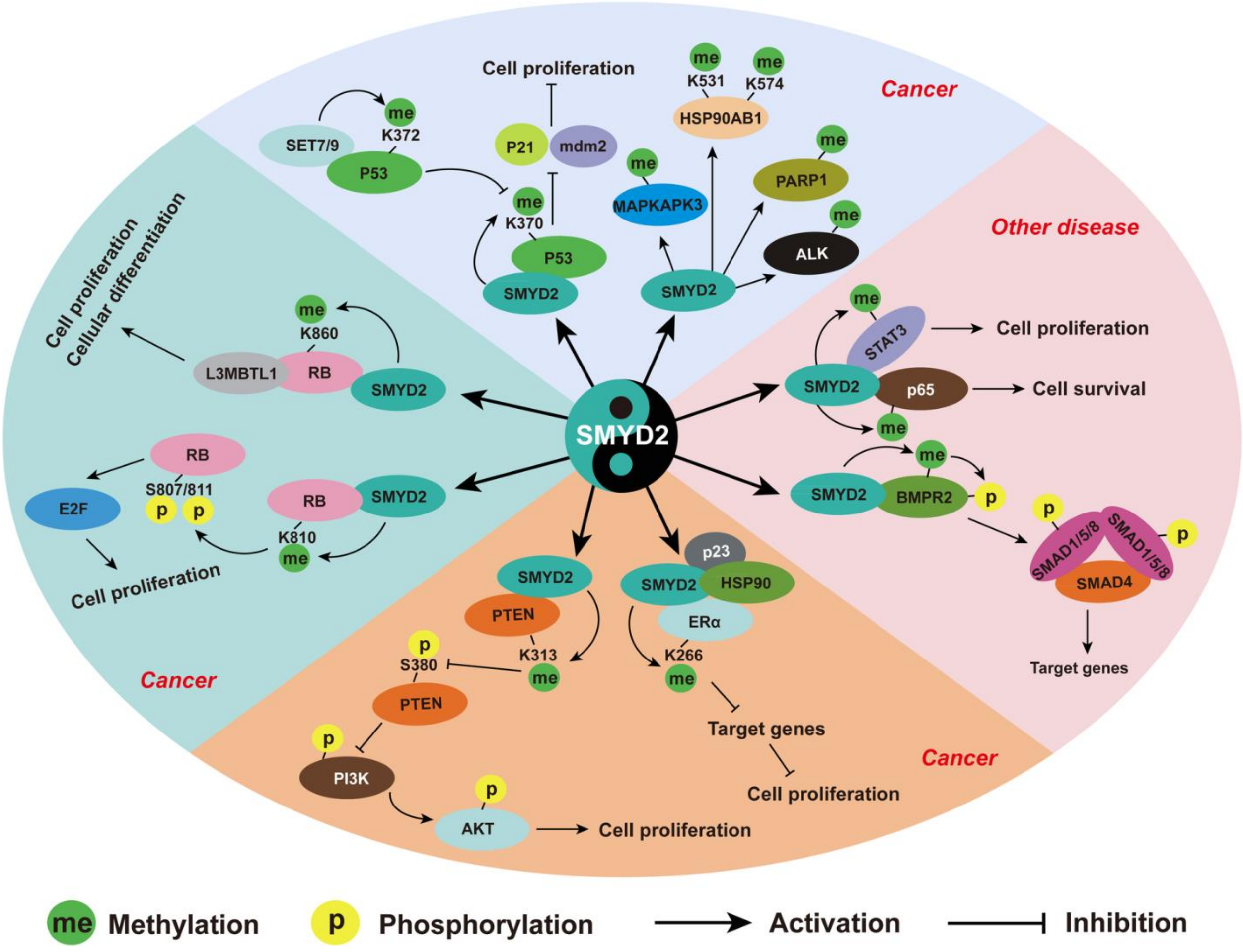


Figure 1.8: The SMYD2 targeted nonhistone substrates for methylation that associated with cancer and other diseases. (Adopted from [10]).

SMYD2 help to methylate RB, at Lysine 860 that establishes a binding site for the L3MBTL1 transcriptional repressor. These proteins are involved in the regulation of cellular differentiation, DNA damage response, and cell cycle progression [13]. SMYD2 methylates Lysine 810 of RB in addition to Lysine 860, promoting phosphorylation of Serine 807/811 of RB and mediating SMYD2's role in cell proliferation and bladder cancer [175]. Methylated RB also promotes cell cycle advancement by increasing E2F transcriptional activity (Figure 1.8). Most importantly, SMYD2 expression in human bladder carcinoma tissues is substantially higher than in non-neoplastic bladder tissues, implying that SMYD2 inhibitors could be employed to treat bladder cancer [175]. SMYD2 mRNA levels were found to be higher in renal cell tumors as compared to their normal counterpart, and levels of SMYD2 mRNA were found to differentiate between tumors and normal tissues of renal cells with the specificity of 100% and 82.1% sensitivity, as well as specificity with 73.3% and sensitivity with 71.0 in chromophobe subtype renal cell carcinoma (chRCC) from oncocytoma [263]. Acute lymphoblastic leukemia (ALL) is the more frequently causing cancer among children. Despite the fact that chimeric antigen receptor (CAR) T-cell therapy is the most standard treatment for a few types of leukemia, chemoradiotherapy remains the primary therapeutic option because the clinical trials for CAR T-cell therapy do not complete yet. Patients who have high SMYD2 expression have a bad prognosis, whereas those who have lower expression of SMYD2 are much more willing to chemotherapy [233]. The fusion oncogene MLL-AF9 promotes the development of leukemia, which is inhibited by SMYD2 ablation [264]. The transcription factor MYC directly activates SMYD2, and knocking down SMYD2 prevents the fusion oncogene MLL-AF9 from causing leukemia. Furthermore, in human acute myeloid leukemia cells, SMYD2 knockdown results in relative tolerance to a variety of agents that are responsible for DNA damage, which is accompanied by higher expression of SET7/9 [265]. The higher expression of SMYD2 and SMYD3 was also observed in patients with chronic lymphocytic leukemia, which is considered by the complex karyotype and higher count of white blood cells [266].

Higher expression of SMYD2 protein was observed in the esophageal squamous cell carcinoma (ESCC) in the sample of the primary tumor. The ESCC patients with higher amounts of SMYD2 expression represent a lower rate of overall survival as compared to the patient having low express SMYD2 [174]. The knockdown of SMYD2 substantially decreases cell growth of ESCC (KYSE790 and KYSE150) [174]. The overexpression of SMYD2 is noticed in 60–70% of cases in a patient with head and neck squamous cell carcinoma (HNSCC). As compared to patients with lower expression of SMYD2, the patients with higher expression of SMYD2 showed a poor rate of overall survival [186]. SMYD2 levels are high in pancreatic ductal adenocarcinoma (PDAC) [267], hepatocellular carcinoma [268], Gastric cancer [243], colon cancer [269], and papillary thyroid carcinoma [270]. Higher expression of the SMYD2 protein has also been linked to larger tumors and has a poorer overall survival rate [268], [270]. To achieve these effects, SMYD2 methylates a range of nonhistone proteins in cancers, including ALK in non-small-cell lung cancer (NSCLC) and MAPKAPK3 in PDAC [267], [271]. HSP90AB1 gets methylated by SMYD2 at lysines 531 and 574 to promote the process of cancer cell proliferation [135]. Even though many studies have been published on the involvement of SMYD2 in cancer, the bulk of these studies has used publically accessible data to examine the SMYD2 expression pattern in tumors. The significant association of SMYD2 in cancer and cardiovascular disease has been widely studied, but the impact of SMYD2 on other diseases is still unclear. According to Li et al., a higher expression level of SMYD2 was found in kidney tissues and renal epithelial cells in patients with autosomal dominant polycystic kidney disease (ADPKD) and Pkd1 mutant mice [272]. SMYD2 has recently been discovered to modulate the bone morphogenic protein (BMP) signaling pathway [273]. Phosphorylation of SMAD1/5 induced by BMP, nuclear localization, and association with SMAD4 are triggered by SMYD2 methylating the kinase domain of BMP type II receptor-2 [273]. As a result, fresh information on the structure and function of SMYD2 underscores its importance as a cancer regulator. To understand the precise role of SMYD2 in carcinogenesis, more research is needed. Because higher expression of SMYD2 is present in varied cancers, it might be considered a potential therapeutic target as well as a biomarker for diagnosis.

1.7.5 Inhibitors of SMYD2

This target class of enzymes has a growing body of evidence indicating they play major harmful roles in various human diseases. The enzymatic processes and structures of PMTs support the idea that small-molecule inhibitors can be used to pharmacologically modulate these enzymes and thus be an effective therapeutic intervention in cancer. Finding PMTs-specific small-molecule inhibitors as a starting point of therapeutic development would be a top priority for researchers. Protein crystal structure analysis is critical for medication research and development. Pharmaceutical companies have been developing SMYD2 inhibitors since the first crystal structure was published in 2011 [239]. AZ505 is the first discovered SMYD2-specific inhibitor with the help of high-throughput chemical screening, that has higher SMYD2 selectivity with efficient inhibitory power [237]. AZ505 bind with SMYD2's peptide-binding groove of SMYD2 and competes for binding with substrates like p53 [237]. Another SMYD2-specific small inhibitor LLY-507 was developed by Eli Lilly and Company in 2015 [274], having 100-fold selectivity for SMYD2. LLY-507 inhibits the capacity of SMYD2 *via* the methylation of p53 with an IC50 of less than 15nM, according to the crystal structure of SMYD2. Furthermore, LLY-507 also inhibits the growth of several tumor cell lines [274]. A-893 is another cell-active benzoxazine inhibitor for SMYD2 with highly selectivity [216]. BAY-598 is an aminopyrazolines based another SMYD2-specific inhibitor. It is a substrate-competitive SMYD2 inhibitor that has more than 100-fold selectivity to SMYD2 in a specific group of 32 types of methyltransferases. BAY-598 can increase doxorubicin efficiency in xenograft cancers in-vivo [242]. SMYD2 inhibitors AZ506 and EPZ033294 can also prevent cancer cell lines from p53 methylation and proliferation [276], [277]. SMYD2 inhibitors come in a range of shapes and sizes, but they're all predicated on the role of SMYD2 in cancer and p53 methylation. Though, SMYD2's function is not restricted to these two areas. As a result, these SMYD2-specific inhibitors could have some drawbacks and unidentified side effects. Moreover, because maximum inhibitors have only studied in *in-vitro*, thus it is necessary to investigate if they affect xenograft tumours, carcinoma in situ, or other illnesses, as well as the underlying mechanisms.

1.8 Prediction of methyltransferases using bioinformatics approaches

Methyltransferase enzymes catalyze the methylation reaction by using SAM donor belongs to distinct classes [278]. It possesses number of enzymes that modify the lysine and arginine residues in histone and non-histone proteins and play a key role in epigenetics [26]. Only little number of potential methyltransferases has been identified and well characterized. In public sequence databases diverse methyltransferase sequences are present, but most of these are assigned as hypothetical, putative, or probable functions based on sequence similarity. Thus, there was a need to identify the correct methyltransferase in several organisms. However, various existing experimental approaches for the identification of these enzymes are costly, time-consuming, labor-intensive, and require specialized equipment. Due to these obstacles, computational techniques emerged as a powerful alternative approach to overcome these obstacles. Previously, some studies have been conducted for the identification of new methyltransferases using bioinformatics methods. Those studied used sequence-based, motif-based, and Hidden Markov Model (HMM) based search approaches and transmit the information into the identification of new methyltransferases [279]. Initially, computational predictions of methyltransferases were performed using sequence comparisons by BLAST search against the protein databases. The protein arginine methyltransferase such as Hmt1/Rmt1 were discovered successfully by applying this approach [280]. The seven beta strands methyltransferases were identified by motif-based search [281], [282]. Further, the identification of methyltransferase domain was performed using the secondary structure information through sequences [283]. In their search, authors used HMM profiles that considered the frequency of odds amino acids along with the frequency of deletions and insertions to justify for the gaps in the alignment. The reference set was considered from the superfamily of methyltransferases to create the HMM profiles for the identification of general proteins. The O-, N- and C- methyltransferases were identified from non-redundant database using HMM profiles that used the methyltransferase domain from nonribosomal peptide synthetase and polyketide synthase [283]. These global search profiles, which cover the full methyltransferases domain, penalise mismatches between motifs, which may prevent the detection of real, previously undiscovered methyltransferases. Furthermore, a new computational approach HHpred [284] and multiple motifs scanning (MMS) [120] were introduced that increased the searching power as compare to BLAST methods.

The methyltransferases have a topologically-distinct family of proteins [279]. Thus, the similarity among primary sequences was found in only a small region of the protein. The search approaches based on sequence alignment are time-consuming and low-sensitive that can lead to problems in the prediction of low similarity proteins. Therefore, we need an advanced method that is based on sequence information rather than simple similarity of amino acids. To overcome these problems, now a day's machine learning becomes a popular approach in the field of bioinformatics. By using extensive sequence-based feature techniques, machine learning has been successfully applied to solve such type of problems in bioinformatics [285]–[289]. In example, determining the family to which a recently sequenced protein belongs is frequently of interest to biologists [290]. This enables research into the protein's evolutionary history and the identification of its biological roles. The use of sequence data in categorization and prediction tasks has seen widespread application of machine learning techniques [291]. A classification problem arises when we need to categories two things that are already known, build a model for those using a classifier, and then predict more unknown data using features and the model we've established. Various supervised machine learning methods can be used to classify new observations based on previously learned information [292]. We have summarized below about the machine-learning techniques used for various enzyme identification/classification purpose. These machine-learning techniques are widely used for the classification of other varieties of biological data and provide important insights from huge biological datasets.

1.9 Machine-learning techniques used in enzyme identification

Proteins known as enzymes function as a biological catalyst that fast ups the biochemical reactions. Different enzymes perform biological processes and interact with various things. The relevant biological function can be determined and the catalytic mechanism of an enzyme can be inferred using the enzyme family to which it belongs. The annotation of the family for an enzyme is crucial given the massive influx of protein sequences into databanks in the post-genomics era. Because experimental approaches are quite expensive, it is essential to create computational methods that will greatly aid in correctly identifying the family of enzymes. Such a computational tool's purpose would be to provide guidance and an approximate idea of what type of enzyme it would be, but the outcome would undoubtedly need to be verified through experimental research.

A wide range of machine learning methods such as SVM, DT, NB, Random Forest, and Neural Network have been used for enzyme classification [292]. Cai *et al.* proposed a SVM-based technique for enzyme family classification [293]. Whereas Dobson and Doig [294] proposed a SVM-based method to predict enzyme class. Lu *et al.* developed an SVM-based method to investigate the enzyme family [295]. Nasibov *et al.* used the frequency of the amino acid residue to represent the enzyme sequences to classify the enzyme families by adopting the KNN classifier with minimum distanced-based classifier [296]. Qie *et al.* proposed an integrated method of SVM with discrete wavelet transform for the classification of enzyme family by using the hydrophobicity of amino acid from PseAAC [297]. LacSubPred was proposed to classify the subtypes of laccases using sequence-based descriptor and help to characterize the laccase protein sequences [298]. Furthermore, Amidi *et al*. [299] explores the usage of nearest neighbour and SVM fusion approaches for the classification of proteins in the PDB. ECPred predicted enzyme class using an ensemble of machine learning models in a hierarchy [300]. Tao *et al.* [301] used four different machine learning algorithms to classify proteins into seven different enzyme classes using various sequence-based features. A prediction method name as eCAMI proposed to classify 390 carbohydrate-active enzymes (CAZymes) family into thousands of subfamilies using k-mer based feature [302]. Wu *et al.* adopted the SVM to identify human enzyme family classes by incorporating rigidity, flexibility and irreplaceability of amino acids into pseudo amino acid composition (PseAAC) [303]. Further, Zhang et al. developed a SVM-based predictor to classify human enzymes using the amino acid composition (AAC) and composition of *k*-spaced amino acid pairs (CKSAAP) [304]. Recently, Wang *et al.* [305] also developed a predictor called IHEC_RAAC, which has the capability to identify whether a protein is a human enzyme and distinguish the function of the human enzyme. For the classification of secretory and non-secretory enzymes, Garg and Raghava proposed a hybrid method by integrating the SVM module with PSIBLAST module using amino acid and dipeptide composition [306]. Recently, Zhang *et al.* proposed a reduced amino acid composition based method called iSP-RAAC using SVM method for the identification of secretory proteins of malaria parasite [307].

All these studies did achieve encouraging results, which may offer crucial hints for categorizing enzymes and annotating their functions. Based on prior research, such machine-learning based classification techniques enhanced the classification performance. However, each of these studies focused on categorizing various enzyme subtypes using such techniques. Predictions for fresh unclassified data can be made using a characterization of the classes. Classes can be a simple binary partition (such as a pair of proteins "PMT" or "non-PMT" for the problem we face in this thesis), or multi-class classification. In this thesis, we have utilized various machine-learning techniques to categorize PMTs and non-PMT sequences, which is similar to the research mentioned above. The detail description of all used machine-learning techniques in this thesis work is described in materials & methods of Chapter 2.

1.10 HYPOTHESIS AND OBJECTIVE

The objective of my Ph.D. work was to build a machine learning-based PMT prediction method, the oncogenic role of PKMT protein SMYD2, and polymorphism analysis along with novel inhibitor identification of SMYD2 protein from the class of PKMT. The specific objectives have been summarized as four major points mentioned here:

1. PMTs are essential for the regulation of epigenetic and gene expression *via* the methylation of several histone and nonhistone substrates. PMTs are associated with the progression of numerous cancer types and have a variety of applications in various cellular processes. As a result, they are widely recognized as potential cancer therapeutic targets. Therefore, in light of the well-recognized significance of PMTs, it is important to have a sophisticated and accurate prediction method for the identification of PMTs. So, we designed our first objective as *"Development of machine-learning-based prediction method for identifying protein methyltransferases"*.
2. The role of SMYD2 in cancer progression is mostly unknown, and more research into the gene's role in cancer is required. Thus, we designed our second objective *"Integrative analysis of the oncogenic role of SMYD2 in multiple human*

malignancies" to use computational methods to investigate the link between SMYD2 and cancer.

3. Clarifying the physiological and pharmacological implications of methylation on unknown nonhistone substrates in cancer is critical for the development of effective anticancer medicines. There are many mutations in SMYD2, and some of them are thought to have a significant impact on the enzymatic activity of methyltransferase. Still, the biological functions of maximum mutations have yet to be investigated. To create anticancer treatments that target the PKMT's SMYD2 protein, more biological research into individual mutations is needed. Thus, we designed our third objective as *"sequence and structure-based nsSNPs analysis of SMYD2"*.
4. Any research-based study or objective has to be accomplished with the development of some crucial product for the problem under consideration. In contrast to the significant research into PKMT-specific inhibitors, there has been limited effort has been made. With respect to small-molecule inhibitors, the development of high-quality PKMT inhibitors has received increased attention. Additionally, new PKMTs specific inhibitors are still in great demand to combat cancer progression. We plan to provide some solutions to the scientific community and therefore we worked on the development of novel inhibitors for PKMT-specific SMYD2 protein. The overexpression of PKMT SMYD2 protein in multiple cancers suggests that it is a potential candidate for the development of anticancer therapy. Thus, we designed our fourth and last objective as *"Structural investigation and simulation studies to design novel inhibitors for SMYD2"*.

REFERENCES

[1] P. A. Boriack-Sjodin and K. K. Swinger, “Protein Methyltransferases: A Distinct, Diverse, and Dynamic Family of Enzymes,” *Biochemistry*, vol. 55, no. 11, pp. 1557–1569, Mar. 2016, doi: 10.1021/acs.biochem.5b01129.

[2] H. Ü. Kaniskan and J. Jin, “Recent progress in developing selective inhibitors of protein methyltransferases,” *Curr Opin Chem Biol*, vol. 39, pp. 100–108, Aug. 2017, doi: 10.1016/j.cbpa.2017.06.013.

[3] R. Hamamoto, V. Saloura, and Y. Nakamura, “Critical roles of non-histone protein lysine methylation in human tumorigenesis,” *Nat Rev Cancer*, vol. 15, no. 2, pp. 110–124, Feb. 2015, doi: 10.1038/nrc3884.

[4] N. Spellmon, J. Holcomb, L. Trescott, N. Sirinupong, and Z. Yang, “Structure and Function of SET and MYND Domain-Containing Proteins,” *Int J Mol Sci*, vol. 16, no. 1, pp. 1406–1428, Jan. 2015, doi: 10.3390/ijms16011406.

[5] J. Wesche, S. Kühn, B. M. Kessler, M. Salton, and A. Wolf, “Protein arginine methylation: a prominent modification and its demethylation,” *Cell Mol Life Sci*, vol. 74, no. 18, pp. 3305–3315, Sep. 2017, doi: 10.1007/s00018-017-2515-z.

[6] X. Yi, X.-J. Jiang, X.-Y. Li, and D.-S. Jiang, “Histone methyltransferases: novel targets for tumor and developmental defects,” *Am J Transl Res*, vol. 7, no. 11, pp. 2159–2175, 2015.

[7] C. Martin and Y. Zhang, “The diverse functions of histone lysine methylation,” *Nat. Rev. Mol. Cell Biol.*, vol. 6, no. 11, pp. 838–849, Nov. 2005, doi: 10.1038/nrm1761.

[8] R. A. Copeland, M. E. Solomon, and V. M. Richon, “Protein methyltransferases as a target class for drug discovery,” *Nat Rev Drug Discov*, vol. 8, no. 9, pp. 724–732, Sep. 2009, doi: 10.1038/nrd2974.

[9] R. Hamamoto and Y. Nakamura, “Dysregulation of protein methyltransferases in human cancer: An emerging target class for anticancer therapy,” *Cancer Sci*, vol. 107, no. 4, pp. 377–384, Apr. 2016, doi: 10.1111/cas.12884.

[10] X. Yi, X.-J. Jiang, and Z.-M. Fang, “Histone methyltransferase SMYD2: ubiquitous regulator of disease,” *Clin Epigenetics*, vol. 11, no. 1, p. 112, Aug. 2019, doi: 10.1186/s13148-019-0711-4.

[11] C. Tracy *et al.*, “The Smyd Family of Methyltransferases: Role in Cardiac and Skeletal Muscle Physiology and Pathology,” *Curr Opin Physiol*, vol. 1, pp. 140–152, Feb. 2018, doi: 10.1016/j.cophys.2017.10.001.

[12] Z. Che, H. Sun, W. Yao, B. Lu, and Q. Han, "Role of post-translational modifications in regulation of tumor suppressor p53 function," *Frontiers of Oral and Maxillofacial Medicine*, vol. 2, no. 0, Jan. 2020, doi: 10.21037/fomm.2019.12.02.

[13] L. A. Saddic *et al.*, "Methylation of the retinoblastoma tumor suppressor by SMYD2," *J Biol Chem*, vol. 285, no. 48, pp. 37733–37740, Nov. 2010, doi: 10.1074/jbc.M110.137612.

[14] J. Huang *et al.*, "Repression of p53 activity by Smyd2-mediated methylation," *Nature*, vol. 444, no. 7119, pp. 629–632, Nov. 2006, doi: 10.1038/nature05287.

[15] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/caac.21492.

[16] H. Sung *et al.*, "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[17] I. Aier, R. Semwal, A. Dhara, N. Sen, and P. K. Varadwaj, "An integrated epigenome and transcriptome analysis identifies PAX2 as a master regulator of drug resistance in high grade pancreatic ductal adenocarcinoma," *PLOS ONE*, vol. 14, no. 10, p. e0223554, Oct. 2019, doi: 10.1371/journal.pone.0223554.

[18] R. A. Copeland, "Molecular pathways: protein methyltransferases in cancer," *Clin Cancer Res*, vol. 19, no. 23, pp. 6344–6350, Dec. 2013, doi: 10.1158/1078-0432.CCR-13-0223.

[19] R. A. Copeland, M. P. Moyer, and V. M. Richon, "Targeting genetic alterations in protein methyltransferases for personalized cancer therapeutics," *Oncogene*, vol. 32, no. 8, pp. 939–946, Feb. 2013, doi: 10.1038/onc.2012.552.

[20] R. A. Copeland, "Protein methyltransferase inhibitors as personalized cancer therapeutics," *Drug Discovery Today: Therapeutic Strategies*, vol. 9, no. 2, pp. e83–e90, Sep. 2012, doi: 10.1016/j.ddstr.2011.08.001.

[21] H. Ü. Kaniskan, M. L. Martini, and J. Jin, "Inhibitors of Protein Methyltransferases and Demethylases," *Chem. Rev.*, vol. 118, no. 3, pp. 989–1068, Feb. 2018, doi: 10.1021/acs.chemrev.6b00801.

[22] Y. Shi *et al.*, "Histone demethylation mediated by the nuclear amine oxidase homolog LSD1," *Cell*, vol. 119, no. 7, pp. 941–953, Dec. 2004, doi: 10.1016/j.cell.2004.12.012.

[23] X. Shi *et al.*, “Modulation of p53 function by SET8-mediated methylation at lysine 382,” *Mol Cell*, vol. 27, no. 4, pp. 636–646, Aug. 2007, doi: 10.1016/j.molcel.2007.07.012.

[24] S. Chuikov *et al.*, “Regulation of p53 activity through lysine methylation,” *Nature*, vol. 432, no. 7015, pp. 353–360, Nov. 2004, doi: 10.1038/nature03117.

[25] T. Kouzarides, “Chromatin modifications and their function,” *Cell*, vol. 128, no. 4, pp. 693–705, Feb. 2007, doi: 10.1016/j.cell.2007.02.005.

[26] B. C. Smith and J. M. Denu, “Chemical mechanisms of histone lysine and arginine modifications,” *Biochim Biophys Acta*, vol. 1789, no. 1, pp. 45–57, Jan. 2009, doi: 10.1016/j.bbagrm.2008.06.005.

[27] C. H. Arrowsmith, C. Bountra, P. V. Fish, K. Lee, and M. Schapira, “Epigenetic protein families: a new frontier for drug discovery,” *Nature Reviews Drug Discovery*, vol. 11, no. 5, Art. no. 5, May 2012, doi: 10.1038/nrd3674.

[28] M. Jansson *et al.*, “Arginine methylation regulates the p53 response,” *Nat Cell Biol*, vol. 10, no. 12, pp. 1431–1439, Dec. 2008, doi: 10.1038/ncb1802.

[29] R. Schneider, A. J. Bannister, and T. Kouzarides, “Unsafe SETs: histone lysine methyltransferases and cancer,” *Trends Biochem Sci*, vol. 27, no. 8, pp. 396–402, Aug. 2002, doi: 10.1016/s0968-0004(02)02141-2.

[30] J. H. Kim, B. C. Yoo, W. S. Yang, E. Kim, S. Hong, and J. Y. Cho, “The Role of Protein Arginine Methyltransferases in Inflammatory Responses,” *Mediators Inflamm*, vol. 2016, p. 4028353, 2016, doi: 10.1155/2016/4028353.

[31] C. Milite *et al.*, “The emerging role of lysine methyltransferase SETD8 in human diseases,” *Clinical Epigenetics*, vol. 8, no. 1, p. 102, Sep. 2016, doi: 10.1186/s13148-016-0268-4.

[32] J.-F. Couture, G. Hauk, M. J. Thompson, G. M. Blackburn, and R. C. Trievel, “Catalytic Roles for Carbon-Oxygen Hydrogen Bonding in SET Domain Lysine Methyltransferases *,” *Journal of Biological Chemistry*, vol. 281, no. 28, pp. 19280–19287, Jul. 2006, doi: 10.1074/jbc.M602257200.

[33] X. Zhang and T. C. Bruice, “Enzymatic mechanism and product specificity of SET-domain protein lysine methyltransferases,” *PNAS*, vol. 105, no. 15, pp. 5728–5732, Apr. 2008, doi: 10.1073/pnas.0801788105.

[34] C. Wang *et al.*, “Structural determinants for the strict monomethylation activity by trypanosoma brucei protein arginine methyltransferase 7,” *Structure*, vol. 22, no. 5, pp. 756–768, May 2014, doi: 10.1016/j.str.2014.03.003.

[35] M. T. Bedford and S. G. Clarke, “Protein arginine methylation in mammals: who, what, and why,” *Mol Cell*, vol. 33, no. 1, pp. 1–13, Jan. 2009, doi: 10.1016/j.molcel.2008.12.013.

[36] L. C. Boffa, J. Karn, G. Vidali, and V. G. Allfrey, “Distribution of NG, NG,-dimethylarginine in nuclear protein fractions,” *Biochem Biophys Res Commun*, vol. 74, no. 3, pp. 969–976, Feb. 1977, doi: 10.1016/0006-291x(77)91613-8.

[37] R. M. Hughes and M. L. Waters, “Arginine methylation in a beta-hairpin peptide: implications for Arg-pi interactions, DeltaCp(o), and the cold denatured state,” *J Am Chem Soc*, vol. 128, no. 39, pp. 12735–12742, Oct. 2006, doi: 10.1021/ja061656g.

[38] M. T. Bedford and S. Richard, “Arginine methylation an emerging regulator of protein function,” *Mol Cell*, vol. 18, no. 3, pp. 263–272, Apr. 2005, doi: 10.1016/j.molcel.2005.04.003.

[39] S. Benhenda *et al.*, “Methyltransferase PRMT1 Is a Binding Partner of HBx and a Negative Regulator of Hepatitis B Virus Transcription,” *Journal of Virology*, vol. 87, no. 8, pp. 4360–4371, Apr. 2013, doi: 10.1128/JVI.02574-12.

[40] S. Pal and S. Sif, “Interplay between chromatin remodelers and protein arginine methyltransferases,” *J Cell Physiol*, vol. 213, no. 2, pp. 306–315, Nov. 2007, doi: 10.1002/jcp.21180.

[41] Z. Simandi *et al.*, “PRMT1 and PRMT8 regulate retinoic acid-dependent neuronal differentiation with implications to neuropathology,” *Stem Cells*, vol. 33, no. 3, pp. 726–741, Mar. 2015, doi: 10.1002/stem.1894.

[42] S. K. Tewary, Y. G. Zheng, and M.-C. Ho, “Protein arginine methyltransferases: insights into the enzyme structure and mechanism at the atomic level,” *Cell Mol Life Sci*, vol. 76, no. 15, pp. 2917–2932, Aug. 2019, doi: 10.1007/s00018-019-03145-x.

[43] F. Lv *et al.*, “Structural basis for Sfm1 functioning as a protein arginine methyltransferase,” *Cell Discov*, vol. 1, no. 1, pp. 1–14, Dec. 2015, doi: 10.1038/celldisc.2015.37.

[44] Y. Yang and M. T. Bedford, “Protein arginine methyltransferases and cancer,” *Nature Reviews Cancer*, vol. 13, no. 1, Art. no. 1, Jan. 2013, doi: 10.1038/nrc3409.

[45] Y. Yang *et al.*, “PRMT9 is a type II methyltransferase that methylates the splicing factor SAP145,” *Nat Commun*, vol. 6, p. 6428, Mar. 2015, doi: 10.1038/ncomms7428.

[46] L. Niu, F. Lu, Y. Pei, C. Liu, and X. Cao, “Regulation of flowering time by the protein arginine methyltransferase AtPRMT10,” *EMBO Rep*, vol. 8, no. 12, pp. 1190–1195, Dec. 2007, doi: 10.1038/sj.embor.7401111.

[47] F. Scebba, M. De Bastiani, G. Bernacchia, A. Andreucci, A. Galli, and L. Pitto, “PRMT11: a new Arabidopsis MBD7 protein partner with arginine methyltransferase activity,” *Plant J*, vol. 52, no. 2, pp. 210–222, Oct. 2007, doi: 10.1111/j.1365-313X.2007.03238.x.

[48] L. M. Webb and M. Guerau-de-Arellano, “Emerging Role for Methylation in Multiple Sclerosis: Beyond DNA,” *Trends Mol Med*, vol. 23, no. 6, pp. 546–562, Jun. 2017, doi: 10.1016/j.molmed.2017.04.004.

[49] J. Kzhyshkowska, E. Kremmer, M. Hofmann, H. Wolf, and T. Dobner, “Protein arginine methylation during lytic adenovirus infection,” *Biochem J*, vol. 383, no. Pt 2, pp. 259–265, Oct. 2004, doi: 10.1042/BJ20040210.

[50] W. J. Friesen, S. Massenet, S. Paushkin, A. Wyce, and G. Dreyfuss, “SMN, the Product of the Spinal Muscular Atrophy Gene, Binds Preferentially to Dimethylarginine-Containing Protein Targets,” *Molecular Cell*, vol. 7, no. 5, pp. 1111–1117, May 2001, doi: 10.1016/S1097-2765(01)00244-1.

[51] A. Couto E Silva *et al.*, “Protein Arginine Methyltransferases in Cardiovascular and Neuronal Function,” *Mol Neurobiol*, vol. 57, no. 3, pp. 1716–1732, Mar. 2020, doi: 10.1007/s12035-019-01850-z.

[52] X. Zhang, L. Zhou, and X. Cheng, “Crystal structure of the conserved core of protein arginine methyltransferase PRMT3,” *EMBO J*, vol. 19, no. 14, pp. 3509–3519, Jul. 2000, doi: 10.1093/emboj/19.14.3509.

[53] X. Zhang and X. Cheng, “Structure of the predominant protein arginine methyltransferase PRMT1 and analysis of its binding to substrate peptides,” *Structure*, vol. 11, no. 5, pp. 509–520, May 2003, doi: 10.1016/s0969-2126(03)00071-6.

[54] J. Lee and M. T. Bedford, “PABP1 identified as an arginine methyltransferase substrate using high-density protein arrays,” *EMBO Rep*, vol. 3, no. 3, pp. 268–273, Mar. 2002, doi: 10.1093/embo-reports/kvf052.

[55] D. Cheng, J. Côté, S. Shaaban, and M. T. Bedford, “The arginine methyltransferase CARM1 regulates the coupling of transcription and mRNA processing,” *Mol Cell*, vol. 25, no. 1, pp. 71–83, Jan. 2007, doi: 10.1016/j.molcel.2006.11.019.

[56] H. Wei, R. Mundade, K. C. Lange, and T. Lu, “Protein arginine methylation of non-histone proteins and its role in diseases,” *Cell Cycle*, vol. 13, no. 1, pp. 32–41, 2014, doi: 10.4161/cc.27353.

[57] A. J. Bannister and T. Kouzarides, “Regulation of chromatin by histone modifications,” *Cell Res*, vol. 21, no. 3, pp. 381–395, Mar. 2011, doi: 10.1038/cr.2011.22.

[58] Z. Zhao and A. Shilatifard, “Epigenetic modifications of histones in cancer,” *Genome Biology*, vol. 20, no. 1, p. 245, Nov. 2019, doi: 10.1186/s13059-019-1870-5.

[59] M. Litt, Y. Qiu, and S. Huang, “Histone arginine methylations: their roles in chromatin dynamics and transcriptional regulation,” *Biosci Rep*, vol. 29, no. 2, pp. 131–141, Apr. 2009, doi: 10.1042/BSR20080176.

[60] S. Huang, M. Litt, and G. Felsenfeld, “Methylation of histone H4 by arginine methyltransferase PRMT1 is essential in vivo for many subsequent histone modifications,” *Genes Dev*, vol. 19, no. 16, pp. 1885–1893, Aug. 2005, doi: 10.1101/gad.1333905.

[61] Y.-H. Lee and M. R. Stallcup, “Minireview: protein arginine methylation of nonhistone proteins in transcriptional regulation,” *Mol Endocrinol*, vol. 23, no. 4, pp. 425–433, Apr. 2009, doi: 10.1210/me.2008-0380.

[62] J. W. Hwang, Y. Cho, G.-U. Bae, S.-N. Kim, and Y. K. Kim, “Protein arginine methyltransferases: promising targets for cancer therapy,” *Exp Mol Med*, vol. 53, no. 5, pp. 788–808, May 2021, doi: 10.1038/s12276-021-00613-y.

[63] M. D. Hebert, K. B. Shpargel, J. K. Ospina, K. E. Tucker, and A. G. Matera, “Coilin methylation regulates nuclear body formation,” *Dev Cell*, vol. 3, no. 3, pp. 329–337, Sep. 2002, doi: 10.1016/s1534-5807(02)00222-8.

[64] E. Guccione and S. Richard, “The regulation, functions and clinical relevance of arginine methylation,” *Nat Rev Mol Cell Biol*, vol. 20, no. 10, pp. 642–657, Oct. 2019, doi: 10.1038/s41580-019-0155-x.

[65] J. Tang *et al.*, “PRMT1 is the predominant type I protein arginine methyltransferase in mammalian cells,” *J Biol Chem*, vol. 275, no. 11, pp. 7723–7730, Mar. 2000, doi: 10.1074/jbc.275.11.7723.

[66] S. Majumder, Y. Liu, O. H. Ford, J. L. Mohler, and Y. E. Whang, “Involvement of arginine methyltransferase CARM1 in androgen receptor function and prostate cancer cell viability,” *Prostate*, vol. 66, no. 12, pp. 1292–1301, Sep. 2006, doi: 10.1002/pros.20438.

[67] H. Hong *et al.*, “Aberrant expression of CARM1, a transcriptional coactivator of androgen receptor, in the development of prostate carcinoma and androgen-independent status,” *Cancer*, vol. 101, no. 1, pp. 83–89, Jul. 2004, doi: 10.1002/cncr.20327.

[68] S. E. Messaoudi *et al.*, “Coactivator-associated arginine methyltransferase 1 (CARM1) is a positive regulator of the Cyclin E1 gene,” *PNAS*, vol. 103, no. 36, pp. 13351–13356, Sep. 2006, doi: 10.1073/pnas.0605692103.

[69] Y.-R. Kim *et al.*, “Differential CARM1 expression in prostate and colorectal cancers,” *BMC Cancer*, vol. 10, p. 197, May 2010, doi: 10.1186/1471-2407-10-197.

[70] D. Chen, S. M. Huang, and M. R. Stallcup, “Synergistic, p160 coactivator-dependent enhancement of estrogen receptor function by CARM1 and p300,” *J Biol Chem*, vol. 275, no. 52, pp. 40810–40816, Dec. 2000, doi: 10.1074/jbc.M005459200.

[71] M. A. Powers, M. M. Fay, R. E. Factor, A. L. Welm, and K. S. Ullman, “Protein arginine methyltransferase 5 accelerates tumor growth by arginine methylation of the tumor suppressor programmed cell death 4,” *Cancer Res*, vol. 71, no. 16, pp. 5579–5587, Aug. 2011, doi: 10.1158/0008-5472.CAN-11-0458.

[72] K. Chiang *et al.*, “PRMT5 Is a Critical Regulator of Breast Cancer Stem Cell Function via Histone Methylation and FOXP1 Expression,” *Cell Rep*, vol. 21, no. 12, pp. 3498–3513, Dec. 2017, doi: 10.1016/j.celrep.2017.11.096.

[73] S. Pal, S. N. Vishwanath, H. Erdjument-Bromage, P. Tempst, and S. Sif, “Human SWI/SNF-associated PRMT5 methylates histone H3 arginine 8 and negatively regulates expression of ST7 and NM23 tumor suppressor genes,” *Mol Cell Biol*, vol. 24, no. 21, pp. 9630–9645, Nov. 2004, doi: 10.1128/MCB.24.21.9630-9645.2004.

[74] L.-M. Liu *et al.*, “Arginine Methyltransferase PRMT1 Regulates p53 Activity in Breast Cancer,” *Life (Basel)*, vol. 11, no. 8, p. 789, Aug. 2021, doi: 10.3390/life11080789.

[75] M. Li, W. An, L. Xu, Y. Lin, L. Su, and X. Liu, “The arginine methyltransferase PRMT5 and PRMT1 distinctly regulate the degradation of anti-apoptotic protein CFLARL in human lung cancer cells,” *Journal of Experimental & Clinical Cancer Research*, vol. 38, no. 1, p. 64, Feb. 2019, doi: 10.1186/s13046-019-1064-8.

[76] Y. Zhao *et al.*, “PRMT1 regulates the tumour-initiating properties of esophageal squamous cell carcinoma through histone H4 arginine methylation coupled with transcriptional activation,” *Cell Death Dis*, vol. 10, no. 5, Art. no. 5, May 2019, doi: 10.1038/s41419-019-1595-0.

[77] J. Zhong *et al.*, “Identification and characterization of novel spliced variants of PRMT2 in breast carcinoma,” *FEBS J*, vol. 279, no. 2, pp. 316–335, Jan. 2012, doi: 10.1111/j.1742-4658.2011.08426.x.

[78] F. Dong *et al.*, “PRMT2 links histone H3R8 asymmetric dimethylation to oncogenic activation and tumorigenesis of glioblastoma,” *Nat Commun*, vol. 9, no. 1, p. 4552, Oct. 2018, doi: 10.1038/s41467-018-06968-7.

[79] J. Zhong *et al.*, “Nuclear loss of protein arginine N-methyltransferase 2 in breast carcinoma is associated with tumor grade and overexpression of cyclin D1 protein,” *Oncogene*, vol. 33, no. 48, pp. 5546–5558, Nov. 2014, doi: 10.1038/onc.2013.500.

[80] M.-C. Hsu *et al.*, “Protein arginine methyltransferase 3-induced metabolic reprogramming is a vulnerable target of pancreatic cancer,” *J Hematol Oncol*, vol. 12, no. 1, p. 79, Jul. 2019, doi: 10.1186/s13045-019-0769-7.

[81] L. Wang *et al.*, “CARM1 methylates chromatin remodeling factor BAF155 to enhance tumor progression and metastasis,” *Cancer Cell*, vol. 25, no. 1, pp. 21–36, Jan. 2014, doi: 10.1016/j.ccr.2013.12.007.

[82] J. Liu *et al.*, “Arginine methylation-dependent LSD1 stability promotes invasion and metastasis of breast cancer,” *EMBO Rep*, vol. 21, no. 2, p. e48597, Feb. 2020, doi: 10.15252/embr.201948597.

[83] C.-Y. Ou *et al.*, “A coactivator role of CARM1 in the dysregulation of β-catenin activity in colorectal cancer cell growth and gene expression,” *Mol Cancer Res*, vol. 9, no. 5, pp. 660–670, May 2011, doi: 10.1158/1541-7786.MCR-10-0223.

[84] S. Karakashev *et al.*, “CARM1-expressing ovarian cancer depends on the histone methyltransferase EZH2 activity,” *Nat Commun*, vol. 9, no. 1, p. 631, Feb. 2018, doi: 10.1038/s41467-018-03031-3.

[85] L. P. Vu *et al.*, “PRMT4 blocks myeloid differentiation by assembling a methyl-RUNX1-dependent repressor complex,” *Cell Rep*, vol. 5, no. 6, pp. 1625–1638, Dec. 2013, doi: 10.1016/j.celrep.2013.11.025.

[86] M. Al-Dhaheri *et al.*, "CARM1 is an important determinant of ERα-dependent breast cancer cell differentiation and proliferation in breast cancer cells," *Cancer Res*, vol. 71, no. 6, pp. 2118–2128, Mar. 2011, doi: 10.1158/0008-5472.CAN-10-2426.

[87] X.-Y. Zhong *et al.*, "CARM1 Methylates GAPDH to Regulate Glucose Metabolism and Is Suppressed in Liver Cancer," *Cell Rep*, vol. 24, no. 12, pp. 3207–3223, Sep. 2018, doi: 10.1016/j.celrep.2018.08.066.

[88] A.-V. Hartley *et al.*, "PRMT5-mediated methylation of YBX1 regulates NF-κB activity in colorectal cancer," *Sci Rep*, vol. 10, no. 1, p. 15934, Sep. 2020, doi: 10.1038/s41598-020-72942-3.

[89] X. Liu *et al.*, "Protein arginine methyltransferase 5-mediated epigenetic silencing of IRX1 contributes to tumorigenicity and metastasis of gastric cancer," *Biochim Biophys Acta Mol Basis Dis*, vol. 1864, no. 9 Pt B, pp. 2835–2844, Sep. 2018, doi: 10.1016/j.bbadis.2018.05.015.

[90] L. Ge *et al.*, "PRMT5 promotes epithelial-mesenchymal transition via EGFR-β-catenin axis in pancreatic cancer cells," *J Cell Mol Med*, vol. 24, no. 2, pp. 1969–1979, Jan. 2020, doi: 10.1111/jcmm.14894.

[91] G. Hu, X. Wang, Y. Han, and P. Wang, "Protein arginine methyltransferase 5 promotes bladder cancer growth through inhibiting NF-kB dependent apoptosis," *EXCLI J*, vol. 17, pp. 1157–1166, Nov. 2018, doi: 10.17179/excli2018-1719.

[92] F. Zhu *et al.*, "PRMT5 is upregulated by B cell receptor signaling and forms a positive feedback loop with PI3K/AKT in lymphoma cells," *Leukemia*, vol. 33, no. 12, pp. 2898–2911, Dec. 2019, doi: 10.1038/s41375-019-0489-6.

[93] H. Tamiya *et al.*, "SHARPIN-mediated regulation of protein arginine methyltransferase 5 controls melanoma growth," *J Clin Invest*, vol. 128, no. 1, pp. 517–530, Jan. 2018, doi: 10.1172/JCI95410.

[94] F. Yan *et al.*, "Genetic validation of the protein arginine methyltransferase PRMT5 as a candidate therapeutic target in glioblastoma," *Cancer Res*, vol. 74, no. 6, pp. 1752–1765, Mar. 2014, doi: 10.1158/0008-5472.CAN-13-0884.

[95] L. Wang, S. Pal, and S. Sif, "Protein arginine methyltransferase 5 suppresses the transcription of the RB family of tumor suppressors in leukemia and lymphoma cells," *Mol Cell Biol*, vol. 28, no. 20, pp. 6262–6277, Oct. 2008, doi: 10.1128/MCB.00923-08.

[96] A. Radzisheuskaya *et al.*, “PRMT5 methylome profiling uncovers a direct link to splicing regulation in acute myeloid leukemia,” *Nat Struct Mol Biol*, vol. 26, no. 11, pp. 999–1012, Nov. 2019, doi: 10.1038/s41594-019-0313-z.

[97] M. Rengasamy *et al.*, “The PRMT5/WDR77 complex regulates alternative splicing through ZNF326 in breast cancer,” *Nucleic Acids Res*, vol. 45, no. 19, pp. 11106–11120, Nov. 2017, doi: 10.1093/nar/gkx727.

[98] T. Fu, X. Lv, Q. Kong, and C. Yuan, “A novel SHARPIN-PRMT5-H3R2me1 axis is essential for lung cancer cell invasion,” *Oncotarget*, vol. 8, no. 33, pp. 54809–54820, Jul. 2017, doi: 10.18632/oncotarget.18957.

[99] E. Beketova *et al.*, “Protein Arginine Methyltransferase 5 Promotes pICln-Dependent Androgen Receptor Transcription in Castration-Resistant Prostate Cancer,” *Cancer Res*, vol. 80, no. 22, pp. 4904–4917, Nov. 2020, doi: 10.1158/0008-5472.CAN-20-1228.

[100] M. Kanda *et al.*, “Protein arginine methyltransferase 5 is associated with malignant phenotype and peritoneal metastasis in gastric cancer,” *Int J Oncol*, vol. 49, no. 3, pp. 1195–1202, Sep. 2016, doi: 10.3892/ijo.2016.3584.

[101] H. Jiang *et al.*, “PRMT5 promotes cell proliferation by inhibiting BTG2 expression via the ERK signaling pathway in hepatocellular carcinoma,” *Cancer Med*, vol. 7, no. 3, pp. 869–882, Mar. 2018, doi: 10.1002/cam4.1360.

[102] K. Okuno *et al.*, “Asymmetric dimethylation at histone H3 arginine 2 by PRMT6 in gastric cancer progression,” *Carcinogenesis*, vol. 40, no. 1, pp. 15–26, Mar. 2019, doi: 10.1093/carcin/bgy147.

[103] N. Jiang *et al.*, “PRMT6 promotes endometrial cancer via AKT/mTOR signaling and indicates poor prognosis,” *Int J Biochem Cell Biol*, vol. 120, p. 105681, Mar. 2020, doi: 10.1016/j.biocel.2019.105681.

[104] S. Avasarala *et al.*, “PRMT6 Promotes Lung Tumor Progression via the Alternate Activation of Tumor-Associated Macrophages,” *Mol Cancer Res*, vol. 18, no. 1, pp. 166–178, Jan. 2020, doi: 10.1158/1541-7786.MCR-19-0204.

[105] L. H. Chan *et al.*, “PRMT6 Regulates RAS/RAF Binding and MEK/ERK-Mediated Cancer Stemness Activities in Hepatocellular Carcinoma through CRAF Methylation,” *Cell Rep*, vol. 25, no. 3, pp. 690-701.e8, Oct. 2018, doi: 10.1016/j.celrep.2018.09.053.

[106] D. Cheng, Z. He, L. Zheng, D. Xie, S. Dong, and P. Zhang, “PRMT7 contributes to the metastasis phenotype in human non-small-cell lung cancer cells possibly through the

interaction with HSPA5 and EEF2," *Onco Targets Ther*, vol. 11, pp. 4869–4876, 2018, doi: 10.2147/OTT.S166412.

[107] F. Liu, L. Wan, H. Zou, Z. Pan, W. Zhou, and X. Lu, "PRMT7 promotes the growth of renal cell carcinoma through modulating the β-catenin/C-MYC axis," *Int J Biochem Cell Biol*, vol. 120, p. 105686, Mar. 2020, doi: 10.1016/j.biocel.2020.105686.

[108] H. Jiang *et al.*, "PRMT9 promotes hepatocellular carcinoma invasion and metastasis via activating PI3K/Akt/GSK-3β/Snail signaling," *Cancer Sci*, vol. 109, no. 5, pp. 1414–1427, May 2018, doi: 10.1111/cas.13598.

[109] R. Alvarez-Venegas, "Bacterial SET domain proteins and their role in eukaryotic chromatin modification," *Frontiers in Genetics*, vol. 5, p. 65, 2014, doi: 10.3389/fgene.2014.00065.

[110] X. Cheng, R. E. Collins, and X. Zhang, "Structural and sequence motifs of protein (histone) methylation enzymes," *Annu Rev Biophys Biomol Struct*, vol. 34, pp. 267–294, 2005, doi: 10.1146/annurev.biophys.34.040204.144452.

[111] S. A. Jacobs, J. M. Harp, S. Devarakonda, Y. Kim, F. Rastinejad, and S. Khorasanizadeh, "The active site of the SET domain is constructed on a knot," *Nat Struct Mol Biol*, vol. 10, no. 7, pp. 578–578, Jul. 2003, doi: 10.1038/nsb0703-578.

[112] M. Schapira, "Structural Chemistry of Human SET Domain Protein Methyltransferases," *Curr Chem Genomics*, vol. 5, no. Suppl 1, pp. 85–94, 2011, doi: 10.2174/1875397301005010085.

[113] L. O. Baumbusch *et al.*, "The Arabidopsis thaliana genome contains at least 29 active genes encoding SET domain proteins that can be assigned to four evolutionarily conserved classes," *Nucleic Acids Res*, vol. 29, no. 21, pp. 4319–4333, Nov. 2001.

[114] T. Kouzarides, "Histone methylation in transcriptional control," *Curr Opin Genet Dev*, vol. 12, no. 2, pp. 198–209, Apr. 2002, doi: 10.1016/s0959-437x(02)00287-3.

[115] C. D. Allis *et al.*, "New nomenclature for chromatin-modifying enzymes," *Cell*, vol. 131, no. 4, pp. 633–636, Nov. 2007, doi: 10.1016/j.cell.2007.10.039.

[116] X. Zhang *et al.*, "Structure of the Neurospora SET domain protein DIM-5, a histone H3 lysine methyltransferase," *Cell*, vol. 111, no. 1, pp. 117–127, Oct. 2002, doi: 10.1016/s0092-8674(02)00999-6.

[117] S. D. Taverna, H. Li, A. J. Ruthenburg, C. D. Allis, and D. J. Patel, "How chromatin-binding modules interpret histone modifications: lessons from professional pocket

pickers," *Nat Struct Mol Biol*, vol. 14, no. 11, pp. 1025–1040, Nov. 2007, doi: 10.1038/nsmb1338.

[118] Z. Zhang and B. F. Pugh, "High-resolution genome-wide mapping of the primary structure of chromatin," *Cell*, vol. 144, no. 2, pp. 175–186, Jan. 2011, doi: 10.1016/j.cell.2011.01.003.

[119] L. Morera, M. Lübbert, and M. Jung, "Targeting histone methyltransferases and demethylases in clinical trials for cancer therapy," *Clinical Epigenetics*, vol. 8, no. 1, p. 57, May 2016, doi: 10.1186/s13148-016-0223-4.

[120] T. C. Petrossian and S. G. Clarke, "Uncovering the human methyltransferasome," *Mol Cell Proteomics*, vol. 10, no. 1, p. M110.000976, Jan. 2011, doi: 10.1074/mcp.M110.000976.

[121] S. C. Dillon, X. Zhang, R. C. Trievel, and X. Cheng, "The SET-domain protein superfamily: protein lysine methyltransferases," *Genome Biol*, vol. 6, no. 8, p. 227, 2005, doi: 10.1186/gb-2005-6-8-227.

[122] P. Rathert, X. Zhang, C. Freund, X. Cheng, and A. Jeltsch, "Analysis of the Substrate Specificity of the Dim-5 Histone Lysine Methyltransferase Using Peptide Arrays," *Chem Biol*, vol. 15, no. 1, pp. 5–11, Jan. 2008, doi: 10.1016/j.chembiol.2007.11.013.

[123] Y. Chang *et al.*, "MPP8 mediates the interactions between DNA methyltransferase Dnmt3a and H3K9 methyltransferase GLP/G9a," *Nat Commun*, vol. 2, p. 533, Nov. 2011, doi: 10.1038/ncomms1549.

[124] D. C. Schultz, K. Ayyanathan, D. Negorev, G. G. Maul, and F. J. Rauscher, "SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins," *Genes Dev*, vol. 16, no. 8, pp. 919–932, Apr. 2002, doi: 10.1101/gad.973302.

[125] P. F. South, K. M. Harmeyer, N. D. Serratore, and S. D. Briggs, "H3K4 methyltransferase Set1 is involved in maintenance of ergosterol homeostasis and resistance to Brefeldin A," *PNAS*, vol. 110, no. 11, pp. E1016–E1025, Mar. 2013, doi: 10.1073/pnas.1215768110.

[126] B. J. Bernard, N. Nigam, K. Burkitt, and V. Saloura, "SMYD3: a regulator of epigenetic and signaling pathways in cancer," *Clinical Epigenetics*, vol. 13, no. 1, p. 45, Feb. 2021, doi: 10.1186/s13148-021-01021-9.

[127] J. Tan, Y. Yan, X. Wang, Y. Jiang, and H. E. Xu, "EZH2: biology, disease, and structure-based drug discovery," *Acta Pharmacol Sin*, vol. 35, no. 2, pp. 161–174, Feb. 2014, doi: 10.1038/aps.2013.161.

[128] E. L. Greer and Y. Shi, "Histone methylation: a dynamic mark in health, disease and inheritance," *Nat Rev Genet*, vol. 13, no. 5, pp. 343–357, Apr. 2012, doi: 10.1038/nrg3173.

[129] P. Ø. Falnes, M. E. Jakobsson, E. Davydova, A. Ho, and J. Małecki, "Protein lysine methylation by seven-β-strand methyltransferases," *Biochem J*, vol. 473, no. 14, pp. 1995–2009, Jul. 2016, doi: 10.1042/BCJ20160117.

[130] M. Luo, "Chemical and Biochemical Perspectives of Protein Lysine Methylation," *Chem Rev*, vol. 118, no. 14, pp. 6656–6705, Jul. 2018, doi: 10.1021/acs.chemrev.8b00008.

[131] J. C. Black, C. Van Rechem, and J. R. Whetstine, "Histone lysine methylation dynamics: establishment, regulation, and biological impact," *Mol Cell*, vol. 48, no. 4, pp. 491–507, Nov. 2012, doi: 10.1016/j.molcel.2012.11.006.

[132] P. Colón-Bolea and P. Crespo, "Lysine methylation in cancer: SMYD3-MAP3K2 teaches us new lessons in the Ras-ERK pathway," *Bioessays*, vol. 36, no. 12, pp. 1162–1169, Dec. 2014, doi: 10.1002/bies.201400120.

[133] R. Hamamoto *et al.*, "SMYD3 encodes a histone methyltransferase involved in the proliferation of cancer cells," *Nat Cell Biol*, vol. 6, no. 8, pp. 731–740, Aug. 2004, doi: 10.1038/ncb1151.

[134] M. Nakakido, Z. Deng, T. Suzuki, N. Dohmae, Y. Nakamura, and R. Hamamoto, "PRMT6 increases cytoplasmic localization of p21CDKN1A in cancer cells through arginine methylation and makes more resistant to cytotoxic agents," *Oncotarget*, vol. 6, no. 31, pp. 30957–30967, Sep. 2015.

[135] R. Hamamoto, G. Toyokawa, M. Nakakido, K. Ueda, and Y. Nakamura, "SMYD2-dependent HSP90 methylation promotes cancer cell proliferation by regulating the chaperone complex formation," *Cancer Lett*, vol. 351, no. 1, pp. 126–133, Aug. 2014, doi: 10.1016/j.canlet.2014.05.014.

[136] A. Karytinos *et al.*, "A Novel Mammalian Flavin-dependent Histone Demethylase," *J Biol Chem*, vol. 284, no. 26, pp. 17775–17782, Jun. 2009, doi: 10.1074/jbc.M109.003087.

[137] R. J. Klose, E. M. Kallin, and Y. Zhang, “JmjC-domain-containing proteins and histone demethylation,” *Nat Rev Genet*, vol. 7, no. 9, pp. 715–727, Sep. 2006, doi: 10.1038/nrg1945.

[138] S. C. Kim *et al.*, “A High-Dimensional, Deep-Sequencing Study of Lung Adenocarcinoma in Female Never-Smokers,” *PLOS ONE*, vol. 8, no. 2, p. e55596, Feb. 2013, doi: 10.1371/journal.pone.0055596.

[139] P. Kim, P. Jia, and Z. Zhao, “Kinase impact assessment in the landscape of fusion genes that retain kinase domains: a pan-cancer study,” *Brief Bioinform*, vol. 19, no. 3, pp. 450–460, May 2018, doi: 10.1093/bib/bbw127.

[140] O. Ostersetzer-Biran *et al.*, “The First Mitochondrial Genomics and Evolution SMBE-Satellite Meeting: A New Scientific Symbiosis,” *Genome Biol Evol*, vol. 9, no. 11, pp. 3054–3058, Nov. 2017, doi: 10.1093/gbe/evx227.

[141] S. Kudithipudi and A. Jeltsch, “Role of somatic cancer mutations in human protein lysine methyltransferases,” *Biochim Biophys Acta*, vol. 1846, no. 2, pp. 366–379, Dec. 2014, doi: 10.1016/j.bbcan.2014.08.002.

[142] A. Bhardwaj, “Personalized cancer medicines,” *Science Translational Medicine*, vol. 7, no. 280, pp. 280ec48-280ec48, Mar. 2015, doi: 10.1126/scitranslmed.aaa9872.

[143] A. J. Brookes, “The essence of SNPs,” *Gene*, vol. 234, no. 2, pp. 177–186, Jul. 1999, doi: 10.1016/s0378-1119(99)00219-x.

[144] L. B. Barreiro, G. Laval, H. Quach, E. Patin, and L. Quintana-Murci, “Natural selection has driven population differentiation in modern humans,” *Nat Genet*, vol. 40, no. 3, pp. 340–345, Mar. 2008, doi: 10.1038/ng.78.

[145] J. C. Lee *et al.*, “Human SNP links differential outcomes in inflammatory and infectious disease to a FOXO3-regulated pathway,” *Cell*, vol. 155, no. 1, pp. 57–69, Sep. 2013, doi: 10.1016/j.cell.2013.08.034.

[146] J. Zeron-Medina *et al.*, “A polymorphic p53 response element in KIT ligand influences cancer risk and has undergone natural selection,” *Cell*, vol. 155, no. 2, pp. 410–422, Oct. 2013, doi: 10.1016/j.cell.2013.09.017.

[147] M. K. Halushka *et al.*, “Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis,” *Nat Genet*, vol. 22, no. 3, pp. 239–247, Jul. 1999, doi: 10.1038/10297.

[148] M. Cargill *et al.*, “Characterization of single-nucleotide polymorphisms in coding regions of human genes,” *Nat Genet*, vol. 22, no. 3, pp. 231–238, Jul. 1999, doi: 10.1038/10290.

[149] S. Sunyaev, V. Ramensky, and P. Bork, “Towards a structural basis of human non-synonymous single nucleotide polymorphisms,” *Trends Genet*, vol. 16, no. 5, pp. 198–200, May 2000, doi: 10.1016/s0168-9525(00)01988-0.

[150] C. Van Rechem and J. R. Whetstine, “Examining the Impact of Gene Variants on Histone Lysine Methylation,” *Biochim Biophys Acta*, vol. 1839, no. 12, pp. 1463–1476, Dec. 2014, doi: 10.1016/j.bbagrm.2014.05.014.

[151] Z. Wang and J. Moult, “SNPs, protein structure, and disease,” *Hum Mutat*, vol. 17, no. 4, pp. 263–270, Apr. 2001, doi: 10.1002/humu.22.

[152] C. George Priya Doss, R. Rajasekaran, and R. Sethumadhavan, “Computational identification and structural analysis of deleterious functional SNPs in MLL gene causing acute leukemia,” *Interdiscip Sci*, vol. 2, no. 3, pp. 247–255, Sep. 2010, doi: 10.1007/s12539-010-0007-z.

[153] N. Gautam, H. Verma, S. Choudhary, S. Kaur, and O. Silakari, “Functional relationship of SNP (Ala490Thr) of an epigenetic gene EZH2 results in the progression and poor survival of ER+/tamoxifen treated breast cancer patients,” *J Genet*, vol. 100, p. 86, 2021.

[154] N. D. Heintzman *et al.*, “Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome,” *Nat Genet*, vol. 39, no. 3, pp. 311–318, Mar. 2007, doi: 10.1038/ng1966.

[155] A. Barski *et al.*, “High-resolution profiling of histone methylations in the human genome,” *Cell*, vol. 129, no. 4, pp. 823–837, May 2007, doi: 10.1016/j.cell.2007.05.009.

[156] Y. B. Schwartz and V. Pirrotta, “Polycomb silencing mechanisms and the management of genomic programmes,” *Nat Rev Genet*, vol. 8, no. 1, pp. 9–22, Jan. 2007, doi: 10.1038/nrg1981.

[157] S. Jørgensen, G. Schotta, and C. S. Sørensen, “Histone H4 lysine 20 methylation: key player in epigenetic regulation of genomic integrity,” *Nucleic Acids Res*, vol. 41, no. 5, pp. 2797–2806, Mar. 2013, doi: 10.1093/nar/gkt012.

[158] M. Takawa *et al.*, “Histone lysine methyltransferase SETD8 promotes carcinogenesis by deregulating PCNA expression,” *Cancer Res*, vol. 72, no. 13, pp. 3217–3227, Jul. 2012, doi: 10.1158/0008-5472.CAN-11-3701.

[159] H.-S. Cho *et al.*, “Demethylation of RB regulator MYPT1 by histone demethylase LSD1 promotes cell cycle progression in cancer cells,” *Cancer Res*, vol. 71, no. 3, pp. 655–660, Feb. 2011, doi: 10.1158/0008-5472.CAN-10-2446.

[160] H. Kontaki and I. Talianidis, “Lysine methylation regulates E2F1-induced cell death,” *Mol Cell*, vol. 39, no. 1, pp. 152–160, Jul. 2010, doi: 10.1016/j.molcel.2010.06.006.

[161] M. Tachibana, K. Sugimoto, T. Fukushima, and Y. Shinkai, “Set domain-containing protein, G9a, is a novel lysine-preferring mammalian histone methyltransferase with hyperactivity and specific selectivity to lysines 9 and 27 of histone H3,” *J Biol Chem*, vol. 276, no. 27, pp. 25309–25317, Jul. 2001, doi: 10.1074/jbc.M101914200.

[162] J. Yang *et al.*, “Reversible methylation of promoter-bound STAT3 by histone-modifying enzymes,” *Proc Natl Acad Sci U S A*, vol. 107, no. 50, pp. 21499–21504, Dec. 2010, doi: 10.1073/pnas.1016147107.

[163] H. Hu *et al.*, “Set9, NF-κB, and microRNA-21 mediate berberine-induced apoptosis of human multiple myeloma cells,” *Acta Pharmacol Sin*, vol. 34, no. 1, pp. 157–166, Jan. 2013, doi: 10.1038/aps.2012.161.

[164] K. Subramanian *et al.*, “Regulation of estrogen receptor alpha by the SET7 lysine methyltransferase,” *Mol Cell*, vol. 30, no. 3, pp. 336–347, May 2008, doi: 10.1016/j.molcel.2008.03.022.

[165] H.-S. Cho *et al.*, “Enhanced HSP70 lysine methylation promotes proliferation of cancer cells through activation of Aurora kinase B,” *Nat Commun*, vol. 3, p. 1072, 2012, doi: 10.1038/ncomms2074.

[166] T. Salz, G. Li, F. Kaye, L. Zhou, Y. Qiu, and S. Huang, “hSETD1A regulates Wnt target genes and controls tumor growth of colorectal cancer cells,” *Cancer Res*, vol. 74, no. 3, pp. 775–786, Feb. 2014, doi: 10.1158/0008-5472.CAN-13-1400.

[167] R. Hamamoto *et al.*, “Enhanced SMYD3 expression is essential for the growth of breast cancer cells,” *Cancer Sci*, vol. 97, no. 2, pp. 113–118, Feb. 2006, doi: 10.1111/j.1349-7006.2006.00146.x.

[168] M. Kunizaki *et al.*, “The lysine 831 of vascular endothelial growth factor receptor 1 is a novel target of methylation by SMYD3,” *Cancer Res*, vol. 67, no. 22, pp. 10759–10765, Nov. 2007, doi: 10.1158/0008-5472.CAN-07-1132.

[169] P. K. Mazur *et al.*, “SMYD3 links lysine methylation of MAP3K2 to Ras-driven cancer,” *Nature*, vol. 510, no. 7504, pp. 283–287, Jun. 2014, doi: 10.1038/nature13320.

[170] F. P. Silva, R. Hamamoto, M. Kunizaki, M. Tsuge, Y. Nakamura, and Y. Furukawa, “Enhanced methyltransferase activity of SMYD3 by the cleavage of its N-terminal region in human cancer cells,” *Oncogene*, vol. 27, no. 19, pp. 2686–2692, Apr. 2008, doi: 10.1038/sj.onc.1210929.

[171] K. W. Foreman *et al.*, “Structural and functional profiling of the human histone methyltransferase SMYD3,” *PLoS One*, vol. 6, no. 7, p. e22290, 2011, doi: 10.1371/journal.pone.0022290.

[172] S.-W. Dong, H. Zhang, B.-L. Wang, P. Sun, Y.-G. Wang, and P. Zhang, “Effect of the downregulation of SMYD3 expression by RNAi on RIZ1 expression and proliferation of esophageal squamous cell carcinoma,” *Oncol Rep*, vol. 32, no. 3, pp. 1064–1070, Sep. 2014, doi: 10.3892/or.2014.3307.

[173] S. Wang, X. Luo, J. Shen, J. Zou, Y. Lu, and T. Xi, “Knockdown of SMYD3 by RNA interference inhibits cervical carcinoma cell growth and invasion in vitro,” *BMB Rep*, vol. 41, no. 4, pp. 294–299, Apr. 2008, doi: 10.5483/bmbrep.2008.41.4.294.

[174] S. Komatsu *et al.*, “Overexpression of SMYD2 relates to tumor cell proliferation and malignant outcome of esophageal squamous cell carcinoma,” *Carcinogenesis*, vol. 30, no. 7, pp. 1139–1146, Jul. 2009, doi: 10.1093/carcin/bgp116.

[175] H.-S. Cho *et al.*, “RB1 Methylation by SMYD2 Enhances Cell Cycle Progression through an Increase of RB1 Phosphorylation,” *Neoplasia*, vol. 14, no. 6, pp. 476–486, Jun. 2012.

[176] L. Piao *et al.*, “The histone methyltransferase SMYD2 methylates PARP1 and promotes poly(ADP-ribosyl)ation activity in cancer cells,” *Neoplasia*, vol. 16, no. 3, pp. 257–264, 264.e2, Mar. 2014, doi: 10.1016/j.neo.2014.03.002.

[177] X. Zhang *et al.*, “Regulation of estrogen receptor α by histone methyltransferase SMYD2-mediated protein methylation,” *Proc Natl Acad Sci U S A*, vol. 110, no. 43, pp. 17284–17289, Oct. 2013, doi: 10.1073/pnas.1307959110.

[178] M. Abu-Farha, J.-P. Lambert, A. S. Al-Madhoun, F. Elisma, I. S. Skerjanc, and D. Figeys, “The tale of two domains: proteomics and genomics analysis of SMYD2, a new histone methyltransferase,” *Mol Cell Proteomics*, vol. 7, no. 3, pp. 560–572, Mar. 2008, doi: 10.1074/mcp.M700271-MCP200.

[179] M. A. Brown, R. J. Sims, P. D. Gottlieb, and P. W. Tucker, "Identification and characterization of Smyd2: a split SET/MYND domain-containing histone H3 lysine 36-specific methyltransferase that interacts with the Sin3 histone deacetylase complex," *Mol Cancer*, vol. 5, p. 26, Jun. 2006, doi: 10.1186/1476-4598-5-26.

[180] D. Kang *et al.*, "The histone methyltransferase Wolf-Hirschhorn syndrome candidate 1-like 1 (WHSC1L1) is involved in human carcinogenesis," *Genes Chromosomes Cancer*, vol. 52, no. 2, pp. 126–139, Feb. 2013, doi: 10.1002/gcc.22012.

[181] Z. Zhou, R. Thomsen, S. Kahns, and A. L. Nielsen, "The NSD3L histone methyltransferase regulates cell cycle and cell invasion in breast cancer cells," *Biochem Biophys Res Commun*, vol. 398, no. 3, pp. 565–570, Jul. 2010, doi: 10.1016/j.bbrc.2010.06.119.

[182] S. M. Kim *et al.*, "Characterization of a novel WHSC1-associated SET domain protein with H3K4 and H3K27 methyltransferase activity," *Biochem Biophys Res Commun*, vol. 345, no. 1, pp. 318–323, Jun. 2006, doi: 10.1016/j.bbrc.2006.04.095.

[183] G. Toyokawa *et al.*, "Histone lysine methyltransferase Wolf-Hirschhorn syndrome candidate 1 is involved in human carcinogenesis through regulation of the Wnt pathway," *Neoplasia*, vol. 13, no. 10, pp. 887–898, Oct. 2011, doi: 10.1593/neo.11048.

[184] T. Ezponda *et al.*, "The histone methyltransferase MMSET/WHSC1 activates TWIST1 to promote an epithelial-mesenchymal transition and invasive properties of prostate cancer," *Oncogene*, vol. 32, no. 23, pp. 2882–2890, Jun. 2013, doi: 10.1038/onc.2012.297.

[185] J.-Y. Kim *et al.*, "Multiple-myeloma-related WHSC1/MMSET isoform RE-IIBP is a histone methyltransferase with transcriptional repression activity," *Mol Cell Biol*, vol. 28, no. 6, pp. 2023–2034, Mar. 2008, doi: 10.1128/MCB.02130-07.

[186] G. G. Wang, L. Cai, M. P. Pasillas, and M. P. Kamps, "NUP98-NSD1 links H3K36 methylation to Hox-A gene activation and leukaemogenesis," *Nat Cell Biol*, vol. 9, no. 7, pp. 804–812, Jul. 2007, doi: 10.1038/ncb1608.

[187] I. H. I. M. Hollink *et al.*, "NUP98/NSD1 characterizes a novel poor prognostic group in acute myeloid leukemia with a distinct HOX gene expression pattern," *Blood*, vol. 118, no. 13, pp. 3645–3656, Sep. 2011, doi: 10.1182/blood-2011-04-346643.

[188] M. Takawa *et al.*, "Validation of the histone methyltransferase EZH2 as a therapeutic target for various types of human cancer and as a prognostic marker," *Cancer Sci*, vol. 102, no. 7, pp. 1298–1305, Jul. 2011, doi: 10.1111/j.1349-7006.2011.01958.x.

[189] M. Kogure *et al.*, “The oncogenic polycomb histone methyltransferase EZH2 methylates lysine 120 on histone H2B and competes ubiquitination,” *Neoplasia*, vol. 15, no. 11, pp. 1251–1261, Nov. 2013, doi: 10.1593/neo.131436.

[190] J. M. Lee *et al.*, “EZH2 generates a methyl degron that is recognized by the DCAF1/DDB1/CUL4 E3 ubiquitin ligase complex,” *Mol Cell*, vol. 48, no. 4, pp. 572–586, Nov. 2012, doi: 10.1016/j.molcel.2012.09.004.

[191] E. Kim *et al.*, “Phosphorylation of EZH2 activates STAT3 signaling via STAT3 methylation and promotes tumorigenicity of glioblastoma stem-like cells,” *Cancer Cell*, vol. 23, no. 6, pp. 839–852, Jun. 2013, doi: 10.1016/j.ccr.2013.04.008.

[192] H.-S. Cho *et al.*, “Enhanced expression of EHMT2 is involved in the proliferation of cancer cells through negative regulation of SIAH1,” *Neoplasia*, vol. 13, no. 8, pp. 676–684, Aug. 2011, doi: 10.1593/neo.11512.

[193] B. Lehnertz *et al.*, “The methyltransferase G9a regulates HoxA9-dependent transcription in AML,” *Genes Dev*, vol. 28, no. 4, pp. 317–327, Feb. 2014, doi: 10.1101/gad.236794.113.

[194] X. Zhong *et al.*, “Overexpression of G9a and MCM7 in oesophageal squamous cell carcinoma is associated with poor prognosis,” *Histopathology*, vol. 66, no. 2, pp. 192–200, Jan. 2015, doi: 10.1111/his.12456.

[195] G. C. Issa *et al.*, “Predictors of outcomes in adults with acute myeloid leukemia and KMT2A rearrangements,” *Blood Cancer J.*, vol. 11, no. 9, Art. no. 9, Sep. 2021, doi: 10.1038/s41408-021-00557-6.

[196] T. G. Natarajan *et al.*, “Epigenetic regulator MLL2 shows altered expression in cancer cell lines and tumors from human breast and colon,” *Cancer Cell Int*, vol. 10, p. 13, Apr. 2010, doi: 10.1186/1475-2867-10-13.

[197] M. Xia *et al.*, “Downregulation of MLL3 in esophageal squamous cell carcinoma is required for the growth and metastasis of cancer cells,” *Tumour Biol*, vol. 36, no. 2, pp. 605–613, Feb. 2015, doi: 10.1007/s13277-014-2616-3.

[198] B. Li, H. Y. Liu, S. H. Guo, P. Sun, F. M. Gong, and B. Q. Jia, “Association of MLL3 expression with prognosis in gastric cancer,” *Genet Mol Res*, vol. 13, no. 3, pp. 7513–7518, Sep. 2014, doi: 10.4238/2014.September.12.18.

[199] M. Ruault, M. E. Brun, M. Ventura, G. Roizès, and A. De Sario, “MLL3, a new human member of the TRX/MLL gene family, maps to 7q36, a chromosome region frequently

deleted in myeloid leukaemia," *Gene*, vol. 284, no. 1–2, pp. 73–81, Feb. 2002, doi: 10.1016/s0378-1119(02)00392-x.

[200] D. O'Carroll *et al.*, "Isolation and characterization of Suv39h2, a second histone H3 methyltransferase gene that displays testis-specific expression," *Mol Cell Biol*, vol. 20, no. 24, pp. 9423–9433, Dec. 2000, doi: 10.1128/MCB.20.24.9423-9433.2000.

[201] K. Sone *et al.*, "Critical role of lysine 134 methylation on histone H2AX for γ-H2AX production and DNA repair," *Nat Commun*, vol. 5, no. 1, Art. no. 1, Dec. 2014, doi: 10.1038/ncomms6691.

[202] A. T. Nguyen, O. Taranova, J. He, and Y. Zhang, "DOT1L, the H3K79 methyltransferase, is required for MLL-AF9-mediated leukemogenesis," *Blood*, vol. 117, no. 25, pp. 6912–6922, Jun. 2011, doi: 10.1182/blood-2011-02-334359.

[203] K. M. Bernt and S. A. Armstrong, "A role for DOT1L in MLL-rearranged leukemias," *Epigenomics*, vol. 3, no. 6, pp. 667–670, Dec. 2011, doi: 10.2217/epi.11.98.

[204] J. Gao *et al.*, "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal," *Sci Signal*, vol. 6, no. 269, p. pl1, Apr. 2013, doi: 10.1126/scisignal.2004088.

[205] E. Cerami *et al.*, "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data," *Cancer Discov*, vol. 2, no. 5, pp. 401–404, May 2012, doi: 10.1158/2159-8290.CD-12-0095.

[206] J. S. Butler, E. Koutelou, A. C. Schibler, and S. Y. R. Dent, "Histone-modifying enzymes: regulators of developmental decisions and drivers of human disease," *Epigenomics*, vol. 4, no. 2, pp. 163–177, Apr. 2012, doi: 10.2217/epi.12.3.

[207] M. J. Campbell and B. M. Turner, "Altered histone modifications in cancer," *Adv Exp Med Biol*, vol. 754, pp. 81–107, 2013, doi: 10.1007/978-1-4419-9967-2_4.

[208] B. Xhemalce, "From histones to RNA: role of methylation in cancer," *Briefings in Functional Genomics*, vol. 12, no. 3, pp. 244–253, May 2013, doi: 10.1093/bfgp/els064.

[209] C. Zagni, U. Chiacchio, and A. Rescifina, "Histone methyltransferase inhibitors: novel epigenetic agents for cancer treatment," *Curr Med Chem*, vol. 20, no. 2, pp. 167–185, 2013, doi: 10.2174/092986713804806667.

[210] M.-W. Chen *et al.*, "H3K9 histone methyltransferase G9a promotes lung cancer invasion and metastasis by silencing the cell adhesion molecule Ep-CAM," *Cancer Res*, vol. 70, no. 20, pp. 7830–7840, Oct. 2010, doi: 10.1158/0008-5472.CAN-10-0833.

[211] J. Huang *et al.*, “G9a and Glp Methylate Lysine 373 in the Tumor Suppressor p53,” *J Biol Chem*, vol. 285, no. 13, pp. 9636–9641, Mar. 2010, doi: 10.1074/jbc.M109.062588.

[212] X.-G. Luo *et al.*, “Effects of SMYD3 overexpression on transformation, serum dependence, and apoptosis sensitivity in NIH3T3 cells,” *IUBMB Life*, vol. 61, no. 6, pp. 679–684, Jun. 2009, doi: 10.1002/iub.216.

[213] F. S. Poke, A. Qadi, and A. F. Holloway, “Reversing aberrant methylation patterns in cancer,” *Curr Med Chem*, vol. 17, no. 13, pp. 1246–1254, 2010, doi: 10.2174/092986710790936329.

[214] J. Huang, C. Plass, and C. Gerhauser, “Cancer chemoprevention by targeting the epigenome,” *Curr Drug Targets*, vol. 12, no. 13, pp. 1925–1956, Dec. 2011, doi: 10.2174/138945011798184155.

[215] R. A. Varier and H. T. M. Timmers, “Histone lysine methylation and demethylation pathways in cancer,” *Biochim Biophys Acta*, vol. 1815, no. 1, pp. 75–89, Jan. 2011, doi: 10.1016/j.bbcan.2010.10.002.

[216] Y. He, I. Korboukh, J. Jin, and J. Huang, “Targeting protein lysine methylation and demethylation in cancers,” *Acta Biochim Biophys Sin (Shanghai)*, vol. 44, no. 1, pp. 70–79, Jan. 2012, doi: 10.1093/abbs/gmr109.

[217] I. Hoffmann *et al.*, “The role of histone demethylases in cancer therapy,” *Mol Oncol*, vol. 6, no. 6, pp. 683–703, Dec. 2012, doi: 10.1016/j.molonc.2012.07.004.

[218] S. Hayami *et al.*, “Overexpression of LSD1 contributes to human carcinogenesis through chromatin regulation in various cancers,” *Int J Cancer*, vol. 128, no. 3, pp. 574–586, Feb. 2011, doi: 10.1002/ijc.25349.

[219] S. Hayami *et al.*, “Overexpression of the JmjC histone demethylase KDM5B in human carcinogenesis: involvement in the proliferation of cancer cells through the E2F/RB pathway,” *Molecular Cancer*, vol. 9, no. 1, p. 59, Mar. 2010, doi: 10.1186/1476-4598-9-59.

[220] M. Luo, “Inhibitors of protein methyltransferases as chemical tools,” *Epigenomics*, vol. 7, no. 8, pp. 1327–1338, 2015, doi: 10.2217/epi.15.87.

[221] T. E. McAllister, K. S. England, R. J. Hopkinson, P. E. Brennan, A. Kawamura, and C. J. Schofield, “Recent Progress in Histone Demethylase Inhibitors,” *J Med Chem*, vol. 59, no. 4, pp. 1308–1329, Feb. 2016, doi: 10.1021/acs.jmedchem.5b01758.

[222] M. Luo, “Current Chemical Biology Approaches to Interrogate Protein Methyltransferases,” *ACS Chem. Biol.*, vol. 7, no. 3, pp. 443–463, Mar. 2012, doi: 10.1021/cb200519y.

[223] R. A. Copeland, “Protein methyltransferase inhibitors as precision cancer therapeutics: a decade of discovery,” *Philos Trans R Soc Lond B Biol Sci*, vol. 373, no. 1748, p. 20170080, Jun. 2018, doi: 10.1098/rstb.2017.0080.

[224] T. Vougiouklakis, B. J. Bernard, N. Nigam, K. Burkitt, Y. Nakamura, and V. Saloura, “Clinicopathologic significance of protein lysine methyltransferases in cancer,” *Clin Epigenetics*, vol. 12, no. 1, p. 146, Oct. 2020, doi: 10.1186/s13148-020-00897-3.

[225] S. L. Berger, “The complex language of chromatin regulation during transcription,” *Nature*, vol. 447, no. 7143, pp. 407–412, May 2007, doi: 10.1038/nature05915.

[226] D. Jaiswal *et al.*, “Function of the MYND Domain and C-Terminal Region in Regulating the Subcellular Localization and Catalytic Activity of the SMYD Family Lysine Methyltransferase Set5,” *Mol Cell Biol*, vol. 40, no. 2, pp. e00341-19, Jan. 2020, doi: 10.1128/MCB.00341-19.

[227] N. Sirinupong, J. Brunzelle, J. Ye, A. Pirzada, L. Nico, and Z. Yang, “Crystal structure of cardiac-specific histone methyltransferase SmyD1 reveals unusual active site architecture,” *J Biol Chem*, vol. 285, no. 52, pp. 40635–40644, Dec. 2010, doi: 10.1074/jbc.M110.168187.

[228] Y. Jiang, N. Sirinupong, J. Brunzelle, and Z. Yang, “Crystal Structures of Histone and p53 Methyltransferase SmyD2 Reveal a Conformational Flexibility of the Autoinhibitory C-Terminal Domain,” *PLoS One*, vol. 6, no. 6, p. e21640, Jun. 2011, doi: 10.1371/journal.pone.0021640.

[229] N. Sirinupong, J. Brunzelle, E. Doko, and Z. Yang, “Structural insights into the autoinhibition and posttranslational activation of histone methyltransferase SmyD3,” *J Mol Biol*, vol. 406, no. 1, pp. 149–159, Feb. 2011, doi: 10.1016/j.jmb.2010.12.014.

[230] X. Tan, J. Rotllant, H. Li, P. DeDeyne, and S. J. Du, “SmyD1, a histone methyltransferase, is required for myofibril organization and muscle contraction in zebrafish embryos,” *PNAS*, vol. 103, no. 8, pp. 2713–2718, Feb. 2006, doi: 10.1073/pnas.0509503103.

[231] L. Wang *et al.*, “Structure of human SMYD2 protein reveals the basis of p53 tumor suppressor methylation,” *J Biol Chem*, vol. 286, no. 44, pp. 38725–38737, Nov. 2011, doi: 10.1074/jbc.M111.262410.

[232] Y. Jiang *et al.*, “Structural insights into estrogen receptor α methylation by histone methyltransferase SMYD2, a cellular event implicated in estrogen signaling regulation,” *J Mol Biol*, vol. 426, no. 20, pp. 3413–3425, Oct. 2014, doi: 10.1016/j.jmb.2014.02.019.

[233] L. H. T. Sakamoto, R. V. de Andrade, M. S. S. Felipe, A. B. Motoyama, and F. Pittella Silva, “SMYD2 is highly expressed in pediatric acute lymphoblastic leukemia and constitutes a bad prognostic factor,” *Leuk Res*, vol. 38, no. 4, pp. 496–502, Apr. 2014, doi: 10.1016/j.leukres.2014.01.013.

[234] X.-G. Luo *et al.*, “Histone methyltransferase SMYD3 promotes MRTF-A-mediated transactivation of MYL9 and migration of MCF-7 breast cancer cells,” *Cancer Letters*, vol. 344, no. 1, pp. 129–137, Mar. 2014, doi: 10.1016/j.canlet.2013.10.026.

[235] K. Gambetta, M. K. Al-Ahdab, M. N. Ilbawi, N. Hassaniya, and M. Gupta, “Transcription repression and blocks in cell cycle progression in hypoplastic left heart syndrome,” *Am J Physiol Heart Circ Physiol*, vol. 294, no. 5, pp. H2268-2275, May 2008, doi: 10.1152/ajpheart.91494.2007.

[236] J. Li *et al.*, “SMYD3 overexpression indicates poor prognosis and promotes cell proliferation, migration and invasion in non-small cell lung cancer,” *Int J Oncol*, vol. 57, no. 3, pp. 756–766, Jul. 2020, doi: 10.3892/ijo.2020.5095.

[237] A. D. Ferguson *et al.*, “Structural basis of substrate methylation and inhibition of SMYD2,” *Structure*, vol. 19, no. 9, pp. 1262–1273, Sep. 2011, doi: 10.1016/j.str.2011.06.011.

[238] L. T. Donlin *et al.*, “Smyd2 controls cytoplasmic lysine methylation of Hsp90 and myofilament organization,” *Genes Dev*, vol. 26, no. 2, pp. 114–119, Jan. 2012, doi: 10.1101/gad.177758.111.

[239] E. Fabini, E. Manoni, C. Ferroni, A. D. Rio, and M. Bartolini, “Small-molecule inhibitors of lysine methyltransferases SMYD2 and SMYD3: current trends,” *Future Med Chem*, vol. 11, no. 8, pp. 901–921, Apr. 2019, doi: 10.4155/fmc-2018-0380.

[240] A. Sajjad *et al.*, “Lysine methyltransferase Smyd2 suppresses p53-dependent cardiomyocyte apoptosis,” *Biochim Biophys Acta*, vol. 1843, no. 11, pp. 2556–2562, Nov. 2014, doi: 10.1016/j.bbamcr.2014.06.019.

[241] J. Wu *et al.*, “Biochemical characterization of human SET and MYND domain-containing protein 2 methyltransferase,” *Biochemistry*, vol. 50, no. 29, pp. 6488–6497, Jul. 2011, doi: 10.1021/bi200725p.

[242] E. Eggert *et al.*, “Discovery and Characterization of a Highly Potent and Selective Aminopyrazoline-Based in Vivo Probe (BAY-598) for the Protein Lysine Methyltransferase SMYD2,” *J. Med. Chem.*, vol. 59, no. 10, pp. 4578–4600, May 2016, doi: 10.1021/acs.jmedchem.5b01890.

[243] S. Komatsu *et al.*, “Overexpression of SMYD2 contributes to malignant outcome in gastric cancer,” *Br J Cancer*, vol. 112, no. 2, pp. 357–364, Jan. 2015, doi: 10.1038/bjc.2014.543.

[244] R. Ohtomo-Oda *et al.*, “SMYD2 overexpression is associated with tumor cell proliferation and a worse outcome in human papillomavirus-unrelated nonmultiple head and neck carcinomas,” *Hum Pathol*, vol. 49, pp. 145–155, Mar. 2016, doi: 10.1016/j.humpath.2015.08.025.

[245] S. Xu, C. Zhong, T. Zhang, and J. Ding, “Structure of human lysine methyltransferase Smyd2 reveals insights into the substrate divergence in Smyd proteins,” *J Mol Cell Biol*, vol. 3, no. 5, pp. 293–300, Oct. 2011, doi: 10.1093/jmcb/mjr015.

[246] N. Spellmon, X. Sun, N. Sirinupong, B. Edwards, C. Li, and Z. Yang, “Molecular Dynamics Simulation Reveals Correlated Inter-Lobe Motion in Protein Lysine Methyltransferase SMYD2,” *PLoS One*, vol. 10, no. 12, p. e0145758, 2015, doi: 10.1371/journal.pone.0145758.

[247] G. L. Blatch and M. Lässle, “The tetratricopeptide repeat: a structural motif mediating protein-protein interactions,” *Bioessays*, vol. 21, no. 11, pp. 932–939, Nov. 1999, doi: 10.1002/(SICI)1521-1878(199911)21:11<932::AID-BIES5>3.0.CO;2-N.

[248] K. Yamamoto *et al.*, “SMYD3 interacts with HTLV-1 Tax and regulates subcellular localization of Tax,” *Cancer Sci*, vol. 102, no. 1, pp. 260–266, Jan. 2011, doi: 10.1111/j.1349-7006.2010.01752.x.

[249] M. Abu-Farha *et al.*, “Proteomic analyses of the SMYD family interactomes identify HSP90 as a novel target for SMYD2,” *J Mol Cell Biol*, vol. 3, no. 5, pp. 301–308, Oct. 2011, doi: 10.1093/jmcb/mjr025.

[250] T. Voelkel, C. Andresen, A. Unger, S. Just, W. Rottbauer, and W. A. Linke, “Lysine methyltransferase Smyd2 regulates Hsp90-mediated protection of the sarcomeric titin springs and cardiac function,” *Biochim Biophys Acta*, vol. 1833, no. 4, pp. 812–822, Apr. 2013, doi: 10.1016/j.bbamcr.2012.09.012.

[251] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide

for 36 cancers in 185 countries," *CA Cancer J Clin*, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/caac.21492.

[252] L. X. Li, J. X. Zhou, J. P. Calvet, A. K. Godwin, R. A. Jensen, and X. Li, "Lysine methyltransferase SMYD2 promotes triple negative breast cancer progression," *Cell Death Dis*, vol. 9, no. 3, pp. 1–17, Feb. 2018, doi: 10.1038/s41419-018-0347-x.

[253] D. Basile *et al.*, "Androgen receptor in estrogen receptor positive breast cancer: Beyond expression," *Cancer Treat Rev*, vol. 61, pp. 15–22, Dec. 2017, doi: 10.1016/j.ctrv.2017.09.006.

[254] A. Makrariya and K. R. Pardasani, "Numerical study of the effect of non-uniformly perfused tumor on heat transfer in women's breast during menstrual cycle under cold environment," *Netw Model Anal Health Inform Bioinforma*, vol. 8, no. 1, p. 9, May 2019, doi: 10.1007/s13721-019-0189-1.

[255] J.-F. Arnal *et al.*, "Membrane and Nuclear Estrogen Receptor Alpha Actions: From Tissue Specificity to Medical Implications," *Physiol Rev*, vol. 97, no. 3, pp. 1045–1087, Jul. 2017, doi: 10.1152/physrev.00024.2016.

[256] K. Sakata, I. Yu, R. Katam, and T. Saito, "A Generalized Entropy-Production Consistent with Perturbation Response in Biological Regulatory Systems," vol. 24, pp. 29–34, 2021, doi: 10.24643/maebit.24.0_29.

[257] W. M. J. Obermann, "A motif in HSP90 and P23 that links molecular chaperones to efficient estrogen receptor α methylation by the lysine methyltransferase SMYD2," *J Biol Chem*, vol. 293, no. 42, pp. 16479–16487, Oct. 2018, doi: 10.1074/jbc.RA118.003578.

[258] M. Nakakido, Z. Deng, T. Suzuki, N. Dohmae, Y. Nakamura, and R. Hamamoto, "Dysregulation of AKT Pathway by SMYD2-Mediated Lysine Methylation on PTEN," *Neoplasia*, vol. 17, no. 4, pp. 367–373, Apr. 2015, doi: 10.1016/j.neo.2015.03.002.

[259] A. Saini *et al.*, "Analysis of Multimerin 1 (MMRN1) expression in ovarian cancer," *Mol Biol Rep*, vol. 47, no. 12, pp. 9459–9468, Dec. 2020, doi: 10.1007/s11033-020-06027-9.

[260] A. Kukita *et al.*, "Histone methyltransferase SMYD2 selective inhibitor LLY-507 in combination with poly ADP ribose polymerase inhibitor has therapeutic potential against high-grade serous ovarian carcinomas," *Biochem Biophys Res Commun*, vol. 513, no. 2, pp. 340–346, May 2019, doi: 10.1016/j.bbrc.2019.03.155.

[261] M. Robson *et al.*, “Olaparib for Metastatic Breast Cancer in Patients with a Germline BRCA Mutation,” *N Engl J Med*, vol. 377, no. 6, pp. 523–533, Aug. 2017, doi: 10.1056/NEJMoa1706450.

[262] A. Poveda *et al.*, “Olaparib tablets as maintenance therapy in patients with platinum-sensitive relapsed ovarian cancer and a BRCA1/2 mutation (SOLO2/ENGOT-Ov21): a final analysis of a double-blind, randomised, placebo-controlled, phase 3 trial,” *Lancet Oncol*, vol. 22, no. 5, pp. 620–631, May 2021, doi: 10.1016/S1470-2045(21)00073-5.

[263] A. S. Pires-Luís *et al.*, “Expression of histone methyltransferases as novel biomarkers for renal cell tumor diagnosis and prognostication,” *Epigenetics*, vol. 10, no. 11, pp. 1033–1043, 2015, doi: 10.1080/15592294.2015.1103578.

[264] S. Bagislar *et al.*, “Smyd2 is a Myc-regulated gene critical for MLL-AF9 induced leukemogenesis,” *Oncotarget*, vol. 7, no. 41, pp. 66398–66415, Oct. 2016, doi: 10.18632/oncotarget.12012.

[265] A. Zipin-Roitman *et al.*, “SMYD2 lysine methyltransferase regulates leukemia cell growth and regeneration after genotoxic stress,” *Oncotarget*, vol. 8, no. 10, pp. 16712–16727, Mar. 2017, doi: 10.18632/oncotarget.15147.

[266] W. Oliveira-Santos *et al.*, “Residual expression of SMYD2 and SMYD3 is associated with the acquisition of complex karyotype in chronic lymphocytic leukemia,” *Tumour Biol*, vol. 37, no. 7, pp. 9473–9481, Jul. 2016, doi: 10.1007/s13277-016-4846-z.

[267] N. Reynoird *et al.*, “Coordination of stress signals by the lysine methyltransferase SMYD2 promotes pancreatic cancer,” *Genes Dev*, vol. 30, no. 7, pp. 772–785, Apr. 2016, doi: 10.1101/gad.275529.115.

[268] S.-R. Zuo *et al.*, “Positive Expression of SMYD2 is Associated with Poor Prognosis in Patients with Primary Hepatocellular Carcinoma,” *J Cancer*, vol. 9, no. 2, pp. 321–330, Jan. 2018, doi: 10.7150/jca.22218.

[269] H. Ren *et al.*, “SMYD2-OE promotes oxaliplatin resistance in colon cancer through MDR1/P-glycoprotein via MEK/ERK/AP1 pathway,” *Onco Targets Ther*, vol. 12, pp. 2585–2594, 2019, doi: 10.2147/OTT.S186806.

[270] W. Xu, F. Chen, X. Fei, X. Yang, and X. Lu, “Overexpression of SET and MYND Domain-Containing Protein 2 (SMYD2) Is Associated with Tumor Progression and Poor Prognosis in Patients with Papillary Thyroid Carcinoma,” *Med Sci Monit*, vol. 24, pp. 7357–7365, Oct. 2018, doi: 10.12659/MSM.910168.

[271] R. Wang *et al.*, "Effects of SMYD2-mediated EML4-ALK methylation on the signaling pathway and growth in non-small-cell lung cancer cells," *Cancer Sci*, vol. 108, no. 6, pp. 1203–1209, Jun. 2017, doi: 10.1111/cas.13245.

[272] L. X. Li *et al.*, "Lysine methyltransferase SMYD2 promotes cyst growth in autosomal dominant polycystic kidney disease," *J Clin Invest*, vol. 127, no. 7, pp. 2751–2764, Jun. 2017, doi: 10.1172/JCI90921.

[273] S. Gao *et al.*, "The lysine methyltransferase SMYD2 methylates the kinase domain of type II receptor BMPR2 and stimulates bone morphogenetic protein signaling," *J Biol Chem*, vol. 292, no. 30, pp. 12702–12712, Jul. 2017, doi: 10.1074/jbc.M117.776278.

[274] H. Nguyen *et al.*, "LLY-507, a Cell-active, Potent, and Selective Inhibitor of Protein-lysine Methyltransferase SMYD2," *J Biol Chem*, vol. 290, no. 22, pp. 13641–13653, May 2015, doi: 10.1074/jbc.M114.626861.

[275] R. F. Sweis *et al.*, "Discovery of A-893, A New Cell-Active Benzoxazinone Inhibitor of Lysine Methyltransferase SMYD2," *ACS Med Chem Lett*, vol. 6, no. 6, pp. 695–700, Jun. 2015, doi: 10.1021/acsmedchemlett.5b00124.

[276] S. D. Cowen *et al.*, "Design, Synthesis, and Biological Activity of Substrate Competitive SMYD2 Inhibitors," *J Med Chem*, vol. 59, no. 24, pp. 11079–11097, Dec. 2016, doi: 10.1021/acs.jmedchem.6b01303.

[277] M. J. Thomenius *et al.*, "Small molecule inhibitors and CRISPR/Cas9 mutagenesis demonstrate that SMYD2 and SMYD3 activity are dispensable for autonomous cancer cell proliferation," *PLoS One*, vol. 13, no. 6, p. e0197372, 2018, doi: 10.1371/journal.pone.0197372.

[278] C. Zhang, S. A. Sultan, R. T, and X. Chen, "Biotechnological applications of S-adenosyl-methionine-dependent methyltransferases for natural products biosynthesis and diversification," *Bioresources and Bioprocessing*, vol. 8, no. 1, p. 72, Aug. 2021, doi: 10.1186/s40643-021-00425-y.

[279] T. Petrossian and S. Clarke, "Bioinformatic Identification of Novel Methyltransferases," *Epigenomics*, vol. 1, no. 1, pp. 163–175, Oct. 2009, doi: 10.2217/epi.09.3.

[280] J. D. Gary, W. J. Lin, M. C. Yang, H. R. Herschman, and S. Clarke, "The predominant protein-arginine methyltransferase from Saccharomyces cerevisiae," *J Biol Chem*, vol. 271, no. 21, pp. 12585–12594, May 1996, doi: 10.1074/jbc.271.21.12585.

[281] A. Niewmierzycka and S. Clarke, “S-Adenosylmethionine-dependent methylation in Saccharomyces cerevisiae. Identification of a novel protein arginine methyltransferase,” *J Biol Chem*, vol. 274, no. 2, pp. 814–824, Jan. 1999, doi: 10.1074/jbc.274.2.814.

[282] R. M. Kagan and S. Clarke, “Widespread occurrence of three sequence motifs in diverse S-adenosylmethionine-dependent methyltransferases suggests a common structure for these enzymes,” *Arch Biochem Biophys*, vol. 310, no. 2, pp. 417–427, May 1994, doi: 10.1006/abbi.1994.1187.

[283] M. Z. Ansari, J. Sharma, R. S. Gokhale, and D. Mohanty, “In silico analysis of methyltransferase domains involved in biosynthesis of secondary metabolites,” *BMC Bioinformatics*, vol. 9, no. 1, p. 454, Oct. 2008, doi: 10.1186/1471-2105-9-454.

[284] J. Söding, A. Biegert, and A. N. Lupas, “The HHpred interactive server for protein homology detection and structure prediction,” *Nucleic Acids Res*, vol. 33, no. Web Server issue, pp. W244-248, Jul. 2005, doi: 10.1093/nar/gki408.

[285] F. Li, X. Guo, D. Xiang, M. E. Pitt, A. Bainomugisa, and L. J. M. Coin, “Computational analysis and prediction of PE_PGRS proteins using machine learning,” *Computational and Structural Biotechnology Journal*, vol. 20, pp. 662–674, Jan. 2022, doi: 10.1016/j.csbj.2022.01.019.

[286] R. Wang, Z. Wang, H. Wang, Y. Pang, and T.-Y. Lee, “Characterization and identification of lysine crotonylation sites based on machine learning method on both plant and mammalian,” *Sci Rep*, vol. 10, no. 1, Art. no. 1, Nov. 2020, doi: 10.1038/s41598-020-77173-0.

[287] C. Meng, Y. Hu, Y. Zhang, and F. Guo, “PSBP-SVM: A Machine Learning-Based Computational Identifier for Predicting Polystyrene Binding Peptides,” *Front Bioeng Biotechnol*, vol. 8, p. 245, Mar. 2020, doi: 10.3389/fbioe.2020.00245.

[288] X. Liu, L. Wang, J. Li, J. Hu, and X. Zhang, “Mal-Prec: computational prediction of protein Malonylation sites via machine learning based feature integration,” *BMC genomics*, 2020, doi: 10.1186/s12864-020-07166-w.

[289] Z. Lv, S. Jin, H. Ding, and Q. Zou, “A Random Forest Sub-Golgi Protein Classifier Optimized via Dipeptide and Amino Acid Composition Features,” *Front Bioeng Biotechnol*, vol. 7, p. 215, 2019, doi: 10.3389/fbioe.2019.00215.

[290] R. Saidi, M. Maddouri, and E. Mephu Nguifo, “Protein sequences classification by means of feature extraction with substitution matrices,” *BMC Bioinformatics*, vol. 11, p. 175, Apr. 2010, doi: 10.1186/1471-2105-11-175.

[291] C. Xu and S. A. Jackson, “Machine learning and complex biological data,” *Genome Biol*, vol. 20, no. 1, p. 76, Apr. 2019, doi: 10.1186/s13059-019-1689-0.

[292] J.-X. Tan, H. Lv, F. Wang, F.-Y. Dao, W. Chen, and H. Ding, “A Survey for Predicting Enzyme Family Classes Using Machine Learning Methods,” *Curr Drug Targets*, vol. 20, no. 5, pp. 540–550, 2019, doi: 10.2174/1389450119666181002143355.

[293] C. Z. Cai, L. Y. Han, Z. L. Ji, and Y. Z. Chen, “Enzyme family classification by support vector machines,” *Proteins*, vol. 55, no. 1, pp. 66–76, Apr. 2004, doi: 10.1002/prot.20045.

[294] P. D. Dobson and A. J. Doig, “Predicting enzyme class from protein structure without alignments,” *J Mol Biol*, vol. 345, no. 1, pp. 187–199, Jan. 2005, doi: 10.1016/j.jmb.2004.10.024.

[295] L. Lu, Z. Qian, Y.-D. Cai, and Y. Li, “ECS: an automatic enzyme classifier based on functional domain composition,” *Comput Biol Chem*, vol. 31, no. 3, pp. 226–232, Jun. 2007, doi: 10.1016/j.compbiolchem.2007.03.008.

[296] E. Nasibov and C. Kandemir-Cavas, “Efficiency analysis of KNN and minimum distance-based classifiers in enzyme family prediction,” *Computational Biology and Chemistry*, vol. 33, no. 6, pp. 461–464, Dec. 2009, doi: 10.1016/j.compbiolchem.2009.09.002.

[297] J.-D. Qiu, J.-H. Huang, S.-P. Shi, and R.-P. Liang, “Using the concept of Chou’s pseudo amino acid composition to predict enzyme family classes: an approach with support vector machine based on discrete wavelet transform,” *Protein Pept Lett*, vol. 17, no. 6, pp. 715–722, Jun. 2010, doi: 10.2174/092986610791190372.

[298] T. Weirick, S. S. Sahu, R. Mahalingam, and R. Kaundal, “LacSubPred: predicting subtypes of Laccases, an important lignin metabolism-related enzyme class, using in silico approaches,” *BMC Bioinformatics*, vol. 15, no. 11, p. S15, Oct. 2014, doi: 10.1186/1471-2105-15-S11-S15.

[299] S. Amidi, A. Amidi, D. Vlachakis, N. Paragios, and E. I. Zacharaki, “Automatic single- and multi-label enzymatic function prediction by machine learning,” *PeerJ*, vol. 5, p. e3095, 2017, doi: 10.7717/peerj.3095.

[300] A. Dalkiran, A. S. Rifaioglu, M. J. Martin, R. Cetin-Atalay, V. Atalay, and T. Doğan, "ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature," *BMC Bioinformatics*, vol. 19, no. 1, p. 334, Sep. 2018, doi: 10.1186/s12859-018-2368-y.

[301] Z. Tao, B. Dong, Z. Teng, and Y. Zhao, "The Classification of Enzymes by Deep Learning," *IEEE Access*, vol. 8, pp. 89802–89811, 2020, doi: 10.1109/ACCESS.2020.2992468.

[302] J. Xu, H. Zhang, J. Zheng, P. Dovoedo, and Y. Yin, "eCAMI: simultaneous classification and motif identification for enzyme annotation," *Bioinformatics*, vol. 36, no. 7, pp. 2068–2075, Apr. 2020, doi: 10.1093/bioinformatics/btz908.

[303] Y. Wu, H. Tang, W. Chen, and H. Lin, "Predicting Human Enzyme Family Classes by Using Pseudo Amino Acid Composition," *Current Proteomics*, vol. 13, no. 2, pp. 99–104.

[304] L. Zhang, B. Dong, Z. Teng, Y. Zhang, and L. Juan, "Identification of Human Enzymes Using Amino Acid Composition and the Composition of k-Spaced Amino Acid Pairs," *BioMed Research International*, vol. 2020, p. e9235920, May 2020, doi: 10.1155/2020/9235920.

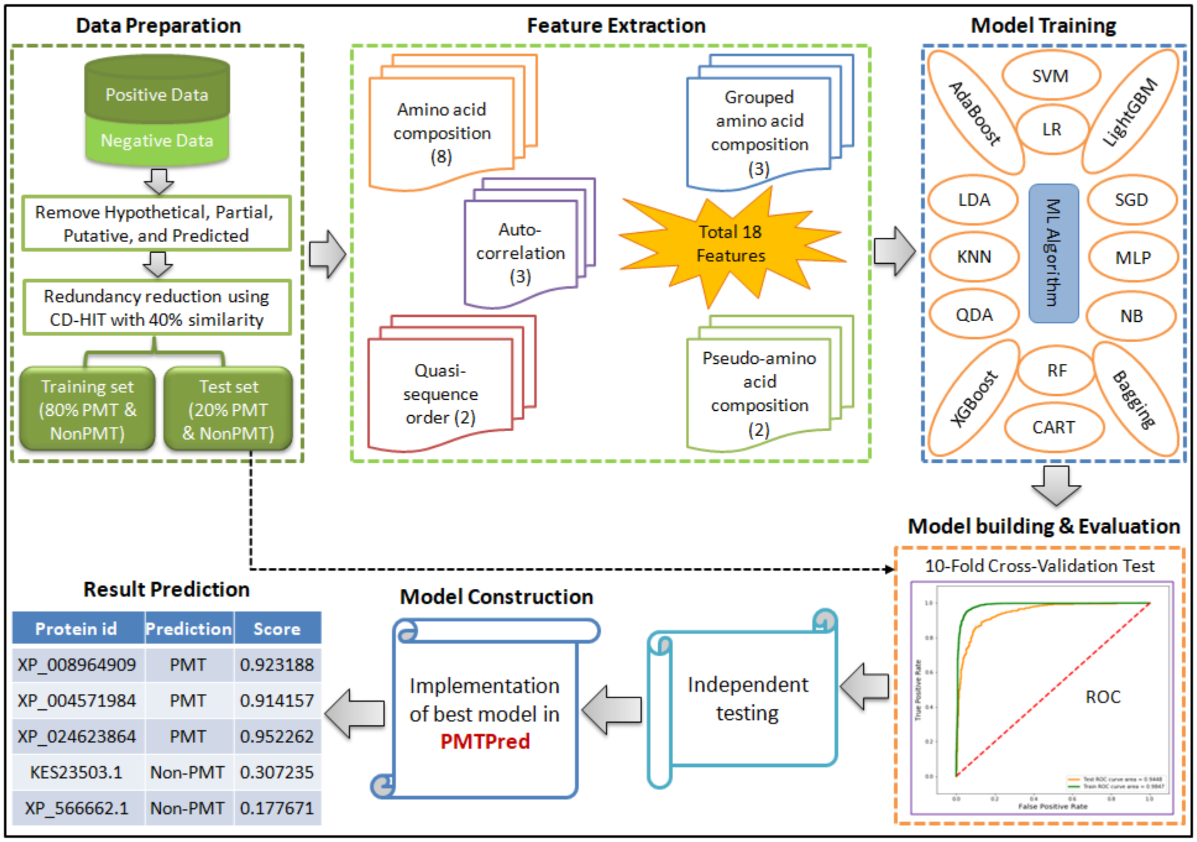
[305] H. Wang, Q. Xi, P. Liang, L. Zheng, Y. Hong, and Y. Zuo, "IHEC_RAAC: a online platform for identifying human enzyme classes via reduced amino acid cluster strategy," *Amino Acids*, vol. 53, no. 2, pp. 239–251, Feb. 2021, doi: 10.1007/s00726-021-02941-9.

[306] A. Garg and G. P. S. Raghava, "A machine learning based method for the prediction of secretory proteins using amino acid composition, their order and similarity-search," *In Silico Biol*, vol. 8, no. 2, pp. 129–140, 2008.

[307] H. Zhang, Q. Xi, S. Huang, L. Zheng, W. Yang, and Y. Zuo, "iSP-RAAC: Identify Secretory Proteins of Malaria Parasite Using Reduced Amino Acid Composition," *Comb Chem High Throughput Screen*, vol. 23, no. 6, pp. 536–545, 2020, doi: 10.2174/1386207323666200402084518.

CHAPTER- 2

Development of Machine-Learning Based Prediction Server for identifying Methyltransferases


Data Preparation
Positive Data
Negative Data
Remove Hypothetical, Partial, Putative, and Predicted
Redundancy reduction using CD-HIT with 40% similarity
Training set (80% PMT & NonPMT)
Test set (20% PMT & NonPMT)
Feature Extraction
Amino acid composition (8)
Grouped amino acid composition (3)
Auto-correlation (3)
Total 18 Features
Quasi-sequence order (2)
Pseudo-amino acid composition (2)
Model Training
SVM
AdaBoost
LR
LightGBM
LDA
SGD
ML Algorithm
KNN
MLP
QDA
NB
XGBoost
RF
Bagging
CART
Model building & Evaluation
10-Fold Cross-Validation Test
ROC
Independent testing
Model Construction
Implementation of best model in PMTPred
Result Prediction

Protein id	Prediction	Score
XP_008964909	PMT	0.923188
XP_004571984	PMT	0.914157
XP_024623864	PMT	0.952262
KES23503.1	Non-PMT	0.307235
XP_566662.1	Non-PMT	0.177671

ABSTRACT

PMTs are the groups of enzymes that help to catalyze the allocation of a methyl group to its substrates. These enzymes play a significant role in epigenetic regulation and are able to methylate various substrates with DNA, RNA, protein, and small-molecule secondary metabolites. Dysregulation of methyltransferases is intricate in different types of human cancers. This group of enzymes is also used commercially in different industries. However, in light of the well-recognized significance of PMTs, it becomes crucial to have reliable and fast methods for identifying these proteins. In this study, we developed a machine-learning-based method for the identification of PMTs. Various sequence-based features was calculated and model training was performed by using several machine-learning algorithms. A ten-fold cross-validation method was applied to train the models. The proposed SVM-based CKSAAP model was identified as the best model for the prediction of PMTs. Our optimal model achieved the highest accuracy of 87.94% with balance Sensitivity (88.8%) and Specificity (87.11%) with MCC of 0.759 and AUROC of 0.945. Finally, the best model was implemented in standalone software of PMTPred that will facilitate to predict PMTs. The PMTPred achieved 88.66% prediction accuracy with 85.00% sensitivity, and 92.33% specificity on the blind dataset. The standalone PMTPred software is available for download at http://www.bioinfoindia.org/PMTPred/ and https://github.com/ArvindYadav7/PMTPred for research and academic use.

2.1 INTRODUCTION

PMTs are enzymes that help to transfer a methyl group to their substrates with the help of universal methyl donor SAM. These enzyme groups play a significant role in the regulation of epigenetic mechanisms via the methylation of various substrates. PMTs have mainly been classified into two major classes known as 1) PRMTs that methylate arginine residues, and 2) PKMTs that methylate lysine residues of the protein substrate. Many diseases such as cancers, pulmonary disorders, and cardiovascular disease are associated with the dysregulation of PRMTs [1]. PKMTs are significantly associated with normal physiology and disease situations [2]. It is a type of serious and vigorous PTM that can control the protein function and stability [3].

PMTs play a crucial role in epigenetic regulation and transcriptional events through the methylation of arginine or lysine residues of histone and non-histone proteins. It has been reported that overexpression of PRMTs and PKMTs is associated with various types of human cancers [4], [5]. Thus, PMTs have begun as favorable and unique anti-cancer targets. Many drug development programs are aimed at the improvement of PMT-based small molecular inhibitors [6]. These molecule inhibitors have enormous importance in the interpretation of disease mechanisms and other biological functions of targeted enzymes. Many inhibitors have been discovered based on PRMTs, and PKMTs [7]–[9]. Up to now, seven PMTs inhibitors have reached human clinical trials as investigative cancer therapeutics [10]. PMTs are also having potential applications in biotechnology, chemical biology, and synthetic biology to produce a variety of synthetic and natural compounds [11], [12].

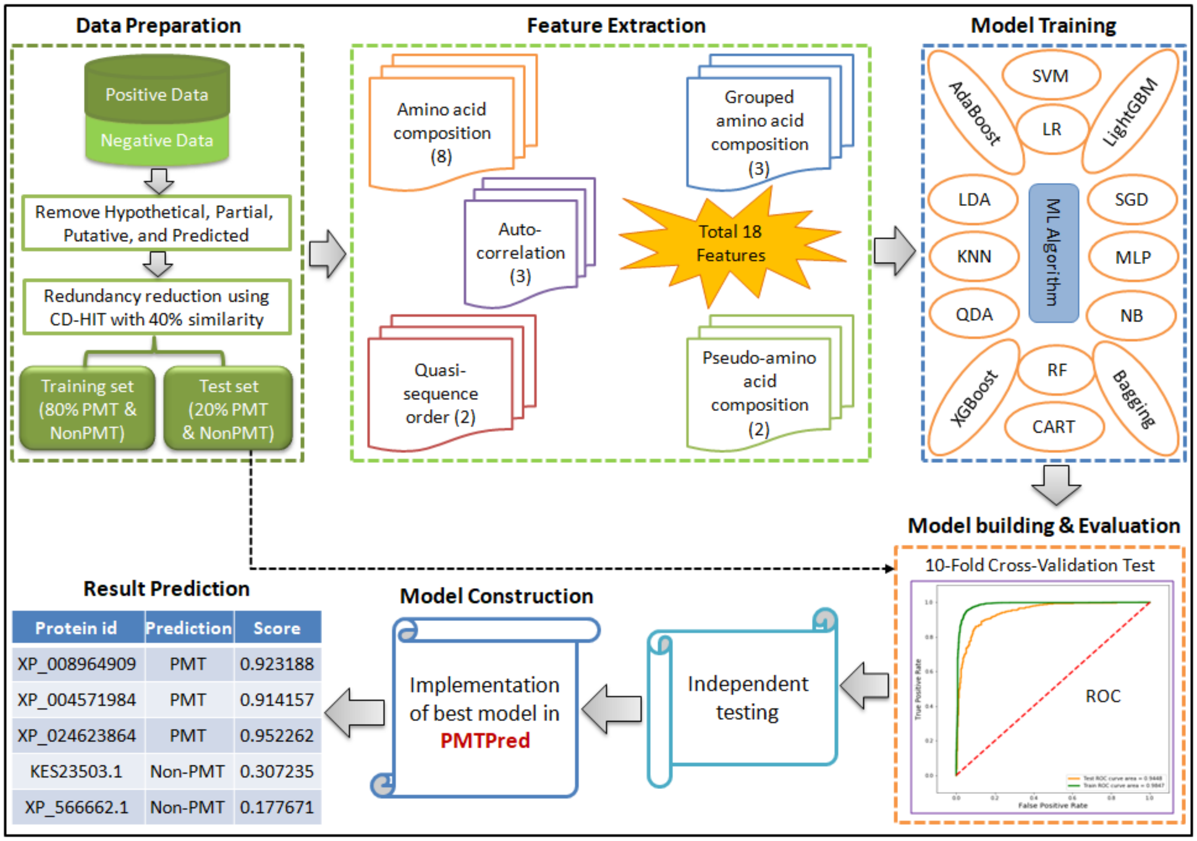
In the light of the well-recognized role of PMTs in anti-cancer therapy, pharmacological and biotechnological applications, it becomes crucial to have superficial methods for the identification of PMTs. Many diverse PMTs sequences are present, but most of these are assigned as hypothetical, putative, or probable functions based on sequence similarity. However, various existing experimental approaches for the identification of this protein are costly, time-consuming, labor-intensive, and require specialized equipment. Due to these obstacles, computational techniques for the prediction of various proteins emerged as a powerful alternative approach. Previously, some studies were conducted for the computational identification of methyltransferase reviewed by Petrossian and Clark [13]. Those studied used sequence-based, motif-based, and Hidden Markov Model (HMM) based search approaches. The methyltransferases have a topologically-distinct family of proteins. Thus, the similarity among primary sequences was found in only a small region of the protein. The search approaches based on sequence alignment are time-consuming and low-sensitive that can lead to problems in the prediction of low similarity proteins. Therefore, we need an advanced method that is based on sequence information rather than simple similarity of amino acids. There is no machine-learning-based prediction method to best our knowledge available for PMTs to date. So, in this study, we have developed a sequence-based new computational method using machine learning techniques that would be helpful for the efficient and accurate prediction of PMTs. We have used different amino acid-based compositions and physicochemical-based features to train the classifiers to facilitate the method. In the obtained results, the composition of k-spaced amino acid pairs (CKSAAP) represents maximum prediction accuracy with the Support Vector Machine (SVM) classifier. The CKSAAP feature has been previously used in several protein prediction problems associated with PTMs [14]–[18]. So, in this study, we offer to utilize a 1-space CKSAAP feature to construct the model for the prediction of PMTs. The final model-based standalone tool named PMTPred is available for academic and research use. Thus, PMTPred could assist as a potent tool for the prediction of PMTs.

2.2 METHODOLOGY

2.2.1 Dataset preparation

The used dataset in the present study was extracted from the public database NCBI (https://www.ncbi.nlm.nih.gov/), which includes all protein methyltransferase. In this manner, 41146 methyltransferase sequences were retrieved from different organisms. Sequences having a length of <50 amino acids were removed. Hypothetical, partial, putative, predicted, and sequences having non-amino acid characters were also removed, and finally identified a total of 17280 methyltransferase protein sequences. Negative data is a prerequisite to developing a supervised machine-learning-based model. Therefore, to create a negative dataset, a total of 33061 sequences of non-methyltransferase proteins were retrieved from the NCBI database. A similar protocol used for methyltransferase was followed and identified 3009 non-methyltransferase sequences. To obtain a non-redundant set of protein sequences CD-HIT program [19] was used at a 40% sequence identity threshold. Thus, CD-HIT results in 2862 methyltransferase and 3009 non-methyltransferase protein sequences. To construct a balanced dataset, we randomly selected 2862 proteins out of 3009 non-methyltransferase proteins. Therefore final dataset has 2862 methyltransferase and 2862 non-methyltransferase proteins. Hereafter, we called instances of the final dataset a positive and negative dataset, respectively.

Further, to create training and test sets, the 80:20 ratios were considered. We have used 80% of the total number of sequences as a training set and the remaining 20% as a test set from both positive and negative datasets respectively. Here, 2290 and 572 sequences were used in training and testing datasets respectively. We have selected an equal number of positive and negative sets in both training and testing data because machine learning can produce an unbiased result on a balanced dataset [20]. The flow diagram of the used methodology is represented in Figure 2.1.


Data Preparation
Positive Data
Negative Data
Remove Hypothetical, Partial, Putative, and Predicted
Redundancy reduction using CD-HIT with 40% similarity
Training set (80% PMT & NonPMT)
Test set (20% PMT & NonPMT)
Feature Extraction
Amino acid composition (8)
Grouped amino acid composition (3)
Auto-correlation (3)
Total 18 Features
Quasi-sequence order (2)
Pseudo-amino acid composition (2)
Model Training
SVM
AdaBoost
LR
LightGBM
LDA
SGD
KNN
ML Algorithm
MLP
QDA
NB
XGBoost
RF
Bagging
CART
Model building & Evaluation
10-Fold Cross-Validation Test
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC
Test ROC curve area = 0.9448
Train ROC curve area = 0.9847
Independent testing
Model Construction
Implementation of best model in PMTPred
Result Prediction

Protein id	Prediction	Score
XP_008964909	PMT	0.923188
XP_004571984	PMT	0.914157
XP_024623864	PMT	0.952262
KES23503.1	Non-PMT	0.307235
XP_566662.1	Non-PMT	0.177671

Figure 2.1: Diagram of the used methodology of the present work. The main steps contained: data preparation, feature extraction, model training, model building and evaluation, independent test, and model construction.

2.2.2 Blind dataset

A blind dataset of PMTs is created from the UniProtKB database (https://www.uniprot.org/). We selected well-annotated PMTs sequences from various organisms whose 3D structures are present. Sequence comparison analysis was performed through the CD-HIT-2D program [19] to ensure any similar sequence was not present in the main dataset (training and test data). Then, randomly selected a set of 300 PMT sequences with similar number of non-PMT sequences and created a blind dataset to check the performance of PMTPred.

2.2.3 Feature extraction

To train the prediction model, we have extracted various types of sequence-based feature descriptors. Different descriptor representation techniques have been used to convert protein sequences into numerical features. The feature extraction techniques are very important to build computational predictors [21]–[23]. In this paper, we have extracted five groups of descriptors such as amino acid composition (AAC), grouped amino acid composition (GAAC), autocorrelation, pseudo-amino acid composition (PAAC), and quasi-sequence-order (QSOrder). Each group may also include several feature extraction techniques. So, in the present study, a total of 18 descriptors were calculated based on above mentioned five feature groups (Table 2.1) through a standalone Python-based toolkit of iFeature [24]. A brief description of the five feature groups is described below.

2.2.3.1 Amino acid composition

The simplest protein feature is amino acid composition (AAC) which is mostly used in protein structure and function prediction. In a sequence, it calculates the amino acid occurrence of each type of residue [25]. Another important composition feature is dipeptide composition (DPC). It calculates the dimer composition of amino acids. We have also calculated some other composition based features such as dipeptide deviations from the expected mean (DDE) and k-spaced amino acid pairs (CKSAAP). We also calculated the amino acid distribution patterns of PMTs sequences using Composition/Transition/Distribution (C/T/D) of different physicochemical properties. These descriptors are Conjoint triad (CTriad), Composition (CTDC), Distribution (CTDD), and Transition (CTDT).

2.2.3.2 Autocorrelation

Moreau and Broto presented autocorrelation-based features which are depend on the distribution of various properties of amino acid [26]. Here, we have utilized three different types of feature descriptors such as schemes Geary, Moran, and Normalized Moreau-Broto (NMBroto).

2.2.3.3 Grouped amino acid composition

In this feature descriptor, the sequence feature is categorized into five groups such as, aromatic, aliphatic, neutral, positive, negative charged amino acid based on their physicochemical properties then calculate the Grouped amino acid composition (GAAC) descriptors [27]. We also determined the GAAC, as well as the Grouped dipeptide composition (GDPC) and the Composition of k-spaced amino acid group pairs (CKSAAGP). These characteristics can provide a thorough result relating to charge, hydrophobicity, and other factors.

2.2.3.4 Quasi-sequence-order

This auspicious descriptor can pass over the enormous exertion of peptide/protein sequences due to the better approach of covariant discriminant (permutations and combinations). It also enables us to make better predictions by utilizing a variety of protein features [28]. In this group, we have calculated the descriptors, Quasi-sequence-order (QSOrder) and Sequence-order-coupling number (SOCNumber).

2.2.3.5 Pseudo-amino acid composition

The pseudo amino acid composition (PseAAC) is presented by the K. C. Chou in the year of 2001 [29]. In comparison to the original AAC technique, this method also describes the protein by using an amino-acid occurrence matrix of the protein sequence, which has very low sequence homology. In this group, we have not only calculated the PseAAC but we have also calculated Amphiphilic PAAC (APAAC) feature.

Table 2.1: Various feature descriptors and the number of descriptors in each group is calculated by iFeature.

Feature group	**Descriptor name**	**Number of attributes**	**Reference**
AAC	AAC	20	[30]
	CKSAAP	2400	[31], [32]
	DDE	400	[33]
	DPC	400	[30], [33]
	CTDC	39	[34]–[38]
	CTDD	39	[34]–[38]
	CTDT	343	[34]–[38]
	CTriad	195	[39]
Autocorrelation	Geary	240	[40]
	Moran	240	[41], [42]
	NMBroto	240	[43]
GAAC	CKSAAGP	150	[22], [44]
	GDPC	25	[22], [44]

	GAAC	5	[22], [44]
QSOrder	QSOrder	100	[45]–[47]
	SOCNumber	60	[45]–[47]
PAAC	PAAC	50	[48], [49]
	APAAC	80	[48], [49]

In the present study, we have used the CKSAAP feature vector to represent the protein sequence in our proposed model. CKSAAP is an extensively used feature in bioinformatics problems [15], [50]–[53]. It calculates the occurrence of pairs of amino acids separated by any k number of amino acid residues. Here, ranges for k were chosen from 0 to 5; meanwhile, CKSAAP gives the same result as DPC when k equals 0; therefore, k ranges from 1 to 5 are considered. For example, if k=0, the residue pair for 0-spaced can be expressed as AA, AC, AD, ..., YY (provide 400 amino acid residue pairs), and if k=1, then 1-spaced residue pair can be represented as AxA, AxC, AxD, ..., YxY (provide 800 amino acid residue pairs). The CKSAAP feature is defined as:

$$k = 0\left(\frac{N[AA]}{N_0}, \frac{N[AC]}{N_0}, \frac{N[AD]}{N_0}, \ldots\ldots, \frac{N[YY]}{N_0}\right)400$$

$$k = 1\left(\frac{N[AxA]}{N_1}, \frac{N[AxC]}{N_1}, \frac{N[AxD]}{N_1}, \ldots\ldots, \frac{N[YxY]}{N_1}\right)400$$

$$k = 2\left(\frac{N[AxxA]}{N_2}, \frac{N[AxxC]}{N_2}, \frac{N[AxxD]}{N_2}, \ldots\ldots, \frac{N[YxxY]}{N_2}\right)400$$

where ‘x’ stands for any of the 20 amino acids; N_k is calculated as $N_k = L - (k + 1)$, k = 1, 2, 3..., where L represent the protein sequence length. The final feature vector was calculated by concatenating the specific feature vectors. At k = 0, 1, 2, 3, 4, and 5, it generates a 2400-dimensional feature vector.

2.2.4 Machine-learning method and Model construction

We used different machine-learning algorithms executed with the Python library scikit-learn (https://scikit-learn.org/). Scikit-learn library allows to development of models by using various machine-learning algorithms. Here, we used 14 machine-learning algorithms for classification, namely Support Vector Machine (SVM) [54], Latent Dirichlet allocation (LDA) [55], k-Nearest

Neighbors (KNN) [56], Logistic Regression (LR) [57], Classification and Regression Tree (CART) [58], Naive Bayes (NB) [58], Random Forest (RF) [59], Multilayer Perceptron (MLP) [59], Adaptive Boosting (AdaBoost) [61], light gradient boosting machine (LightGBM) [60], extreme gradient boosting (XGBoost) [61], Stochastic Gradient Descent (SGD) [62], bootstrap aggregating (Bagging) [63], and Quadratic Discriminant Analysis (QDA) [55]. The hyperparameter tuning was applied on training set to find the best possible combination of the parameter using the 10-fold cross-validation technique for each classification algorithm to obtain the best model.

2.2.5 Model description

The detail descriptions about the algorithms with tuned parameters are provided below:

- **Support vector machine:** The position of the data in regard to a boundary between the positive and negative classes is used by SVM to classify the data. This border line is called as the hyperplane that help to maximize the distance between both classes. By extending the idea of constructional risk minimization, SVM addresses the issue of overfitting and looks at the ideal hyperplane between the two classes. The interaction between dependent and independent variables is described by this method [54]. It is widely used machine learning method in the field of bioinformatics [64]–[68]. The radial basis function (RBF) kernel has been used in the SVM method. We have used the grid search method, and optimization has been performed for penalty parameter C (from 1.0 to 15.0 at step 1), and the kernel parameter gamma (from 0.00001 to 5.0 at step 0.1).
- **Latent dirichlet allocation:** One of the most widely used techniques for topic modelling is latent Dirichlet allocation. Each document has a variety of words, and certain words can be connected to particular topics. The LDA's goal is to identify the themes to which the document belongs based on the terms that are present in it. It makes the assumption that texts on related subjects will employ the same set of vocabulary [55]. In order to map the probability distribution over latent themes and topics that are probability distribution, this is necessary. The hyperparameters tuning were performed for alpha 0.1 to 0.5 and beta 0.01 to 0.05.

- **k-nearest neighbours:** It is a non-parametric supervised learning technique that groups together previously known data and uses that information to categorise new data based on similarity. The linear decomposition approach, which retrieves neighbour data using the Euclidean distance between data points, operates on the coordinate plane [56]. KNN method is primarily utilised for classification issues. The hyperparameter tuning was performed for n_neighbors (from 1 to 10 at step 1), leaf size (from 20 to 40 at step 1) and p (1, 2).
- **Logistic regression:** The powerful linear combination of variables that is most likely to affect the observed outcome is presented iteratively in LR. In this approach, a logistic function is used to simulate the probability describing the potential outcomes of a single track. The sigmoid function is used to calculate the likelihood of a label [57]. It is frequently employed for binary categorization issues like true or false, positive or negative, etc. The hyperparameters tuning was performed for penalty (l1 and l2) and C (0.01, 0.1, 1.0, 10, 15).
- **Classification and regression tree:** A predictive method called CARD shows how the values of a target variable can be anticipated based on the values of other variables. It is a decision tree, and each fork is divided into a predictor, with predictions for the target variable at each node at the end. A classification algorithm known as CARD is necessary to construct a decision tree using Gini's impurity index [58]. When the dataset needs to be divided into classes that correspond to the response variable, CARD is utilised. Positive or negative classifications might apply in numerous situations. The hyperparameter tuning was performed for max_depth (from 2 to 10 at step 1).
- **Naïve bayes:** The naive assumption that each feature is independent of the others was employed in the naive Bayes technique, which applied Bayes' Theorem to determine the conditional probability based on prior knowledge [59]. For each new classification operation, the nave Bayes algorithm must re-scan the entire dataset, which could slow it down. As a result, it functions well with less training data. It was run at default parameters.
- **Random forest:** Random forest is a popular algorithm that belongs to the supervised learning technique. It is being developed and tested as a collection of decision trees. With

several estimators and automatic variable selection, RF can handle big datasets. According to reports, it offers fair estimates [59]. The hyperparameter tuning was performed for tree depth as, trees range from 50 to 500 at step 50.

- **Multilayer perceptron:** The multilayer perceptron is feed forward artificial neural network (ANN). From a set of input layers, it creates a set of output layers. The real computational engine of the MLP consists of an arbitrary number of hidden layers that are sandwiched between the input and output layers. Data in MLP moves forward from the input layer to the output layer. Backpropagation was a supervised learning method employed by MLP [59]. The hyperparameters tuning was performed for hidden layer sizes (100-150), epochs (300) and activation (relu).
- **Adaptive boosting:** One of the earliest boosting algorithms to employ the ensemble method to address the binary classification issue was AdaBoost. Adaboost aids in fusing a number of poor classifiers into one powerful classifier. The basic idea behind boosting methods is that after creating a model using the training dataset, we create a second model to fix any mistakes in the original one. This process is repeated until the mistakes are reduced and the dataset can be accurately forecasted [61]. The base estimator hyperparameter was tuned in AdaBoost and used at 10, 50, and 100.
- **Light gradient boosting machine:** Gradient boosting has a framework and a variation called LightGBM. Light GBM is based on Decision tree methods, just as another gradient boosting method. We can decrease memory utilisation and boost efficiency with the aid of Light GBM. LightGBM expands leaf-wise but other gradient boosting frameworks increase level-wise, which is the primary distinction between them [60]. The hyperparameters tuning were performed for leaves range (from 20 to 100 at step 10), depth range (from 15 to 55 at step 10) and learning rate range (from 0.01 to 0.15 at step 0.02).
- **Extreme gradient boosting:** When training data is utilised to forecast a target variable, supervised learning problems are addressed with XGBoost [61]. It is built on the principles of gradient boosting framework and designed to "push the extreme of the computation limits of machines to provide a *scalable*, *portable* and *accurate* library." The

hyperparameters tuning were performed for depth range (from 3 to 10 at step 1) and learning rate range (from 0.01 to 0.3 at step 0.05).

- **Stochastic gradient descent:** It is a straightforward but incredibly effective method for optimising linear classifiers. It employs a straightforward stochastic gradient descent learning procedure that supports various classification loss functions and penalties [61]. The optimization procedure stochastic gradient descent is frequently used in machine learning applications to identify the model parameters that best match the expected and actual outputs. It is a crude but effective method. The hyperparameters tuning were performed for learning rate (0.1, 0.01, 0.001), estimators (10, 50, 100) and max depth (3, 5).
- **Bootstrap aggregating:** A common ensemble technique called bootstrap aggregation, also known as bagging, fits a decision tree to various bootstrap samples of the training dataset. In ensemble machine learning, bagging uses a number of weak models to aggregate the predictions and choose the best one [63]. The hyperparameter tuning were performed for estimator range (from 10 to 100 at step 10).
- **Quadratic discriminant analysis:** QDA is similar to LDA based on the fact that there is an assumption of the observations being drawn from a normal distribution. The difference is that QDA assumes that each class has its own covariance matrix, while LDA does not [55]. The hyperparameters tuning were performed for solver (‘svd’) and shrinkage (from 0 to 1 at step 0.01).

2.2.6 Model performance evaluation

The k-fold cross-validation technique was employed to measure the performance of classifiers. In this technique, each instance of the dataset is used to be tested once for prediction. Therefore, it is a purely unbiased method for the testing model efficiency. In this study, a 10-fold cross-validation technique has been used. In this procedure, the dataset is distributed into 10-equal subsets one subset at a time considered a test set, and the rest of the nine subsets were combined and used as a training set. This whole procedure is repeated for 10-times so that every subset can be considered as a test set for at least one time. Finally, the scores of evaluation metrics of these

10 groups are averaged that evaluates the performance of the trained model. Then, independent test set (data that is not present in training set) was also used for validate the models performance.

Further, to estimate the model's performance, we have calculated various metrics like sensitivity (Sen), specificity (Spe), accuracy (Acc), precision (Pre), and Matthews correlation coefficient (MCC). All these parameters are computed by the utilization of the following formulas.

$$Sen = \frac{TP}{(TP + FN)} \times 100 \tag{1}$$

$$Spe = \frac{TN}{(TN + FP)} \times 100 \tag{2}$$

$$Acc = \frac{(TP + TN)}{(TP + FP + TN + FN)} \times 100 \tag{3}$$

$$\mathrm{Pr}\,e = \frac{TP}{(TP + FP)} \times 100 \tag{4}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{5}$$

where, TP, TN, FP, and FN are the True Positives (correctly classified PMT), True Negatives (correctly classified non-PMT), False Positives (incorrect classification of non-PMT as PMT), and False Negatives (incorrect classification of PMT as non-PMT), respectively. Besides that, the area under receiver operating characteristics (ROC) curve (AUROC) and area under the precision-recall curve (AUPRC) were also calculated. Graph for AUROC and AUPRC has been plotted to estimate the visual assessment of the model. The ROC curve represents the false positive rate vs. true-positive rate, and the PR curve represents precision vs. recall. AUROC and AUPRC also present the performance measure, and an AUROC and AUPRC are close to 1, which signifies the best prediction of the model.

2.3 RESULTS AND DISCUSSION

2.3.1 Performance of different machine-learning algorithms

The biological activity of any protein depends upon its basic sequence [69]; therefore we have used the sequence-based feature for model development. Various machine-learning algorithms were applied with different features on training data to train the models. It is because machine-learning methods are problem-specific [70]. Therefore, method selection requires exploring different methods on the same dataset for the selection of the best one. Here, we have considered 18 features calculated from five different groups of descriptors. Then, each feature is further trained with 14 machine-learning algorithms and measured the performance accuracy using a 10-fold cross-validation technique, and independent testing. The parameters of all classification algorithms have been carefully optimized to achieve the best result. The performance of all 14 models with 18 features in term of accuracy using training dataset is shown in Annexure table 2.1. Then independent test dataset was applied to evaluate the performance of all models and result is shown in Annexure table 2.2 and Figure 2.2.

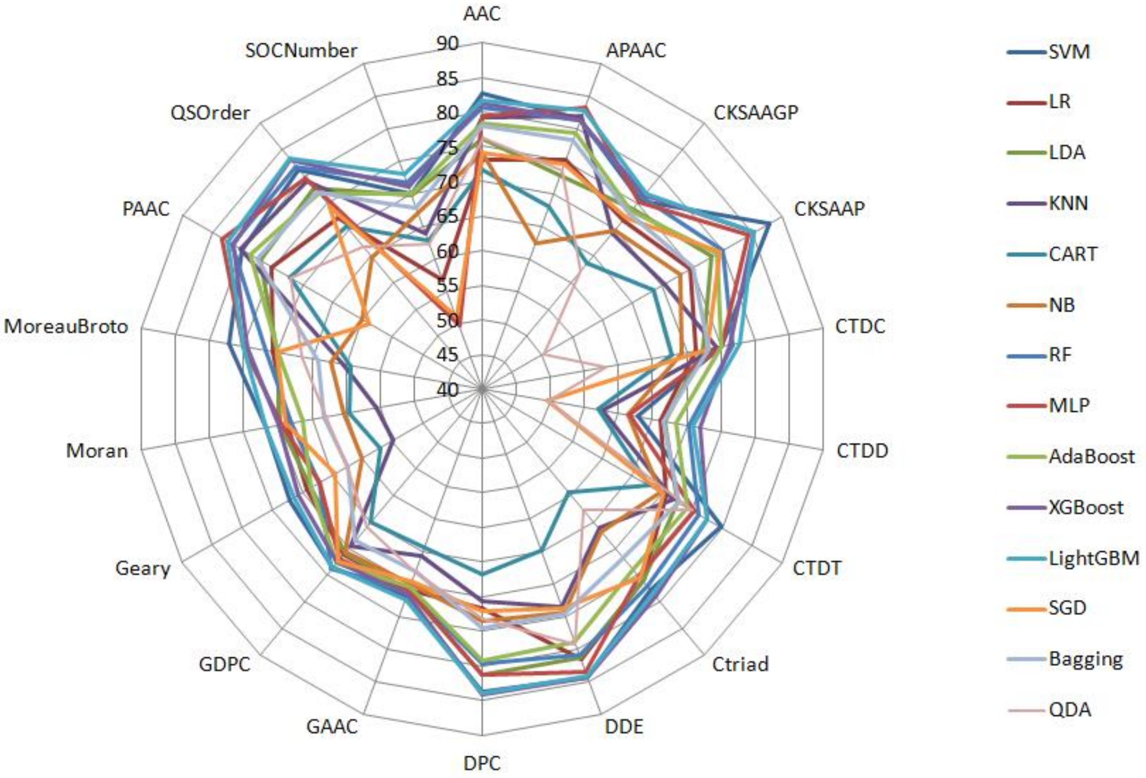


Figure 2.2: Performance of different machine learning algorithms trained with various feature sets in terms of accuracy on the 10-fold cross-validation test.

Comparative analysis of all 14 methods using 18 features represents that the top classification proficiency was achieved using the 5-spaced CKSAAP feature with SVM classifier, followed by XGBoost, LightGBM, and MLP as shown in Figure 2.2. It achieved the prediction accuracy of 87.02 and 87.94% on the train and test set respectively. CKSAAP also outperform with XGBoost (85.32% accuracy) and LightGBM (85.24% accuracy). In various studies, SVM found most suitable machine-learning method for binary classification problems [64]–[68]. Additionally, it was observed that–CKSAAP feature has been used in various protein prediction problems associated with PTMs [14]–[18]. The SVM (c= 1 and gamma= scales) provides maximum performance using CKSAAP with Sen of 87.38%, Spe of 88.48%, Acc of 87.94% with MCC of 0.759, and AUROC of 0.904. Although the AUROC of some features (DDE, DPC, and PAAC) is higher than CKSAAP, other metrics like Sen, Spe, Acc, and MCC are lower (Table 2.2). Our results revealed that SVM is a robust classification algorithm for the prediction of PMTs. Thus, based on performance, SVM is adopted as a classifier for this study.

Table 2.2: Top performance of each feature set on independent test set. For each row, we list the feature name, ML algorithm, and the number of features and performance evaluation of 10-fold cross-validation.

Feature	**ML Algorithm**	**Performance evaluation**					
		Acc	**Sen**	**Spe**	**Pre**	**MCC**	**AUROC**
AAC	SVM	82.62	84.90	80.41	80.74	0.653	0.904
APAAC	MLP	83.39	81.17	77.49	77.72	0.664	0.877
CKSAAGP	LightGBM	76.85	78.68	75.08	75.34	0.538	0.841
CKSAAP	SVM	87.94	87.38	88.48	88.01	0.759	0.904
CTDC	LightGBM	77.70	78.89	76.50	77.41	0.554	0.847
CTDD	XGBoost	71.94	73.87	69.96	71.52	0.439	0.785
CTDT	SVM	79.80	85.64	73.85	76.98	0.599	0.859
Ctriad	XGBoost	79.02	77.16	80.91	80.50	0.581	0.864
DDE	XGBoost	84.45	84.36	84.53	84.07	0.689	0.912
DPC	XGBoost	84.10	85.61	82.64	82.67	0.683	0.923
GAAC	SVM	72.05	78.33	65.97	69.01	0.446	0.784
GDPC	RF	73.88	76.02	71.99	72.41	0.480	0.811
Geary	SVM	72.20	71.43	72.43	72.72	0.444	0.801
Moran	SVM	71.85	70.76	72.96	72.77	0.437	0.799
MoreauBroto	SVM	77.18	78.54	75.79	76.81	0.544	0.848

PAAC	MLP	83.49	85.96	83.67	83.59	0.696	0.912
QSOrder	LightGBM	83.40	84.36	82.47	82.32	0.668	0.908
SOCNumber	LightGBM	73.01	73.00	73.02	72.35	0.460	0.802

2.3.2 Performance of different CKSAAP feature subset

The high-dimensional features eventually increase the computational complexity and may also contain irrelevant or redundant attributes that affect accuracy reduction [71], [72]. Also over-fitting will be increased exponentially with the increase of k-value in CKSAAP features due to the large feature dimension. So, to avoid the risk of over-fitting with a high dimensional vector in SVM, we have performed a test with reduced feature dimension of CKSAAP. The dimension reduction of CKSAAP has been carried out by changing the value of k-spaced. We have decreased the k-spaced value from 5 to 1, and reduced the dimensionality of CKSAAP from 2400 to 800 attributes. So, by decreasing the k-value we have encoded four more CKSAAP feature subsets. Then, performance was evaluated with 14 machine-learning algorithms using 10-fold cross-validation test (Figure 2.3).

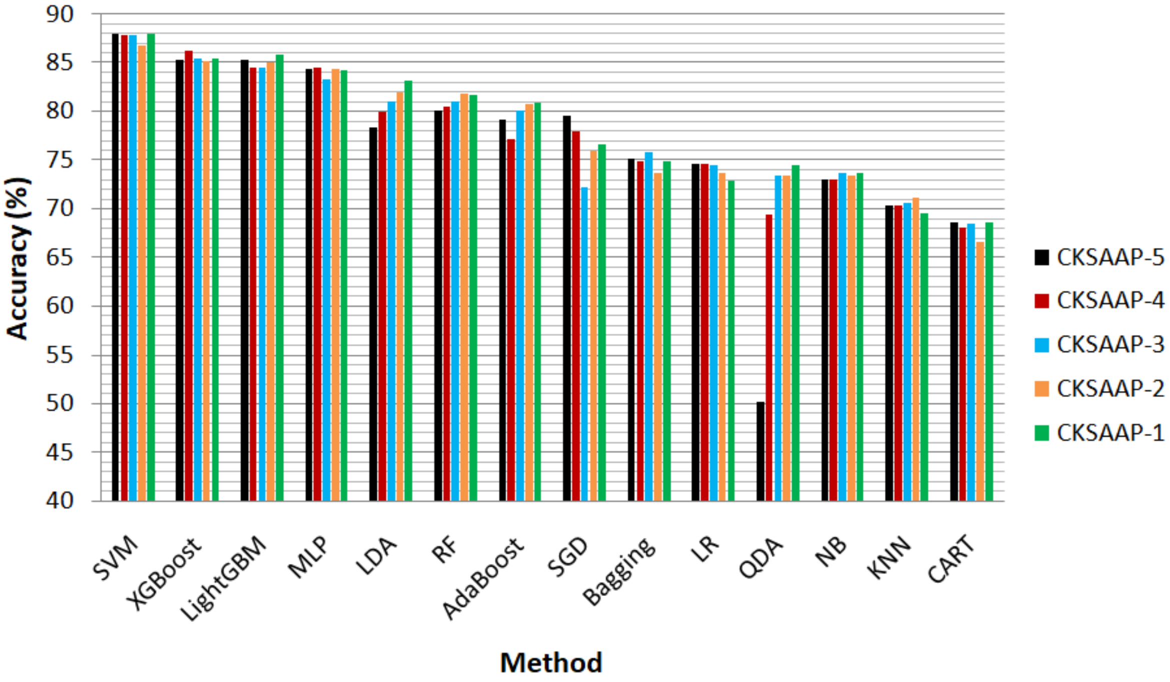


Figure 2.3: Performance comparison of CKSAAP features with a different set of descriptors using various machine-learning methods.

The accuracy performance of $CKSAAP_{k=1}$ with vector size 800 is similar to the performance of $CKSAAP_{k=5}$ with vector size 2400. The maximum performance was achieved by SVM on c=1 and g = 'scales' in each feature set. The performance of CKSAAP at different k-spaced values does not change much. Both models of $CKSAAP_{k=5}$ and $CKSAAP_{k=1}$ achieved the maximum accuracy of 87.94%. Thus, the $CKSAAP_{k=1}$ model can effectively avoid the over-fitting problems compared with $CKSAAP_{k=5}$. This model will also speed up the prediction and improves efficiency. $CKSAAP_{k=1}$ had the highest Sen of 88.80% and MCC of 0.759 amongst all variants of CKSAAP with AUROC of 0.945. The accuracy of $CKSAAP_{k=4}$ is almost similar to $CKSAAP_{k=1}$ with an improved AUROC of 0.952 (Table 2.3). The value of Sen is more considerable because it can improve the identification accuracy of the positive sample by reducing its scope. In the case of the $CKSAAP_{k=1}$ model, a little change in AUROC may help us reduce the computation and lower the risk of over-fitting due to the large feature dimension.

Table 2.3: The feature selection process is driven by the performance of SVM on different k-spaced values of CKSAAP.

Feature	k-spaced value	Vector size	Acc	Sen	Spe	Pre	MCC	ROC
$CKSAAP_{k=5}$	5	2400	87.94	87.38	88.48	88.01	0.759	0.904
$CKSAAP_{k=4}$	4	2000	87.77	87.74	87.8	87.43	0.755	0.952
$CKSAAP_{k=3}$	3	1600	87.86	88.27	87.45	87.19	0.757	0.949
$CKSAAP_{k=2}$	2	1200	86.81	87.21	86.42	86.14	0.736	0.946
$CKSAAP_{k=1}$	1	800	87.94	88.8	87.11	86.95	0.759	0.945

Figure 2.4A shows the performance of the $CKSAAP_{k=1}$ model in terms of the ROC curve on training and testing datasets. The highest AUROC values achieved by the training and testing datasets model were 0.9847 and 0.9448, respectively. The model had an AUPRC value of 0.9415 (Figure 2.4B). The ROC and PRC values towards one suggested that our selected model has better prediction ability. We have performed the trade-off between precision and recall to find the best threshold value. The best threshold value was obtained at 0.5 (Figure 2.4C). Therefore, we have selected the $CKSAAP_{k=1}$ model based on best performance with a low-dimension feature

vector as a final model for the implementation purpose in our proposed PMTPred tool. Further, we have used the blind dataset to validate the prediction performance of PMTPred.

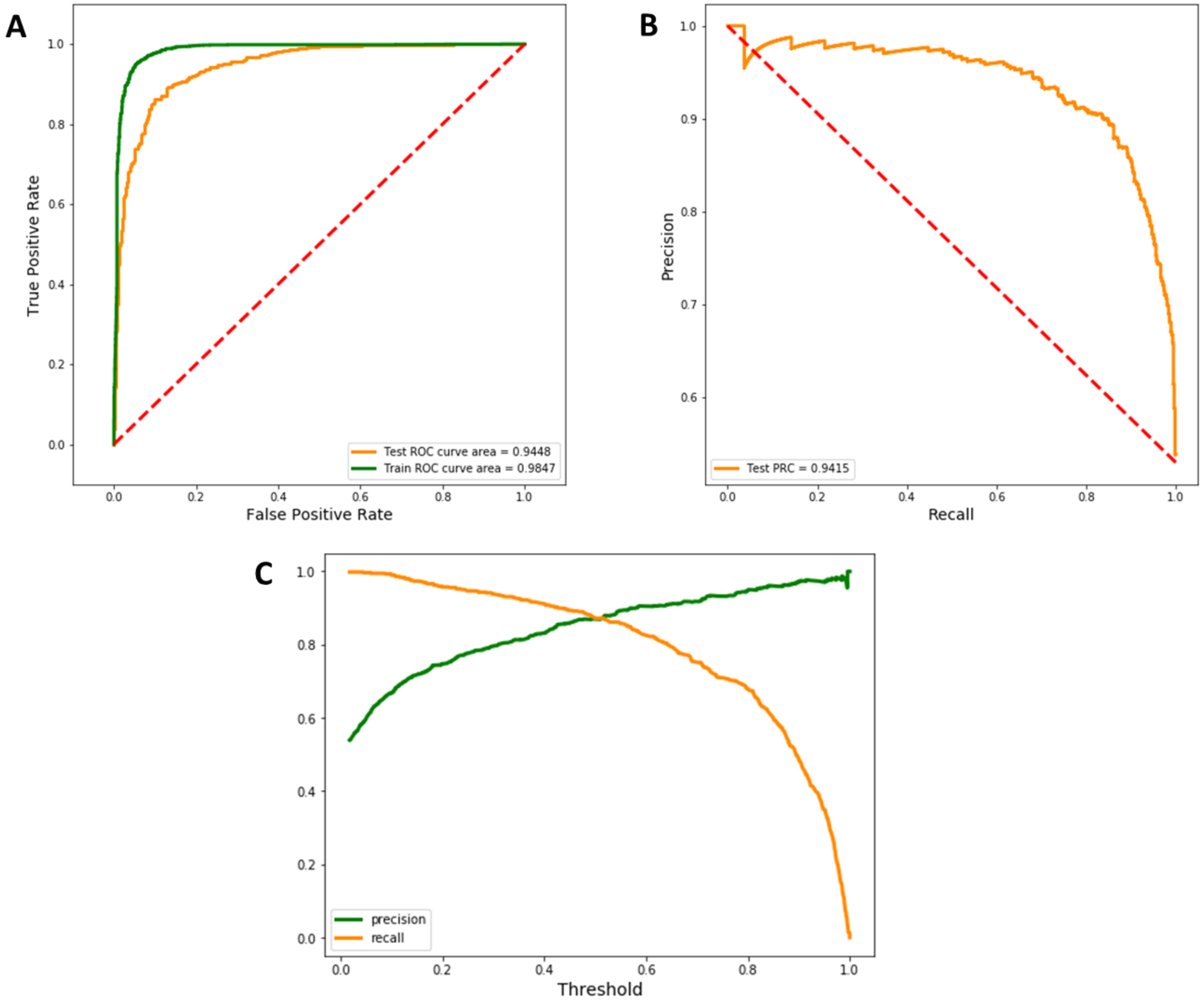


Figure 2.4: ROC and PRC curve of CKSAAP model. (A) ROC and (B) PR curve for the proposed model (CKSAAP$_{k=1}$) (C) Graph for precision and recall curve vs. threshold.

2.3.3 Validation of PMTPred on a blind dataset

To evaluate the unbiased performance of the CKSAAP-based SVM classifier (PMTPred), a blind test evaluation has been carried out on the blind test dataset consisting of 300 PMT and 300 non-PMT protein sequences. The prediction accuracy of PMTPred was found to be 88.66% with 85.00% sensitivity, and 92.33% specificity. The performance metrics is shown in Table 2.4.

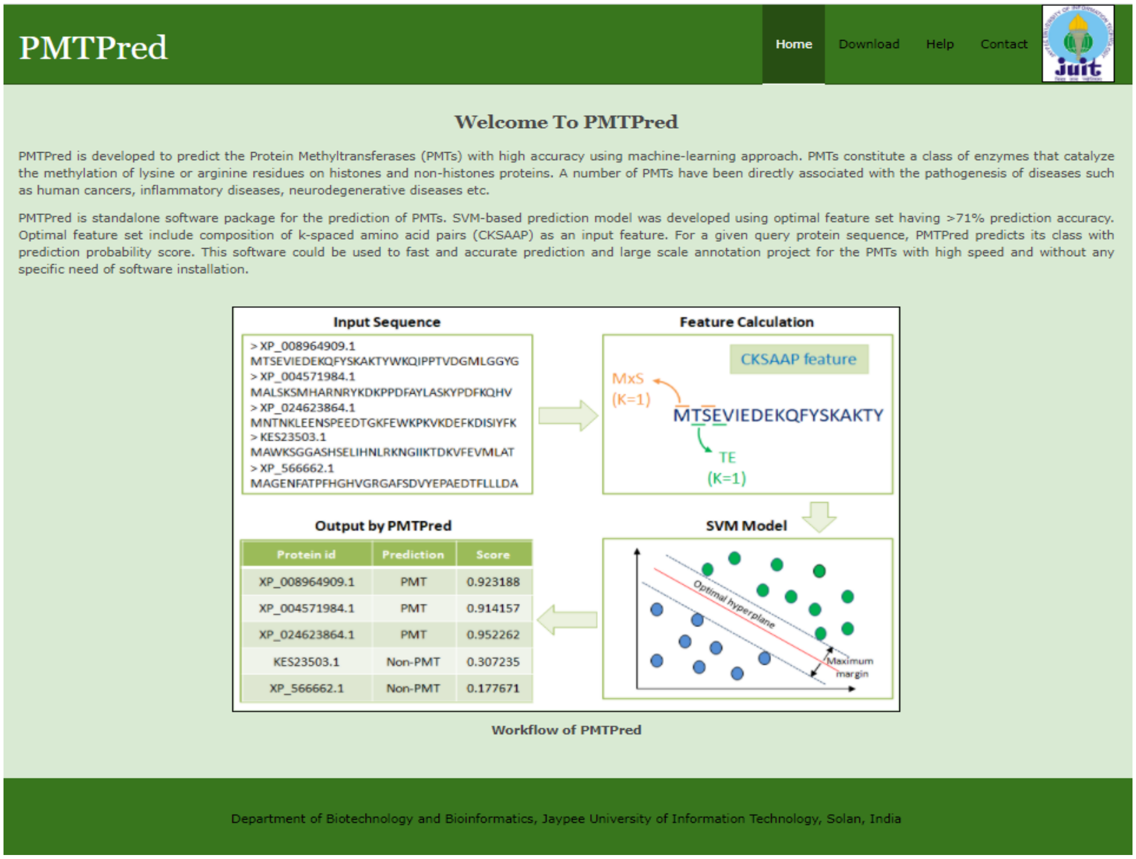
Therefore, the blind-test performance established that CKSAAP based SVM model has good prediction capacity.

Table 2.4: Performance of PMTPred on blind dataset.

Parameter	Performance
Accuracy	88.67%
Sensitivity	85.00%
Specificity	92.33%
Precision	91.73%
MCC	0.7754
F1	0.8824
AUROC	0.9504
AUPRC	0.9505

2.3.4 Standalone PMTPred

To serve the research and academic community, we have developed the standalone PMTPred software in Python. A snapshot of the PMTPred home page is shown in Figure 2.5. The designed code and other useful files can be downloaded from http://www.bioinfoindia.org/PMTPred/ and https://github.com/ArvindYadav7/PMTPred. Users can provide input sequence files in FASTA format, and the result can be saved as CSV file format. The output result has sequence ID, prediction class, and probability score in the CSV file.



Welcome To PMTPred

PMTPred is developed to predict the Protein Methyltransferases (PMTs) with high accuracy using machine-learning approach. PMTs constitute a class of enzymes that catalyze the methylation of lysine or arginine residues on histones and non-histones proteins. A number of PMTs have been directly associated with the pathogenesis of diseases such as human cancers, inflammatory diseases, neurodegenerative diseases etc.

PMTPred is standalone software package for the prediction of PMTs. SVM-based prediction model was developed using optimal feature set having >71% prediction accuracy. Optimal feature set include composition of k-spaced amino acid pairs (CKSAAP) as an input feature. For a given query protein sequence, PMTPred predicts its class with prediction probability score. This software could be used to fast and accurate prediction and large scale annotation project for the PMTs with high speed and without any specific need of software installation.

Input Sequence

```
>XP_008964909.1
MTSEVIEDEKQFYSKAKTYWKQIPPTVDGMLGGYG
>XP_004571984.1
MALSKSMHARNRYKDKPPDFAYLASKYPDFKQHV
>XP_024623864.1
MNTNKLEENSPEEDTGKFEWKPKVKDEFKDISIYFK
>KES23503.1
MAWKSGGASHSELIHNLRKNGIIKTDKVFEVMLAT
>XP_566662.1
MAGENFATPFHGHVGRGAFSDVYEPAEDTFLLLDA
```

Output by PMTPred

Protein id	Prediction	Score
XP_008964909.1	PMT	0.923188
XP_004571984.1	PMT	0.914157
XP_024623864.1	PMT	0.952262
KES23503.1	Non-PMT	0.307235
XP_566662.1	Non-PMT	0.177671

Workflow of PMTPred

Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Solan, India

Figure 2.5: Home page of PMTPred.

2.4 CONCLUSION

In this work, a computational approach for constructing the machine learning-based model was proposed and successfully utilized for the prediction of PMTs. Based on performance, it is concluded that SVM offers the best possible model for the prediction of PMTs. 10-fold cross-validation method revealed that composition-based CKSAAP feature given top performance. A low-dimensional CKSAAP feature model with better performance has been obtained through the iteration of the k-spaced value. The AUROC and AUPRC values also demonstrated that the SVM-based CKSAAP$_{k=1}$ model had a better predictive performance. Thus, we implemented the best and low-dimensional CKSAAP$_{k=1}$ model in the PMTPred tool for the prediction of PMTs. We believe that PMTPred would be very useful for the efficient prediction of PMTs to the scientific community and will have myriad applications.

REFERENCES

[1] R. A. Copeland, M. E. Solomon, and V. M. Richon, “Protein methyltransferases as a target class for drug discovery,” *Nat Rev Drug Discov*, vol. 8, no. 9, pp. 724–732, Sep. 2009, doi: 10.1038/nrd2974.

[2] C. Martin and Y. Zhang, “The diverse functions of histone lysine methylation,” *Nat. Rev. Mol. Cell Biol.*, vol. 6, no. 11, pp. 838–849, Nov. 2005, doi: 10.1038/nrm1761.

[3] D. Han *et al.*, “Lysine methylation of transcription factors in cancer,” *Cell Death Dis*, vol. 10, no. 4, p. 290, Mar. 2019, doi: 10.1038/s41419-019-1524-2.

[4] R. Hamamoto and Y. Nakamura, “Dysregulation of protein methyltransferases in human cancer: An emerging target class for anticancer therapy,” *Cancer Sci.*, vol. 107, no. 4, pp. 377–384, Apr. 2016, doi: 10.1111/cas.12884.

[5] A. K. Yadav and T. R. Singh, “Novel structural and functional impact of damaging single nucleotide polymorphisms (SNPs) on human SMYD2 protein using computational approaches,” *Meta Gene*, vol. 28, p. 100871, Jun. 2021, doi: 10.1016/j.mgene.2021.100871.

[6] V. Saloura *et al.*, “The role of protein methyltransferases as potential novel therapeutic targets in squamous cell carcinoma of the head and neck,” *Oral Oncol*, vol. 81, pp. 100–108, Jun. 2018, doi: 10.1016/j.oraloncology.2018.04.014.

[7] H. Ü. Kaniskan, K. D. Konze, and J. Jin, “Selective inhibitors of protein methyltransferases,” *J. Med. Chem.*, vol. 58, no. 4, pp. 1596–1629, Feb. 2015, doi: 10.1021/jm501234a.

[8] H. Ü. Kaniskan and J. Jin, “Chemical probes of histone lysine methyltransferases,” *ACS Chem. Biol.*, vol. 10, no. 1, pp. 40–50, Jan. 2015, doi: 10.1021/cb500785t.

[9] A. K. Yadav and T. R. Singh, “Novel inhibitors design through structural investigations and simulation studies for human PKMTs (SMYD2) involved in cancer,” *Molecular Simulation*, vol. 47, no. 14, pp. 1149–1158, Sep. 2021, doi: 10.1080/08927022.2021.1957882.

[10] R. A. Copeland, “Protein methyltransferase inhibitors as precision cancer therapeutics: a decade of discovery,” *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, vol. 373, no. 1748, 05 2018, doi: 10.1098/rstb.2017.0080.

[11] S. Malla, M. A. G. Koffas, R. J. Kazlauskas, and B.-G. Kim, “Production of 7-O-methyl aromadendrin, a medicinally valuable flavonoid, in Escherichia coli,” *Appl. Environ. Microbiol.*, vol. 78, no. 3, pp. 684–694, Feb. 2012, doi: 10.1128/AEM.06274-11.

[12] P. Nawabi, S. Bauer, N. Kyrpides, and A. Lykidis, “Engineering Escherichia coli for biodiesel production utilizing a bacterial fatty acid methyltransferase,” *Appl. Environ. Microbiol.*, vol. 77, no. 22, pp. 8052–8061, Nov. 2011, doi: 10.1128/AEM.05046-11.

[13] T. Petrossian and S. Clarke, “Bioinformatic Identification of Novel Methyltransferases,” *Epigenomics*, vol. 1, no. 1, pp. 163–175, Oct. 2009, doi: 10.2217/epi.09.3.

[14] H. Wang, X. Chen, C. Li, Y. Liu, F. Yang, and C. Wang, “Sequence-Based Prediction of Cysteine Reactivity Using Machine Learning,” *Biochemistry*, vol. 57, no. 4, pp. 451–460, Jan. 2018, doi: 10.1021/acs.biochem.7b00897.

[15] M. M. Hasan, S. Yang, Y. Zhou, and M. N. H. Mollah, “SuccinSite: a computational tool for the prediction of protein succinylation sites by exploiting the amino acid patterns and properties,” *Mol. BioSyst.*, vol. 12, no. 3, pp. 786–795, Feb. 2016, doi: 10.1039/C5MB00853K.

[16] A. N. Nilamyani, F. N. Auliah, M. A. Moni, W. Shoombuatong, M. M. Hasan, and H. Kurata, “PredNTS: Improved and Robust Prediction of Nitrotyrosine Sites by Integrating Multiple Sequence Features,” *International Journal of Molecular Sciences*, vol. 22, no. 5, Art. no. 5, Jan. 2021, doi: 10.3390/ijms22052704.

[17] S. Li *et al.*, “Deep learning based prediction of species-specific protein S-glutathionylation sites,” *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, vol. 1868, no. 7, p. 140422, Jul. 2020, doi: 10.1016/j.bbapap.2020.140422.

[18] X. Liu, L. Wang, J. Li, J. Hu, and X. Zhang, “Mal-Prec: computational prediction of protein Malonylation sites via machine learning based feature integration,” *BMC genomics*, 2020, doi: 10.1186/s12864-020-07166-w.

[19] Y. Huang, B. Niu, Y. Gao, L. Fu, and W. Li, “CD-HIT Suite: a web server for clustering and comparing biological sequences,” *Bioinformatics*, vol. 26, no. 5, pp. 680–682, Mar. 2010, doi: 10.1093/bioinformatics/btq003.

[20] L. Wei, C. Zhou, H. Chen, J. Song, and R. Su, “ACPred-FL: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides,” *Bioinformatics*, Jun. 2018, doi: 10.1093/bioinformatics/bty451.

[21] Z. Chen *et al.*, “iLearnPlus: a comprehensive and automated machine-learning platform for nucleic acid and protein sequence analysis, prediction and visualization,” *Nucleic Acids Research*, no. gkab122, Feb. 2021, doi: 10.1093/nar/gkab122.

[22] Z. Chen *et al.*, “iLearn: an integrated platform and meta-learner for feature engineering, machine-learning analysis and modeling of DNA, RNA and protein sequence data,” *Brief Bioinform*, vol. 21, no. 3, pp. 1047–1057, May 2020, doi: 10.1093/bib/bbz041.

[23] A. Pande *et al.*, “Computing wide range of protein/peptide features from their sequence and structure,” *bioRxiv*, p. 599126, Apr. 2019, doi: 10.1101/599126.

[24] Z. Chen *et al.*, “iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences,” *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, Jul. 2018, doi: 10.1093/bioinformatics/bty140.

[25] M. Bhasin and G. P. S. Raghava, “Classification of nuclear receptors based on amino acid composition and dipeptide composition,” *J Biol Chem*, vol. 279, no. 22, pp. 23262–23266, May 2004, doi: 10.1074/jbc.M401932200.

[26] J. Shen *et al.*, “Predicting protein-protein interactions based only on sequences information,” *Proc Natl Acad Sci U S A*, vol. 104, no. 11, pp. 4337–4341, Mar. 2007, doi: 10.1073/pnas.0607879104.

[27] Z. Chen *et al.*, “iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences,” *Bioinformatics*, vol. 34, no. 14, pp. 2499–2502, Jul. 2018, doi: 10.1093/bioinformatics/bty140.

[28] K. C. Chou, “Prediction of protein subcellular locations by incorporating quasi-sequence-order effect,” *Biochem Biophys Res Commun*, vol. 278, no. 2, pp. 477–483, Nov. 2000, doi: 10.1006/bbrc.2000.3815.

[29] K. C. Chou, “Prediction of protein cellular attributes using pseudo-amino acid composition,” *Proteins*, vol. 43, no. 3, pp. 246–255, May 2001, doi: 10.1002/prot.1035.

[30] M. Bhasin and G. P. S. Raghava, “Classification of Nuclear Receptors Based on Amino Acid Composition and Dipeptide Composition *,” *Journal of Biological Chemistry*, vol. 279, no. 22, pp. 23262–23266, May 2004, doi: 10.1074/jbc.M401932200.

[31] K. Chen, Y. Jiang, L. Du, and L. Kurgan, “Prediction of integral membrane protein type by collocated hydrophobic amino acid pairs,” *J Comput Chem*, vol. 30, no. 1, pp. 163–172, Jan. 2009, doi: 10.1002/jcc.21053.

[32] K. Chen, L. A. Kurgan, and J. Ruan, “Prediction of flexible/rigid regions from protein sequences using k-spaced amino acid pairs,” *BMC Struct Biol*, vol. 7, p. 25, Apr. 2007, doi: 10.1186/1472-6807-7-25.

[33] V. Saravanan and N. Gautham, “Harnessing Computational Biology for Exact Linear B-Cell Epitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor,” *OMICS*, vol. 19, no. 10, pp. 648–658, Oct. 2015, doi: 10.1089/omi.2015.0095.

[34] C. Z. Cai, L. Y. Han, Z. L. Ji, and Y. Z. Chen, “Enzyme family classification by support vector machines,” *Proteins*, vol. 55, no. 1, pp. 66–76, Apr. 2004, doi: 10.1002/prot.20045.

[35] C. Z. Cai, L. Y. Han, Z. L. Ji, X. Chen, and Y. Z. Chen, “SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence,” *Nucleic Acids Res*, vol. 31, no. 13, pp. 3692–3697, Jul. 2003, doi: 10.1093/nar/gkg600.

[36] I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk, and S. H. Kim, “Recognition of a protein fold in the context of the Structural Classification of Proteins (SCOP) classification,” *Proteins*, vol. 35, no. 4, pp. 401–407, Jun. 1999.

[37] I. Dubchak, I. Muchnik, S. R. Holbrook, and S. H. Kim, “Prediction of protein folding class using global description of amino acid sequence.,” *Proc Natl Acad Sci U S A*, vol. 92, no. 19, pp. 8700–8704, Sep. 1995.

[38] L. Y. HAN, C. Z. CAI, S. L. LO, M. C. M. CHUNG, and Y. Z. CHEN, “Prediction of RNA-binding proteins from primary sequence by a support vector machine approach,” *RNA*, vol. 10, no. 3, pp. 355–368, Mar. 2004, doi: 10.1261/rna.5890304.

[39] J. Shen *et al.*, “Predicting protein–protein interactions based only on sequences information,” *PNAS*, vol. 104, no. 11, pp. 4337–4341, Mar. 2007, doi: 10.1073/pnas.0607879104.

[40] R. R. Sokal and B. A. Thomson, “Population structure inferred by local spatial autocorrelation: an example from an Amerindian tribal population,” *Am J Phys Anthropol*, vol. 129, no. 1, pp. 121–131, Jan. 2006, doi: 10.1002/ajpa.20250.

[41] Z.-P. Feng and C.-T. Zhang, “Prediction of Membrane Protein Types Based on the Hydrophobic Index of Amino Acids,” *J Protein Chem*, vol. 19, no. 4, pp. 269–275, May 2000, doi: 10.1023/A:1007091128394.

[42] G. Pollastri, A. J. Martin, C. Mooney, and A. Vullo, “Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information,” *BMC Bioinformatics*, vol. 8, no. 1, p. 201, Jun. 2007, doi: 10.1186/1471-2105-8-201.

[43] D. S. Horne, “Prediction of protein helix content from an autocorrelation analysis of sequence hydrophobicities,” *Biopolymers*, vol. 27, no. 3, pp. 451–477, Mar. 1988, doi: 10.1002/bip.360270308.

[44] C. Zhou *et al.*, “Identification and analysis of adenine N 6 -methylation sites in the rice genome,” *Nature Plants*, vol. 4, no. 8, Art. no. 8, Aug. 2018, doi: 10.1038/s41477-018-0214-x.

[45] K. C. Chou, “Prediction of protein subcellular locations by incorporating quasi-sequence-order effect,” *Biochem Biophys Res Commun*, vol. 278, no. 2, pp. 477–483, Nov. 2000, doi: 10.1006/bbrc.2000.3815.

[46] K.-C. Chou and Y.-D. Cai, “Prediction of protein subcellular locations by GO-FunD-PseAA predictor,” *Biochem Biophys Res Commun*, vol. 320, no. 4, pp. 1236–1239, Aug. 2004, doi: 10.1016/j.bbrc.2004.06.073.

[47] G. Schneider and P. Wrede, “The rational design of amino acid sequences by artificial neural networks and simulated molecular evolution: de novo design of an idealized leader peptidase cleavage site.,” *Biophys J*, vol. 66, no. 2 Pt 1, pp. 335–344, Feb. 1994.

[48] K.-C. Chou, “Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes,” *Bioinformatics*, vol. 21, no. 1, pp. 10–19, Jan. 2005, doi: 10.1093/bioinformatics/bth466.

[49] K. C. Chou, “Prediction of protein cellular attributes using pseudo-amino acid composition,” *Proteins*, vol. 43, no. 3, pp. 246–255, May 2001, doi: 10.1002/prot.1035.

[50] M. M. Hasan, Y. Zhou, X. Lu, J. Li, J. Song, and Z. Zhang, “Computational Identification of Protein Pupylation Sites by Using Profile-Based Composition of k-Spaced Amino Acid Pairs,” *PLOS ONE*, vol. 10, no. 6, p. e0129635, Jun. 2015, doi: 10.1371/journal.pone.0129635.

[51] M. M. Hasan and H. Kurata, “GPSuc: Global Prediction of Generic and Species-specific Succinylation Sites by aggregating multiple sequence features,” *PLOS ONE*, vol. 13, no. 10, p. e0200283, Oct. 2018, doi: 10.1371/journal.pone.0200283.

[52] M. Usman and J. A. Lee, "AFP-CKSAAP: Prediction of Antifreeze Proteins Using Composition of k-Spaced Amino Acid Pairs with Deep Neural Network," in *2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE)*, Oct. 2019, pp. 38–43. doi: 10.1109/BIBE.2019.00016.

[53] C. White, H. D. Ismail, H. Saigo, and D. B. KC, "CNN-BLPred: a Convolutional neural network based predictor for β-Lactamases (BL) and their classes," *BMC Bioinformatics*, vol. 18, no. 16, p. 577, Dec. 2017, doi: 10.1186/s12859-017-1972-6.

[54] C. Cortes and V. Vapnik, "Support-Vector Networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1023/A:1022627411411.

[55] McLachlan, G.J., "Discrimination via Normal Models," in *Discriminant Analysis and Statistical Pattern Recognition*, John Wiley & Sons, Ltd, 1992, pp. 52–100. doi: 10.1002/0471725293.ch3.

[56] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, Aug. 1992, doi: 10.1080/00031305.1992.10475879.

[57] D. A. Freedman, *Statistical Models: Theory and Practice*. Cambridge University Press, 2009.

[58] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification And Regression Trees*. Routledge, 2017. doi: 10.1201/9781315139470.

[59] L. Breiman, "Random Forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[60] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," p. 9.

[61] T. Chen and T. He, "xgboost: eXtreme Gradient Boosting," p. 4.

[62] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *MACHINE LEARNING IN PYTHON*, p. 6.

[63] L. Breiman, "Bagging predictors," *Mach Learn*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi: 10.1007/BF00058655.

[64] L. Gonzalez-Abril, C. Angulo, H. Nuñez, and Y. Leal, "Handling binary classification problems with a priority class by using Support Vector Machines," *Applied Soft Computing*, vol. 61, pp. 661–669, Dec. 2017, doi: 10.1016/j.asoc.2017.08.023.

[65] S. HUANG, N. CAI, P. P. PACHECO, S. NARANDES, Y. WANG, and W. XU, “Applications of Support Vector Machine (SVM) Learning in Cancer Genomics,” *Cancer Genomics Proteomics*, vol. 15, no. 1, pp. 41–51, Dec. 2017, doi: 10.21873/cgp.20063.

[66] A. K. Yadav and D. Singla, “VacPred: Sequence-based prediction of plant vacuole proteins using machine-learning techniques,” *J Biosci*, vol. 45, no. 1, p. 106, Aug. 2020, doi: 10.1007/s12038-020-00076-9.

[67] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, “Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes,” *BMC Medical Informatics and Decision Making*, vol. 10, no. 1, p. 16, Mar. 2010, doi: 10.1186/1472-6947-10-16.

[68] N. Sharma, S. Patiyal, A. Dhall, A. Pande, C. Arora, and G. P. S. Raghava, “AlgPred 2.0: an improved method for predicting allergenic proteins and mapping of IgE epitopes,” *Briefings in Bioinformatics*, vol. 22, no. 4, p. bbaa294, Jul. 2021, doi: 10.1093/bib/bbaa294.

[69] J. M. Berg, J. L. Tymoczko, and L. Stryer, “Protein Structure and Function,” *Biochemistry. 5th edition*, 2002, Accessed: May 06, 2021. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK21177/

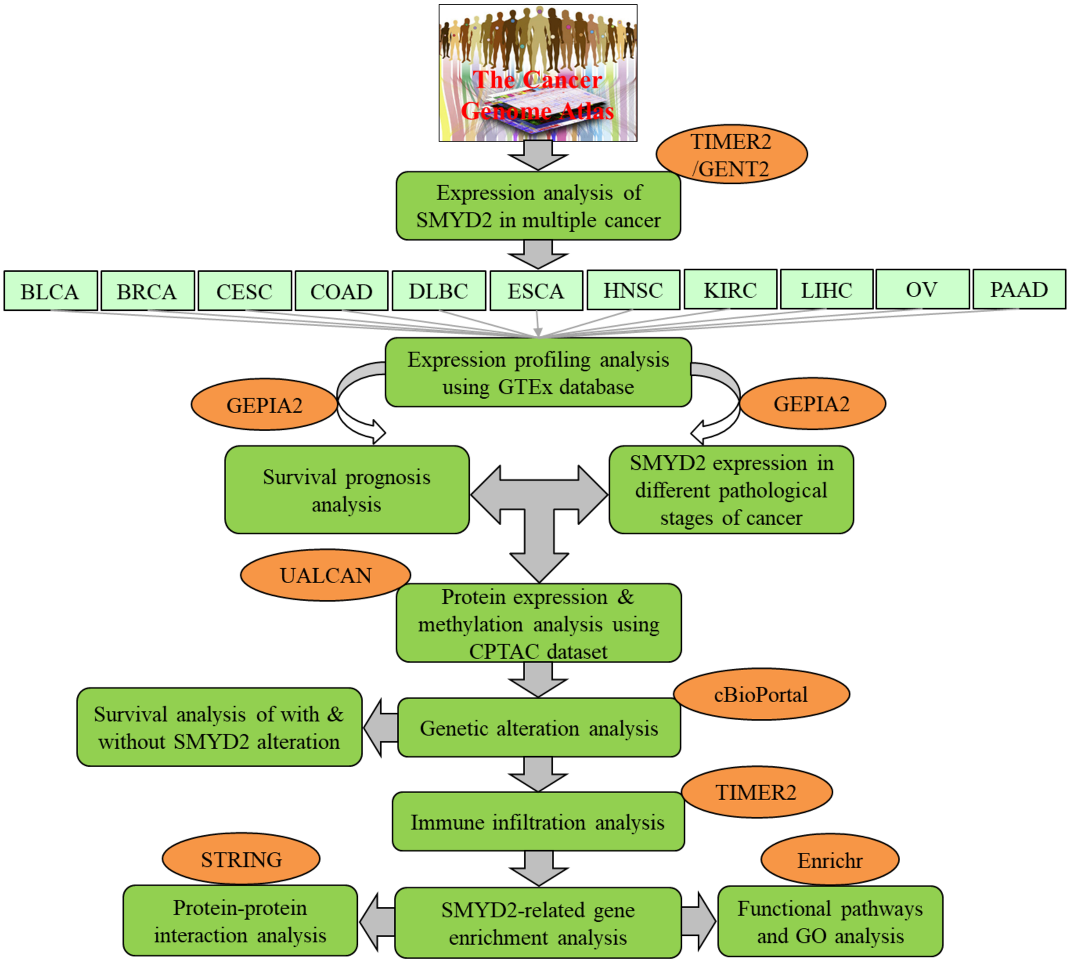
[70] D. Zhang and J. J. P. Tsai, *Machine Learning Applications In Software Engineering (Series on Software Engineering and Knowledge Engineering)*. River Edge, NJ, USA: World Scientific Publishing Co., Inc., 2005.

[71] M. M. Hasan, M. A. Alam, W. Shoombuatong, and H. Kurata, “IRC-Fuse: improved and robust prediction of redox-sensitive cysteine by fusing of multiple feature representations,” *J Comput Aided Mol Des*, vol. 35, no. 3, pp. 315–323, Mar. 2021, doi: 10.1007/s10822-020-00368-0.

[72] L. Wei, J. Hu, F. Li, J. Song, R. Su, and Q. Zou, “Comparative analysis and prediction of quorum-sensing peptides using feature representation learning and machine learning algorithms,” *Brief Bioinform*, Oct. 2018, doi: 10.1093/bib/bby107.

CHAPTER- 3

Integrative Analysis of the Oncogenic Role of SMYD2 in Multiple Human Malignancies


The Cancer Genome Atlas
TIMER2 /GENT2
Expression analysis of SMYD2 in multiple cancer
BLCA
BRCA
CESC
COAD
DLBC
ESCA
HNSC
KIRC
LIHC
OV
PAAD
Expression profiling analysis using GTEx database
GEPIA2
GEPIA2
Survival prognosis analysis
SMYD2 expression in different pathological stages of cancer
UALCAN
Protein expression & methylation analysis using CPTAC dataset
cBioPortal
Survival analysis of with & without SMYD2 alteration
Genetic alteration analysis
TIMER2
Immune infiltration analysis
STRING
Enrichr
Protein-protein interaction analysis
SMYD2-related gene enrichment analysis
Functional pathways and GO analysis

ABSTRACT

SMYD2 is a protein from the family SMYD of PKMT. It has the ability to regulate the transcription of the gene by catalyzing the methylation of lysine residues in protein substrates, which show a critical role in tumorigenesis. Although emerging evidence supports the association of SMYD2 in the progression of cancers but mostly are unexplored. Therefore further study of the gene about cancer progression needs to be conducted. Based on the TCGA data, we computationally analyzed the potential carcinogenic effect of SMYD2 in multiple tumors. The transcriptional expression, survival rate, mutations, enriched pathways, and gene ontology of the SMYD2 was determined using different bioinformatics servers. In addition, the protein-protein interaction (PPI) analysis was performed and examined the correlation between genes co-expressed with SMYD2 and immunocyte infiltration in multiple cancer types. It was observed that a higher expression of SMYD2 was present in tumor tissues as compared to normal tissues in queried cancer types. The prognostic analysis showed a strong association with cancers. We detected 15 missenses, 4 truncating mutations, and 5 others. In PPI analysis, out of 50 protein partners, most of them indicated significant co-expression. The gene GNPAT was found to be the most highly associated. Gene ontological properties and pathways were found to be significantly linked to the development of cancer. Collectively, our data-driven results provide a relatively comprehensive insight into the understanding of the carcinogenic effect of SMYD2. It is projected that SMYD2 could be a significant target for biomarker identification in different human tumors.

3.1 INTRODUCTION

Cancer is considered as most death-causing disease around the world and has become a serious health burden on our society. Cancer is an imperative cause of morbidity and mortality in every country of the world. It has surpassed cardiovascular diseases as the next largest reason of death globally [1]. According to the latest press release by the WHO in 2020, there were 19.3 million new cancer patients diagnosed and nearly 10 million cancer deaths occurred globally. It is expected that the burden of cancer will be reached 28.4 million patients by the year 2040 [2]. The increasing changes that occur in genetic and epigenetic parameters boost tumorogenesis, which is positively correlated with the diagnosis of cancer patients. Differentially expressed genes linked to cancer patient survival could be exploited as diagnostic markers to aid early cancer detection [3]. Therefore, cancer investigation, identification of related biomarkers, and development methods for active prevention are important requirements for early cancer diagnosis and treatment.

PMT is a catalytic enzyme that helps to transfer the methyl group to their substrate from the SAM. It shows a significant role in epigenetic mechanisms and is involved in the methylation of various substrates for example DNA, RNA, protein, and secondary metabolites [4]. PMTs play a crucial role in transcriptional events through histone methylation and non-histone methylation at the position of arginine or lysine residues. PKMT is a type of PMT that helps to transfer a methyl group to the lysine residue of the substrate protein. It has been described that overexpression of PKMTs is linked with different types of human cancers [5], [6].

SMYD2 is a protein from PKMTs that is implicated in tumorigenesis and can influence gene transcription through lysine methylation [7]. Numerous studies have uncovered the activity of SMYD2 methylation to non-histone protein [8], [9]. The SMYD2-specific non-histone substrates are significantly associated with the processes related to cancer [10], [11]. Numerous tumor-causing proteins for example [12], HSP90 [13], RB [14], [15], ERα [16], PTEN [17], PARP1 [18], and STAT3 [19] are get methylate through SMYD2. Therefore, it has been evidenced that SMYD2 is an onco-related protein that can affect the function of cancer suppressor proteins. The various analysis represented that overexpression of SMYD2 is present in various human cancers, like breast, bladder, colorectal, esophageal, lymphoma, ovarian, head, and neck, cervical, and pancreatic cancer [12], [14], [20].

In this analysis, we scientifically explored the SMYD2 expression and its clinical outcomes to evaluate the SMYD2 as a potential marker for cancer treatment. Various expression and patient survival datasets available on several online platforms were used for this analysis. We measured multiple factors such as the difference in gene expression, survival value, gene mutations, phosphorylation, methylation, immune infiltration, and functional enrichment analysis to explore the prospective molecular mechanism of the oncogenic role of SMYD2 on the pathogenesis.

3.2 METHODOLOGY

3.2.1 Gene expression analysis

The *SMYD2* expression level in tumor tissue and corresponding normal tissue for different cancers across all tumors of the TCGA project was analyzed using the GEPIA2 [21] (http://gepia2.cancer-pku.cn/#analysis) and TIMER2 (http://timer.cistrome.org/) [22].

GEPIA2: It is an online platform that used RNA sequencing expression data of 9736 tumors and 8587 normal samples from the TCGA and Genotype-Tissue Expression (GTEx) project using a standard protocol [23]. We analysed the expression differences of SMYD2 gene in all TCGA tumors and normal tissues. Additionally, we also analyzed the SMYD2 gene expression difference in tumors and their corresponding normal tissues using GTEx database as control. Statistical method analysis of variance (ANOVA) was used to calculate the differential gene expression in between tumor and control tissues. Parameters for the assessment method were under the setting of log2FC (fold change) of 1, *q*-value cutoff of 0.01, and "Match TCGA normal and GTEx data". Moreover, the violin plots for the expression level of *SMYD2* in diverse pathological stages (i.e. stage I-V) of all TCGA tumors were also explored. The log2 [TPM (Transcripts per million) +1] transformed data were applied for the construction of violin plots. Statistical significance was set at P< 0.05. The complete workflow of the present study is shown in Figure 3.1.

TIMER2: It is a public database that for differential expression between tumor and adjacent normal tissues at transcript level across all TCGA tumors. The data was normalized using log2 TPM transformation. Box plots were created to represent the level of expression distributions. The statistical significance between tumors and normal tissues was calculated by Wilcoxon test was annotated by the number of stars (*: p-value < 0.05; **: p-value <0.01; ***: p-value <0.001).

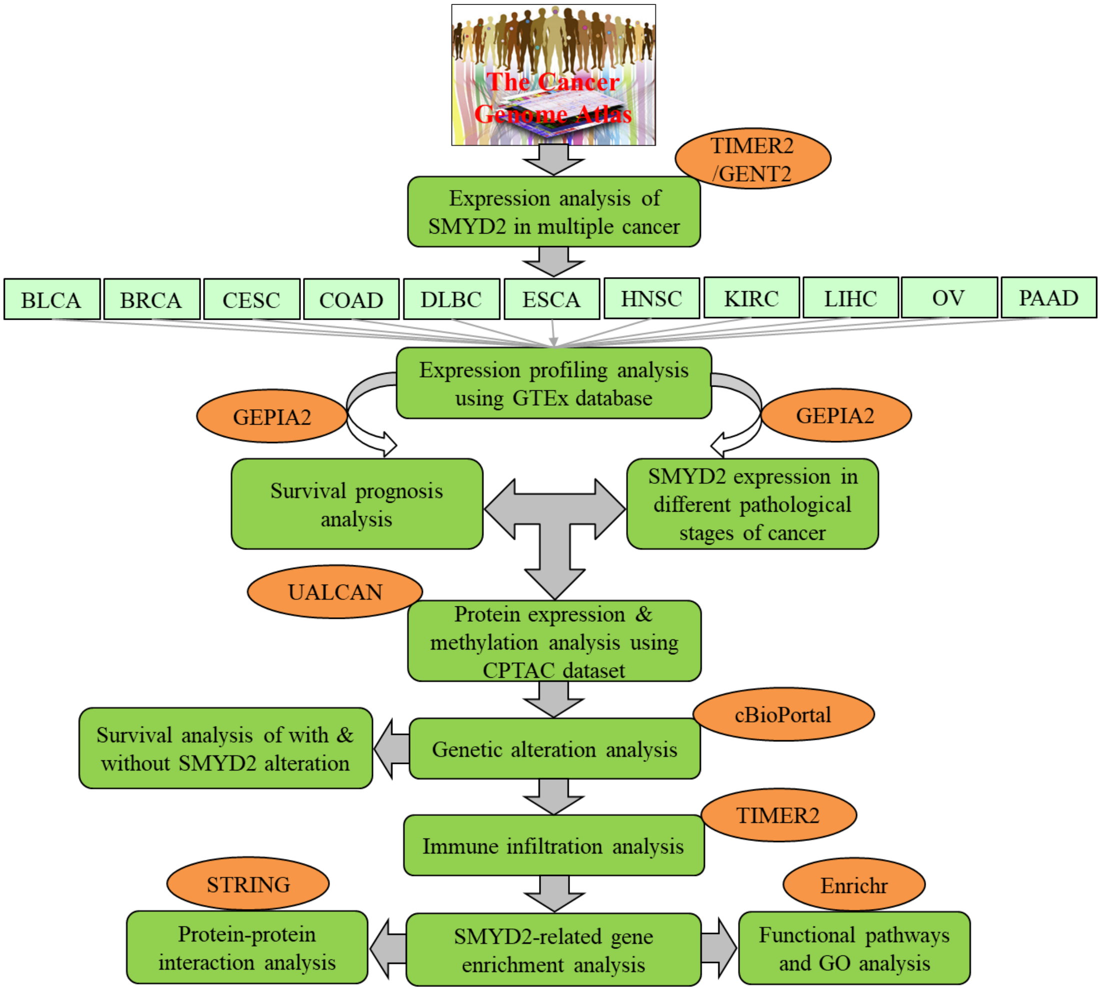

The Cancer Genome Atlas
TIMER2 /GENT2
Expression analysis of SMYD2 in multiple cancer
BLCA
BRCA
CESC
COAD
DLBC
ESCA
HNSC
KIRC
LIHC
OV
PAAD
Expression profiling analysis using GTEx database
GEPIA2
GEPIA2
Survival prognosis analysis
SMYD2 expression in different pathological stages of cancer
UALCAN
Protein expression & methylation analysis using CPTAC dataset
cBioPortal
Survival analysis of with & without SMYD2 alteration
Genetic alteration analysis
TIMER2
Immune infiltration analysis
STRING
Enrichr
Protein-protein interaction analysis
SMYD2-related gene enrichment analysis
Functional pathways and GO analysis

Figure 3.1: Methodology of the proposed work.

3.2.2 Survival prognosis analysis

The cancer patient survival analysis with the expression of SMYD2 gene was performed using TCGA cancer patient data. To calculate the overall survival (OS) and disease-free survival (DFS) of cancer patient the Kaplan-Meier (K-M) plots were calculated by using online resource GEPIA2 [21]. The median score was used as cutoff the division of high-expression and low-expression cohorts, the 50% cutoff value was used as the low and high expression thresholds. For the hypothesis test, the log-rank test also called Mentel-Cox test was utilized. Additionally, the hazard ratio (HR) with 95% confidence intervals was computed. The *P*< 0.05 was considered statistically significant for all survival analysis.

3.2.3 Promoter methylation analysis

The promoter methylation analysis in multiple cancer patients has been studied in TCGA dataset by using UALCAN (http://ualcan.path.uab.edu/analysis-prot.html) web portal [24]. It is an online resource to evaluate the omics data associated with cancer and helped to analyze the expression of protein datasets from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) and TGCA databases. To generate the results, the parameters were left at their default values. The database utilized TPM to normalize the methylation expression value of row data from TCGA. The statistical method Wilcoxon rank sum test was utilized for the methylation differential analysis. The methylation level indicates in term of Beta value that ranging from 0 (non-methylated) to 1 (fully methylated). The Beta value cut-off between 0.25 to 0.3 and 0.5 to 0.7 has been considered as hypo-methylation and hyper-methylation respectively.

3.2.4 Genetic alteration analysis

Alteration analysis for SMYD2 was investigated using the online resource of cBioPortal (https://www.cbioportal.org/) [25], [26] by selecting the “TCGA Pan-Cancer Atlas Studies”. All TCGA tumors were examined for the alteration frequency, mutation type, and copy number alteration by using 4617 samples. The mutation frequency chart was generated to see the distributions of mutations. We also analyzed the differences between TCGA cancer patients with and without SMYD2 mutations in terms of disease-specific, disease-free, overall, and progression-free survival. To create K-M graphs, the log rank of the *P*-value was utilized. A log-rank *P*-value < 0.05 was considered significant.

3.2.5 Immune infiltration analysis

Using six cutting-edge algorithms, TIMER2.0 provides a more reliable estimation of immune infiltration levels for TCGA or user-provided tumor profiles [22]. By choosing CD8+ T-cells and cancer-associated fibroblasts in “Immune” module using “Gene Symbol: SMYD2” generate a heatmap with purity-adjusted spearman’s rho value across the cancer types. Then by clicking on a particular cell, we generated the immune cell infiltration and *SMYD2* expression plot across all tumor types. TIMER2 utilized the following algorithms for the estimation of immune infiltration.

CIBERSORT (https://cibersortx.stanford.edu/): In this algorithm, single cell RNA-seq or bulk-sorted reference profiles are used to estimate cell-type frequencies from bulk gene expression data. Cell-type-specific gene expression profiles of n different cell types are inferred at the sample level, and platform-specific variance is overcome [27].

TIMER (http://cistrome.org/TIMER/): In this algorithm, analysis of six immune cell types across various cancer types is focused on quantifying immune cell infiltration from TCGA Project [28].

QUANTISEQ (http://icbi.at/quantiseq): In this algorithm, calculates the ratios of 10 different immune cell types using data from bulk RNA sequencing [29].

EPIC (http://epic.gfellerlab.org/): It uses transcriptome data to calculate the absolute fraction of several immune and cancer cell types [30].

XCELL (https://xcell.ucsf.edu/): It utilises gene signature enrichment from 64 different cell types to infer cell abundance [31].

MCPCOUNTER (https://github.com/ebecht/MCPcounter): It employed an enrichment method to infer the presence of eight immune and two stromal cell types in bulk tissues [32].

Mostly, immune cell types have a negative correlation with tumor purity, thus tumor purity considered as a significant confounding factor. Therefore, we used "Purity Adjustment" option, which will carry out this association study using the partial spearman's correlation. The P-values and partial correlation values were obtained via the purity-adjusted Spearman’s rank correlation test. A P-value < 0.05 was considered statistically significant. Correlation outputs data were represented with the help of scatter plots and heatmaps.

3.2.6 SMYD2-associated gene enrichment analysis

The protein name “SMYD2” was searched on the STRING (https://string-db.org/) database with the organism *“Homo sapiens”*. Consequently, other parameters were considered as the score for the minimum required interaction [“Low confidence (0.150)”], the meaning of network edges (“evidence”), maximum interactors to display (“no more than 50 interactors”), and active interaction sources (“experiments”). Lastly, the SMYD2-associated proteins were determined experimentally, and the interactions of the protein-protein network were generated through Cytoscape [33].

The datasets for selected normal and tumor tissues were used in the GEPIA2 server the finding the top 100 targeted genes associated with SMYD2. In this study, Enrichr web (https://maayanlab.cloud/Enrichr/) [34] was used for pathways and gene ontology (GO) analysis. Reactome 2022 and the Kyoto encyclopedia of genes and genomes (KEGG) 2021 databases were utilized to define the signaling pathways. The SMYD2-correlated genes were classified into biological processes, cellular components, and molecular functions using GO terms. Statistical significance was defined as a *p*-value of less than 0.05. The *p*-values and q-values were calculated for significant terms. The q-value (adjusted *p*-value) was calculated using the Benjamini-Hochberg method. The top 10 enriched terms for input gene set were displayed on bar charts based on the -log10(*p*-value).

3.3 RESULTS AND DISCUSSION

3.3.1 Analysis of gene expression data

SMYD2 has been found to be associated with several biological and cellular processes in numerous species [35]. Several studies have been reported about the functional association between SMYD2 and tumors [36]–[39]. There, here to gain an overall oncogenic role of the *SMYD2* in cancers, the gene expression profile was analyzed in several normal and cancer types of tissues by using the TIMER2 database. The *SMYD2* is highly expressed in nearly all types of cancer (*P*<0.001) as compared to their equivalent normal tissues (Figure 3.2A). The x-axis shows the *SMYD2* gene expression in log2 fold change values whereas, the y-axis shows the tissue types where *SMYD2 is* expressed. Analysis of TCGA datasets by using the GEPIA2 database also showed similar *SMYD2* expression in bladder carcinoma, colon adenocarcinoma, cervical squamous cell carcinoma, diffuse large B-cell lymphoma, liver hepatocellular carcinoma, pancreatic adenocarcinoma, rectum adenocarcinoma, skin cutaneous melanoma, thymoma, uterine corpus endometrial carcinoma and uterine carcinosarcoma (Figure 3.2B).

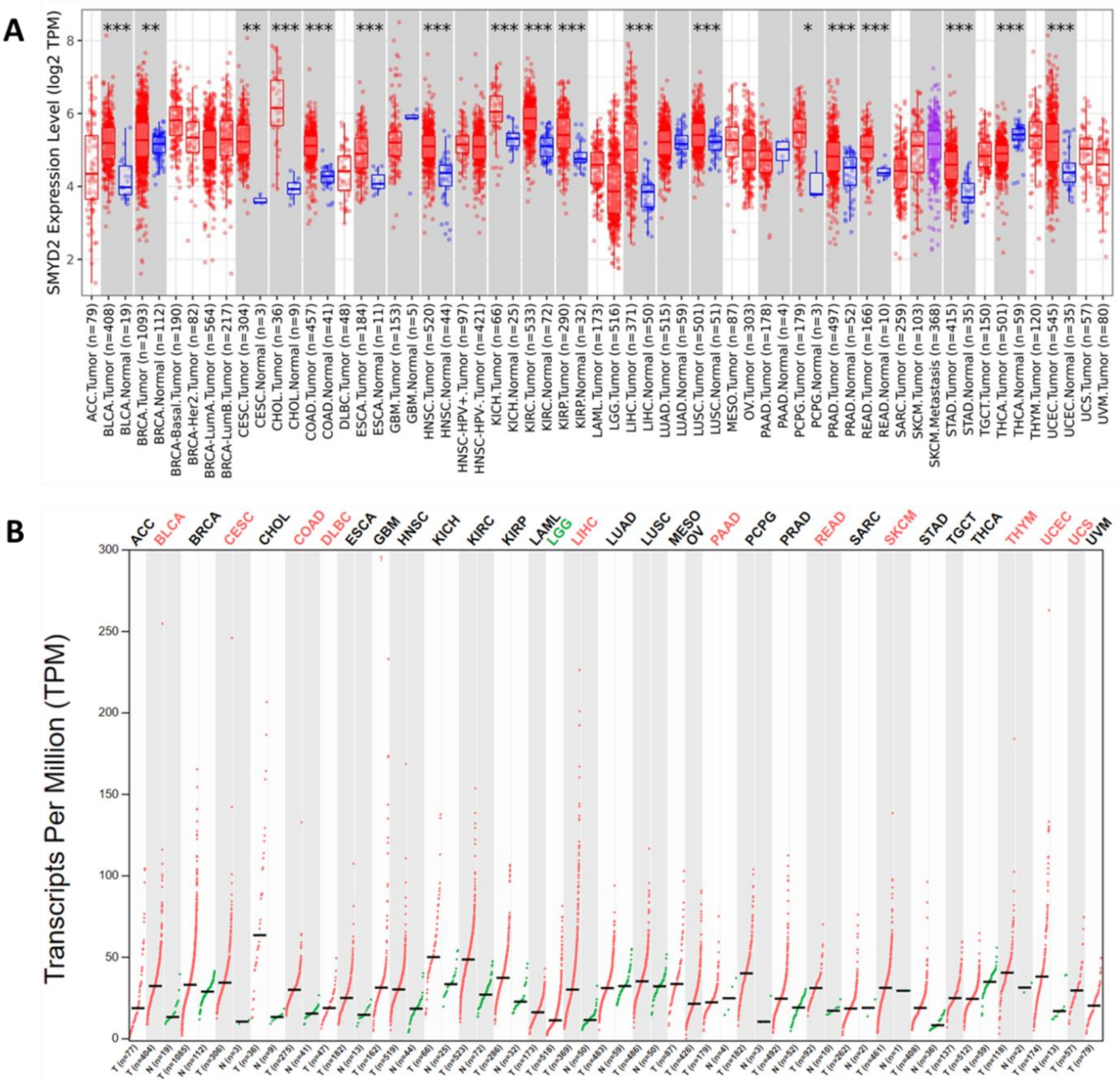


Figure 3.2: The difference in *SMYD2* expression level between the normal and tissues in all TCGA cancer types. (A) *SMYD2* expression levels were analyzed by using TIMER2 in 33 types of human cancers. The tumor expression is shown in red color and normal expression is shown in blue color. The stars represent the statistical significance level (* *P*<0.05; ** *P*<0.01; *** *P*<0.001) (B) Gene expression profile of *SMYD2* analyzed in tumors (red dashed line) and normal (green dashed line) tissue samples using GEPIA2. The cancers names with significantly expressed are marked with red color and low expressed are marked with green color.

From expression analysis, we selected only those cancer types in which the higher expression of *SMYD2* is reported in the progression of cancer. Those cancer types are bladder, cervical, lymphoid, head and neck, liver, ovarian, kidney, colon, breast, pancreatic cancer, and esophageal.

We next used GEPIA2 to compare the expression difference of *SMYD2* in the GTEx dataset as a control. Tumor and normal tissues of 11 cancer types such as breast invasive carcinoma (BRCA), bladder urothelial carcinoma (BLCA), colon adenocarcinoma (COAD), cervical squamous cell carcinoma, and endocervical adenocarcinoma (CESC), lymphoid neoplasm diffuse large B-cell lymphoma (DLBC), head and neck squamous cell carcinoma (HNSC), esophageal carcinoma (ESCA), kidney renal clear cell carcinoma (KIRC), liver hepatocellular carcinoma (LIHC), ovarian serous cystadenocarcinoma (OV), and pancreatic adenocarcinoma (PAAD) were used to evaluate the expression difference of SMYD2. The expression comparison between normal and tumor tissues is represented by the box plot. As compared to normal tissues, all tumor types showed higher expression. A significant expression difference (P<0.01) was observed for the tumors BLCA, COAD, CESC, DLBC, and PAAD (Figure 3.3). Henceforth, the analyzed result indicates that the overexpression of *SMYD2* plays a vital role in the various cancer progression.

Furthermore, we used the GEPIA2 tool to examine the correlation between the *SMYD2* expression and cancer pathological stages (stage I-V). The violin plots for the pathological stages of all 11 cancer types are shown in Figure 3.4. The expression of *SMYD2* varied more than six orders of magnitude in maximum cancer types. This analysis suggested that *SMYD2* has a promoting role in the incidence and progression of cancer.

3.3.2 Analysis of survival data

To explore the critical efficiency of *SMYD2* in the survival of various cancer cases, we used GEPIA2 to evaluate survival data and create a link between miRNA expression and cancer patient survival. The K-M plots for OS and DFS analysis for all types of cancer were analyzed**.** A high level of *SMYD2* in CESC (HR 2.3, P = 0.00045) and a lower level of *SMYD2* in KIRC (HR 0.49, P = 6e-06) were expressively correlated with the overall survival of cancer patients (Figure 3.5). The lower level of *SMYD2* in BRCA (HR 0.68, P = 0.046) and KIRC (HR 0.63, P = 0.013) and a higher level of *SMYD2* in COAD (HR 2, P = 0.0061) and HNSC (HR 1.6, P = 0.008) were expressively correlated with disease-free survival (Figure 3.6). The low *SMYD2* expression group had a greater survival rate as compared to the higher expression group in maximum cancer types for both the OS and DFS.

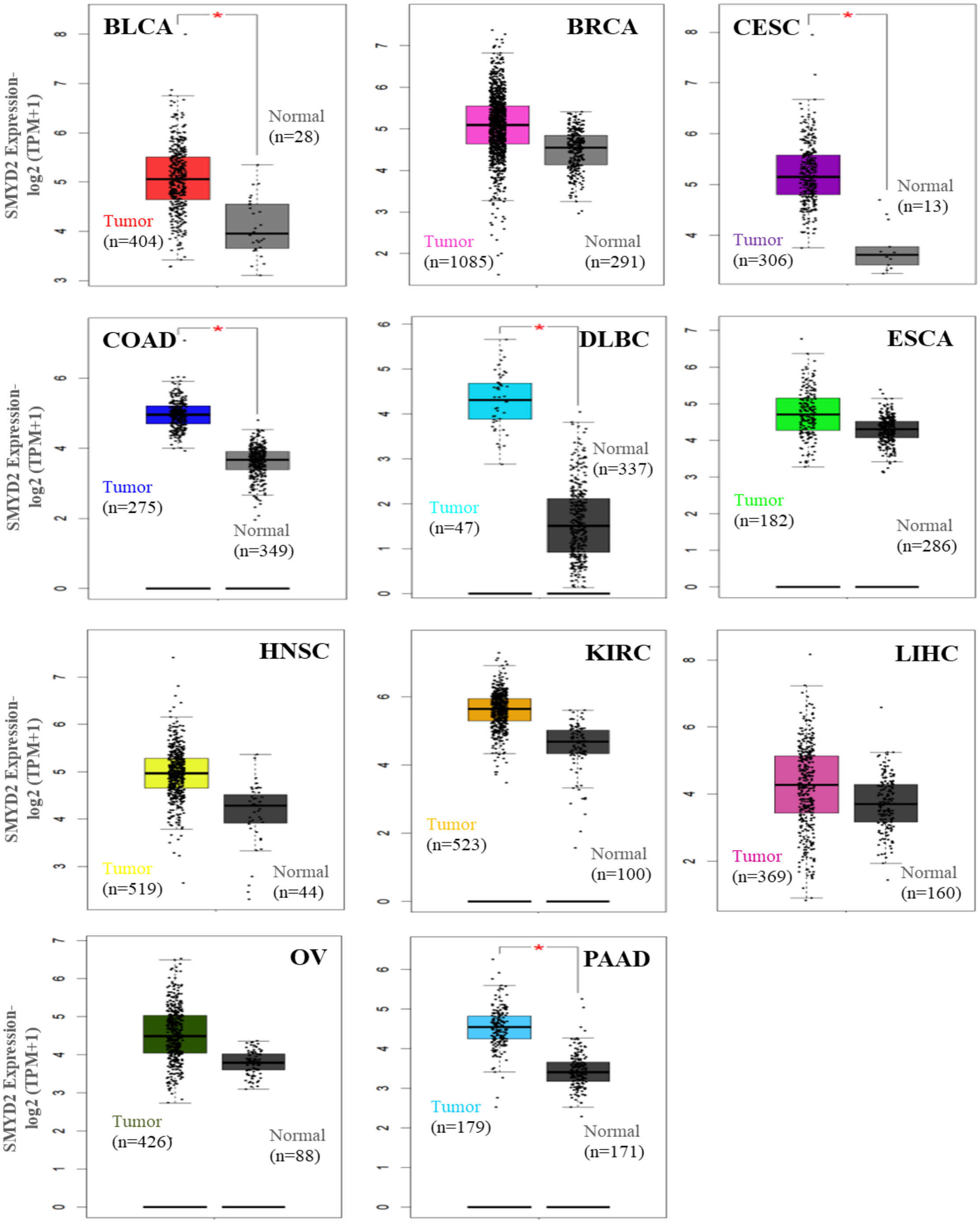


Figure 3.3: *SMYD2* expression in different cancers analyzed by GEPIA2 using the TCGA dataset. The expression of *SMYD2* in all cancer tissues was compared to the equivalent normal

tissues using the GTEx as a control database. The box plots depicting the expression of *SMYD2* in the tumor tissues are shown in different colors while corresponding normal tissues are gray. The statistically significant difference (*P*<0.01) was marked with a red color asterisk (*).

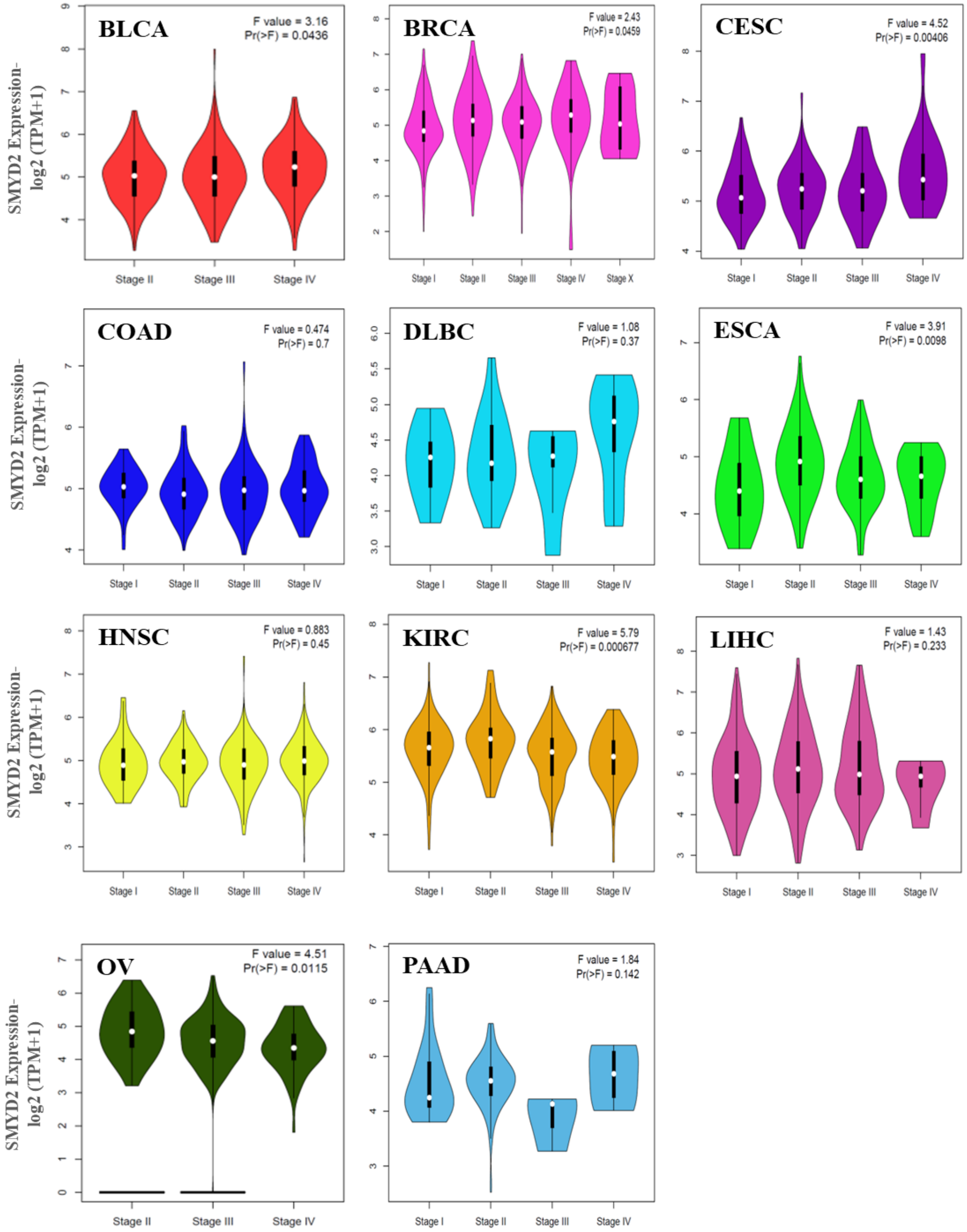


Figure 3.4: The levels of *SMYD2* expression between different pathological stages (stage I-V) based on the TCGA dataset by applying the log-scale as Log2 (TPM+1).

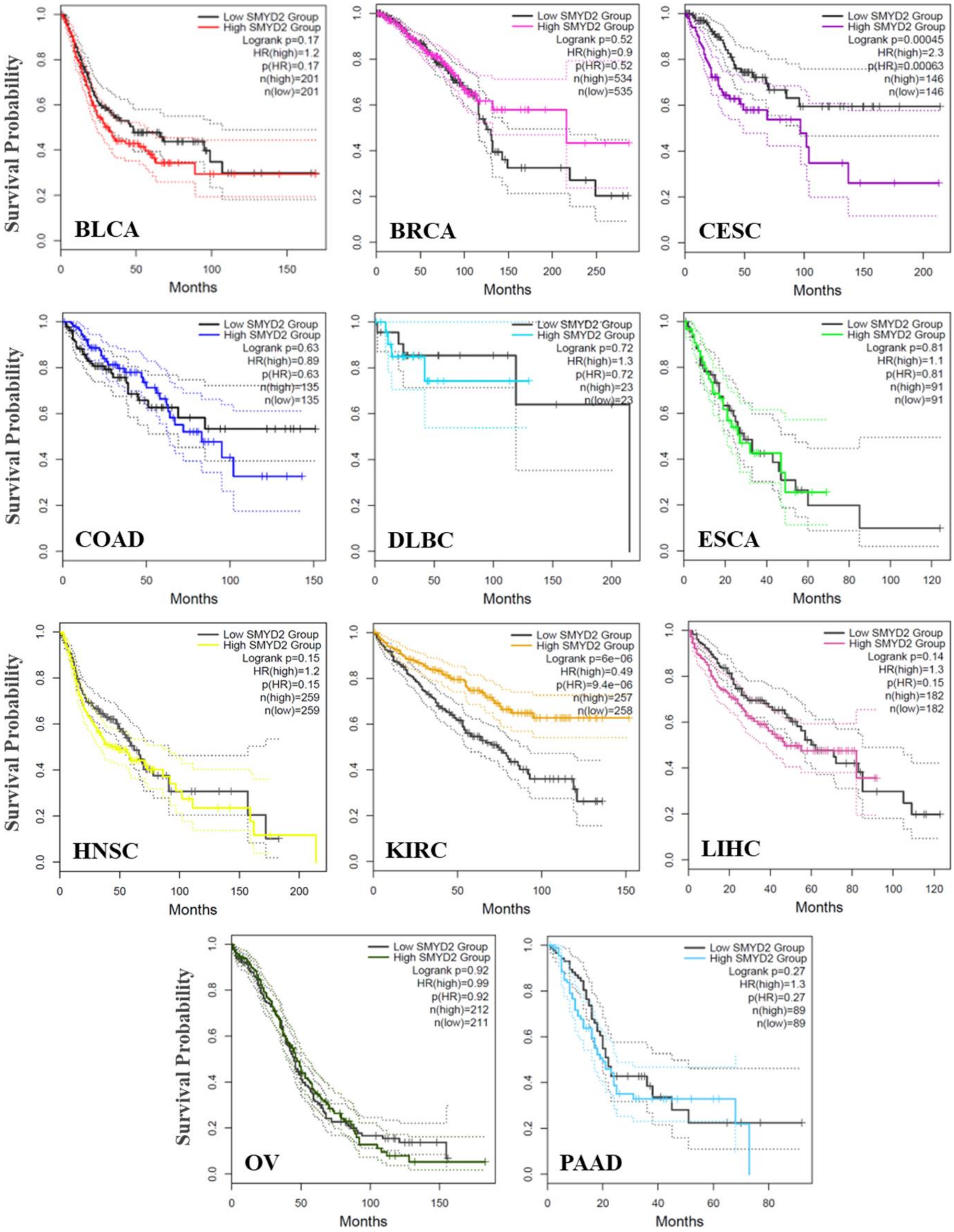


Figure 3.5: K-M plots represented the relationship between high (color) and low (black) *SMYD2* gene expression with overall survival of patients in multiple cancers.

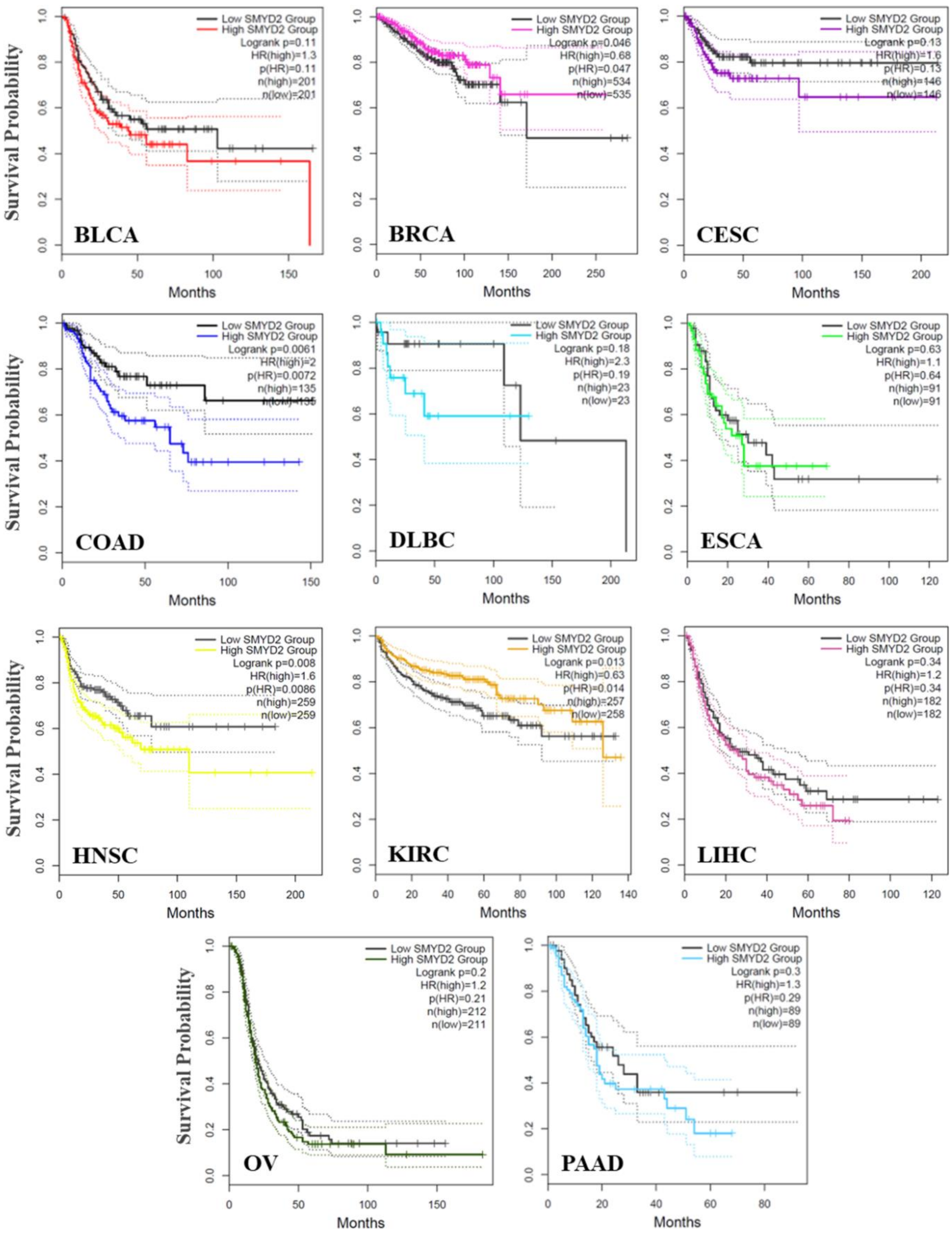


Figure 3.6: K-M plots represented the relationship between high (color) and low (black) *SMYD2* gene expression with disease-free survival of patients in multiple cancers.

3.3.3 Analysis of DNA methylation data

DNA methylation at the promoter level is more important epigenetic regulators of gene expression. Multiple malignancies have been found to have distinct and abnormal hypermethylation of CpG-rich regions (called CpG islands) or whole-genome hypermethylation [40], [41]. Hence, we discovered a possible correlation between *SMYD2* expression and methylation in a variety of cancer types. The beta value shows the degree of DNA methylation, which ranges from 0 (unmethylated) to 1 (completely methylated). In BRCA, CESC, COAD, ESCA, and KIRC cancers, the promoter methylation level was shown to be higher. The reduced methylation level was observed in BLCA, HNSC, LIHC, and PAAD tumors in comparison to their counterparts of normal tissues in the UALCAN analysis using the TCGA dataset (Figure 3.7).

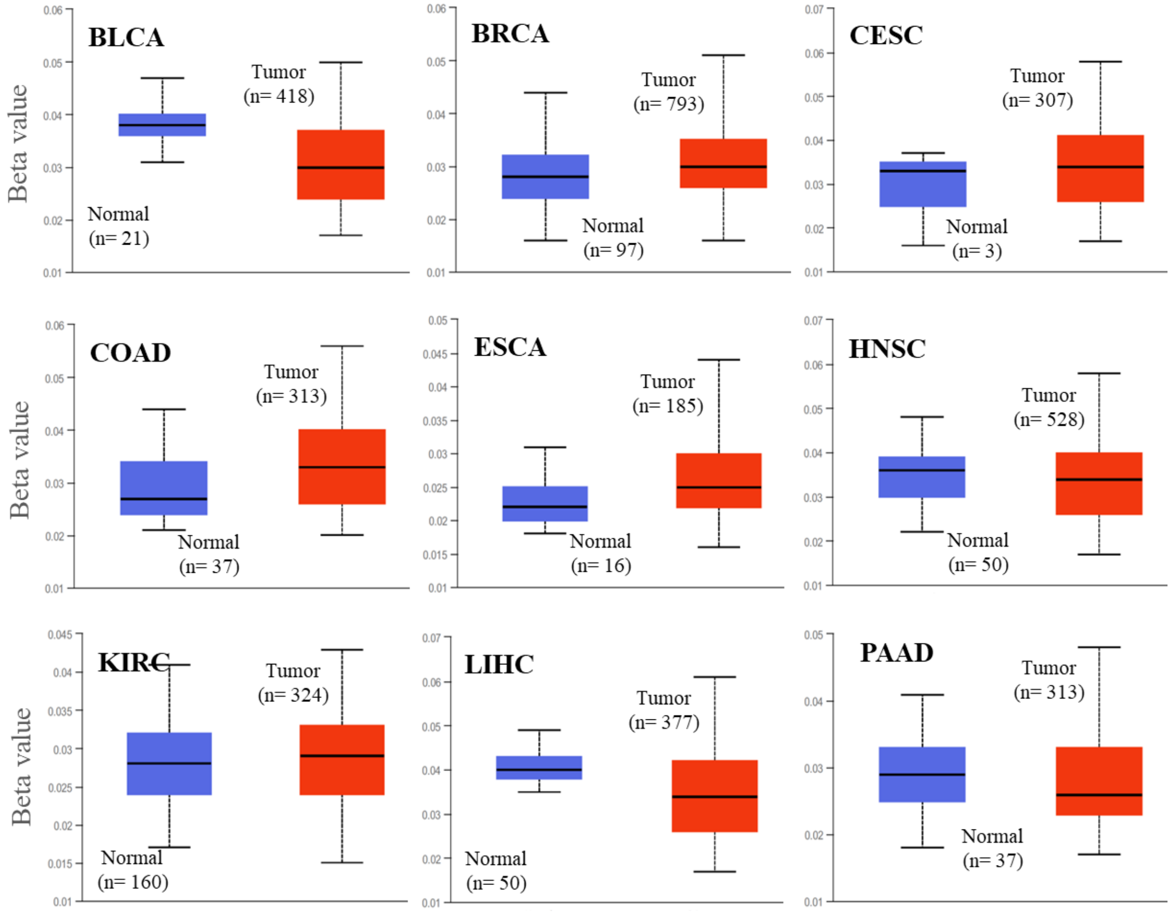


Figure 3.7: Promoter methylation of the *SMYD2* gene in different tumor tissues and their corresponding normal tissues. The promoter methylation level was expressed as a box plot, the blue color represents the normal tissues and the red color represents the tumor tissues. DNA methylation level has been shown in terms of beta value.

3.3.4 Analysis of genetic alteration data

The overexpression of the *SMYD2* gene in tumor tissues was induced by a genetic mutation, copy number alterations, and epigenetic control. Here, using the TCGA tumor sample via the cBioPortal database, the genetic alteration analysis of *SMYD2* in several cancers was performed. The *SMYD2* gene mutations were searched in 4617 samples from 11 different cancer studies, including breast, colon, bladder, esophageal, kidney, head and neck, cervical, liver, lymphoid, pancreatic, and ovarian. In SMYD2 0-433 AA sequences, 24 mutations were detected, with which missense mutation (15 mutations) being the most common form of genetic alteration (Figure 3.8A). The mutational analysis from TCGA data suggested that the highest alteration frequency of *SMYD2* (>10%) seems in lymphoid cancer patients. The amplification type of copy number alteration (CNA) was the key type in the lymphoid, breast, and liver cancer cases (Figure 3.8B**).** CNA works as raw materials for the expansion of gene family and affect the level of gene transcription and translation and associated with cancer [42]. The deletion type of mutation was detected in several cancer patients. Additionally, as seen in the TCGA dataset, amplification and gain were more common (Figure 3.8C). Moreover, the potential link between genetic variants of *SMYD2* and survival prognosis across all TCGA cancer was also investigated. The survival analysis in comparison to SMYD2 with and without alteration was analyzed in overall, disease-specific, disease-free, and progression-free survival (Figure 3.8D). The maximum survival difference was observed in disease-free survival analysis.

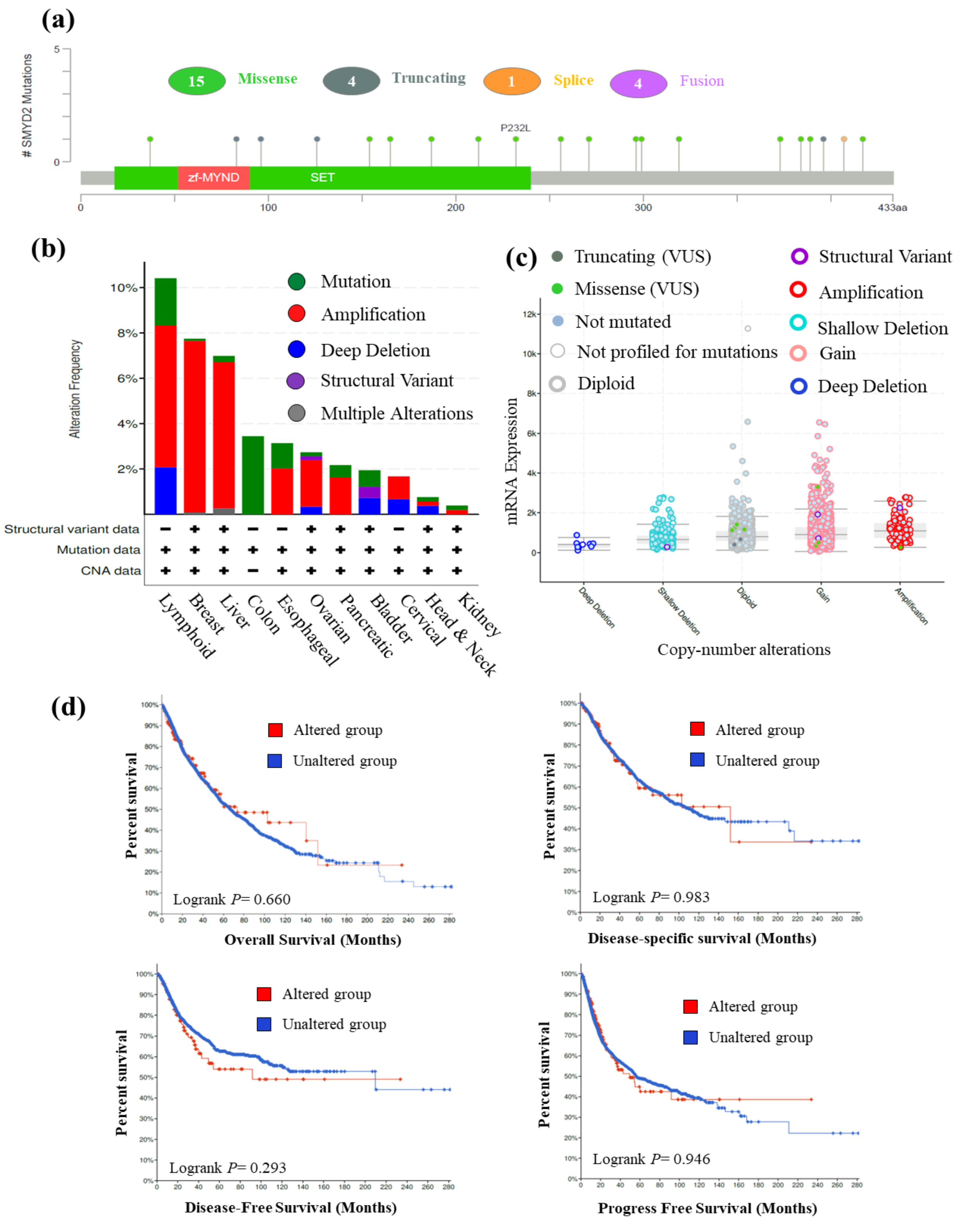


Figure 3.8: Genetic alteration and mutation of SMYD2. (A) The lollipop diagram depicts the alteration types within the SMYD2 protein sequence (1-433 AA) (B) The bar plot depicts

alteration frequencies and genome alteration in the *SMYD2* gene (C) a correlation between the CNAs of *SMYD2* in the TCGA dataset (D) a potential association between mutation status and various conditions of survival analysis.

3.3.5 Analysis of immune-infiltration data

In the tumor microenvironment, tumor-infiltrating immune cells are the key constituent and they play a crucial role in cancer progression, invasion, and metastasis [43], [44]. Cancer-related fibroblasts related to cancer present in the stroma of the tumor microenvironment have been discovered to play a role in the functional regulation of immune cells infiltrating malignancies [45]–[47]. Here, we used various algorithms such as CIBERSORT, TIMER, XCELL, MCPCOUNTER, CIBERSORT-ABS, EPIC, and QUANTISEQ to examine the correlation between immune cells infiltration and *SMYD2* expression in multiple types of cancers.

Analysis revealed that a statistically significant positive correlation (p<0.05 and Rho>0) was detected for BLCA (Rho= 0.208, *p*= 5.37E-05), BRCA (Rho= 0.115, *p*= 0.0002), DLBC (Rho= 0.340, *p*= 0.031), KIRC (Rho= 0.262, *p*= 1.12E-08), and LIHC (Rho= 0.129, *p*= 0.016) with CD8+ T-cell. The statistically significant negative correlation (p<0.05 and Rho<0) was observed between the immune cell infiltration of CD8+ T-cell and expression of *SMYD2* in BRCA-Basal (Rho= -0.160, *p*= 0.034), CESC (Rho= -0.137, *p*= 0.023), COAD (Rho= -0.171, *p*= 0.004), ESCA (Rho= -0.287, *p*= 9.50E-05), HNSC (Rho= -0.220, *p*= 7.69E-07), and PAAD (Rho= -0.195, *p*= 0.010) cancers of TCGA based on most or at least one algorithm. Furthermore, for BLCA (Rho= 0.208, *p*= 5.37E-05), BRCA-LumA (Rho= 0.208, *p*= 5.37E-05), BRCA-LumB (Rho= 0.237, *p*= 0.0009), CESC (Rho= 0.217, *p*= 0.0002), ESCA (Rho= 0.449, *p*= 2.67E-10), HNSC (Rho= 0.162, *p*= 0.0002), HNSC-HPV- (Rho= 0.240, *p*= 1.08E-06), and LIHC (Rho= 0.129, *p*= 0.016) a positive significant correlation was found between the value of fibroblast of cancer-associated infiltration and *SMYD2* expression. The scatter plot for these tumors was created by using one algorithm are shown in Figure 3.9. The negative association was detected for DLBC (Rho= -0.336, *p*= 0.031), KIRC (Rho= -0.135, *p*= 0.004), and PAAD (Rho= -0.200, *p*= 0.008) of TCGA tumors with cancer-associated cells based on most or at least one algorithms. The significant correlation of cancers with CD8+ T-cell and cancer-associated fibroblast cells both positively and negatively are tabulated in Annexure Table 3.1.

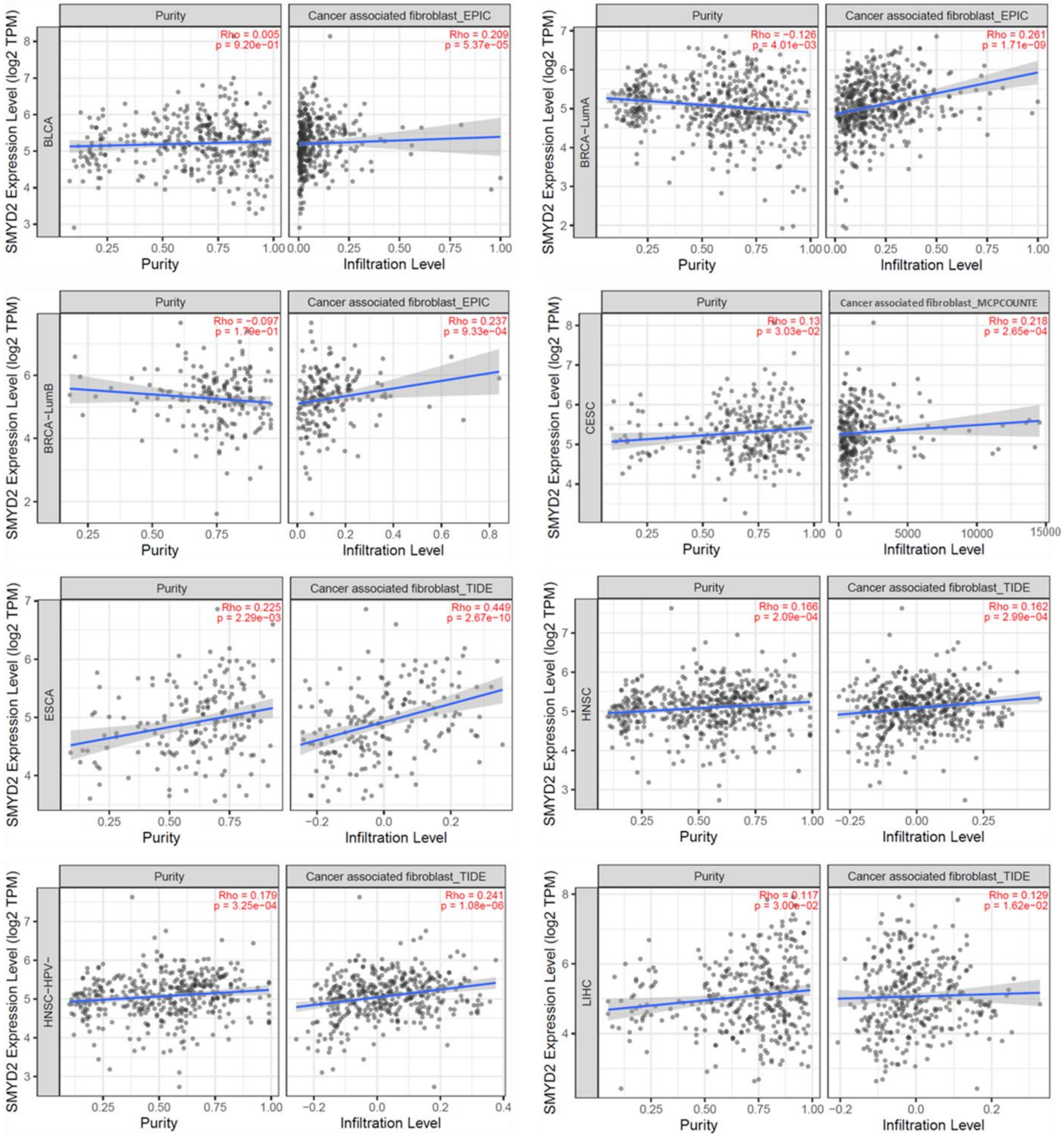


Figure 3.9: Correlation analysis between the immune infiltrations of fibroblasts associated with cancer and expression of *SMYD2* across multiple cancer types of TCGA data.

3.3.6 SMYD2-associated genes enrichment analysis

Further, we discover the *SMYD2*-associated proteins and co-expressed genes with SMYD2 for the analysis of pathways enrichment to learn more about the mechanism of SMYD2 in cancer. A total of 50 SMYD2-associated proteins were found using the STRING database based on experimental evidence. The protein-protein interaction network visualized in Cytoscape is shown in Figure 3.10.

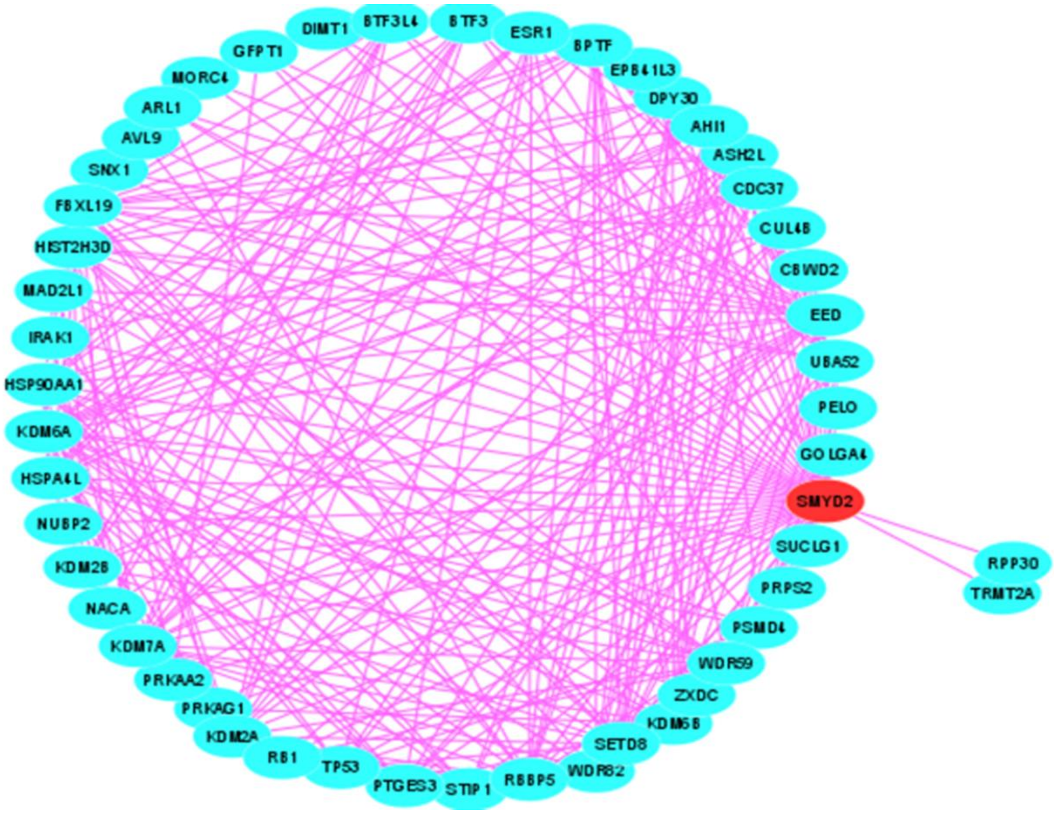


Figure 3.10: Protein-protein interaction network of experimentally determined partners of SMYD2 proteins through STRING database.

Furthermore, the GEPIA2 server was utilized to identify the first 100 correlated genes with SMYD2 by combining all TCGA tumor expression data. The highest correlation was found in Glyceronephosphate O-acyltransferase (GNPAT) (r= 0.42), Insulin-induced gene 2 (INSIG2) (r= 0.41), and Egl-9 family hypoxia-inducible factor 1 (EGLN1) (r= 0.40). Finally, we used the list of associated genes with SMYD2 in various cancers to do an ontology analysis to identify the putative signaling pathways.

The top 10 pathways from REACTOME and their interrelated genes were associated with the activation of Arylsulfatases, RAB Geranylgeranylation, Metabolism of proteins, Post-chaperonin tubulin folding pathway, Metal ion SLC transporters, Cargo concentration in ER, Protein folding, Gamma Carboxylation & Hypusin formation, Post-translation protein modification and Glycosphingolipid metabolism. All these pathways showed a significant correlation with *SMYD2* (Figure 3.11A). These pathways were associated with carcinogenesis [48]–[53]. The top 10 KEGG pathways are mainly associated with a HIF-1 signaling pathway, glycolysis/ gluconeogenesis, central carbon metabolism in cancer, thiamine metabolism, selenocompund metabolism, glycosaminoglycan degradation, one carbon pool by folate, histidine metabolism, renin-angiotensin system, and beta-alanine metabolism (Figure 3.11B). A significant association was observed in the HIF-1 signaling pathway. It is significantly associated with the progression and metastasis of cancer [54]–[56]. Additionally, we also calculated the GO terms for genes associated with *SMYD2* to see what functions they have in biological processes, molecular functions, and cellular components. The recommended GO features were mostly involved in mitochondrial transport and oxaloacetate metabolic process in biological process (Figure 3.11C), guanosine diphosphate in molecular function (Figure 3.11D), an integral component of the mitochondrial membrane in the process of cellular component (Figure 3.11E).

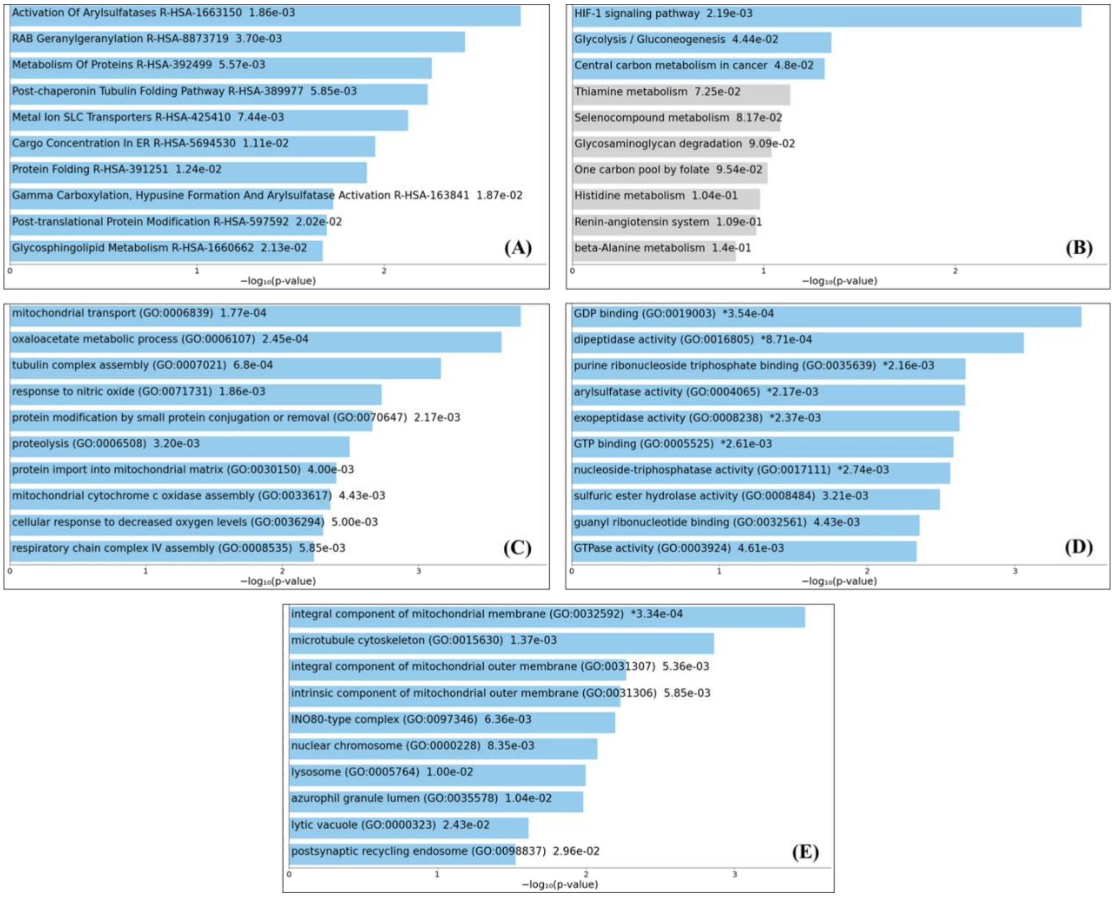


Figure 3.11: The bar charts shows top 10 *SMYD2*-associated enriched terms. The best 100 *SMYD2*-associated genes obtained from GEPIA2 in TCGA projects were used for ontology analysis. The top 10 enriched terms displayed based on –log10 (p-value), with the actual p-value shown next to each term that involved in the pathways and gene ontology (GO) functions. (A) RECTOME pathways 2022 (B) KEGG pathways 2021 (C) Enrichment analysis of GO biological process 2021 (D) Enrichment of GO molecular function 2021 (E) Enrichment of GO cellular component 2021. The more significant term present on the top. Colored bars correspond to term with significant p-values (<0.05). An asterisk (*) next to the p-value represented the term also has a significant adjusted p-value (<0.05). All significant terms involved in pathways and GO functions along with p-value and q-value (adjusted p-value) are shown in Annexure Table 3.2.

3.4 CONCLUSION

This data mining study used various bioinformatics web tools to elicit the *SMYD2* expression, prognostics value, DNA methylation, mutation, CNAs, and correlated genes of *SMYD2* in various human cancers. This multi-omics study shows that *SMYD2* is highly expressed in multiple cancers. Additionally, our findings will contribute to an enhanced understanding of the role of *SMYD2* in the process of tumorigenesis and metastasis. It provides a potential mechanism that suggested that the expression of *SMYD2* might modulate tumors. However, these results were based on data analysis, additional experimental verification will be needed to understand the complete molecular analysis of *SMYD2*. In conclusion, *SMYD2* would be a probable biomarker and a significant drug target for the prevention of human cancers.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/caac.21492.

[2] H. Sung *et al.*, “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, 2021, doi: 10.3322/caac.21660.

[3] C. Goswami *et al.*, “Molecular signature comprising 11 platelet-genes enables accurate blood-based diagnosis of NSCLC,” *BMC Genomics*, vol. 21, no. 1, p. 744, Oct. 2020, doi: 10.1186/s12864-020-07147-z.

[4] R. A. Copeland, M. E. Solomon, and V. M. Richon, “Protein methyltransferases as a target class for drug discovery,” *Nat Rev Drug Discov*, vol. 8, no. 9, pp. 724–732, Sep. 2009, doi: 10.1038/nrd2974.

[5] R. Hamamoto and Y. Nakamura, “Dysregulation of protein methyltransferases in human cancer: An emerging target class for anticancer therapy,” *Cancer Sci.*, vol. 107, no. 4, pp. 377–384, Apr. 2016, doi: 10.1111/cas.12884.

[6] A. K. Yadav and T. R. Singh, “Novel structural and functional impact of damaging single nucleotide polymorphisms (SNPs) on human SMYD2 protein using computational approaches,” *Meta Gene*, vol. 28, p. 100871, Jun. 2021, doi: 10.1016/j.mgene.2021.100871.

[7] E. L. Greer and Y. Shi, “Histone methylation: a dynamic mark in health, disease and inheritance,” *Nature Reviews Genetics*, vol. 13, no. 5, Art. no. 5, May 2012, doi: 10.1038/nrg3173.

[8] E. Eggert *et al.*, “Discovery and Characterization of a Highly Potent and Selective Aminopyrazoline-Based in Vivo Probe (BAY-598) for the Protein Lysine Methyltransferase SMYD2,” *J. Med. Chem.*, vol. 59, no. 10, pp. 4578–4600, 26 2016, doi: 10.1021/acs.jmedchem.5b01890.

[9] R. Hamamoto, V. Saloura, and Y. Nakamura, “Critical roles of non-histone protein lysine methylation in human tumorigenesis,” *Nat. Rev. Cancer*, vol. 15, no. 2, pp. 110–124, Feb. 2015, doi: 10.1038/nrc3884.

[10] H. Nguyen *et al.*, “LLY-507, a Cell-active, Potent, and Selective Inhibitor of Protein-lysine Methyltransferase SMYD2,” *J Biol Chem*, vol. 290, no. 22, pp. 13641–13653, May 2015, doi: 10.1074/jbc.M114.626861.

[11] A. K. Yadav and T. R. Singh, “Novel inhibitors design through structural investigations and simulation studies for human PKMTs (SMYD2) involved in cancer,” *Molecular Simulation*, vol. 47, no. 14, pp. 1149–1158, Sep. 2021, doi: 10.1080/08927022.2021.1957882.

[12] J. Huang *et al.*, “Repression of p53 activity by Smyd2-mediated methylation,” *Nature*, vol. 444, no. 7119, pp. 629–632, Nov. 2006, doi: 10.1038/nature05287.

[13] T. Voelkel, C. Andresen, A. Unger, S. Just, W. Rottbauer, and W. A. Linke, “Lysine methyltransferase Smyd2 regulates Hsp90-mediated protection of the sarcomeric titin springs and cardiac function,” *Biochim. Biophys. Acta*, vol. 1833, no. 4, pp. 812–822, Apr. 2013, doi: 10.1016/j.bbamcr.2012.09.012.

[14] L. A. Saddic *et al.*, “Methylation of the Retinoblastoma Tumor Suppressor by SMYD2,” *J Biol Chem*, vol. 285, no. 48, pp. 37733–37740, Nov. 2010, doi: 10.1074/jbc.M110.137612.

[15] H.-S. Cho *et al.*, “RB1 Methylation by SMYD2 Enhances Cell Cycle Progression through an Increase of RB1 Phosphorylation,” *Neoplasia*, vol. 14, no. 6, pp. 476–486, Jun. 2012.

[16] X. Zhang *et al.*, “Regulation of estrogen receptor α by histone methyltransferase SMYD2-mediated protein methylation,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 110, no. 43, pp. 17284–17289, Oct. 2013, doi: 10.1073/pnas.1307959110.

[17] M. Nakakido, Z. Deng, T. Suzuki, N. Dohmae, Y. Nakamura, and R. Hamamoto, “Dysregulation of AKT Pathway by SMYD2-Mediated Lysine Methylation on PTEN,” *Neoplasia*, vol. 17, no. 4, pp. 367–373, Apr. 2015, doi: 10.1016/j.neo.2015.03.002.

[18] L. Piao *et al.*, “The histone methyltransferase SMYD2 methylates PARP1 and promotes poly(ADP-ribosyl)ation activity in cancer cells,” *Neoplasia*, vol. 16, no. 3, pp. 257–264, 264.e2, Mar. 2014, doi: 10.1016/j.neo.2014.03.002.

[19] L. X. Li *et al.*, “Lysine methyltransferase SMYD2 promotes cyst growth in autosomal dominant polycystic kidney disease,” *J. Clin. Invest.*, vol. 127, no. 7, pp. 2751–2764, Jun. 2017, doi: 10.1172/JCI90921.

[20] D. K. Jarrell, K. N. Hassell, D. C. Crans, S. Lanning, and M. A. Brown, “Characterizing the Role of SMYD2 in Mammalian Embryogenesis—Future Directions,” *Vet Sci*, vol. 7, no. 2, May 2020, doi: 10.3390/vetsci7020063.

[21] Z. Tang, B. Kang, C. Li, T. Chen, and Z. Zhang, "GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis," *Nucleic Acids Research*, vol. 47, no. W1, pp. W556–W560, Jul. 2019, doi: 10.1093/nar/gkz430.

[22] T. Li *et al.*, "TIMER2.0 for analysis of tumor-infiltrating immune cells," *Nucleic Acids Res*, vol. 48, no. W1, pp. W509–W514, Jul. 2020, doi: 10.1093/nar/gkaa407.

[23] Z. Tang, C. Li, B. Kang, G. Gao, C. Li, and Z. Zhang, "GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses," *Nucleic Acids Res*, vol. 45, no. Web Server issue, pp. W98–W102, Jul. 2017, doi: 10.1093/nar/gkx247.

[24] D. S. Chandrashekar *et al.*, "UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses," *Neoplasia*, vol. 19, no. 8, pp. 649–658, Aug. 2017, doi: 10.1016/j.neo.2017.05.002.

[25] E. Cerami *et al.*, "The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data," *Cancer Discov*, vol. 2, no. 5, pp. 401–404, May 2012, doi: 10.1158/2159-8290.CD-12-0095.

[26] J. Gao *et al.*, "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal," *Sci Signal*, vol. 6, no. 269, p. pl1, Apr. 2013, doi: 10.1126/scisignal.2004088.

[27] A. M. Newman *et al.*, "Robust enumeration of cell subsets from tissue expression profiles," *Nat Methods*, vol. 12, no. 5, pp. 453–457, May 2015, doi: 10.1038/nmeth.3337.

[28] B. Li *et al.*, "Comprehensive analyses of tumor immunity: implications for cancer immunotherapy," *Genome Biol*, vol. 17, no. 1, p. 174, Aug. 2016, doi: 10.1186/s13059-016-1028-7.

[29] F. Finotello *et al.*, "Molecular and pharmacological modulators of the tumor immune contexture revealed by deconvolution of RNA-seq data," *Genome Med*, vol. 11, no. 1, p. 34, May 2019, doi: 10.1186/s13073-019-0638-6.

[30] J. Racle, K. de Jonge, P. Baumgaertner, D. E. Speiser, and D. Gfeller, "Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data," *Elife*, vol. 6, p. e26476, Nov. 2017, doi: 10.7554/eLife.26476.

[31] D. Aran, Z. Hu, and A. J. Butte, "xCell: digitally portraying the tissue cellular heterogeneity landscape," *Genome Biol*, vol. 18, no. 1, p. 220, Nov. 2017, doi: 10.1186/s13059-017-1349-1.

[32] E. Becht *et al.*, “Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression,” *Genome Biol*, vol. 17, no. 1, p. 218, Oct. 2016, doi: 10.1186/s13059-016-1070-5.

[33] P. Shannon *et al.*, “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks,” *Genome Res*, vol. 13, no. 11, pp. 2498–2504, Nov. 2003, doi: 10.1101/gr.1239303.

[34] M. V. Kuleshov *et al.*, “Enrichr: a comprehensive gene set enrichment analysis web server 2016 update,” *Nucleic Acids Res*, vol. 44, no. Web Server issue, pp. W90–W97, Jul. 2016, doi: 10.1093/nar/gkw377.

[35] X. Yi, X.-J. Jiang, and Z.-M. Fang, “Histone methyltransferase SMYD2: ubiquitous regulator of disease,” *Clin Epigenetics*, vol. 11, no. 1, p. 112, Aug. 2019, doi: 10.1186/s13148-019-0711-4.

[36] L. X. Li, J. X. Zhou, J. P. Calvet, A. K. Godwin, R. A. Jensen, and X. Li, “Lysine methyltransferase SMYD2 promotes triple negative breast cancer progression,” *Cell Death Dis*, vol. 9, no. 3, pp. 1–17, Feb. 2018, doi: 10.1038/s41419-018-0347-x.

[37] R. Ohtomo-Oda *et al.*, “SMYD2 overexpression is associated with tumor cell proliferation and a worse outcome in human papillomavirus-unrelated nonmultiple head and neck carcinomas,” *Hum Pathol*, vol. 49, pp. 145–155, Mar. 2016, doi: 10.1016/j.humpath.2015.08.025.

[38] J.-J. Sun, H.-L. Li, H. Ma, Y. Shi, L.-R. Yin, and S.-J. Guo, “SMYD2 promotes cervical cancer growth by stimulating cell proliferation,” *Cell & Bioscience*, vol. 9, no. 1, p. 75, Sep. 2019, doi: 10.1186/s13578-019-0340-9.

[39] L. Wu *et al.*, “SMYD2 promotes tumorigenesis and metastasis of lung adenocarcinoma through RPS7,” *Cell Death Dis*, vol. 12, no. 5, p. 439, May 2021, doi: 10.1038/s41419-021-03720-w.

[40] M. Kulis and M. Esteller, “DNA methylation and cancer,” *Adv Genet*, vol. 70, pp. 27–56, 2010, doi: 10.1016/B978-0-12-380866-0.60002-2.

[41] S. A. Wajed, P. W. Laird, and T. R. DeMeester, “DNA Methylation: An Alternative Pathway to Cancer,” *Ann Surg*, vol. 234, no. 1, pp. 10–20, Jul. 2001.

[42] A. Salmi *et al.*, “CNV-LDC: an optimised method for copy number variation discovery in low depth of coverage data,” *International Journal of Data Mining and Bioinformatics*, vol. 21, no. 2, pp. 169–181, Jan. 2018, doi: 10.1504/IJDMB.2018.096408.

[43] S. I. Grivennikov, F. R. Greten, and M. Karin, “Immunity, Inflammation, and Cancer,” *Cell*, vol. 140, no. 6, pp. 883–899, Mar. 2010, doi: 10.1016/j.cell.2010.01.025.

[44] T. L. Whiteside, “The tumor microenvironment and its role in promoting tumor growth,” *Oncogene*, vol. 27, no. 45, pp. 5904–5912, Oct. 2008, doi: 10.1038/onc.2008.271.

[45] T. Liu *et al.*, “Cancer-associated fibroblasts: an emerging target of anti-cancer immunotherapy,” *Journal of Hematology & Oncology*, vol. 12, no. 1, p. 86, Aug. 2019, doi: 10.1186/s13045-019-0770-1.

[46] Q. Ping *et al.*, “Cancer-associated fibroblasts: overview, progress, challenges, and directions,” *Cancer Gene Ther*, pp. 1–16, Mar. 2021, doi: 10.1038/s41417-021-00318-4.

[47] E. Sahai *et al.*, “A framework for advancing our understanding of cancer-associated fibroblasts,” *Nat Rev Cancer*, vol. 20, no. 3, pp. 174–186, Mar. 2020, doi: 10.1038/s41568-019-0238-1.

[48] Y. Shen *et al.*, “Arylsulfatase I is a prognostic biomarker for head and neck squamous cell carcinoma and Pan-cancer,” *Journal of Clinical Laboratory Analysis*, vol. 36, no. 9, p. e24600, 2022, doi: 10.1002/jcla.24600.

[49] C. Recchi and M. C. Seabra, “Novel functions for Rab GTPases in multiple aspects of tumour progression,” *Biochem Soc Trans*, vol. 40, no. Pt 6, pp. 1398–1403, Dec. 2012, doi: 10.1042/BST20120199.

[50] Z. Wei, X. Liu, C. Cheng, W. Yu, and P. Yi, “Metabolism of Amino Acids in Cancer,” *Frontiers in Cell and Developmental Biology*, vol. 8, 2021, Accessed: Oct. 08, 2022. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fcell.2020.603837

[51] Y. Zhang, Y. Zhang, K. Sun, Z. Meng, and L. Chen, “The SLC transporter in nutrient and metabolic sensing, regulation, and drug development,” *Journal of Molecular Cell Biology*, vol. 11, no. 1, pp. 1–13, Jan. 2019, doi: 10.1093/jmcb/mjy052.

[52] S. A. Oakes, “Endoplasmic Reticulum Stress Signaling in Cancer Cells,” *Am J Pathol*, vol. 190, no. 5, pp. 934–946, May 2020, doi: 10.1016/j.ajpath.2020.01.010.

[53] B. Ogretmen, “Sphingolipid metabolism in cancer signalling and therapy,” *Nat Rev Cancer*, vol. 18, no. 1, Art. no. 1, Jan. 2018, doi: 10.1038/nrc.2017.96.

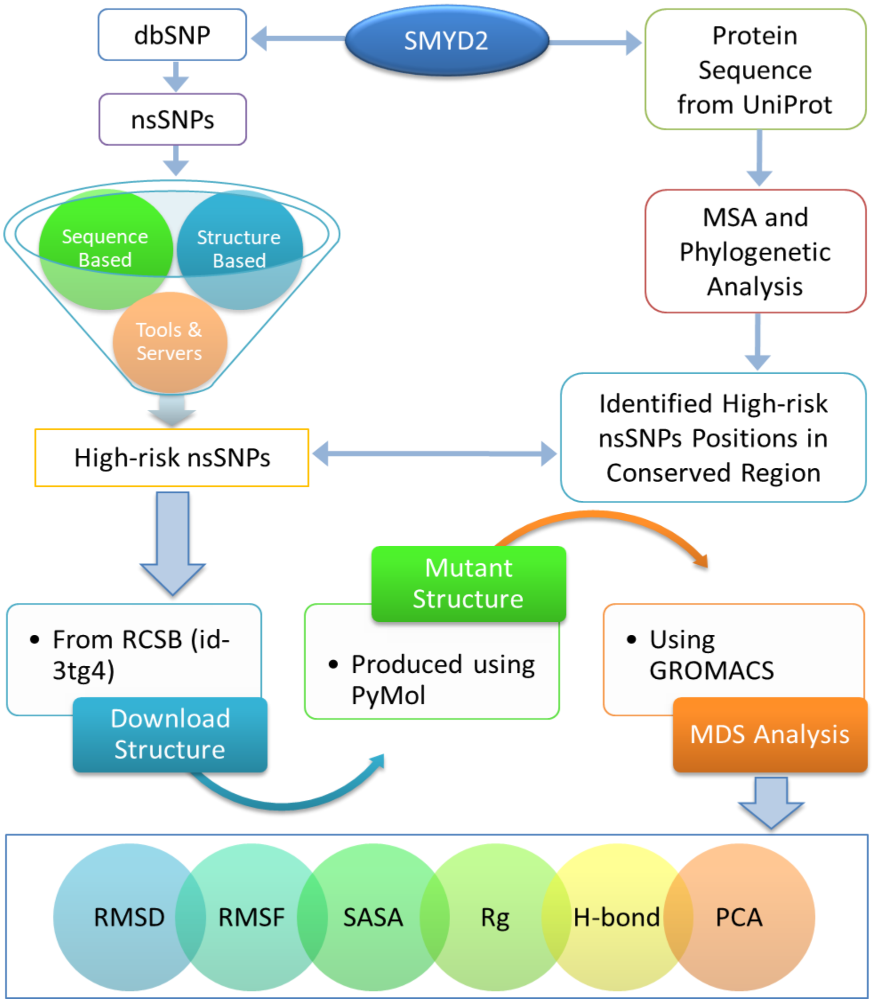
[54] G. N. Masoud and W. Li, “HIF-1α pathway: role, regulation and intervention for cancer therapy,” *Acta Pharm Sin B*, vol. 5, no. 5, pp. 378–389, Sep. 2015, doi: 10.1016/j.apsb.2015.05.007.

[55] X. Jin, L. Dai, Y. Ma, J. Wang, and Z. Liu, “Implications of HIF-1α in the tumorigenesis and progression of pancreatic cancer,” *Cancer Cell International*, vol. 20, no. 1, p. 273, Jun. 2020, doi: 10.1186/s12935-020-01370-0.

[56] S. K. Shaji, G. Drishya, D. Sunilkumar, P. Suravajhala, G. B. Kumar, and B. G. Nair, “Systematic understanding of anti-tumor mechanisms of Tamarixetin through network and experimental analyses,” *Sci Rep*, vol. 12, no. 1, Art. no. 1, Mar. 2022, doi: 10.1038/s41598-022-07087-6.

CHAPTER- 4

Sequence and Structure Level SNPs Analysis for SMYD2



ABSTRACT

Single-nucleotide polymorphisms (SNPs) associated with complex diseases can originate, remove, or change protein-coding sites. SNPs in cancer-related genes alter gene expression in a variety of ways. SMYD2 is a protein methyltransferase from the family of SMYD. Methylation inhibits the functional activity of some tumor suppressor proteins, which is important for tumor development and metastasis, by upregulating SMYD2 expression. Furthermore, epigenetic changes caused by genetic polymorphisms add to the elementary complication of cancer susceptibility linked to SNPs. As a result, SNP is a crucial genetic marker for studying the features of various malignancies. The functional impact of SMYD2-related SNPs is yet to be studied. Here, we assessed the genetic variants in SMYD2 and investigated their structural and functional consequences using a rigorous computational method. Out of the 264 nsSNPs, three nsSNPs (H207D, C209W, and C209R) have the most deleterious impact. According to a molecular dynamics simulation study, these mutations have a significant effect on the enzymatic activity of SMYD2. This computational approach is expected to aid in the prioritization of SNPs, clarifying the fundamental genetic pathways of cancer formation. The findings of this study are intended to be used as a starting point to perform the precision-based analysis, and they may be appropriate for the analysis of new genes.

4.1 INTRODUCTION

Cancer has exceeded heart disease as the world's top cause of mortality. Each year, cancer develops as a big health problem in society and the world as a result of rising instances. On a global scale, cancer is the topmost reason of disease and mortality. Cardiovascular illnesses, it is the second greatest cause of mortality [1]. According to the most recent WHO press release, the global estimate for new cancer diagnoses in 2020 is 19.3 million, with 10.0 million cancer deaths. On a global scale, cancer kills around 1 in every 6 people [1]. Cancer has also surpassed heart disease as India's biggest cause of death [2]. The International Agency for Cancer Research (IACR) predicts that by 2040, the worldwide burden of cancer would have increased to 28.4 million new cases. As a result, cancer research, the identification of important biomarkers, and the development of measures for active prevention and management are critical for initial cancer findings and treatment. The progression of human cancers is accelerated by the accumulation of genetic and epigenetic alterations. Because of the link between these increases and cancer, they can be used to diagnose malignancy. The SNPs for a certain gene is used to determine the accumulation amount of epigenetic changes [3]. As a result, SNPs are thought to be potential markers and are crucial for early cancer detection and therapy [4].

SMYD2 methylates the variety of nonhistone and histone proteins and is important for epigenetic control [5]. It acts as an oncogene by methylating tumor suppressor proteins, repressing their activities, and acting as a tumor suppressor [5]–[7]. SMYD2 is a protein having a length of 433 amino acid residues. The overall structure of SMYD2 consists of five different types of domains [8], [9]. SMYD2 has recently attracted a lot of attention due to its molecular mechanism and association with various human cancers [5]. Higher SMYD2 expression has been discovered in breast, bladder cervical, colorectal, esophageal, head and neck, pancreatic, lymphoma, and ovarian cancers [10], [11]. SMYD2 overexpression has been linked to tumor growth and metastasis in several studies [12], [13]. As a result of its substantial relationship with human tumors, SMYD2 might be a probable target for therapeutics in cancer.

SNPs are considered useful genetic markers for determining the cause of complicated disorders [14], [15]. SNPs found in the coding areas of physiologically essential genes have been shown to change the function of protein products [16]. Non-synonymous SNPs (nsSNPs) are responsible to perform the alteration of protein function, which can contribute to a variety of human disorders [17]–[19]. nsSNPs in cancer-related genes have recently attracted a lot of attention [20]. There have been several in-silico screening and molecular dynamics simulation (MDS) analyses present for nsSNPs associated with cancer [16], [17], [21], [22]. Despite the fact that SMYD2 is an important gene, no computational or experimental investigations have been published that identify SMYD2's detrimental nsSNPs.

As a result, the current study focused on the SMYD2 gene's nsSNPs which have a significant impact on a range of cancer characteristics. Given the huge number of SNPs accessible, it is necessary to identify the human SMYD2 gene's high-risk SNPs. SIFT, PROVEAN, PolyPhen-2, PhD-SNP, MutPred, iMutant, SNAP2, SNP&GO, Mutation Assessor, PON-P2, CUPSAT, and other computational tools were used to prioritize the risk associated nsSNPs present in dbSNP database. MDS analysis, on the other hand, was used to determine the influence of mutations on the atomic level of protein structure. MDS was used to examine the structural properties of mutant and native proteins. The three mutants such as H207R, C209R, and C209W have the most deleterious effect on the SMYD2 protein, according to the findings. These three mutations were studied further to see if they could have a function in human cancer regulation.

4.2 METHODOLOGY

4.2.1 Data collection

A total of 13944 SNPs of the SMYD2 gene were present in dbSNP (http://www.ncbi.nlm.nih.gov/snp/) [23]. 12821 (92%) of them were present in non-coding regions, 388 (3%) were in coding regions, and 735 (5%) were present in other regions of the gene. Among the 388 non-coding SNPs, non-synonymous SNPs account for 264 (2%) and synonymous SNPs account for 124 (1%) (Figure 4.1). We considered 264 nsSNPs (access date: December 25, 2019) for this study. The SMYD2 protein sequence (UniProt ID: Q9NRG4) was retrieved from UniProt (https://www.uniprot.org/) database [24], and structure for protein (PDB ID: 3TG4, X-ray resolution: 2.00Å) was downloaded from RCSB protein data bank (http://www.rcsb.org/) [9]. Figure 4.2 depicts the methodology used in the current investigation.

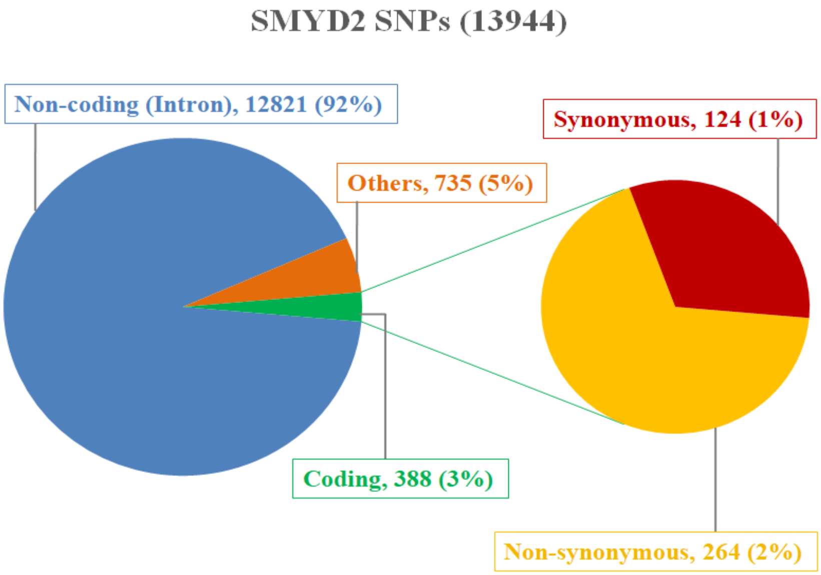


Figure 4.1: SMYD2 SNPs distribution as present in dbSNP.

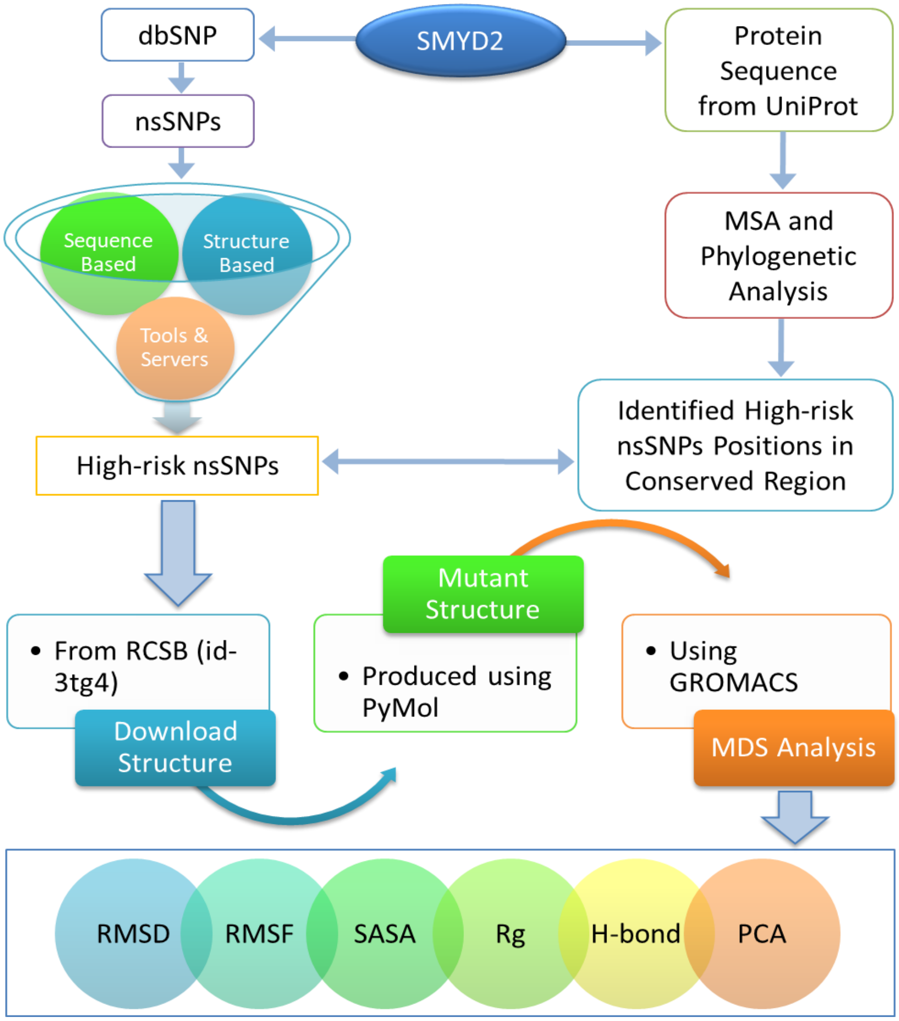


Figure 4.2: The overall strategy systematically used for the SNPs screening.

4.2.2 Analysis of deleterious nsSNPs

Several conventional tools and servers were utilized to analyze the functional influence of nsSNPs on the protein. High-risk nsSNPs were classified on the basis of predictions performed by below described tools and servers.

- SIFT (http://sift.jcvi.org/) predicts the substitution of amino acids that are responsible for change in the function of protein. The prediction is based on the sequence homology and physicochemical properties of amino acids. SIFT predicts mutation as deleterious if score is <0.05 while tolerable if score >0.05 [25].

- PROVEAN (http://provean.jcvi.org/index.php) web-server was used to predict the effect of amino acid variants. Prediction was performed by blast search by taking input query as protein sequence and amino acid variants. Cut-off was set at -2.5, score below the cut-off was predicted as deleterious nsSNPs [26].
- PolyPhen-2 (http://genetics.bwh.harvard.edu/pph2/) predict structural and function impact on protein by using amino acid substitutions and protein sequence. Prediction was performed based on the position-specific independent count (PSIC) score which ranges from 0-1. In Output, prediction score of substitution >0.15 assigned as possibly damaging, >0.85 assigned as probably damaging, and rest assigned as benign [27].
- PANTHER (http://pantherdb.org/tools/csnpScoreForm.jsp) was used to predict the effect of variants on protein function. It utilizes evolutionary information of related protein and calculates Substitution Position Specific evolutionary conservation (subPSEC) score. Using protein sequence, amino acid variants, and human organism as an input query PANTHER classified the variants [28]. The cutoff score subPSEC < –3 indicates that variants are highly associated with disease.
- SNP&GO (http://snps.biofold.org/snps-and-go/) is a web server based on Support Vector Machine (SVM). It was used to predict deleterious nsSNPs using functional annotation of protein. The server uses sequence or 3D structure of protein and also uses Gene Ontology terms. The variants having probability score >0.5 were predicted as disease-associated [29].
- PhD-SNP (http://snps.biofold.org/phd-snp/phd-snp.html) is a machine-learning based method used to predict whether nsSNP is related to disease or not. It distinguishes variants into disease and neutral on the basis of reliability index (cutoff 0.05) [30].
- MUTPred2 (http://mutpred.mutdb.org/) is a machine learning-based method to predict the pathogenicity of amino acid substitutions. A cutoff value 0.5 was used for pathogenicity prediction [31].
- P-Mut (http://mmb.irbbarcelona.org/PMut/) uses neural network algorithm and allows fast and accurate prediction of pathological character of single amino acid mutation. Using protein sequence as input query P-Mut predict neutral or disease-related variants [32]. A

prediction score more than 0.5 indicates nsSNPs having a pathological impact on protein function

- PON-P2 (http://structure.bmc.lu.se/PON-P2/) was used to predict the pathogenicity of amino acid substitutions using machine-learning based method. PON-P2 prediction was used through properties of amino acids, evolutionary conserved sequence, functional annotation, and GO annotations. It classifies nsSNPs in three groups based on cutoff as: pathogenic (≥0.8), unknown (0.3 to 0.8) and natural (below 0.3) [33].

4.2.3 Analysis of protein stability change caused by mutation

To check the stability of protein caused by single amino acid substitution, following tools were used.

- I-mutant 3.0 (http://gpcr2.biocomp.unibo.it/cgi/predictors/I-Mutant3.0/I-Mutant3.0.cgi) predict the change in protein stability caused by single amino acid substitution of structure or sequence of protein. It calculates Gibbs-free energy change (DDG) of native and mutant protein. DDG value can denote that either variants are stabilizing or destabilizing the protein [34]. I-Mutant 3.0 furthermore classifies mutations into three categories: neutral mutation ($-0.5 \leq$ DDG≤ 0.5), large decrease (≤ -0.5), and large increase (>0.5).
- MuPro (http://mupro.proteomics.ics.uci.edu/) is SVM and Neural network based web server that predict the effect of single amino acid substitution on protein stability. The stability score ranges from -1 to 1 [35]. A score less than 0 means the mutation decreases the protein stability a score more than 0 means the mutation increases the protein stability.
- Dyamute (http://biosig.unimelb.edu.au/dynamut/) is a web server used for protein dynamics analysis by sampling conformation. It also helps to predict the impact of mutation on dynamics and protein stability. To generate a consensus prediction of mutation's impact, it incorporates graph-based signature with normal mode dynamics [36]. It defined $\Delta\Delta G \geq 0$ as stabilizing and $\Delta\Delta G < 0$ as destabilizing.
- SDM (http://marid.bioc.cam.ac.uk/sdm2) used knowledge-based approach for the prediction of protein stability caused by mutations. This server calculates stability difference score of wild type and mutant protein by using protein structure and amino acid

substitution as a query [37]. It considers stabilizing (DDG ≥ 0.0) and destabilising (DDG < 0.0).

- CUPSAT (http://cupsat.tu-bs.de/) server used to predicts the protein stability changes upon single amino acid mutation. The prediction method uses amino acid-atom potential and torsion angle distribution to assess the amino acid environment of the mutation site [38]. The negative and positive predicted ΔΔG value indicates the destabilizing and stabilizing effect.
- Mutation Assessor (http://mutationassessor.org/r3/) predicts the effect of variants by entropy differences of aligned conserved sequence positions. The conservation of each residue was calculated based on information of sequence homology of protein families and its subfamilies. Predicted variants were classified into high (>3.5), medium (1.9 to ≤ 3.5), low (0.8 to <1.9), and neutral (≤0.8). For this analysis, high or medium cutoff level was set for deleterious variants [39].

4.2.4 Analysis of SNPs present in conserved regions and phylogeny of SMYD2

ConSurf (http://consurf.tau.ac.il/) [40] was utilized to predict the position of amino acids present in the evolutionarily conserved region. The Bayesian technique was used to define the conservation score by taking the input query as a protein sequence or structure. The conservation score ranges from 1 to 9, with the residues having a score of 1 signifying the highly mutated and a score of 9 indicating that the residues are well conserved. Structure and functional amino acids were also predicted by ConSurf [41], [42].

The SMYD2 protein for humans (UniProt ID: SMYD2 HUMAN) with the protein sequence of seven diverse species, including Rattus norvegius (SMYD2 RAT), Mus musculus (SMYD2 MOUSE), Danio rerio (SMYD2 DANRE), Sus scrofa (SMYD2 PIG), Bos taurus (SMYD2 BOVIN), Xenopus laevis (SMYD2_XENLA) and Gallus gallus (SMYD2_CHICK). The neighbor-joining method was used to produce multiple sequence alignment (MSA) using the MEGA11 tool [43], and the phylogenetic tree was constructed using the bootstrap value 1000.

4.2.5 Mutant protein modeling and analysis of the structural effect

The protein structure for SMYD2 (PDB ID: 3TG4) was utilized to estimate the structural impact of nsSNPs. PyMol software was used to insert the mutations of H207D, C209W, and C209R in the SMYD2 structure [44]. Using the HOPE (https://www3.cmbi.umcn.nl/hope/) service, the structural ramifications of chosen nsSNPs were investigated [45].

4.2.6 Molecular Dynamics simulation

As evidenced by other similar research [46]–[48] MDS analysis is a potent method for revealing the effects of mutations on the protein. As a result, we used MDS study to examine the effect of mutations on the structure of SMYD2 protein. GROMACS 2018.1 [49], [50], software was used to perform the MDS study. MDS analysis was performed for the mutants and native proteins by considering the water molecule as TIP3P and cubic shape box, which were neutralized by the addition of 7 Na + ions. MDS analysis for 100 ns at 300K using the Amber99sb-ildn force field was performed and at every 10 ps time, interwell trajectory was recorded. By using the various GROMACS utilities we determined the root mean square fluctuation (RMSF), root means square deviation (RMSD), a number of hydrogen bonds, solvent accessible surface area (SASA), and radius of gyration (Rg). The associated motions generated by the mutations were also predicted using principal component analysis (PCA). It was also executed to see the difference in the associated motions caused by the mutations.

4.3 RESULTS AND DISCUSSION

4.3.1 Prediction of high-risk nsSNPs

In this study, we focused only on the nsSPNs, and perform the structural and functional effect on the cancer-causing SMYD2 protein. To find relevant nsSNPs for SMYD2, a total of 264 nsSNP datasets were examined. The influence of all nsSNPs was investigated on the protein structure and function of the SMYD2 protein. Initially, we subjected 264 nsSNPs in SIFT, PROVEAN, PhD-SNP, PANTHER, and SNP&GO tools to screen out deleterious nsSNPs. These initial tools produced 21 nsSNPs, therefore we used these 21 nsSNPs for further analysis. SIFT projected that eight nsSNPs would be most damaging that have a prediction score of 0.00, and the 13 nsSNPs considered as damaging have scores ranging from 0.001 to 0.5. The prediction score for PROVEAN ranged from -2.56 to -11.32 (Table 4.1).

Table 4.1: Prediction of deleterious nsSNPs through PROVEAN, SIFT, PhD-SNP, SNP&GO, PANTHER, PolyPhen2, Pmut, MutPred, and PON-P2 tools. The nsSNPs commonly predicted deleterious by all tools are represented in bold text.

nsSNPs (rs id)	**Amino acid substitution**	**PROVEAN**		**SIFT**		**PhD-SNP**	**SNP& GO**	**PANT-HER**	**PolyPhen2**		**Pmut**		**MutPred**		**PON-P2**	
		Score	**Prediction**	**Score**	**Prediction**	**Prediction**			**Score**	**Prediction**	**Score**	**Prediction**	**Score**	**Prediction**	**Score**	**Prediction**
rs201567034	R48G	-2.56	Del	0.025	Dam	Dis	Dis	Dis	0.97	Probably D	0.32	Neutral	0.56	Pathogenic	0.396	UN
rs746777864	C68W	-9.06	Del	0	Dam	Dis	Dis	Dis	1	Probably D	0.88	Disease	0.96	Pathogenic	0.576	UN
rs148351602	**P102L**	**-6.93**	**Del**	**0**	**Dam**	**Dis**	**Dis**	**Dis**	**0.97**	**Probably D**	**0.64**	**Disease**	**0.59**	**Pathogenic**	**0.82**	**PG**
rs770105953	**R111S**	**-4.84**	**Del**	**0.05**	**Dam**	**Dis**	**Dis**	**Dis**	**1**	**Probably D**	**0.72**	**Disease**	**0.79**	**Pathogenic**	**0.86**	**PG**
rs757673085	G195R	-7.58	Del	0	Dam	Dis	Dis	Dis	1	Probably D	0.82	Disease	0.86	Pathogenic	0.746	UN
rs750954495	**H207D**	**-8.5**	**Del**	**0**	**Dam**	**Dis**	**Dis**	**Dis**	**1**	**Probably D**	**0.89**	**Disease**	**0.94**	**Pathogenic**	**0.883**	**PG**
rs199700774	**C209W**	**-10.38**	**Del**	**0.001**	**Dam**	**Dis**	**Dis**	**Dis**	**1**	**Probably D**	**0.9**	**Disease**	**0.94**	**Pathogenic**	**0.865**	**PG**
rs780140828	**C209R**	**-11.32**	**Del**	**0.001**	**Dam**	**Dis**	**Dis**	**Dis**	**1**	**Probably D**	**0.9**	**Disease**	**0.95**	**Pathogenic**	**0.94**	**PG**
rs1268352357	**P211T**	**-7.56**	**Del**	**0**	**Dam**	**Dis**	**Dis**	**Dis**	**0.99**	**Probably D**	**0.84**	**Disease**	**0.78**	**Pathogenic**	**0.819**	**PG**
rs755437613	N212T	-4.78	Del	0.002	Dam	Dis	Dis	Dis	0.88	Possible D	0.73	Disease	0.77	Pathogenic	0.506	UN
rs1489727899	**Y217C**	**-5.32**	**Del**	**0.001**	**Dam**	**Dis**	**Dis**	**Dis**	**1**	**Probably D**	**0.56**	**Disease**	**0.83**	**Pathogenic**	**0.847**	**PG**
rs764794193	E235G	-6.58	Del	0.002	Dam	Dis	Dis	Dis	0.16	Benign	0.82	Disease	0.78	Pathogenic	0.649	UN
rs1425329978	E235K	-3.73	Del	0.002	Dam	Dis	Dis	Dis	0.88	Possible D	0.49	Neutral	0.74	Pathogenic	0.86	PG
rs1355012130	Y245C	-3.23	Del	0.024	Dam	Dis	Dis	Dis	1	Probably D	0.58	Disease	0.5	Unknown	0.948	PG
rs1453879180	**R250T**	**-5.65**	**Del**	**0**	**Dam**	**Dis**	**Dis**	**Dis**	**1**	**Probably D**	**0.84**	**Disease**	**0.93**	**Pathogenic**	**0.828**	**PG**
rs1338877658	**Y258N**	**-8.23**	**Del**	**0.002**	**Dam**	**Dis**	**Dis**	**Dis**	**1**	**Probably D**	**0.87**	**Disease**	**0.85**	**Pathogenic**	**0.944**	**PG**
rs768344398	**C262W**	**-10.37**	**Del**	**0.001**	**Dam**	**Dis**	**Dis**	**Dis**	**1**	**Probably D**	**0.88**	**Disease**	**0.95**	**Pathogenic**	**0.879**	**PG**
rs138767186	C267R	-11.12	Del	0	Dam	Dis	Dis	Dis	1	Possible D	0.88	Disease	0.91	Pathogenic	0.95	PG
rs1468913646	R299C	-3.3	Del	0.04	Dam	Dis	Dis	Dis	0.26	Benign	0.4	Neutral	0.37	Unknown	0.72	UN
rs887231491	R306W	-3.83	Del	0.001	Dam	Dis	Dis	Dis	0.39	Benign	0.4	Neutral	0.62	Pathogenic	0.807	PG
rs535733694	A359S	-2.7	Del	0	Dam	Dis	Dis	Dis	0.99	Probably D	0.63	Disease	0.53	Pathogenic	0.507	UN

Del: Deleterious; Dam: Damaging; Dis: Disease; D: Damaging; UN: Unknown, PG: Pathogenic

Table 4.2: Prediction of change in protein stability caused due to the substitution of amino acid using iMutant, Mupro, Dyamute, Mutation Assessor, SDM, and CUPSAT programs. The nsSNPs that predicted with reduced stability with all tools are presented in bold text.

Amino acid substitution	iMutant		MuPro		Dyamute	Mutation assessor		SDM		CUPSAT		
	ΔΔG	Stability	ΔΔG	Prediction	Prediction	FI Score	Prediction	Score	Prediction	ΔΔG	Torsion	Overall Stability
R48G	-1.49	Decrease	-2.067	Decrease	Destabilizing	1.605	Low	-2.11	RS	-3.44	UF	Destabilizing
C68W	-0.1	Decrease	-1.311	Decrease	Stabilizing	4.51	High	-1.41	RS	-14.48	UF	Destabilizing
P102L	-0.56	Decrease	-0.0302	Decrease	Stabilizing	2.425	Medium	0.69	IS	1.39	UF	Stabilizing
R111S	-1.13	Decrease	-1.597	Decrease	Destabilizing	3.13	Medium	-3.45	RS	4.93	F	Stabilizing
G195R	-0.53	Decrease	-0.643	Decrease	Destabilizing	3.425	Medium	-2.25	RS	-5.62	UF	Destabilizing
H207D	**-0.31**	**Decrease**	**-0.851**	**Decrease**	**Destabilizing**	**3.51**	**High**	**-1.62**	RS	**-2.75**	**UF**	**Destabilizing**
C209W	**-0.31**	**Decrease**	**-0.772**	**Decrease**	**Destabilizing**	**3.555**	**High**	**-1.37**	RS	**-12.98**	**UF**	**Destabilizing**
C209R	**-0.47**	**Decrease**	**-1.012**	**Decrease**	**Destabilizing**	**3.555**	**High**	**-1.54**	RS	**-8.01**	**UF**	**Destabilizing**
P211T	-1.46	Decrease	-1.341	Decrease	Destabilizing	3.495	Medium	-0.84	RS	-4.41	UF	Destabilizing
N212T	-0.01	Increase	-0.581	Decrease	Stabilizing	2.7	Medium	-0.29	RS	-3.03	F	Destabilizing
Y217C	-1.19	Decrease	-1.117	Decrease	Destabilizing	1.995	Medium	-1.42	RS	-6.61	F	Destabilizing
E235G	-1.22	Decrease	-1.443	Decrease	Destabilizing	3.465	Medium	-0.9	RS	-2.48	UF	Destabilizing
E235K	-0.68	Decrease	-0.96	Decrease	Destabilizing	3.465	Medium	-0.93	RS	-2.74	UF	Destabilizing
Y245C	-1.49	Decrease	-1.156	Decrease	Destabilizing	2.565	Medium	-0.15	RS	1.56	F	Stabilizing
R250T	-1.21	Decrease	-0.741	Decrease	Destabilizing	2.62	Medium	-3.49	RS	-1.53	UF	Destabilizing
Y258N	-1.45	Decrease	-1.840	Decrease	De-stabilizing	2.62	Medium	-1.17	RS	-3.65	F	Destabilizing
C262W	-0.06	Increase	-0.565	Decrease	Stabilizing	2.65	Medium	-1.59	RS	0.29	F	Stabilizing
C267R	-0.02	Increase	-1.139	Decrease	Stabilizing	2.65	Medium	-0.93	RS	-0.13	F	Destabilizing
R299C	-1.06	Decrease	-1.408	Decrease	Destabilizing	1.79	Low	-0.76	RS	-5.23	UF	Destabilizing
R306W	-0.64	Decrease	-0.767	Decrease	Destabilizing	2.045	Medium	-0.21	RS	1.32	F	Stabilizing
A359S	-0.61	Decrease	-0.832	Decrease	Destabilizing	2.62	Medium	-2.11	RS	-1.99	UF	Destabilizing

ΔΔG: Delta Delta G (DDG); RS: Reduced Stability; IS: Increased Stability UF: Unfavorable; F: Favorable

A total of 21 initially screened nsSNPs were further analysed by using the Polyphen2, Pmut, MutPred and PON-P2. A total of 15 nsSNPs were predicted as probably damaging out of 21 nsSNPs by Polyphen2. Out of 21 nsSNPs 17 nsSNPs predicted as disease associated by Pmut, 19 nsSNPs were predicted as pathogenic through MutPred and 14 nsSNP were predicted as pathogenic. A total of 10 nsSNPs were predicted as highly disease associated by all tools (Table 4.1). Then these 21 nsSNP were further subject to predict the change in protein stability caused by these mutations. To analyze the change in protein stability due to mutations, we utilized iMutant3.0, MuPro, Dyamute, Mutation assessor, SDM and CUPSAT. We find decreased stability in 18 nsSNPs predicted by iMutant and 21 nsSNPs predicted by the MuPro program that may alter the stability of the protein. The Mutation assessor identified four nsSNPs to be highly disease-associated. Dyamute, SDM, and CUPSAT find reduced stability for 16, 20, and 16 nsSNPs, respectively, out of a total of 21 nsSNPs (Table 4.2). Finally we identified the consensus nsSNPs that are predicted as deleterious from Table 4.1 (represented as bold text) and consensus nsSNPs that reduced the protein stability from Table 4.2 (represented as bold text). A total of 15 computational techniques were utilized in this study to filter out important SNPs in order to achieve reliable results and eliminate the chance of false positives. Three nsSNPs such as H207D, C209W, and C209R have the most damaging effect by aggregating the predictions of all tools. Therefore, three nsSNPs that were most deleterious and reduced the protein stability were selected as final nsSNPs for further analysis.

MutPred2 was used to estimate the molecular mechanism for these three nsSNPs. Metal-binding was altered, catalytic sites were gained or lost, disulfide linkage was gained or lost, and N-linked glycosylation was gained or lost. Table 4.4 shows the details of the molecular mechanisms that were predicted. The HOPE server was used to undertake a further examination of extremely disease-associated nsSNPs, and the sites of the altered amino acids present in the protein structure are indicated in Figure 4.3. From here, the most deleterious three nsSNPs such as H207D, C209W, and C209R were subject to MDS analysis.

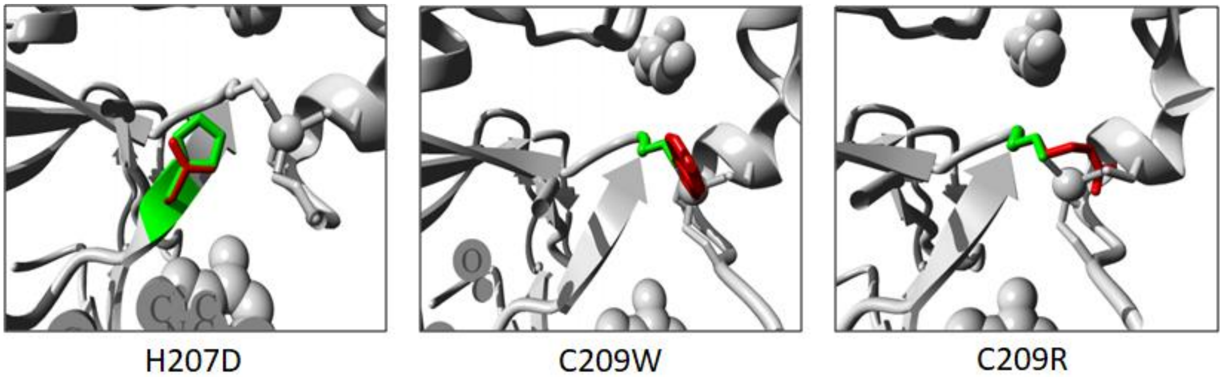


Figure 4.3: Most deleterious nsSNPs predicted by using the server HOPE. In the structure, the red color represented the mutant, and the green color represented the native amino acids.

Table 4.3: Highly disease-associated three nsSNPs and their molecular mechanisms predicted by MutPred2 tool.

nsSNP	rs id	MutPred2 Score	p-value	Molecular Mechanism
H207D	rs750954495	0.94	7.8e-03	Altered Metal binding
			0.03	Loss of N-linked glycosylation at N206
			2.2e-03	Loss of Catalytic site at H207
			0.03	Loss of Disulfide linkage at C209
C209W	rs199700774	0.945	8.8e-03	Altered Metal binding
			0.02	Loss of Disulfide linkage at C209
			4.3e-03	Loss of Catalytic site at H207
			0.02	Gain of N-linked glycosylation at N206
C209R	rs780140828	0.95	8.6e-03	Altered Metal binding
			2.1e-03	Loss of Catalytic site at H207
			0.03	Gain of N-linked glycosylation at N206
			0.02	Loss of Disulfide linkage at C209

4.3.2 Analysis of conservation and evolutionary of sequence

The conserved amino acid residues have an influence on the properties of protein structure and function. Protein evolutionary information is required to identify variants with severe disease-associated mutations. Validating certain mutants through the conservation score for any specific amino acid with the help of the ConSurf service is another level of conformation. The results revealed that certain mutations are conserved residues, as indicated by the orange rectangle box

(Figure 4.4). The core set domain contains the most conserved residues (183-245). The selected mutants have a conservation score of 8 and 9 on a scale of 10.

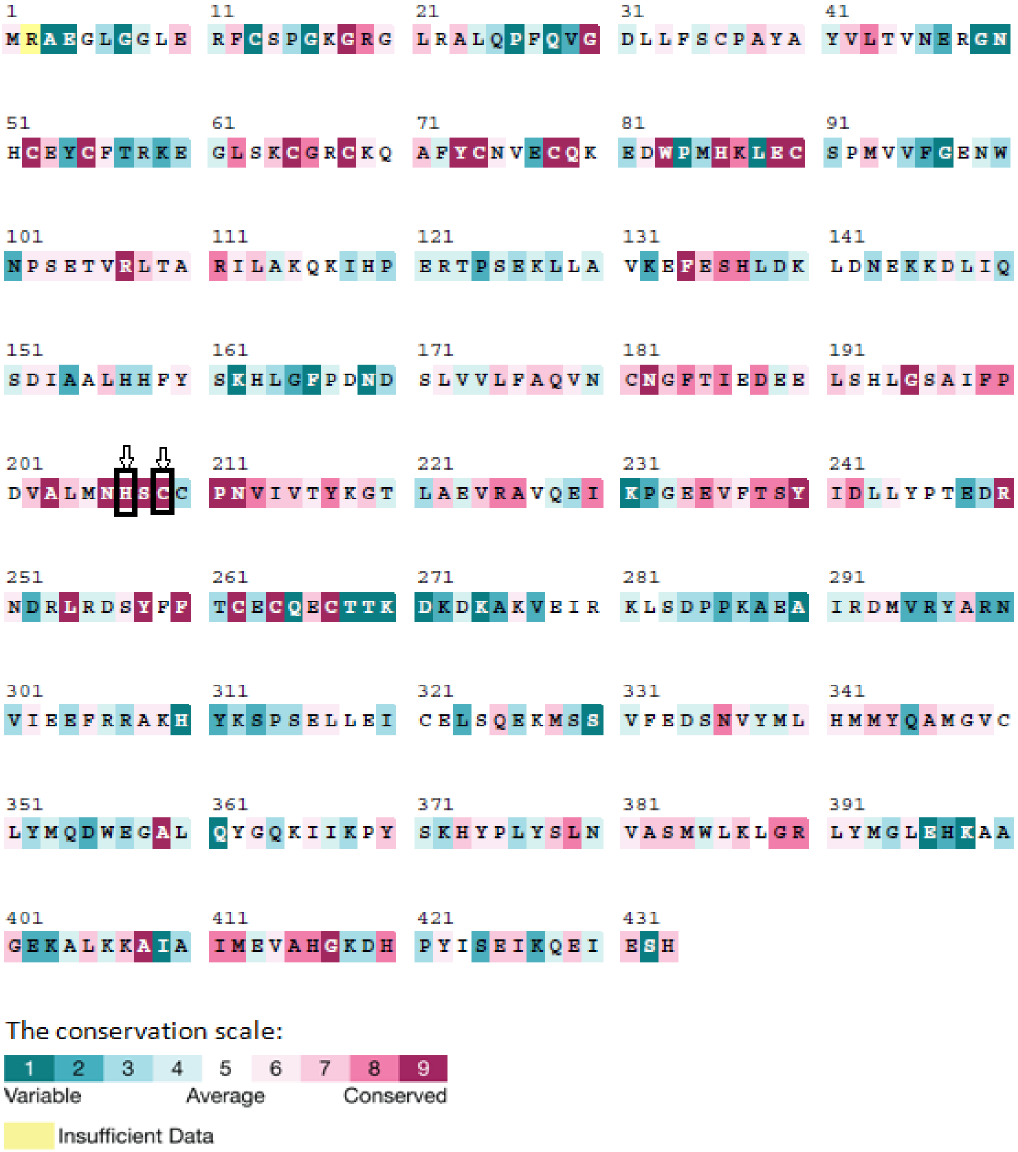

1 MRAEGLGGLE 11 RFCSPGKGRG 21 LRALQPFQVG 31 DLLFSCPAYA 41 YVLTVNERGN
51 HCEYCFTRKE 61 GLSKCGRCKQ 71 AFYCNVECQK 81 EDWPMHKLEC 91 SPMVVFGENW
101 NPSETVRLTA 111 RILAKQKIHP 121 ERTPSEKLLA 131 VKEFESHLDK 141 LDNEKKDLIQ
151 SDIAALHHFY 161 SKHLGFPDND 171 SLVVLFAQVN 181 CNGFTIEDEE 191 LSHLGSAIFP
201 DVALMNHSCC 211 PNVIVTYKGT 221 LAEVRAVQEI 231 KPGEEVFTSY 241 IDLLYPTEDR
251 NDRLRDSYFF 261 TCECQECTTK 271 DKDKAKVEIR 281 KLSDPPKAEA 291 IRDMVRYARN
301 VIEEFRRAKH 311 YKSPSELLEI 321 CELSQEKMSS 331 VFEDSNVYML 341 HMMYQAMGVC
351 LYMQDWEGAL 361 QYGQKIIKPY 371 SKHYPLYSLN 381 VASMWLKLGR 391 LYMGLEHKAA
401 GEKALKKAIA 411 IMEVAHGKDH 421 PYISEIKQEI 431 ESH
The conservation scale:
1 2 3 4 5 6 7 8 9
Variable
Average
Conserved
Insufficient Data

Figure 4.4: Prediction of evolutionarily conserved residues of SMYD2 protein using ConSurf server. The positions of deleterious nsSNPs are labeled in a black rectangle box.

This investigation is also validated by the MSA study. The MEGA7 program was used to do MSA analysis on eight diverse species, including humans. Analysis results revealed that conserved residues in all species are homologous, as indicated by the black rectangular box (Figure 4.5A). The evolutionary relationship between the eight different species was also determined via phylogenetic analysis. It symbolized the strong relationship between people and cattle and pigs (Figure 4.5B).

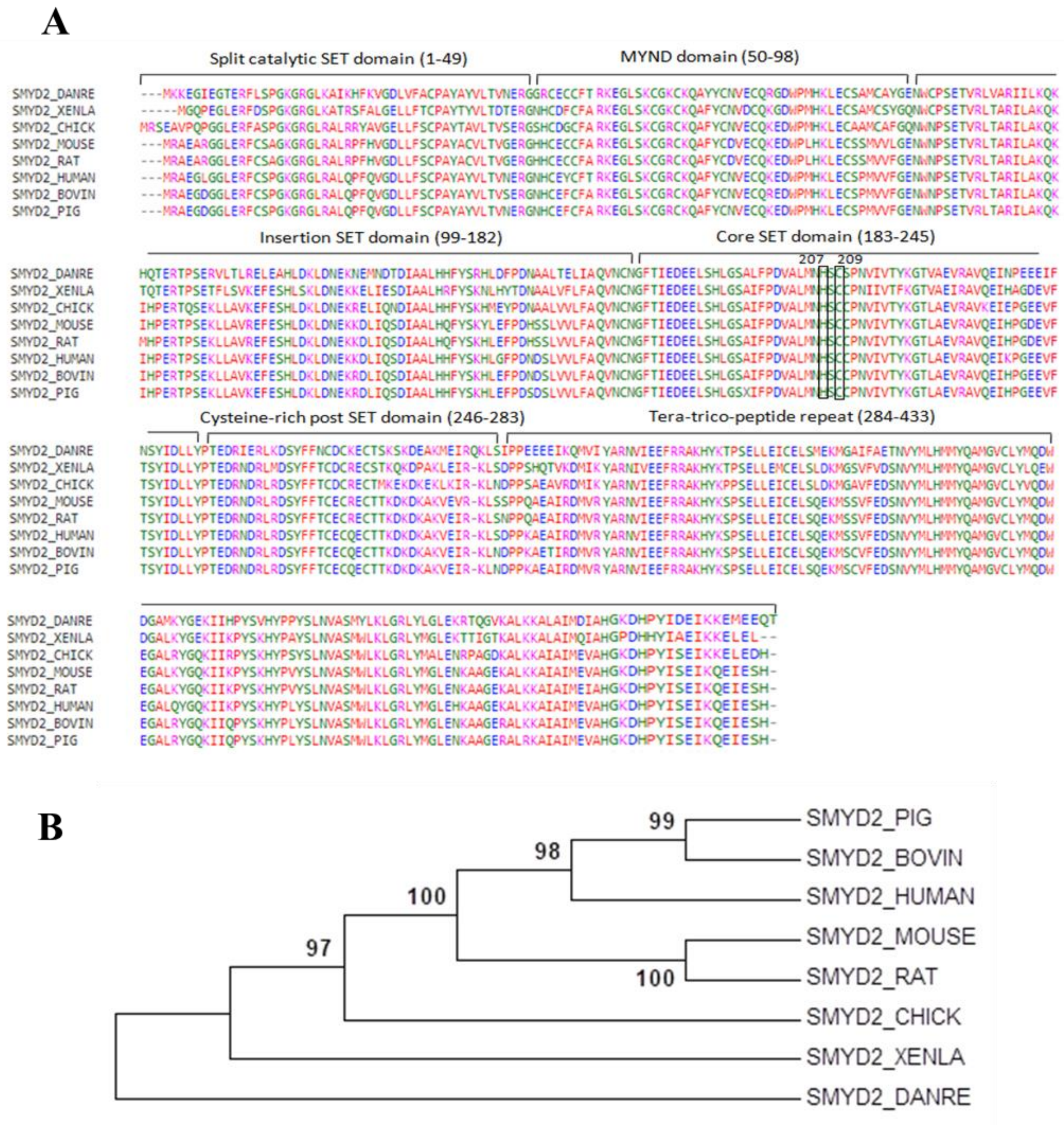


Figure 4.5: (A) Multiple sequence alignment analysis and evolutionarily conserved behavior within SMYD2 protein. Mutant H207 and C209 are found conserved in the core SET domain shown in the black rectangle (B) Phylogenetic analysis of SMYD2 protein.

4.3.3 Analysis of molecular dynamics simulations

MDS study helps to understand the structural conformations of the mutant and native protein in physiological conditions. It's employed to study the kinetics of mutational effects and the protein-ligand complex [51]. MDS analysis helps to study the kinetics of mutational effects and the protein-ligand complex [51]. GROMACS was used to analyze MDS for 100 ns after producing four systems (three mutant proteins for H207D, C209W, C209R, and one native protein). The trajectories were analyzed using a variety of GROMACS programs. Among the all mutants, C209R has a greater fluctuation as compared to native protein, after 60 ns RMSD analysis. All trajectories were well equilibrated after 65 ns. Thus for normal and mutant proteins, the data for the last 35 ns was acquired to perform several other analyses.

4.3.3.1 Stability analysis

The RMSD values are mostly used for protein structure comparison and stability prediction [52]. It is also used to estimate the conformational disturbance in the protein backbone. After 65 ns, all trajectories had reached equilibration, as presented in Figure 4.6A. The average value for RMSD was 0.252, 0.307, and 0.243 nm for mutant types H207D, C209W, and C209R, respectively while for the mutant type was 0.227 nm. At 20 ns, all RMSD values were equal, however, at 60 ns, mutant C209W showed a sharp rise. The RMSD values became constant after 65 ns and showed a constant peak until the completion. The results represented those mutant proteins have less stable as compared to the native protein. Any protein must have a specific structure to perform its proper function. Therefore, on the basis of RMSD analysis, it can conclude that proteins get destabilized due to the mutant therefore unable to perform their respective functions.

4.3.3.2 Flexibility analysis

The structural flexibility was analyzed for the mutant proteins along with native protein after evaluating the RMSD analysis. Structural flexibility is necessary to preserve the protein structures. The structural flexibility of mutant proteins was predicted using RMSF, as shown in Figure 4.6B. Residual mobilities were illustrated with the help of RMSF analysis performed for the last 35 ns equilibrated trajectory. The average value of calculated RMSF is 0.09 nm for native protein, while it is 0.10, 0.10, and 0.11 nm for H207D, C209W, and C209R mutants, respectively. It was discovered that mutant structures are less stable than original protein structures. Two mutants such as H207D and C209W represented a similar pattern for RMSF. We found that all mutants were less stable than the normal protein based on the RMSF value. The RMSF analysis is in respectable agreement with the finding of RMSD.

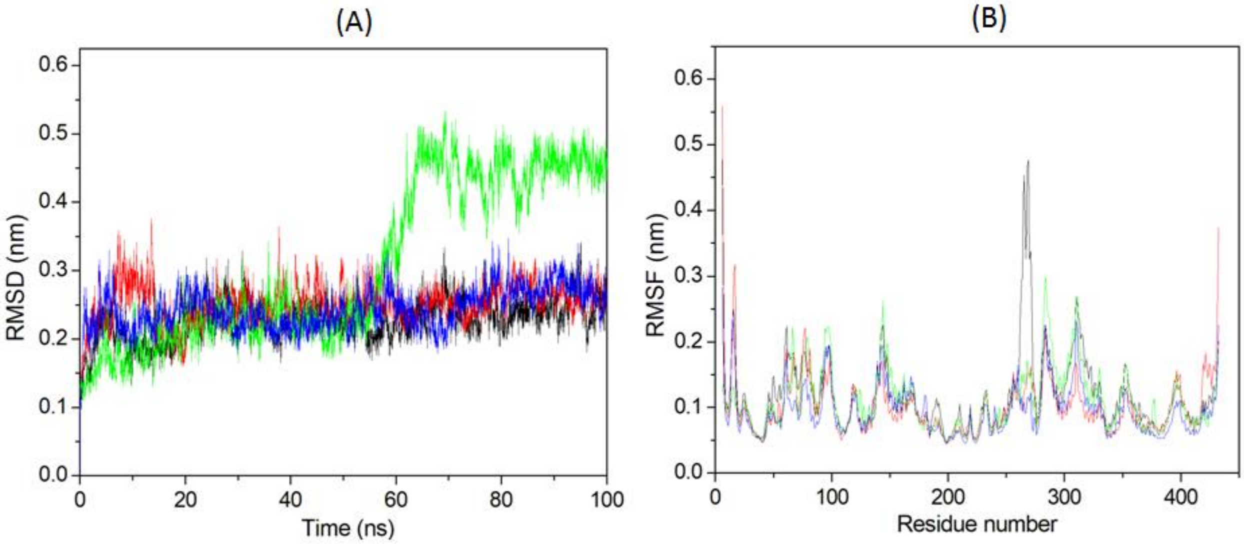


Figure 4.6: Molecular dynamics simulation results. (A) The pattern for RMSD graphs for mutant and native proteins. (B) The RMSF graph for mutant and native proteins over the last 35 ns simulation. The colors used in panels show, black for native, red for H207D, green for C209W, and blue for C209R.

4.3.3.3 Compactness analysis

To describe the molecule's distribution, Rg values are used during the MDS. Rg has been identified as a protein structural compactness marker [53]. Variations in Rg for native and all mutants are plotted against time in Figure 4.7A in this study. Rg is 2.30 nm for native and 2.28 nm for H207D, 2.26 nm for C209W, and 2.31 nm for C209R mutant. It means that two mutants such as H207D and C209W have smaller Rg patterns than the native one. In comparison to others, mutant H207D had the smallest Rg peak. Rg analysis represented that proteins with the mutation of H207D and C209W have less compact than native proteins, which agrees well with the result of RMSD.

4.3.3.4 Hydrogen bonding analysis

To maintain protein stability, hydrogen bonds play a crucial role. The structure having more hydrogen bonds represents a stable structure and with less hydrogen bond represents a less stable structure. We used to calculate the intramolecular hydrogen bonds to estimate their effect on mutation protein structure. Figure 4.7B shows the plot for the total number of hydrogen bond counts that formed during the simulation in native and mutant proteins. The native has 338 hydrogen bonds, while the mutants have hydrogen bonds 327 for H207D, 336 for C209W, and 329 for C209R. The higher or lower number of hydrogen bonds that occurs due to the deleterious nsSNPs can affect protein structure and function [54]. As a result, this study revealed a substantial link between the RMSD, RMSF, and Rg results.

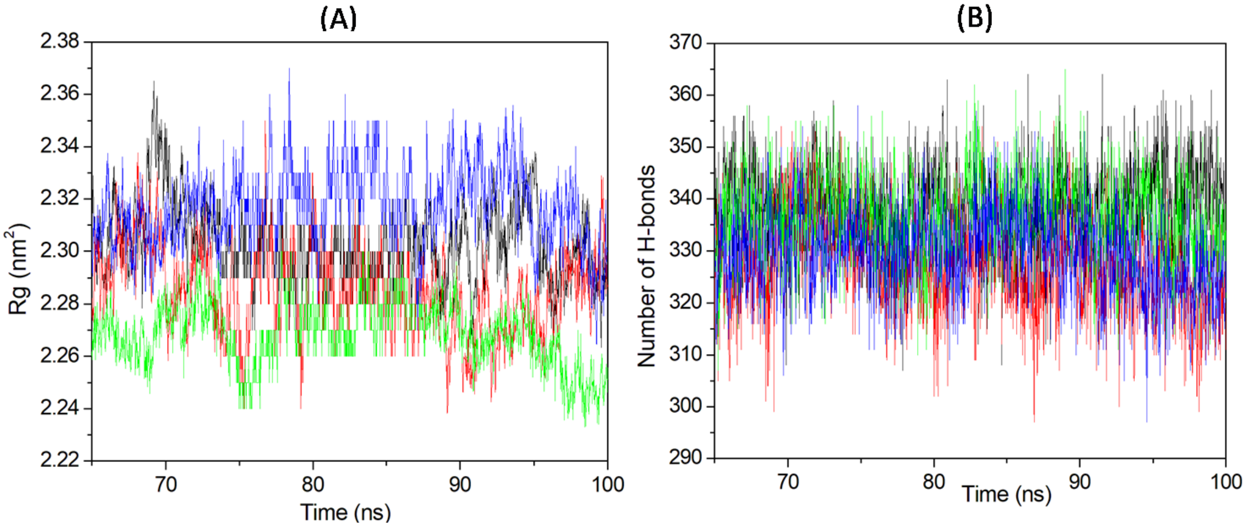


Figure 4.7: Structure stability. (A) Rg plot for protein backbone over the period of simulation of mutant and native proteins (B) The total number of hydrogen bond counts formed during the simulation of mutant and native proteins. The colors used in panels show, black for native, green for C209W, red for H207D, and blue for C209R.

4.3.3.5 Solvent accessible surface area analysis

To describe the accessible area for solvent, a SASA study was done for the protein of native and mutants. The higher the SASA number, the more extended the protein is, and the lower the SASA value, the more compact the protein is. The SASA value was considered for the trajectory of the last 35 ns and Figure 4.8A shows the graph against time. The average value for SASA was 223.99 nm^2 for native while it was 223.36 nm^2 for H207D, 223.47 nm^2 for C209W, and 226.92

nm^2 for C209R. As a result, the SASA value indicated that both the mutants H207D and C209W had structural stability parallel to the native protein. Conformational protein stability was diminished in the mutant C209R. In addition, the SASA value versus residue plot was determined and shown in Figure 4.8B. It supports knowing the conformational changes caused by residues.

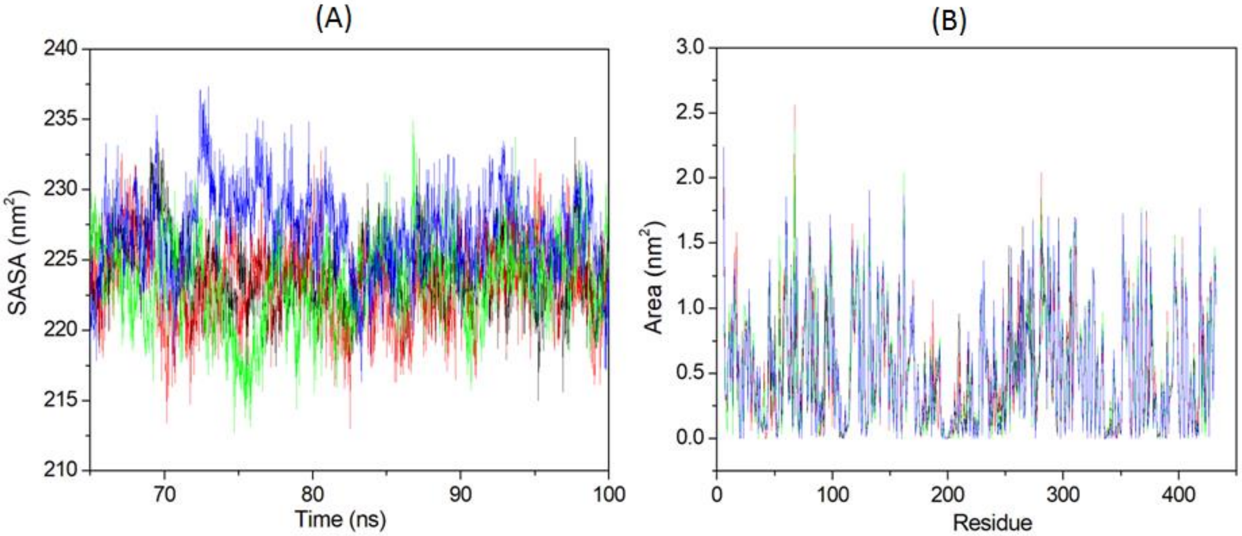


Figure 4.8: SASA patterns for Cα atoms. (A) SASA plots *vs.* time for mutant and native protein structures (B) Residue-based SASA value for native and mutant proteins at 300K. The colors used in both panels show black for native, red for H207D, green for C209W, and blue for C209R.

4.3.3.6 Principal component analysis

To estimate the collective movements for the mutant and native structure of proteins, PCA was performed. Eigenvectors, or principal components, are important in depicting the movements of a protein. For native and mutant proteins, a covariance matrix was identified for an atomic fluctuation to predict the eigenvalue. Figure 4.9A shows the plot for the eigenvalue vs. the first 50 eigenvectors of mutant and native protein structures. The first few eigenvectors, as shown in Figure 4.9A, are contributed significantly to the movement of protein. The percentages of correlated movements in native and mutant proteins were also computed. The principal components account for 71.83% for native and 76.38% for H207D, 70% for C209W, and 79.60% for C209R mutant, in the first 10 components. As compared to native protein, mutant H207D and C209R indicated a lot of correlated movements. In comparison to all expected hits, mutant C209R showed the highest correlation movements in the beginning, whereas mutant C209W showed the least. As a result of the mutation, correlation movements were disturbed, which hampered the structure dynamics of both the native and mutant proteins. Figure 4.9B depicts the residue-by-residue correlation motion expected by computing the eigRMSF value to get variations in native and mutant proteins. The eigRMSF values were 0.03, 0.04, 0.03, and 0.05 nm for native, H207D, C209W, and C209R respectively. These results suggest that protein conformation gets changed due to the mutations. PCA examination also has a resilient connection between the results of RMSD, RMSF, Rg, and SASA.

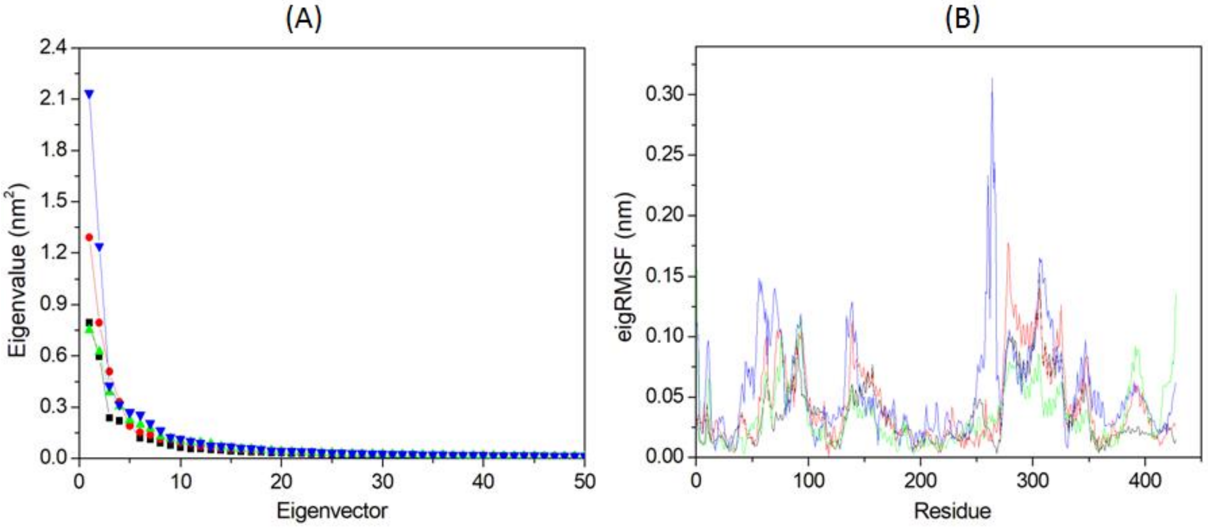


Figure 4.9: Principal component analysis for first 50 eigenvectors. (A) The plot for eigenvector *vs.* eigenvalue (B) The plot for EigRMSF values with residues for native and mutant protein structures. The colors used in both panels show black for native, red for H207D, green for C209W, and blue for C209R

Our computational research has aided in the identification and description of disease-related SNPs at the level of molecules. It also offers a molecular podium that approves alterations in the target protein's activity, stability, binding, and other features in mutants. Mutation, according to our findings, causes structural alterations and, as a result, the loss of native protein functions. Because of its overexpression in a variety of cancer forms, SMYD2 has deliberated a new oncogenic protein [55]. As a result, studying SMYD2 SNPs will provide new understandings of carcinogenesis.

4.4 CONCLUSION

SNP investigation of potential cancer genes in sequencing laboratories is the most expensive and time-consuming approach. The in-silico analysis offered here establishes a fast and economical method for identifying disease-related and highly vulnerable SNPs. The current investigation discovered the three most deleterious nsSNPs such as H207D, C209R, and C209W in the protein of SMYD2. These mutations are likely to cause the native protein's structure to be disrupted. The consequence of these mutations on the SMYD2 structure was investigated using MDS analysis. RMSD, RMSF, and Rg graphs were used to examine the mutants' stability, flexibility, and compactness. These findings show that mutants lose stability when compared to native protein, and the number of hydrogen bonds study supports this conclusion. SASA and PCA analyses were also carried out. In comparison to the native protein, there were additional differences in mutant proteins. These aberrations can cause a break in protein structural conformation, which can lead to a loss of complete protein stability. The findings of extremely related nsSNPs with malignancies are still primary, and they need to be explained and tested further. Similar findings from large-scale investigations in other populations suggest that the nsSNP is an undeniable cancer risk factor. This computational method can assist in determining the order in which genes should be validated. This method may also aid in the discovery of new SNPs that are important in the development of cancer. These findings are expected to aid experimental scientists around the world in their work on limited data for cancer management.

REFERENCES

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/caac.21492.

[2] R. D. Smith and M. K. Mallath, “History of the Growing Burden of Cancer in India: From Antiquity to the 21st Century,” *JGO*, no. 5, pp. 1–15, Nov. 2019, doi: 10.1200/JGO.19.00048.

[3] H. Takeshima and T. Ushijima, “Accumulation of genetic and epigenetic alterations in normal cells and cancer risk,” *npj Precision Oncology*, vol. 3, no. 1, Art. no. 1, Mar. 2019, doi: 10.1038/s41698-019-0079-0.

[4] N. Deng, H. Zhou, H. Fan, and Y. Yuan, “Single nucleotide polymorphisms and cancer susceptibility,” *Oncotarget*, vol. 8, no. 66, Dec. 2017, doi: 10.18632/oncotarget.22372.

[5] X. Yi, X.-J. Jiang, and Z.-M. Fang, “Histone methyltransferase SMYD2: ubiquitous regulator of disease,” *Clinical Epigenetics*, vol. 11, no. 1, p. 112, Aug. 2019, doi: 10.1186/s13148-019-0711-4.

[6] H.-S. Cho *et al.*, “RB1 methylation by SMYD2 enhances cell cycle progression through an increase of RB1 phosphorylation,” *Neoplasia*, vol. 14, no. 6, pp. 476–486, Jun. 2012, doi: 10.1593/neo.12656.

[7] R. F. Sweis *et al.*, “Discovery of A-893, A New Cell-Active Benzoxazinone Inhibitor of Lysine Methyltransferase SMYD2,” *ACS Med. Chem. Lett.*, vol. 6, no. 6, pp. 695–700, Jun. 2015, doi: 10.1021/acsmedchemlett.5b00124.

[8] A. D. Ferguson *et al.*, “Structural Basis of Substrate Methylation and Inhibition of SMYD2,” *Structure*, vol. 19, no. 9, pp. 1262–1273, Sep. 2011, doi: 10.1016/j.str.2011.06.011.

[9] L. Wang *et al.*, “Structure of Human SMYD2 Protein Reveals the Basis of p53 Tumor Suppressor Methylation,” *J. Biol. Chem.*, vol. 286, no. 44, pp. 38725–38737, Nov. 2011, doi: 10.1074/jbc.M111.262410.

[10] R. Hamamoto, V. Saloura, and Y. Nakamura, "Critical roles of non-histone protein lysine methylation in human tumorigenesis," *Nat. Rev. Cancer*, vol. 15, no. 2, pp. 110–124, Feb. 2015, doi: 10.1038/nrc3884.

[11] R. Hamamoto and Y. Nakamura, "Dysregulation of protein methyltransferases in human cancer: An emerging target class for anticancer therapy," *Cancer Sci.*, vol. 107, no. 4, pp. 377–384, Apr. 2016, doi: 10.1111/cas.12884.

[12] S. Komatsu *et al.*, "Overexpression of SMYD2 relates to tumor cell proliferation and malignant outcome of esophageal squamous cell carcinoma," *Carcinogenesis*, vol. 30, no. 7, pp. 1139–1146, Jul. 2009, doi: 10.1093/carcin/bgp116.

[13] L. H. T. Sakamoto, R. V. de Andrade, M. S. S. Felipe, A. B. Motoyama, and F. Pittella Silva, "SMYD2 is highly expressed in pediatric acute lymphoblastic leukemia and constitutes a bad prognostic factor," *Leukemia Research*, vol. 38, no. 4, pp. 496–502, Apr. 2014, doi: 10.1016/j.leukres.2014.01.013.

[14] M. H. Hofker, J. Fu, and C. Wijmenga, "The genome revolution and its role in understanding complex diseases," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol. 1842, no. 10, pp. 1889–1895, Oct. 2014, doi: 10.1016/j.bbadis.2014.05.002.

[15] Y. Yair *et al.*, "Genomics-based epidemiology of bovine Mycoplasma bovis strains in Israel," *BMC Genomics*, vol. 21, no. 1, p. 70, Jan. 2020, doi: 10.1186/s12864-020-6460-0.

[16] M. J. Islam, A. M. Khan, M. R. Parves, M. N. Hossain, and M. A. Halim, "Prediction of Deleterious Non-synonymous SNPs of Human STK11 Gene by Combining Algorithms, Molecular Docking, and Molecular Dynamics Simulation," *Scientific Reports*, vol. 9, no. 1, pp. 1–16, Nov. 2019, doi: 10.1038/s41598-019-52308-0.

[17] K. N. Chitrala and S. Yeguvapalli, "Computational Screening and Molecular Dynamic Simulation of Breast Cancer Associated Deleterious Non-Synonymous Single Nucleotide Polymorphisms in TP53 Gene," *PLOS ONE*, vol. 9, no. 8, p. e104242, Aug. 2014, doi: 10.1371/journal.pone.0104242.

[18] M. Jia, B. Yang, Z. Li, H. Shen, X. Song, and W. Gu, "Computational analysis of functional single nucleotide polymorphisms associated with the CYP11B2 gene," *PLoS ONE*, vol. 9, no. 8, p. e104311, 2014, doi: 10.1371/journal.pone.0104311.

[19] M. Sehgal and T. R. Singh, “Systems biology approach for mutational and site-specific structural investigation of DNA repair genes for xeroderma pigmentosum,” *Gene*, vol. 543, no. 1, pp. 108–117, Jun. 2014, doi: 10.1016/j.gene.2014.03.057.

[20] M. Sehgal and T. R. Singh, “DR-GAS: a database of functional genetic variants and their phosphorylation states in human DNA repair systems,” *DNA Repair (Amst.)*, vol. 16, pp. 97–103, Apr. 2014, doi: 10.1016/j.dnarep.2014.01.004.

[21] A. Kumar and R. Purohit, “Use of Long Term Molecular Dynamics Simulation in Predicting Cancer Associated SNPs,” *PLoS Comput Biol*, vol. 10, no. 4, p. e1003318, Apr. 2014, doi: 10.1371/journal.pcbi.1003318.

[22] Q. Wang, A. Mehmood, H. Wang, Q. Xu, Y. Xiong, and D.-Q. Wei, “Computational Screening and Analysis of Lung Cancer Related Non-Synonymous Single Nucleotide Polymorphisms on the Human Kirsten Rat Sarcoma Gene,” *Molecules*, vol. 24, no. 10, p. 1951, May 2019, doi: 10.3390/molecules24101951.

[23] S. T. Sherry *et al.*, “dbSNP: the NCBI database of genetic variation,” *Nucleic Acids Res*, vol. 29, no. 1, pp. 308–311, Jan. 2001, doi: 10.1093/nar/29.1.308.

[24] R. Apweiler *et al.*, “UniProt: the Universal Protein knowledgebase,” *Nucleic Acids Res*, vol. 32, no. Database issue, pp. D115-119, Jan. 2004, doi: 10.1093/nar/gkh131.

[25] N.-L. Sim, P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng, “SIFT web server: predicting effects of amino acid substitutions on proteins,” *Nucleic Acids Research*, vol. 40, no. W1, pp. W452–W457, Jul. 2012, doi: 10.1093/nar/gks539.

[26] 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, pp. 2745–2747, Aug. 2015, doi: 10.1093/bioinformatics/btv195.

[27] I. A. Adzhubei *et al.*, “A method and server for predicting damaging missense mutations,” *Nat Methods*, vol. 7, no. 4, pp. 248–249, Apr. 2010, doi: 10.1038/nmeth0410-248.

[28] P. D. Thomas *et al.*, “PANTHER: a library of protein families and subfamilies indexed by function,” *Genome Res.*, vol. 13, no. 9, pp. 2129–2141, Sep. 2003, doi: 10.1101/gr.772403.

[29] E. Capriotti, R. Calabrese, P. Fariselli, P. L. Martelli, R. B. Altman, and R. Casadio, “WS-SNPs&GO: a web server for predicting the deleterious effect of human protein variants using functional annotation,” *BMC Genomics*, vol. 14, no. 3, p. S6, May 2013, doi: 10.1186/1471-2164-14-S3-S6.

[30] E. Capriotti, R. Calabrese, and R. Casadio, “Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information,” *Bioinformatics*, vol. 22, no. 22, pp. 2729–2734, Nov. 2006, doi: 10.1093/bioinformatics/btl423.

[31] V. Pejaver *et al.*, “MutPred2: inferring the molecular and phenotypic impact of amino acid variants,” *bioRxiv*, p. 134981, May 2017, doi: 10.1101/134981.

[32] V. López-Ferrando, A. Gazzo, X. de la Cruz, M. Orozco, and J. L. Gelpí, “PMut: a web-based tool for the annotation of pathological variants on proteins, 2017 update,” *Nucleic Acids Res*, vol. 45, no. Web Server issue, pp. W222–W228, Jul. 2017, doi: 10.1093/nar/gkx313.

[33] 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, p. e0117380, Feb. 2015, doi: 10.1371/journal.pone.0117380.

[34] E. Capriotti, P. Fariselli, and R. Casadio, “I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure,” *Nucleic Acids Res.*, vol. 33, no. Web Server issue, pp. W306-310, Jul. 2005, doi: 10.1093/nar/gki375.

[35] J. Cheng, A. Randall, and P. Baldi, “Prediction of protein stability changes for single-site mutations using support vector machines,” *Proteins*, vol. 62, no. 4, pp. 1125–1132, Mar. 2006, doi: 10.1002/prot.20810.

[36] C. H. Rodrigues, D. E. Pires, and D. B. Ascher, “DynaMut: predicting the impact of mutations on protein conformation, flexibility and stability,” *Nucleic Acids Res*, vol. 46, no. W1, pp. W350–W355, Jul. 2018, doi: 10.1093/nar/gky300.

[37] A. P. Pandurangan, B. Ochoa-Montaño, D. B. Ascher, and T. L. Blundell, “SDM: a server for predicting effects of mutations on protein stability,” *Nucleic Acids Res.*, vol. 45, no. W1, pp. W229–W235, 03 2017, doi: 10.1093/nar/gkx439.

[38] V. Parthiban, M. M. Gromiha, and D. Schomburg, “CUPSAT: prediction of protein stability upon point mutations,” *Nucleic Acids Res*, vol. 34, no. Web Server issue, pp. W239–W242, Jul. 2006, doi: 10.1093/nar/gkl190.

[39] 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, p. e118, Sep. 2011, doi: 10.1093/nar/gkr407.

[40] H. Ashkenazy, E. Erez, E. Martz, T. Pupko, and N. Ben-Tal, "ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids," *Nucleic Acids Res.*, vol. 38, no. Web Server issue, pp. W529-533, Jul. 2010, doi: 10.1093/nar/gkq399.

[41] I. Mayrose, D. Graur, N. Ben-Tal, and T. Pupko, "Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior," *Mol. Biol. Evol.*, vol. 21, no. 9, pp. 1781–1791, Sep. 2004, doi: 10.1093/molbev/msh194.

[42] H. Ashkenazy *et al.*, "ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules," *Nucleic acids research*, vol. 44, no. W1, pp. W344–W350, 2016.

[43] B. G. Hall, "Building Phylogenetic Trees from Molecular Data with MEGA," *Mol Biol Evol*, vol. 30, no. 5, pp. 1229–1235, May 2013, doi: 10.1093/molbev/mst012.

[44] H. Patel, B. A. Grüning, S. Günther, and I. Merfort, "PyWATER: a PyMOL plug-in to find conserved water molecules in proteins by clustering," *Bioinformatics*, vol. 30, no. 20, pp. 2978–2980, Oct. 2014, doi: 10.1093/bioinformatics/btu424.

[45] H. Venselaar, T. A. te Beek, R. K. Kuipers, M. L. Hekkelman, and G. Vriend, "Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces," *BMC Bioinformatics*, vol. 11, no. 1, p. 548, Nov. 2010, doi: 10.1186/1471-2105-11-548.

[46] X.-N. Peng, J. Wang, and W. Zhang, "Molecular dynamics simulation analysis of the effect of T790M mutation on epidermal growth factor receptor protein architecture in non-small cell lung carcinoma," *Oncology Letters*, vol. 14, no. 2, pp. 2249–2253, Aug. 2017, doi: 10.3892/ol.2017.6387.

[47] A. Samy, B. E. Suzek, M. K. Ozdemir, and O. Sensoy, "In Silico Analysis of a Highly Mutated Gene in Cancer Provides Insight into Abnormal mRNA Splicing: Splicing Factor 3B Subunit 1K700E Mutant," *Biomolecules*, vol. 10, no. 5, Art. no. 5, May 2020, doi: 10.3390/biom10050680.

[48] M. J. U. Hasnain *et al.*, "Computational analysis of functional single nucleotide polymorphisms associated with SLC26A4 gene," *PLOS ONE*, vol. 15, no. 1, p. e0225368, Jan. 2020, doi: 10.1371/journal.pone.0225368.

[49] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, “GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation,” *J. Chem. Theory Comput.*, vol. 4, no. 3, pp. 435–447, Mar. 2008, doi: 10.1021/ct700301q.

[50] S. Pronk *et al.*, “GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit,” *Bioinformatics*, vol. 29, no. 7, pp. 845–854, Apr. 2013, doi: 10.1093/bioinformatics/btt055.

[51] R. Kumar *et al.*, “In silico screening of deleterious single nucleotide polymorphisms (SNPs) and molecular dynamics simulation of disease associated mutations in gene responsible for oculocutaneous albinism type 6 (OCA 6) disorder,” *Journal of Biomolecular Structure and Dynamics*, vol. 37, no. 13, pp. 3513–3523, Sep. 2019, doi: 10.1080/07391102.2018.1520649.

[52] S. Srivastava *et al.*, “An efficient algorithm for protein structure comparison using elastic shape analysis,” *Algorithms for Molecular Biology*, vol. 11, no. 1, p. 27, Sep. 2016, doi: 10.1186/s13015-016-0089-1.

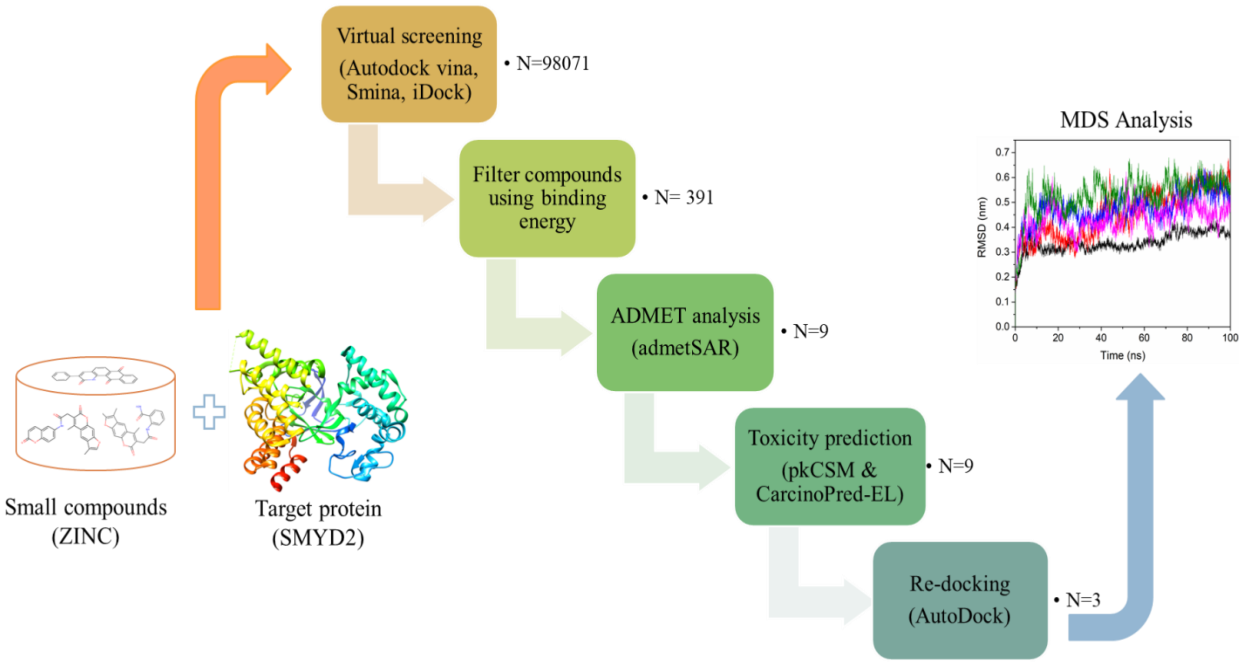
[53] M. I. Lobanov, N. S. Bogatyreva, and O. V. Galzitskaia, “[Radius of gyration is indicator of compactness of protein structure],” *Mol. Biol. (Mosk.)*, vol. 42, no. 4, pp. 701–706, Aug. 2008.

[54] C. N. Pace *et al.*, “Contribution of hydrogen bonds to protein stability,” *Protein Sci.*, vol. 23, no. 5, pp. 652–661, May 2014, doi: 10.1002/pro.2449.

[55] J.-J. Sun, H.-L. Li, H. Ma, Y. Shi, L.-R. Yin, and S.-J. Guo, “SMYD2 promotes cervical cancer growth by stimulating cell proliferation,” *Cell & Bioscience*, vol. 9, no. 1, p. 75, Sep. 2019, doi: 10.1186/s13578-019-0340-9.

CHAPTER- 5

Structural Investigation and Simulation Studies to Design Novel Inhibitors for SMYD2



ABSTRACT

SMYD2 is a member of the SMYD family of PKMTs that can control the regulation of gene transcription by the methylation of lysine residues in substrate proteins. It plays an important role in the progression of various cancers. Inhibition of SMYD2 could be a significant target for the development of anti-cancer therapeutics in which it plays a role. The current work used structure-based virtual screening to find possible SMYD2 inhibitors. A consensus docking technique was used to screen a large set (n=98071) of small natural chemical compounds from the ZINC database, that yielded 391 powerful molecules. These 391 substances were also assessed using a variety of ADMET criteria. Nine compounds were chosen to pursue the re-docking that fit on drug-likeness standards. We identified three compounds (ZINC03844862, ZINC08490711, and ZINC08764231) as probable inhibitors on the basis of docking score and interaction analysis between protein and ligands. Finally, the structural stability of these three compounds with SMYD2 protein was investigated using a 100 ns MDS study. Furthermore, various analyses related to MDS revealed that selected compounds with the SMYD2 structure had a stable binding. The computational methods used in this investigation resulted in a series of novel potential SMYD2 inhibitors. Following *in vivo* and *in vitro* research, these inhibitors could be viable leads for developing cancer treatments.

5.1 INTRODUCTION

PKMTs are a type of PMT that aid in the relocation of a methyl group from a substrate protein to another protein. PKMTs are associated with varieties of biological processes by the methylating of lysine residues in both the non-histone and histone proteins [1]. Various studies have suggested that PKMTs are associated with cancer progression, therefore, nowadays it become an eye-catching target for the development of cancer drugs [2]. PKMT has a SET and MYND domain-containing family SMYD. The SET domain is in charge of transferring methyl groups to target protein lysine residues, while the MYND domain is convoluted in protein-protein interactions [2]. SMYD2 protein from the SMYD family is one of the most studied proteins and is involved in tumor formation via its methylation activity. SMYD2 methylation activity has been revealed to play an important role in non-histone protein methylation in several investigations [3],

[4]. It was found that most of the SMYD2-specific non-histone substrates are associated with cancer cell proliferation and cell apoptosis [5].

A number of oncogenic proteins such as P^{53} [6], PTEN [7], RB [7,8], HSP90 [10], ERα [11], PARP1 [12], and STAT3 [13] have been methylated by SMYD2. The overexpression of SMYD2 has been found in many human malignancies, including cervical, colorectal, esophageal, breast, bladder, head and neck, ovarian, pancreatic, and lymphoma cancer [6], [8], [14], [15]. As a result, the substrates of SMYD2 have carcinogenic nature and are involved in cancer progression advocated for the association of SMYD2 in human cancers. Therefore, SMYD2 has a growing interest in designing SMYD2-specific inhibitors.

Structure-based virtual screening (SBVS) is a promising drug discovery technique based on consensus docking [16]–[18]. Natural compounds were selected as a ligands, because it has the most promising source to develop anticancer drugs [19]. To measure the ligand-binding affinity and orientation, molecular docking employs a scoring function [20]–[22]. As a result, in virtual screening, using more than one approach to reach a consensus improved accuracy in the pose prediction of ligand [16]. To reduce the cost of the drug design process, virtual screening approaches are significantly used [23]. This method has been used to successfully identify novel inhibitors for several epigenetic targets such as SET7, KDM4B, PRMT5, and SIRT2 [24]–[27]. The effectiveness of the SBVS approach in the studied epigenetic targets encouraged us toward the prediction of novel SMYD2 inhibitors.

Here, a total of 391 potential compounds were selected in this investigation using the SBVS technique. The selected set of compounds is further used for the prediction of ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) profiles. Based on ADMET properties, nine compounds were considered those fit on all ADMET criteria, and further docking was done to re-check the binding pose. Finally, the top three compounds based on binding energy were used for MDS analysis to evaluate the stability of receptor-ligand interaction.

5.2 METHODOLOGY

5.2.1 Ligand and protein preparation

The x-ray crystallized structure of SMYD2 that has been present in the complex with LLY-507 inhibitor (PDB ID: 4WUY, 1.6Å) [5] was taken from the protein data bank. UCSF Chimera 1.14 software was used to create the protein structure [28] and for energy minimization with an Amber ff99SB force field [29]. The virtual screening was then performed using the minimized structure. The ZINC database [30] was used to retrieve the natural compound subset (n=98071). Figure 5.1 depicts the overall technique of the current investigation.

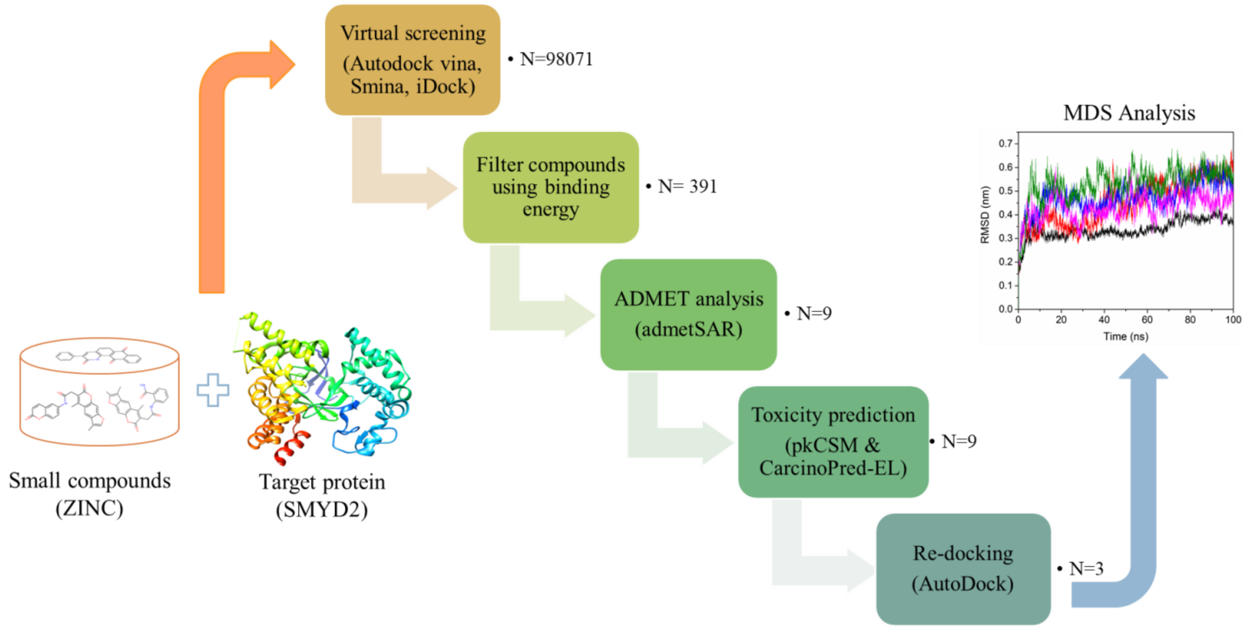


Figure 5.1: Complete methodology used to perform this study.

5.2.2 Virtual screening

A large subset (98071 compounds) of natural compounds, present in the ZINC database were employed to find out the best hits by using SBVS studies. As compared to high-throughput screening (HTS), it is a more efficient and cost-effective method. To perform a consensus analysis in SBVS studies, three docking programs such as Smina [31], AutoDock vina [32], and iDock [33] were used with default parameters. AutoDock vina was developed first and it provides background for Smina and iDock. To search for a docking pose, Vina used both Monte Carlo step and gradient calculation on the basis of scoring function used for the searching and ranking. Smina used similar searching algorithm but used a different set of scoring functions for the generation of ligand docking process and the ranking process. iDoock also used scoring function as Vina but utilized few steps of Monte Carlo instead of parallel search in its search process. Molecular docking is used in SBVS to compute the correct binding pose of the target protein, where ligand molecules fit for accurate binding. In the drug discovery process, the interpretation of binding behavior is critical. Using the above-mentioned three methods, experimentally proved SMYD2 binding site (target site where crystallographic ligand LLY-507 is bound) was considered for the docking of all compounds (98071 compounds). Then, we sorted the highly active compounds with binding energies of >= -11.7 Kcal.mol^{-1} from all three methods we selected. The cutoff value was chosen based on the findings of virtual screening. Out of 98071 compounds, finally, we sorted to 391 compounds. Then, these 391 compounds were used for the ADMET prediction.

5.2.3 ADMET prediction

The ADMET studies are very crucial for candidate drugs and play a key role in lead compound optimization. The admetSAR server [34] was used to predict the ADMET profiles for 391 compounds that were chosen. In admetSAR, approximately 210 000 ADMET annotated data points have been meticulously selected from a wide range of various literatures for more than 96 000 unique chemicals with 45 types of ADMET-associated characteristics, proteins, species, or animals. Machine-learning algorithms such as SVM, RF, and KNN were used for the model building. Various ADMET properties were predicted using the admetSAR server, such as blood-brain barrier (BBB), human intestinal absorption (HIA), biodegradability, caco-2 permeability, plasma protein binding, P-gp substrate/inhibitor, cytochrome-P450 (CYP450) substrate/inhibitor, carcinogenicity, rat acute toxicity, and AMES toxicity. Many compounds failed to pass one, two, or even all of the parameters. Nine compounds were filtered out of 391 that met all of the criteria for being drug-like. Furthermore, pkCSM [35] was used to predict toxicity parameters for the selected compounds. It uses distance-based graph signatures approach to predict a range of ADMET properties for compounds. The CarcinoPred-EL (Carcinogenicity Prediction using Ensemble Learning method) [36] server was used to predict carcinogenicity of the compounds. This server utilized, three ensemble classification models such as, ensemble SVM, ensemble RF and ensemble XGBoost to develop models for the prediction of carcinogenicity of compounds using 1003 compounds with rat carcinogenicity. In CarcinoPred-EL, the XGBoost model was chosen due to better accuracy performance. Further, the selected nine drug-likeness compounds were subject to re-docking to assess previous docking results.

5.2.4 Re-docking of drug-like compounds

The molecular docking study was used to investigate the binding pose of selected compounds on the binding site of the protein structure [37]–[39]. The re-docking was done by AutoDock tool to compare the results of Autodock Vina, Smina, and iDock. For this docking analysis, we used the binding pocket of SMYD2 that was used to perform the virtual screening. The AutoDock tool is extensively employed for calculating IC50 values and ligand-receptor binding affinity. A semi-empirical free energy force field with Lamarckian Genetic algorithm was used to determine the conformation of the ligand-binding site. AutoDock uses a grid-based method to allow rapid evaluation of the binding energy of trial conformations. AutoDock tool was also used to create the SMYD2 protein structure and ligands. Atomic charges and polar hydrogen atoms were added to the structure of SMYD2 before it was converted to the PDBQT file format. The ligand optimization was done with AutoDock and non-polar hydrogens and Gasteiger charges option were merged. AutoDock needs to define a 3D grid box around the active site to confirm that ligand exact binds with the active site of SMYD2. The center grid values were adjusted to a definite coordinate (X=-19.250°, Y= -29.136°, and Z=26.141°) with a spacing of 0.500Å. The number of grid boxes was set as X=62°, Y=46°, and Z=52°. To produce an optimal binding pose for each ligand, the Lamarckian genetic algorithm was used. A total of 100 binding conformations were produced for each ligand. The default values were used for the remaining docking parameters. For the depiction of residue interactions, Discovery Studio Visualizer was used. The top docked complexes having the best binding energy were subjected to an MDS study.

5.2.5 Molecular Dynamics simulation

The best-docked complexes along with control and apo-SMYD2 structures were simulated using GROMACS 2018.1 [40], [41] software on a Fedora PC. The protein-ligand complexes were solvated with the SPC water model [42] and placed in a cubic shape box. To generate the ligand topology file, an automated PRODRG server [43] was used. The topology file was built by using the GROMOS45A7 force field [44]. To neutralize the systems, counter-ions were added. Systems were equilibrated after energy minimization. After that, under the condition of NVT and NPT, simulation was performed for 1 ns with constant pressure and volume at a temperature of 300K. At a 10 ps interval, the trajectory was recorded. By using the various utilities of GROMACS we calculated the RMSD, RMSF, Rg, number of hydrogen bonds, and SASA. PCA was also performed to calculate the correlation motions during the simulation within a protein during MDS. To generate and visualize all plots, Origin 2016 was used.

5.2 RESULTS AND DISCUSSION

5.3.1 Virtual Screening analysis

The ZINC database was used to obtain a batch of natural chemicals (n= 98071) and test them for SMYD2. The crystallographic ligand LLY-507 was utilized as a control molecule in this analysis since it is available with the co-crystal structure of SMYD2. All ligands were docked with SMYD2 using Autodock Vina, Smina, and iDcok for virtual screening. The best binding energy of SMYD2 protein was used to rank all docked molecules. Binding energy ranged from -14.29 to -3.73 Kcal.mol^{-1} for iDock, -14.6 to -4.1 Kcal.mol^{-1} for Smina, and -14.5 to -4.0 Kcal.mol^{-1} for Autodock Vina. On the basis of binding energy cut-off (cut-off value was -11.7 Kcal.mol^{-1}), we selected the consensus compounds from three used docking tools for the ADMET study. Finally, we got 391 compounds (docking score range from -11.7 to -14.6 Kcal.mol^{-1}) for the prediction of the ADMET profile.

5.3.2 Analysis of ADMET properties

Analysis of the ADMET profile is very crucial for the new candidate drug. Without meeting the ADMET criteria, no drug can be approved. As a result, we used the admetSAR server to estimate ADMET parameters using all 391 chemicals picked from the virtual screening. In the human gastrointestinal tract, the absorption of any substance was defined by the HIA. It's a crucial aspect of oral drug administration. The HIA was able to absorb 388 compounds out of 391 analyzed compounds. The drug molecule's absorption in the large intestine was determined using the Caco-2 measure. Out of 391 compounds, 97 compounds have the ability to pass the test for Caco-2 permeability. The BBB is a vital pharmacological factor because it works as a physiological barrier that prevents the compound to move into Sentral Nervous System (CNS) [45]. The admetSAR server displays a (+) symbol with compounds that pass the BBA test and a (−) symbol with compounds that do not pass the BBB test. It was observed that 345 compounds fit on the BBB test that does not cross the BBB. P-gp is involved in xenobiotic efflux in a big way. The P-gp behavior was conformed for 327 compounds as substrates and 64 compounds as non-substrates. The P-gp receptor is targeted by 166 non-inhibitor and 225 inhibitor molecules among the 391 molecules.

In liver cells, CYP-450 plays a substantial role in the detoxification of heterologous drugs. As a result, the behavior of drug metabolism was predicted through CYP-450 inhibition. A total of 135 of the 391 substances tested showed mild inhibition of the CYP450 enzyme, whereas the remaining compounds showed stronger inhibition. The toxicity of the molecule is the most important consideration in drug development. If a drug meets all of the criteria yet has a harmful effect, it cannot be considered a candidate drug. Consequently, toxicity analysis was carried out for 391 compounds. Results revealed that 349 compounds showed non-toxic characteristics. The carcinogenicity of compounds was also analyzed and it was observed that only four compounds have carcinogenic characteristics. The hERG gene produces a channel protein that aids in the transmission of electric current across the cell membrane. The hERG gene inhibition was analyzed for selected compounds and it was observed that 212 compounds are likely hERG inhibitors and 179 compounds are likely hERG non-inhibitors. The *in-silico* approach was used to determine the LD50 value using the rat model. The LD50 values range from 2.04 and 4.03 mol.kg^{-1}.

Table 5.1: The summary of ADMET profiles for selected nine compounds.

	ZINC05 220992	**ZINC03 844862**	**ZINC06 475191**	**ZINC08 764231**	**ZINC11 865353**	**ZINC08 490711**	**ZINC68 589462**	**ZINC08 918302**	**ZINC12 659865**
P-gp inhibitor	NI	NI	NI	NI	NI	NI	I	NI	NI
BBB probability	+/0.9867	+/0.9827	+/0.9523	+/0.9191	+/0.8745	+/0.5279	+/0.9869	+/0.9808	+/0.9198
CYP-2C9 substrate/ inhibitor	NS/NI	NS/NI	NS/I	NS/ NI	NS/NI	NS/NI	NS/ NI	NS/NI	NS/NI
Caco-2	+/0.5258	+/0.5851	-/0.6193	+/0.5148	-/0.5130	-/0.6383	+/0.7017	+/0.5443	+/0.5706

permeability probability									
CYP-2D6 substrate/ inhibitor	NS/NI	NS/NI	NS/NI	NS/ NI	NS/NI	NS/NI	NS/NI	NS/NI	NS/NI
P-gp substrate	NS	NS	NS	NS	NS	NS	NS	NS	NS
HIA-probability	+/0.9810	+/1.0000	+/0.9937	+/0.9315	+/0.9732	+/0.6641	+/1.0000	+/0.9768	+/0.9772
Caco-2 permeability	0.9743	1.5817	0.7400	0.6742	0.8284	0.2599	1.6066	0.8796	1.1875
CYP-1A2 inhibitor	I	I	NI	NI	NI	NI	NI	NI	NI
CYP-3A4 substrate/ inhibitor	NS/ NI	NS/NI	S/NI	S/NI	NS/NI	S/I	S/I	NS/NI	NS/NI
CYP inhibitory promiscuity	Low	Low	High	Low	Low	Low	Low	Low	Low
CYP-2C19 inhibitor	NI	NI	NI	NI	NI	NI	NI	NI	NI
hERG inhibition	NI	NI	NI	NI	NI	NI	NI	NI	NI
AMES Toxicity	NT	NT	NT	NT	NT	NT	NT	NT	NT
Acute Oral Toxicity	III/ 0.4718	III/ 0.5881	II/ 0.5860	III/ 0.6082	III/ 0.5843	III/ 0.6299	III/ 0.6806	III/ 0.5026	III/ 0.6125
Carcinogen	NC	NC	NC	NC	NC	NC	NC	NC	NC
Rat LD_{50} (mol.kg^{-1})	2.1161	2.5965	3.0974	2.4333	2.5842	2.3221	2.2444	2.0506	2.6510

NS: Non-Substrate; NI: Non-Inhibitor; NT: Non-Toxic; NC: Non-carcinogens; I: Inhibitor; S: Substrate

The ADMET predictions indicated that out of 391 compounds, nine compounds have good clearance (Table 5.1) that could be used as a drug candidate. These nine compounds were anticipated to be HIA positive, indicating that they have the ability to absorb orally. The BBB is an important factor in drug distribution. BBB approved all of the substances that were tested. The

Caco-2 parameter demonstrated that the large intestine can absorb the majority of the chemicals. Except for ZINC68589462, the rest of the compounds were characterized as non-inhibitors for P-gp inhibitor and non-substrates for P-gp substrate. Eight compounds had a modest ability for CYP enzyme inhibition, indicating that they can be efficiently removed from the body after ensuring the therapeutic response. The AMES toxicity and carcinogenic analysis indicated that selected compounds could not act as carcinogens not as toxic. The overall ADMET study indicated that the selected nine compounds were found to have good parameters and are appropriate for further investigation.

5.3.3 Toxicity analysis

The drug molecules should be efficiently eliminated from the body after ensuring the beneficial response otherwise they cause a harmful effect. Therefore, toxicity prediction is essential for the candidate drugs. So, we further determine the toxicity of the selected compounds for good clearance (Table 5.2). AMES toxicity, maximum tolerated dosage (MTD), and LD50 values have justified the behavior of drug-likeness for selected compounds. The prediction of AMES toxicity revealed that the substances were not mutagenic and toxic. Any drug can be used at a higher dose that has an MTD value of >0.477 logs (mg/kg/day). The hepatotoxicity had a positive effect that validated the harmful nature of compounds. The activity of skin sensitization shows negative effects proving that compounds can be applied dermally. Furthermore, the CarcinoPred-EL prediction revealed that none of the chemicals tested were carcinogenic.

Table 5.2: Profiles of toxicity and carcinogenicity for selected compounds.

	ZINC 05220 992	**ZINC 03844 862**	**ZINC 06475 191**	**ZINC 08764 231**	**ZINC 11865 353**	**ZINC 08490 711**	**ZINC 08918 302**	**ZINC 12659 865**	**ZINC 68589 462**
hERG II inhibitor	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	Yes
Max. tolerated dose (human) (log mg)	0.439	0.477	0.51	0.675	0.68	0.277	0.444	-0.205	0.244
Oral Rat Acute Toxicity (LD_{50}) (mol/kg)	2.455	2.684	3.089	2.506	2.936	2.252	2.425	2.688	2.518

hERG I inhibitor	No	No	No	No	No	No	No	No	No
AMES toxicity	No	Yes	No	No	Yes	No	Yes	No	Yes
Oral Rat Chronic Toxicity (LOAEL) (log mg/kg)	-1.055	0.718	-0.046	0.754	0.721	1.998	-0.908	1.137	0.658
Skin Sensitisation	No	No	No	No	No	No	No	No	No
Minnow toxicity (log mM)	-3.933	-0.636	1.579	-2.633	-1.021	-0.674	-0.408	-0.319	-1.269
***T. Pyriformis* toxicity** (log ug/L)	0.285	0.288	0.285	0.289	0.29	0.333	0.285	0.304	0.288
Hepatotoxicity	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
CarcinoPred-EL									
Class	NC	NC	NC	NC	NC	NC	NC	NC	NC
Score	0.45	0.44	0.44	0.45	0.38	0.42	0.45	0.40	0.42

NC: Non-Carcinogen

5.3.4 Analysis of molecular docking

The correct binding pose for compounds on the active site of SMYD2 was investigated using molecular docking research. Nine compounds were selected as lead compounds that have good clearance with ADMET profiles (Table 5.1&5.2). The nine lead compounds, in addition to the control compound (LLY-507) were used to perform the re-docking study using AutoDock. The binding energy of the control molecule was -11.29 Kcal.mol^{-1}. Based on their binding energy, three compounds such as ZINC08490711, ZINC08764231, and ZINC03844862 showed good binding energy as compared to the control compound and were selected as the top compounds for further studies (Table 5.3). Other docking applications, including Autodock Vina, Samina, and iDock also found that these compounds had the lowest binding energy as compared to the control molecule (Table 5.3).

Table 5.3: Top compounds chosen from docking analysis with control molecule LLY-507. The ID of the compound, the binding affinity, and the name of hydrogen-making residues discovered using various docking programs is shown.

ZINC ID	**Auto dock (Kcal. mol^{-1})**	**Hydrogen bond forming Residues**	**Autodoc Vina (Kcal. mol^{-1})**	**Hydrogen bond forming Residues**	**Smina (Kcal. mol^{-1})**	**Hydrogen bond forming Residues**	**iDock (Kcal. mol^{-1})**	**Hydrogen bond forming Residues**
ZINC03844862	-11.57	Thr238 and Tyr240	-12.1	-	-12.60	-	-12.32	-
ZINC08490711	-11.99	Val213, Val215, Thr238, and Ser239,	-11.7	Val213, Val215, Thr238, and Ser239	-11.65	Val213, Val215, Thr238, and Ser239	-11.77	Val213, Val215, Thr238,and Ser239
ZINC08764231	-11.62	His137, Cys181, and Tyr240	-12.3	His137, Asn182, and Ala203	-12.37	His137, Asn182, and Ala203	-12.45	His137, Asn182, and Ala203
LLY-507	-11.29	Thr185	-10.8	Thr185 and Tyr258	-10.9	Thr185 and Tyr258	-11.15	Thr185 Tyr258

5.3.5 Description of selected complexes

The docking analysis showed that among the nine selected compounds, three top compounds were reported and interactions were explored with control compounds. These three potential compounds were taken from a large pool of datasets (98071 compounds) using various characteristics. Figure 5.2 depicts the interactions of these three compounds along with the control compound at the active site of SMYD2 protein.

5.3.5.1 LLY-507

In molecular docking analysis, the control compound LLY-507 forms a single hydrogen bond with SMYD2 protein through Thr185 and has -11.29 Kcal.mol^{-1} binding energy. The binding affinity for this complex with Autodock Vina was -10.80 Kcal.mol^{-1} and formed two hydrogen bonds. The binding energy was 11.15 for Kcal.mol^{-1} for iDock and -10.90 Kcal.mol^{-1} for Smina and in both programs two hydrogen bonds were formed with Thr185 and Tyr258 (Table 5.3). The interaction forming residues of SMYD2 is shown in Figure 5.2. The interacting residues of protein and various interaction types with the LLY-507 compound were analyzed by the Autodock tool.

5.3.5.2 ZINC03844862

The molecular docking investigation for this compound with SMYD2 revealed binding energy of -11.57 Kcal.mol^{-1} using Autodock. The two hydrogen bonds were formed with the ligand by two residues such as Thr238 and Tyr240. The binding energy for other programs was -12.60 Kcal.mol^{-1} for Smina, -12.32 Kcal.mol^{-1} for iDock, and -12.10 Kcal.mol^{-1} for Autodock Vina (Table 5.3). These three programs did not show any hydrogen bond formation between compound and protein. Diverse hydrophobic interactions were established by this compound. All interacting residues, analyzed by the Autodock tool are shown in Figure 5.2.

5.3.5.3 ZINC08490711

The binding energy for this compound with SMYD2 was -11.99 Kcal.mol^{-1}. Autodock's binding energy is the highest it has ever shown. Binding affinity for other docking programs was -11.65 Kcal.mol^{-1} for Smina, -11.77 Kcal.mol^{-1} for iDock, and -11.70 Kcal.mol^{-1} Autodock Vina. Four hydrogen bonds were formed with residues Ser239, Val215, Thr238, and Val213. All docking methods identified these hydrogen-forming residues to be prevalent (Table 5.3). ZINC08490711 was also used to make other hydrophobic connections. Various other interactions were also established by this compound. All interacting residues of SMYD2 with this ligand are shown in Figure 5.2. The protein-ligand interactions are analyzed by the Autodock program.

5.3.5.4 ZINC08764231

The binding affinity of the complex generated by Autodock with SMYD2 was -11.62 Kcal.mol^{-1}. In the Autodock analysis, three hydrogen bonds were formed with residues Tyr240, Cys181, and His137. Other programs such as Smina, iDock, and AutodocVina, produced complexes with the binding affinity of -12.37, -12.45, and -12.30 Kcal.mol^{-1}, respectively. The His137, Cys182, and Ala203 residues established three hydrogen bonds with the complex produced by this three software (Table 5.3). All interacting residues produced by Autodock software are shown in Figure 5.2.

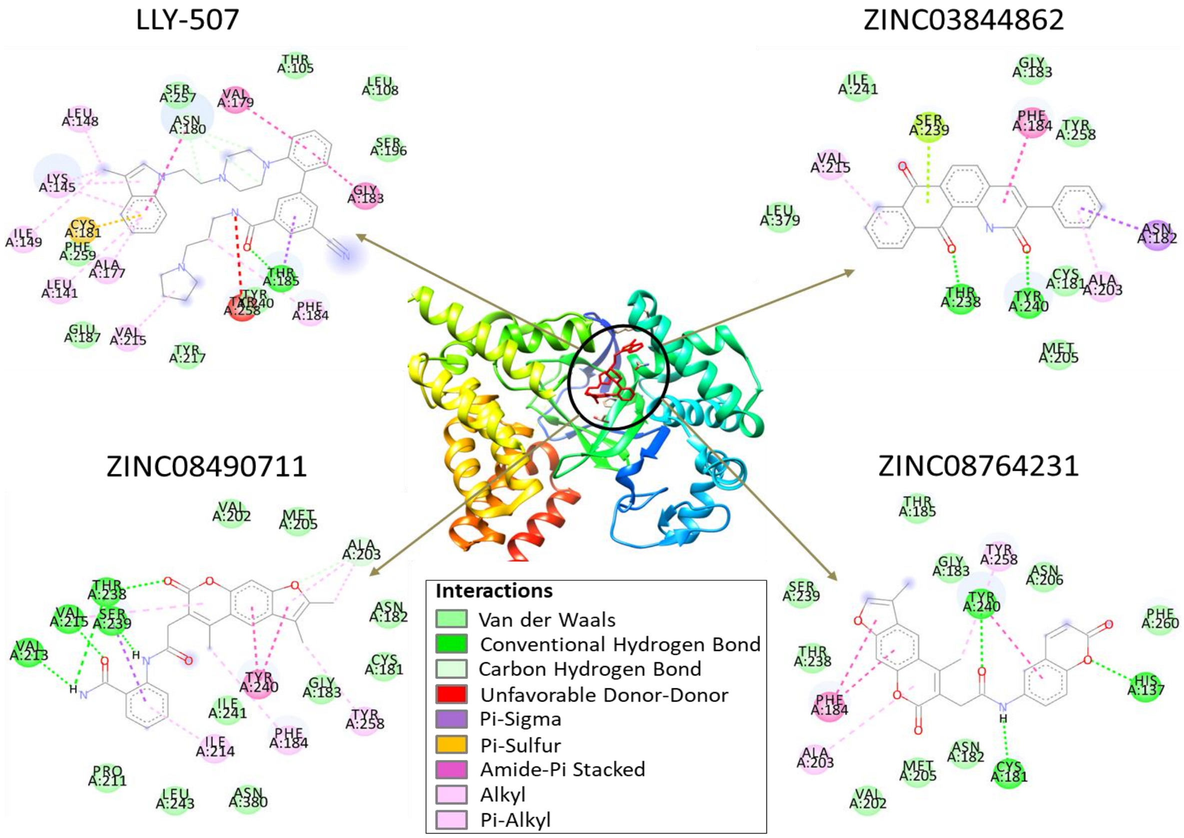


Figure 5.2: The three anticipated compounds' optimal binding postures on the active site of SMYD2 and their interactions along with the control compound LLY-507.

The capacity to predict the best-suited complexes was demonstrated in the redocking experiment. Through cross-docking and redocking tests, three putative SMYD2 inhibitors (ZINC03844862, ZINC08490711, and ZINC08764231) with the highest binding energy were discovered. The interacting residues Asn182, Ala203, Cys181, Gly183, Met205, Phe184, Ser239, Thr238, Tyr240, and Tyr258 were found common in the three complexes that were implicated in small molecule binding. There were multiple hydrogen bond interactions in all three complexes. Residues Thr238 and Tyr240 established two hydrogen bonds with ZINC03844862. ZINC08490711 established a hydrogen bond with the residues Thr213, Val215, Thr238, and Ser239. The residues Cys181, His137, and Tyr240 created three hydrogen bonds with the compound ZINC08764231. Similarly, there were a small number of amide-pi stacking, pi-alkyl, and alkyl interactions in complexes (Figure 5.2). Lastly, MDS analysis was performed on these three compounds.

5.3.6 Analysis of Molecular dynamics simulations

To analyzed the stability and flexibility of protein-ligand complexes are frequently assessed using MDS analysis [17], [46], [47]. MDS was conducted at 100 ns for five systems in this study. The MDS study was used to examine the residual flexibility of the top three docked complexes (SMYD2-ZINC08490711, SMYD2-ZINC08764231, and SMYD2-ZINC03844862) in addition to apo-SMYD2 and control complex (SMYD2-LLY-507).

5.3.6.1 RMSD

RMSD study was used to estimate SMYD2's residual flexibility with targets. The complex stability was investigated using RMSD analysis. Figure 5.3A represented the RMSD fluctuations observed throughout the simulation of 100 ns for all complexes along with apo protein. After 10 nanoseconds, all complexes were found to be in equilibrium. The average RMSD value was 0.34 nm for apo-SMYD2, 0.43 nm for SMYD2-ZINC08490711, 0.53 nm for SMYD2-ZINC08764231, 0.47 nm for SMYD2-ZINC03844862, and 0.45 nm for SMYD2-LLY-507. The selected three complexes show low RMDS values as compared to the control complex. These findings revealed that the selected complexes' stability was consistent.

5.3.6.2 RMSF

The atomic mobility of the protein's alpha carbon atoms was measured using RMSF analysis. The RMSF value was calculated for backbone atoms of all complexes to determine the flexibility of each residue (Figure 5.3B). The average RMSF value was 0.18 nm for apo-SMYD2, 0.21 nm for SMYD2-ZINC08490711, 0.20 nm for SMYD2-ZINC08764231, 0.19 nm for SMYD2-ZINC03844862, and 0.19 nm for SMYD2-LLY-507. These findings demonstrated that during MDS, the molecules can form persistent connections with the protein. The RMSF result corresponded to the RMSD analysis.

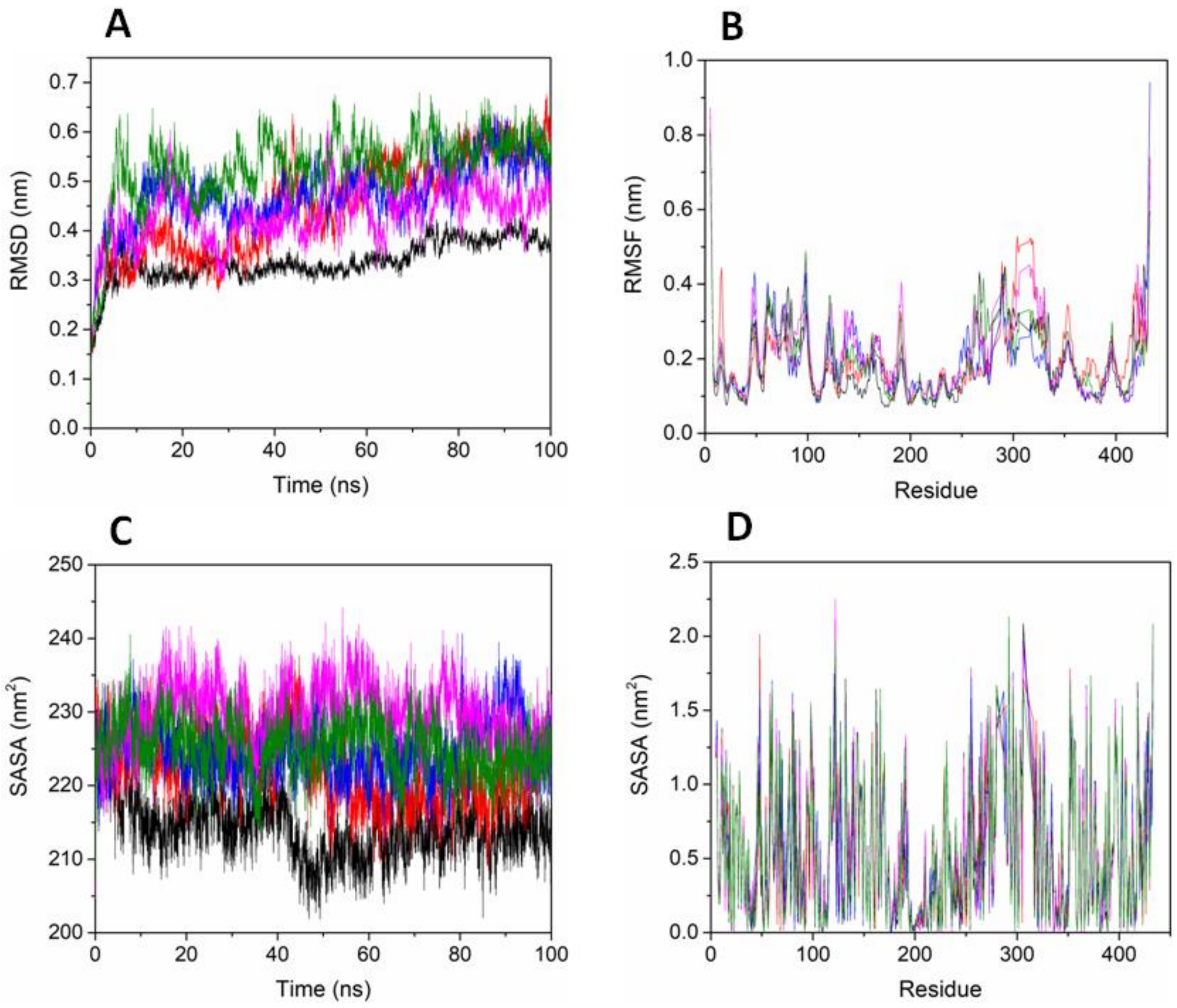


Figure 5.3: Dynamics of ligand binding with SMYD2. (A) Plot for RMSD as a function of time achieved for all systems, (B) Fluctuation of residual backbone (RMSF) plot for all systems, (C) Plot for SASA value as a function of time, and (D) Plot for SASA value with respect to residues. The colors used in the panels are as black for apo-SMYD2, blue for SMYD2-ZINC08490711, and magenta for SMYD2-ZINC08764231, green for SMYD2-ZINC03844862, and red for SMYD2-LLY-507.

5.3.6.4 SASA

SASA stands for the solvent molecule's access to the protein surface area. The increased SASA value indicates that hydrophobic residues play a significant role in ligand binding. As a result, the SASA value was calculated for all complexes in addition to apo-protein, as shown in Figure 5.3C.

The average SASA value was 214.19 nm^2 for apo-SMYD2, 229.68 nm^2 for SMYD2-ZINC08490711, 225.17 nm^2 for SMYD2-ZINC08764231, 224.81 nm^2 for SMYD2-ZINC03844862, and 222.22 nm^2 for SMYD2-LLY-507. The graphs in Figure 5.3D show the results of a residue-wise SASA study. The average value of residue-wise SASA was 0.52 nm^2 for apo-SMYD2, 0.55 nm^2 for SMYD2-ZINC08490711, 0.54 nm^2 for SMYD2-ZINC08764231, 0.54 nm^2 for SMYD2-ZINC03844862, and 0.53 nm^2 for SMYD2-LLY-507. Our complexes have higher SASA values as compared to the control complex indicating that our complexes show induced stability when it binds with the ligand.

5.3.6.4 Hydrogen bonds analysis

The hydrogen bond formation throughout the simulation study was also calculated, and the plot for the number of hydrogen bonds is shown in Figure 5.4A. Throughout the simulation, the total number of estimated hydrogen bonds varied from 1 to 7. The maximum four hydrogen bonds were formed for the complex SMYD2-ZINC08490711. The compounds having a greater average value of hydrogen bonds suggested that compounds expressively inhibit the SMYD2 protein.

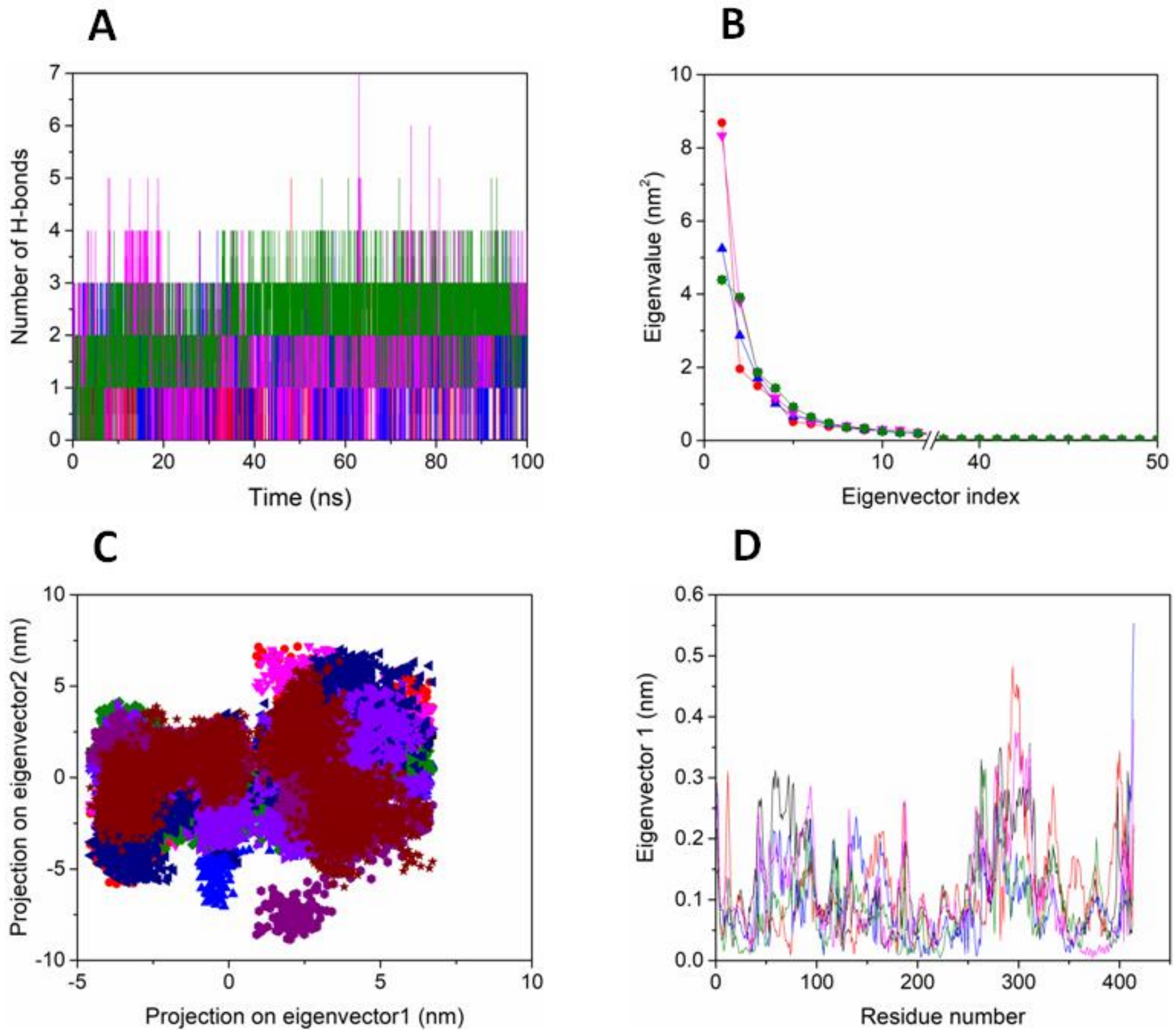


Figure 5.4: Hydrogen bond and PCA analysis. (A) The average number of hydrogen bonds in respect to time, (B) Correlation motion as eigenvector vs. eigenvalues index plot. (C) Projection of protein motion prediction in 2D phase space and (D) Residue-wise EigRMSF value in all complexes. The colors used in the panels are as black for apo-SMYD2, blue for SMYD2-ZINC08490711, and magenta for SMYD2-ZINC08764231, green for SMYD2-ZINC03844862, and red for SMYD2-LLY-507.

5.3.6.5 Principal component analysis

The study of PCA was done to identify the collective motions among all complexes, along with apo-protein. In Figure 5.4B, we displayed the first 50 eigenvectors versus eigenvalue. The first few eigenvectors significantly contribute to the protein's movements. The starting 10 eigenvectors accounted 76.45% for apo-SMYD2, 78.83% for SMYD2-ZINC08490711, 76.51% for SMYD2-ZINC08764231, 75.80% for SMYD2-ZINC03844862 and 79.18% for SMYD2-LLY-507. We observed the correlation motions by plotting the first two eigenvectors against each other in phase space (Figure 5.4C). There were stable clusters in all of the complexes. The residues-wise mobility was also calculated to see the residue-wise correlation motion by using the eignRMSF value as shown in Figure 5.4D. The value of eigRMSF was 0.12 nm for apo-SMYD2, 0.11 nm for SMYD2-ZINC08490711, and 0.08 nm for SMYD2-ZINC08764231 0.09 nm for SMYD2-ZINC03844862, and 0.11 nm for SMYD2-LLY-507. As a consequence of the PCA study, the results indicated that our top complexes have slight fluctuations and are pretty stable.

5.4 CONCLUSION

The SMYD2 protein has an essential role in numerous cancers progression by the methylation of several tumor suppressor proteins; therefore, it can be used as a significant target for new cancer drug development. Using a variety of *in-silico* analyses, we have anticipated possible SMYD2 inhibitors in this study. Molecular docking analysis was used to improve the accuracy of the screening process and to reveal the interactions amongst the compounds that were chosen. Finally, three compounds were identified as possible lead compounds that could be employed as cancer therapeutics: ZINC03844862, ZINC08490711, and ZINC08764231. The prediction of ADMET profiles indicated that our proposed compounds have good clearance on all parameters and would be used as a nontoxic and efficient drug candidate. The MDS study of top compounds in complex with SMYD2 showed stable binding. The anticipated molecules will essential to be tested in the lab before they may be produced as anti-cancer drugs. The hypothesized chemicals could be ideal candidates for cancer prevention and control, according to the researchers.

REFERENCES

[1] M. Luo, “Chemical and Biochemical Perspectives of Protein Lysine Methylation,” *Chem Rev*, vol. 118, no. 14, pp. 6656–6705, Jul. 2018, doi: 10.1021/acs.chemrev.8b00008.

[2] E. L. Greer and Y. Shi, “Histone methylation: a dynamic mark in health, disease and inheritance,” *Nature Reviews Genetics*, vol. 13, no. 5, Art. no. 5, May 2012, doi: 10.1038/nrg3173.

[3] E. Eggert *et al.*, “Discovery and Characterization of a Highly Potent and Selective Aminopyrazoline-Based in Vivo Probe (BAY-598) for the Protein Lysine Methyltransferase SMYD2,” *J. Med. Chem.*, vol. 59, no. 10, pp. 4578–4600, 26 2016, doi: 10.1021/acs.jmedchem.5b01890.

[4] R. Hamamoto, V. Saloura, and Y. Nakamura, “Critical roles of non-histone protein lysine methylation in human tumorigenesis,” *Nat. Rev. Cancer*, vol. 15, no. 2, pp. 110–124, Feb. 2015, doi: 10.1038/nrc3884.

[5] H. Nguyen *et al.*, “LLY-507, a Cell-active, Potent, and Selective Inhibitor of Protein-lysine Methyltransferase SMYD2,” *J Biol Chem*, vol. 290, no. 22, pp. 13641–13653, May 2015, doi: 10.1074/jbc.M114.626861.

[6] J. Huang *et al.*, “Repression of p53 activity by Smyd2-mediated methylation,” *Nature*, vol. 444, no. 7119, pp. 629–632, Nov. 2006, doi: 10.1038/nature05287.

[7] M. Nakakido, Z. Deng, T. Suzuki, N. Dohmae, Y. Nakamura, and R. Hamamoto, “Dysregulation of AKT Pathway by SMYD2-Mediated Lysine Methylation on PTEN,” *Neoplasia*, vol. 17, no. 4, pp. 367–373, Apr. 2015, doi: 10.1016/j.neo.2015.03.002.

[8] L. A. Saddic *et al.*, “Methylation of the Retinoblastoma Tumor Suppressor by SMYD2,” *J Biol Chem*, vol. 285, no. 48, pp. 37733–37740, Nov. 2010, doi: 10.1074/jbc.M110.137612.

[9] H.-S. Cho *et al.*, “RB1 Methylation by SMYD2 Enhances Cell Cycle Progression through an Increase of RB1 Phosphorylation,” *Neoplasia*, vol. 14, no. 6, pp. 476–486, Jun. 2012.

[10] T. Voelkel, C. Andresen, A. Unger, S. Just, W. Rottbauer, and W. A. Linke, “Lysine methyltransferase Smyd2 regulates Hsp90-mediated protection of the sarcomeric titin springs and cardiac function,” *Biochim. Biophys. Acta*, vol. 1833, no. 4, pp. 812–822, Apr. 2013, doi: 10.1016/j.bbamcr.2012.09.012.

[11] X. Zhang *et al.*, “Regulation of estrogen receptor α by histone methyltransferase SMYD2-mediated protein methylation,” *Proc. Natl. Acad. Sci. U.S.A.*, vol. 110, no. 43, pp. 17284–17289, Oct. 2013, doi: 10.1073/pnas.1307959110.

[12] L. Piao *et al.*, “The histone methyltransferase SMYD2 methylates PARP1 and promotes poly(ADP-ribosyl)ation activity in cancer cells,” *Neoplasia*, vol. 16, no. 3, pp. 257–264, 264.e2, Mar. 2014, doi: 10.1016/j.neo.2014.03.002.

[13] L. X. Li *et al.*, “Lysine methyltransferase SMYD2 promotes cyst growth in autosomal dominant polycystic kidney disease,” *J. Clin. Invest.*, vol. 127, no. 7, pp. 2751–2764, Jun. 2017, doi: 10.1172/JCI90921.

[14] D. K. Jarrell, K. N. Hassell, D. C. Crans, S. Lanning, and M. A. Brown, “Characterizing the Role of SMYD2 in Mammalian Embryogenesis—Future Directions,” *Vet Sci*, vol. 7, no. 2, May 2020, doi: 10.3390/vetsci7020063.

[15] A. K. Yadav and T. R. Singh, “Novel structural and functional impact of damaging single nucleotide polymorphisms (SNPs) on human SMYD2 protein using computational approaches,” *Meta Gene*, vol. 28, p. 100871, Jun. 2021, doi: 10.1016/j.mgene.2021.100871.

[16] D. R. Houston and M. D. Walkinshaw, “Consensus Docking: Improving the Reliability of Docking in a Virtual Screening Context,” *J. Chem. Inf. Model.*, vol. 53, no. 2, pp. 384–390, Feb. 2013, doi: 10.1021/ci300399w.

[17] R. Shukla and T. R. Singh, “Virtual screening, pharmacokinetics, molecular dynamics and binding free energy analysis for small natural molecules against cyclin-dependent kinase 5 for Alzheimer’s disease,” *J. Biomol. Struct. Dyn.*, vol. 38, no. 1, pp. 248–262, 2020, doi: 10.1080/07391102.2019.1571947.

[18] Y. Ben Shoshan-Galeczki and M. Y. Niv, “Structure-based screening for discovery of sweet compounds,” *Food Chemistry*, vol. 315, p. 126286, Jun. 2020, doi: 10.1016/j.foodchem.2020.126286.

[19] I. S. Gade *et al.*, “Anticancer Activity of Combretum fragrans F. Hoffm on Glioblastoma and Prostate Cancer Cell Lines,” *Asian Pac J Cancer Prev*, vol. 22, no. 4, pp. 1087–1093, Apr. 2021, doi: 10.31557/APJCP.2021.22.4.1087.

[20] S. Aliebrahimi, S. Montasser Kouhsari, S. N. Ostad, S. S. Arab, and L. Karami, “Identification of Phytochemicals Targeting c-Met Kinase Domain using Consensus

Docking and Molecular Dynamics Simulation Studies," *Cell Biochem. Biophys.*, vol. 76, no. 1–2, pp. 135–145, Jun. 2018, doi: 10.1007/s12013-017-0821-6.

[21] H. M. Ashtawy and N. R. Mahapatra, "Machine-learning scoring functions for identifying native poses of ligands docked to known and novel proteins," *BMC Bioinformatics*, vol. 16, no. 6, p. S3, Apr. 2015, doi: 10.1186/1471-2105-16-S6-S3.

[22] V. Kumar, S. Krishna, and M. I. Siddiqi, "Virtual screening strategies: recent advances in the identification and design of anti-cancer agents," *Methods*, vol. 71, pp. 64–70, Jan. 2015, doi: 10.1016/j.ymeth.2014.08.010.

[23] N. Razzaghi-Asl, S. Mirzayi, K. Mahnam, and S. Sepehri, "Identification of COX-2 inhibitors via structure-based virtual screening and molecular dynamics simulation," *Journal of Molecular Graphics and Modelling*, vol. 83, pp. 138–152, Aug. 2018, doi: 10.1016/j.jmgm.2018.05.010.

[24] C.-H. Chu *et al.*, "KDM4B as a Target for Prostate Cancer: Structural Analysis and Selective Inhibition by a Novel Inhibitor," *J. Med. Chem.*, vol. 57, no. 14, pp. 5975–5985, Jul. 2014, doi: 10.1021/jm500249n.

[25] F. Meng *et al.*, "Discovery and Optimization of Novel, Selective Histone Methyltransferase SET7 Inhibitors by Pharmacophore- and Docking-Based Virtual Screening," *J. Med. Chem.*, vol. 58, no. 20, pp. 8166–8181, Oct. 2015, doi: 10.1021/acs.jmedchem.5b01154.

[26] S. Huang *et al.*, "Discovery of New SIRT2 Inhibitors by Utilizing a Consensus Docking/Scoring Strategy and Structure-Activity Relationship Analysis," *J Chem Inf Model*, vol. 57, no. 4, pp. 669–679, 24 2017, doi: 10.1021/acs.jcim.6b00714.

[27] Q. Wang *et al.*, "Identification of a Novel Protein Arginine Methyltransferase 5 Inhibitor in Non-small Cell Lung Cancer by Structure-Based Virtual Screening," *Front Pharmacol*, vol. 9, p. 173, 2018, doi: 10.3389/fphar.2018.00173.

[28] E. F. Pettersen *et al.*, "UCSF Chimera--a visualization system for exploratory research and analysis," *J Comput Chem*, vol. 25, no. 13, pp. 1605–1612, Oct. 2004, doi: 10.1002/jcc.20084.

[29] J. W. Ponder and D. A. Case, "Force fields for protein simulations," *Adv. Protein Chem.*, vol. 66, pp. 27–85, 2003, doi: 10.1016/s0065-3233(03)66002-x.

[30] J. J. Irwin and B. K. Shoichet, “ZINC – A Free Database of Commercially Available Compounds for Virtual Screening,” *J Chem Inf Model*, vol. 45, no. 1, pp. 177–182, 2005, doi: 10.1021/ci049714.

[31] D. R. Koes, M. P. Baumgartner, and C. J. Camacho, “Lessons learned in empirical scoring with smina from the CSAR 2011 benchmarking exercise,” *J Chem Inf Model*, vol. 53, no. 8, pp. 1893–1904, Aug. 2013, doi: 10.1021/ci300604z.

[32] O. Trott and A. J. Olson, “AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization and multithreading,” *J Comput Chem*, vol. 31, no. 2, pp. 455–461, Jan. 2010, doi: 10.1002/jcc.21334.

[33] H. Li, K.-S. Leung, and M.-H. Wong, “idock: A multithreaded virtual screening tool for flexible ligand docking,” in *2012 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, San Diego, CA, USA, May 2012, pp. 77–84. doi: 10.1109/CIBCB.2012.6217214.

[34] F. Cheng *et al.*, “admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties,” *J Chem Inf Model*, vol. 52, no. 11, pp. 3099–3105, Nov. 2012, doi: 10.1021/ci300367a.

[35] D. E. V. Pires, T. L. Blundell, and D. B. Ascher, “pkCSM: Predicting Small-Molecule Pharmacokinetic and Toxicity Properties Using Graph-Based Signatures,” *J Med Chem*, vol. 58, no. 9, pp. 4066–4072, May 2015, doi: 10.1021/acs.jmedchem.5b00104.

[36] L. Zhang *et al.*, “CarcinoPred-EL: Novel models for predicting the carcinogenicity of chemicals using molecular fingerprints and ensemble learning methods,” *Sci Rep*, vol. 7, no. 1, p. 2118, May 2017, doi: 10.1038/s41598-017-02365-0.

[37] G. M. Morris *et al.*, “AutoDock4 and AutoDockTools4: Automated Docking with Selective Receptor Flexibility,” *J Comput Chem*, vol. 30, no. 16, pp. 2785–2791, Dec. 2009, doi: 10.1002/jcc.21256.

[38] K. Prerna and V. K. Dubey, “Repurposing of FDA-approved drugs as autophagy inhibitors in tumor cells,” *Journal of Biomolecular Structure and Dynamics*, vol. 0, no. 0, pp. 1–12, Jan. 2021, doi: 10.1080/07391102.2021.1873862.

[39] A. L. da Fonseca *et al.*, “Docking, QM/MM, and molecular dynamics simulations of the hexose transporter from Plasmodium falciparum (PfHT),” *J Mol Graph Model*, vol. 66, pp. 174–186, May 2016, doi: 10.1016/j.jmgm.2016.03.015.

[40] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, “GROMACS 4: Algorithms for Highly Efficient, Load-Balanced, and Scalable Molecular Simulation,” *J. Chem. Theory Comput.*, vol. 4, no. 3, pp. 435–447, Mar. 2008, doi: 10.1021/ct700301q.

[41] S. Pronk *et al.*, “GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit,” *Bioinformatics*, vol. 29, no. 7, pp. 845–854, Apr. 2013, doi: 10.1093/bioinformatics/btt055.

[42] H. J. C. Berendsen, J. R. Grigera, and T. P. Straatsma, “The missing term in effective pair potentials,” *J. Phys. Chem.*, vol. 91, no. 24, pp. 6269–6271, Nov. 1987, doi: 10.1021/j100308a038.

[43] S. Aw and van A. Dm, “PRODRG: a tool for high-throughput crystallography of protein-ligand complexes,” *Acta crystallographica. Section D, Biological crystallography*, Aug. 2004. https://pubmed.ncbi.nlm.nih.gov/15272157/ (accessed Nov. 17, 2020).

[44] W. Huang, Z. Lin, and W. F. van Gunsteren, “Validation of the GROMOS 54A7 Force Field with Respect to β-Peptide Folding,” *J. Chem. Theory Comput.*, vol. 7, no. 5, pp. 1237–1243, May 2011, doi: 10.1021/ct100747y.

[45] C. Rawat *et al.*, “Downregulation of peripheral PTGS2/COX-2 in response to valproate treatment in patients with epilepsy,” *Sci Rep*, vol. 10, no. 1, Art. no. 1, Feb. 2020, doi: 10.1038/s41598-020-59259-x.

[46] S. Kachhap and B. Singh, “Role of DNA conformation & energetic insights in Msx-1-DNA recognition as revealed by molecular dynamics studies on specific and nonspecific complexes,” *J Biomol Struct Dyn*, vol. 33, no. 10, pp. 2069–2082, 2015, doi: 10.1080/07391102.2014.995709.

[47] V. Randhawa, S. Pathania, and M. Kumar, “Computational Identification of Potential Multitarget Inhibitors of Nipah Virus by Molecular Docking and Molecular Dynamics,” *Microorganisms*, vol. 10, no. 6, Art. no. 6, Jun. 2022, doi: 10.3390/microorganisms10061181.

ANNEXURE

Table 2.1: Performance of all models on training dataset with respect to all features in term of accuracy percentage.

	SVM	**LR**	**LDA**	**KNN**	**CART**	**NB**	**RF**	**MLP**	**AdaBoost**	**XGBoost**	**LightGBM**	**SGD**	**Bagging**	**QDA**
AAC	81.39	71.39	74.97	76.95	71.26	73.33	81.15	77.7	76.87	80.32	80.38	72.57	77.72	76.45
APAAC	78.88	74.65	74.89	80	70.34	61.25	80.33	81.33	76.98	81.79	81.94	69.6	76.79	72.4
CKSAAGP	74.47	69.55	73.74	66.49	64.44	70.38	73.26	75.65	71.37	75.25	76.45	70.58	69.53	63.83
CKSAAP	87.02	73.31	77.79	67.98	66.91	71.28	79.23	84.71	77.59	83.53	83.22	77.74	73.88	51.97
CTDC	75.43	71.3	73.55	73.09	67.96	68.44	75.19	75.19	73.75	75.23	76.2	71.5	72.96	51.5
CTDD	63.98	65.84	65.64	59.68	59.26	60.03	69.23	58.96	66.93	69.62	69.99	61.05	63.96	50.19
CTDT	75.39	70.93	73.47	70.78	66.04	68.5	74.4	73.79	72.26	73.99	74.56	70.6	7054	72.88
Ctriad	77.88	75.87	75.85	64.42	63.06	68.11	75.72	76.35	73.22	77.73	77.94	74.21	72.35	64.37
DDE	82.24	80.82	80.86	71.19	66.76	71.36	78.72	81.72	76.15	82.39	81.67	76.85	73	76.8
DPC	83.57	70.19	80.27	67.65	66.47	71.65	79.14	80.6	75.67	82.5	82.48	66.34	73.72	71.39
GAAC	70.88	68.09	68.35	67.5	62.06	68.94	69.27	69.42	69.6	68.37	69.66	67.67	66.19	68.66
GDPC	72.89	67.41	70.97	68.85	64.2	70.49	73.57	72.52	69.97	72.02	73.57	67.24	69.9	65.93
Geary	72.79	69.68	69.49	56.51	58.47	58.8	66.47	66.01	66.36	69.58	70.52	63	63.02	60.18
Moran	73.66	70.36	70.27	56.51	57.32	59.17	67.04	69.95	66.84	70.23	70.58	70.65	62.47	60.42
MoreauBroto	76.09	72.2	72.13	59.06	59.74	60.94	71.72	75.61	69.03	74.56	75.45	72	65.8	63.65
PAAC	78.86	74.47	72.63	79.01	70.32	57.67	81.32	82.92	77.74	81.52	81.54	66.56	76.52	70.97
QSOrder	79.97	71.89	77.52	77.35	68.9	65.36	80.82	78.05	76.52	81.45	82.46	72.39	75.91	67.48
SOCNumber	66.87	57.52	67.46	61.36	60.45	66.34	69.31	52.34	66.32	68.72	69.42	51.19	66.52	63.2

Table 2.2: Performance of all models on independent test dataset with respect to all features in term of accuracy percentage.

	SVM	LR	LDA	KNN	CART	NB	RF	MLP	Ada-Boost	XG-Boost	Light-GBM	SGD	Bagging	QDA
AAC	82.62	73.1	76.15	79.21	71.7	74.23	80.61	79.38	78.51	81.31	81.65	74.23	78.07	76.24
APAAC	81.46	75.26	72.9	81.99	68	62.5	81.29	83.39	79.28	81.31	82.86	74.65	78.23	73.68
CKSAAGP	75.54	71.7	73.97	69.25	63.66	69.78	76.15	75.28	73.53	76.5	76.85	72.4	73.01	62.18
CKSAAP	87.94	74.67	78.34	70.39	68.64	73.01	80.08	84.36	79.21	85.32	85.24	79.56	75.19	50.21
CTDC	75.25	71.32	72.37	74.47	67.75	69.31	76.74	75.17	75.17	76.31	77.7	72.98	73.16	58.12
CTDD	62.84	66.25	66.78	57.69	57.08	61.36	70.27	61.53	68.44	71.94	70.97	49.47	66.6	49.73
CTDT	79.8	70.8	72.55	71.85	67.91	69.84	76.22	75.34	74.56	77.44	77.62	70.19	72.9	75.17
Ctriad	77.97	76.13	76.13	66.25	59.44	66.87	77.18	75.26	73.95	79.02	78.4	75.61	71.67	62.76
DDE	84.36	81.65	81.31	73.44	64.89	74.41	80.96	83.49	78.95	84.45	84.19	73.79	74.49	79.38
DPC	83.66	71.7	81.22	70.74	66.81	73.53	79.65	81.22	79.3	84.1	83.93	72.13	74.49	73.18
GAAC	72.05	70.91	71.17	65.67	64.45	70.04	70.91	70.82	70.48	71.61	72.48	69.6	68.12	68.2
GDPC	73.62	71.17	70.91	69.6	65.06	70.82	73.88	72.05	72.75	72.66	73.62	72.48	68.47	65.93
Geary	72.2	69.23	68.53	54.8	56.99	60.13	67.22	67.04	68.61	70.71	71.41	64.51	62.41	62.58
Moran	71.85	68.7	69.75	55.59	59.61	60.4	68	69.49	66.34	69.84	71.67	68.79	63.19	63.02
MoreauBroto	77.18	70.71	70.1	60.4	59.26	62.15	71.76	75	69.93	74.56	75.17	70.1	64.07	66.52
PAAC	80.17	75.1	77.03	80.34	72.22	60.17	81.83	83.49	78.68	81.57	82.44	58.95	77.64	71.96
QSOrder	81.48	72.31	77.99	79.3	70.56	64.8	82	79.65	76.94	83.14	83.4	74.32	77.11	66.81
SOCNumber	69.95	56.85	69.78	64.01	62.88	67.16	71.7	49.78	69.95	71.17	73.01	51	67.94	62.35

Table 3.1: Correlation of SMYD2 with various cancers in immune cell infiltration. Significant correlation (both positively and negatively) with their top performing algorithm is shown.

Cancer name (number of sample)	Immune cell (Algorithm name)	Rho-value	p-value	Adjusted p-value
BLCA (n=408)	Cancer associated fibroblast (EPIC)	0.208921184	5.37E-05	0.000940094
BRCA (n=1100)	Cancer associated fibroblast (XCELL)	-0.084872921	0.007421402	0.046177614
BRCA (n=1100)	T cell CD8+ (MCPCOUNTER)	0.115119323	0.000275847	0.003218216
BRCA-Basal (n=191)	T cell CD8+ central memory (XCELL)	-0.163808446	0.030786747	0.117754565
BRCA-LumA (n=568)	Cancer associated fibroblast (EPIC)	0.2609479	1.71E-09	1.36E-07
BRCA-LumA (n=568)	T cell CD8+ (TIMER)	0.238381553	4.10E-08	1.77E-06
BRCA-LumB (n=219)	Cancer associated fibroblast (EPIC)	0.237009138	0.000933165	0.00926377
CESC (n=306)	Cancer associated fibroblast (MCPCOUNTER)	0.217527063	0.00026452	0.003151727
CESC (n=306)	T cell CD8+ (CIBERSORT-ABS)	-0.136613491	0.02296046	0.10204649
COAD (n=458)	T cell CD8+ central memory (XCELL)	-0.171289316	0.004390426	0.031122005
DLBC (n=48)	Cancer associated fibroblast (XCELL)	-0.336533182	0.031443397	0.118975017
DLBC (n=48)	T cell CD8+ (TIMER)	0.339587863	0.029841558	0.116862044
ESCA (n=185)	Cancer associated fibroblast (TIDE)	0.448698139	2.67E-10	2.49E-08
ESCA (n=185)	T cell CD8+ (EPIC)	-0.286775743	9.50E-05	0.00147793
HNSC (n=522)	Cancer associated fibroblast (TIDE)	0.162352433	0.00029918	0.003419199
HNSC (n=522)	T cell CD8+ (EPIC)	-0.22066232	7.69E-07	2.53E-05
HNSC-HPV- (n=422)	Cancer associated fibroblast (TIDE)	0.240976725	1.08E-06	3.04E-05
HNSC-HPV- (n=422)	T cell CD8+ (EPIC)	-0.280527191	1.14E-08	7.11E-07
KIRC (n=533)	Cancer associated fibroblast (MCPCOUNTER)	-0.134977876	0.003689838	0.02718828
KIRC (n=533)	T cell CD8+ (EPIC)	0.262045061	1.12E-08	7.11E-07
LIHC (n=371)	Cancer associated fibroblast (TIDE)	0.129325836	0.016238763	0.080881159
LIHC (n=371)	T cell CD8+ (EPIC)	0.129227666	0.016320662	0.080881159
PAAD (n=179)	Cancer associated fibroblast (XCELL)	-0.200365854	0.008598782	0.05068756
PAAD (n=179)	T cell CD8+ (CIBERSORT)	-0.195066225	0.010566378	0.059171715

Table 3.2: Significant pathways and functional enrichment analysis list with GO term of SMYD2 related genes.

Term	p-value	q-value
RECTOME Pathways		
Activation Of Arylsulfatases R-HSA-1663150	0.001862428	0.28068574
RAB Geranylgeranylation R-HSA-8873719	0.003704472	0.28068574
Metabolism Of Proteins R-HSA-392499	0.005567543	0.28068574
Post-chaperonin Tubulin Folding Pathway R-HSA-389977	0.00584762	0.28068574
Metal Ion SLC Transporters R-HSA-425410	0.007438995	0.2856574
Cargo Concentration In ER R-HSA-5694530	0.011134522	0.33980545
Protein Folding R-HSA-391251	0.01238874	0.33980545
Gamma Carboxylation, Hypusine Formation And Arylsulfatase Activation R-HSA-163841	0.018714055	0.378438151
Post-translational Protein Modification R-HSA-597592	0.020222681	0.378438151
Glycosphingolipid Metabolism R-HSA-1660662	0.02131096	0.378438151
KEGG Pathways		
HIF-1 signaling pathway	0.002190906	0.111736198
Glycolysis / Gluconeogenesis	0.044353776	0.494592481
Central carbon metabolism in cancer	0.047984787	0.494592481
GO Biological		
Mitochondrial transport (GO:0006839)	0.000177012	0.077815177
Oxaloacetate metabolic process (GO:0006107)	0.000245087	0.077815177
Tubulin complex assembly (GO:0007021)	0.000679555	0.143839177
Response to nitric oxide (GO:0071731)	0.001862428	0.25053564
Protein modification by small protein conjugation or removal (GO:0070647)	0.002165765	0.25053564
Proteolysis (GO:0006508)	0.003202217	0.25053564
Protein import into mitochondrial matrix (GO:0030150)	0.004004046	0.25053564
mitochondrial cytochrome c oxidase assembly (GO:0033617)	0.0044345	0.25053564
Cellular response to decreased oxygen levels (GO:0036294)	0.005004401	0.25053564
Respiratory chain complex IV assembly (GO:0008535)	0.00584762	0.25053564
GO Molecular		
GDP binding (GO:0019003)	0.000354195	0.044628576

Dipeptidase activity (GO:0016805)	0.000870867	0.049261757
Purine ribonucleoside triphosphate binding (GO:0035639)	0.002164601	0.049261757
Arylsulfatase activity (GO:0004065)	0.002165765	0.049261757
Exopeptidase activity (GO:0008238)	0.002370205	0.049261757
GTP binding (GO:0005525)	0.002607179	0.049261757
Nucleoside-triphosphatase activity (GO:0017111)	0.002736764	0.049261757
Sulfuric ester hydrolase activity (GO:0008484)	0.003205288	0.050483291
Guanyl ribonucleotide binding (GO:0032561)	0.004428742	0.058030422
GTPase activity (GO:0003924)	0.004605589	0.058030422
GO Cellular		
Integral component of mitochondrial membrane (GO:0032592)	0.000334323	0.040453096
Microtubule cytoskeleton (GO:0015630)	0.001371658	0.082985311
Integral component of mitochondrial outer membrane (GO:0031307)	0.005356497	0.153876846
Intrinsic component of mitochondrial outer membrane (GO:0031306)	0.00584762	0.153876846
INO80-type complex (GO:0097346)	0.006358547	0.153876846
Nuclear chromosome (GO:0000228)	0.008346007	0.157408952
Lysosome (GO:0005764)	0.010033373	0.157408952
Azurophil granule lumen (GO:0035578)	0.010407203	0.157408952
Lytic vacuole (GO:0000323)	0.024264693	0.264603316
Postsynaptic recycling endosome (GO:0098837)	0.029630963	0.264603316

RESEARCH PUBLICATIONS FROM Ph.D. WORK

1. **A. K. Yadav** and T. R. Singh, “Novel inhibitors design through structural investigations and simulation studies for human PKMTs (SMYD2) involved in cancer,” *Molecular Simulation*, vol. 47, no. 14, pp. 1149–1158, 2021, doi: 10.1080/08927022.2021.1957882.

2. **A. K. Yadav** and T. R. Singh, “Novel structural and functional impact of damaging single nucleotide polymorphisms (SNPs) on human SMYD2 protein using computational approaches,” *Meta Gene*, vol. 28, p. 100871, 2021, doi: 10.1016/j.mgene.2021.100871.

3. **A. K. Yadav**, P. K. Gupta and T. R. Singh, “PMTPred: Machine learning based prediction of protein methyltransferases using the composition of k-spaced amino acid pairs”. *Interdisciplinary Sciences: Computational Life Sciences*, [Under review]

4. **A. K. Yadav** and T. R. Singh. “An integrative analysis for the potential role of SMYD2 protein in the prognosis of human cancers”. *Journal of Genetic Engineering and Biotechnology*, [Under review]

CONFERENCES

1. Poster presentation in the *“17th International Conference on Bioinformatics (INCOB-2018)”* at Jawaharlal Nehru University, New Delhi, India, held on September 26-28, 2018.

2. Oral presentation in the *International Conference on Biotechnology and Bioinformatics* on August 1-3, 2019 at Jaypee University of Information Technology, Solan, India on the topic “Development of machine-learning based prediction method for protein methyltransferases”.

3. Attended *“International conference on Recent trends in Science and Technology”* virtually organized by Sevadal Mahila Mahavidyalaya, Nagpur Sardar Patel

Mahavidyalaya, Chandrapur, and Guru Nanak Collage of Science, Ballarpur, on September 3, 2021.

4. Oral presentation in the *International Conference on Advances in biosciences and Biotechnology* on January 20-22, 2022 at Jaypee Institute of Information Technology, Noida on the topic "A multi-omics analysis to reveal the significant correlation of SMYD2 with different cancers using TCGA dataset".

BOOK CHAPTERS

1. **A. K. Yadav**, R. Shukla, and T. R. Singh, "Chapter 22 - Topological parameters, patterns, and motifs in biological networks," in *Bioinformatics*, D. B. Singh and R. K. Pathak, Eds. Academic Press, 2022, pp. 367–380. doi: 10.1016/B978-0-323-89775-4.00012-2.

2. R. Shukla, **A. K. Yadav**, W. O. Sote, M. C. Junior, and T. R. Singh, "Chapter 25 - Systems biology and big data analytics," in *Bioinformatics*, D. B. Singh and R. K. Pathak, Eds. Academic Press, 2022, pp. 425–442. doi: 10.1016/B978-0-323-89775-4.00005-5.

3. **A. K. Yadav**, R. Shukla, and T. R. Singh, "Chapter 11 - Machine learning in expert systems for disease diagnostics in human healthcare," in *Machine Learning, Big Data, and IoT for Medical Informatics*, P. Kumar, Y. Kumar, and M. A. Tawhid, Eds. Academic Press, 2021, pp. 179–200. doi: 10.1016/B978-0-12-821777-1.00022-7.

4. R. Shukla, **A. K. Yadav**, and T. R. Singh, "Application of Deep Learning in Biological Big Data Analysis," in *Large-Scale Data Streaming, Processing, and Blockchain Security*, IGI Global, 2021, pp. 117-148. DOI: 10.4018/978-1-7998-3444-1.ch006

MACHINE-LEARNING BASED PREDICTION AND COMPUTATIONAL INVESTIGATIONS ON PROTEIN METHYLTRANSFERASES INVOLVED IN HUMAN MALIGNANCIES

Thesis submitted in fulfillment of the requirements for the Degree of

DOCTOR OF PHILOSOPHY

IN

BIOINFORMATICS

BY

ARVIND KUMAR YADAV

Department of Biotechnology and Bioinformatics

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY
WAKNAGHAT, SOLAN, H.P.-173234, INDIA
DECEMBER 2022

6.1 CONCLUSION

The identification of correct PMTs is essential to understanding their precise biological functions as they have significant importance in various ways. However, in this Ph.D. work, a machine-learning based method was developed using the amino acid sequence feature for the prediction of PMTs. The role of SMYD2 in numerous cancer progressions over the methylation of non-histone tumor suppressor proteins is crucial because it can aid in the development of new cancer drugs and therapies. As a result, SMYD2 has become an attractive and possible beneficial target for the development of cancer drugs. Herein, a variety of *in silico* analyses was performed and three novel SMYD2-associated nsSNPs were identified that significantly affect the enzymatic activity of SMYD2. Furthermore, novel SMYD2-specific inhibitors were also proposed in this study. Molecular docking analysis was used to advance the accuracy of the screening process and based on binding energy, the most appropriate compounds were chosen. Finally, three compounds were identified as possible lead compounds that could be employed as cancer therapeutics. The lead molecules were also examined using ADMET, and we discovered that they met the requirements for blood-brain barrier permeability and were not hazardous or carcinogenic. The MDS study of chosen chemicals revealed that the binding of ligand molecules with protein was stable. The anticipated molecules will essential to be tested in the lab before they may be produced as anti-cancer drugs. The hypothesized drug-like molecules could be ideal candidates for cancer prevention and control.

6.2 Important findings from the research work

- Chapter 1 of this thesis deals with the development of a machine-learning-based model that helps to classify the PMTs and non-PMTs. The significant importance of PMTs suggested that there is a need to develop an accurate prediction method for the identification of PMTs. The prediction method is trained using amino acid sequences from a public database. A total of 18 sequence-based numerical features were calculated and model training was performed by using 14 different machine-learning algorithms to find out the best feature and best algorithm in the case of my dataset. Out of 18 features, the CKSAAP feature achieved the highest accuracy with the SVM algorithm. Performance of this model was observed as the accuracy of 87.94% with balance Sensitivity (88.8%) and Specificity (87.11%) with MCC of 0.759 and AUROC of 0.945. The AUROC value towards one also suggested that this model has better prediction ability. Therefore, the SVM-based CKSAAP model was identified as an optimal model for the prediction of PMTs. Finally, the best model was developed as standalone software called PMTPred which is available for download at http://www.bioinfoindia.org/PMTPred/ and https://github.com/ArvindYadav7/PMTPred for research and academic use.

- Chapter 2 of this thesis comprehensively examines the expression level of SMYD2 in various cancer patients available in TCGA and its level of correlation with cancer prognosis. The outcomes suggested that SMYD2 was highly expressed in tumor tissues as compared to their corresponding normal tissues. The analysis of overall survival and disease-free survival of cancer patients proposed a high level of SMYD2 in CESC (*P* = 0.00045) and a lower level of SMYD2 in KIRC (*P* = 6e-06) significantly affected the overall survival of the cancer patients. While a lower level of SMYD2 in BRCA (*P* = 0.046) and KIRC (*P* = 0.013) and a higher level of SMYD2 in COAD (*P* = 0.0061) and HNSC (*P* = 0.008) were significantly correlated with disease-free survival of the cancer patients. A total of 24 mutations were observed in the SMYD2 protein sequence. Gene ontological properties and pathways were found to be significantly linked to the development of cancer. These results will help us to understand how the SMYD2 functions have been associated with the development of tumorigenesis and metastasis. Therefore, SMYD2 would be a probable biomarker and a significant drug target for the prevention of human cancers.
- Chapter 3 of this thesis examined the most deleterious nsSNPs that significantly affect the enzymatic activity of SMYD2. In this study, three significant nsSNPs such as H207D, C209W, and C209R were identified by using a rigorous computational method. The MDS study suggested that these mutations have a greater effect on the SMYD2 protein structure and function. These nsSNPs could be considered as a potential biomarker for SMYD2-associated cancers and might be applied for early diagnosis and personalized medicine for cancer risk. The findings of this study are intended to be used as a starting point to perform the precision-based analysis, therefore experimental validation is needed. Moreover, this computational approach is expected to aid in the prioritization of SNPs, and may also be appropriate for the analysis of new genes that elucidate the fundamental genetic pathways of cancer formation.
- Chapter 5 of this thesis utilized a variety of *in-silico* analyses in order to propose three novel SMYD2-specific inhibitors. Virtual screening was performed for 98071 natural

compounds by using SMYD2 as the target protein. On the basis of binding energy and protein-ligand interaction analysis, we selected three top compounds ZINC08764231, ZINC08490711, and ZINC03844862 as probable inhibitors. These selected inhibitors have good clearance on all parameters according to ADMET analysis and would be used as a nontoxic and efficient drug candidate. MDS study revealed that selected compounds with the SMYD2 structure had a stable binding. Therefore, the outcome of this study proposed three novel SMYD2-specific inhibitors that can be used as targeted therapy in several cancers. Although, the anticipated molecules will essential to be tested experimentally before they may be produced as anti-cancer drugs. The hypothesized compounds could be ideal candidates for cancer prevention and control.

6.3 FUTURE PROSPECT

- We have developed a method using a sequence-based feature. Here, we tested various machine-learning algorithms with multiple features but prediction accuracy was not much better for the used dataset. Therefore, deep learning can be applied to improve the prediction accuracy for the better prediction of PMTs. Additionally, this method can be integrated into the form of a user-friendly web server so that researchers from a non-bioinformatics background can also use it.
- The identified nsSNPs are wanted to be explored further for their clinical translational value. In order to convert these genetic insights into better cancer screening and therapy techniques, the molecular mechanisms of these nsSNPs function must be explored. Therefore, *in vitro* and *in vivo* experiments can be performed for a better understanding of the effect caused by mutations in the regulation of disease and their mechanisms.
- The proposed compounds can be tested by using *in vitro* and *in vivo* experiments to further understand the pharmacological characteristics such as effectiveness, toxicity, and use for cancer patients. Understanding the MDS and interaction analysis of ligand and receptor in association with SMYD2 protein can be a useful link to discovering many more SMYD2-specific drug compounds in the future.

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

CONSENT FORM

FOR THESIS/DISSERTATION/PROJECT REPORT SUBMISSION

Date: 14/12/2022

Type of Document (Tick): [✓ PhD Thesis] MTech/MSc Dissertation BTech Project Report

Name: Arvind kumar Yadav Department: BT&BI Enrolment No 176502

Contact No. 9205629936 E-mail. mbiarvind03@gmail.com

Name of the Supervisor: Dr. Tiratharaj Singh & Dr. Pradeep Kumar Gupta

Title of the Dissertation/Project Report (In Capital letters): MACHINE-LEARNING BASED PREDICTION AND COMPUTATIONAL INVESTIGATIONS ON PROTEIN METHYLTRANSFERASES INVOLVED IN HUMAN MALIGNANCIES

AGREEMENT

I hereby grant to the Jaypee University of Information Technology (JUIT) and its agents the non-exclusive license to archive and make accessible, my project report/dissertation /thesis in all forms of media under the conditions specified below:

Release entire work of my project report/dissertation for 'JUIT' only for following duration and after this time release the work for worldwide access.

Duration	Write 'YES' on your choice
Immediately after the final submission	—
1 Year	yes.
2 Year	—

Ayadav
Signature of the Scholar

14/12/2022 14/12/2022
Signature of the Guide

14/12/2022
Signature and seal of the Dean (A&R)

(For Library Use only)

Form received date: ____________

Full text uploaded on JUIT Repository (date): ________ (Sign of the uploader):________

Full text uploaded on Shodhganga (for thesis only):________ (Name of the uploader):________