

***Diospyros kaki's* Phytochemical Mediated Inhibition of GBM by
targeting PDK-1: A machine learning model.**

A DISSERTATION

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

MASTER OF SCIENCE

IN

BIOTECHNOLOGY

SUBMITTED BY

**NIDA E FALAK
2K21/MSCBIO/61**

Under the supervision of

DR. ASMITA DAS



DEPARTMENT OF BIOTECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi college of Engineering)

Bawana Road, Delhi-110042

MAY,2023

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi college of Engineering)

Bawana Road, Delhi-110042

DECLARATION

I, Nida E Falak, 2K21/MSCBIO/61 of MSc. Biotechnology, hereby declare that the project Dissertation titled ***Diospyros kaki's phytochemical mediated inhibition of GBM by targeting PDK-1: A machine leaning model***, which is submitted by me to the Department of Biotechnology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Science Biotechnology and submitted to the Department of Biotechnology, Delhi Technological University, Delhi, is an authentic record of my own carried out during the period from January-May 2023 under the supervision of **Dr. Asmita Das**.

The matter presented in this report has not been submitted by me for the award for any other degree of this or any other institute/University. The work has been accepted in SCI/SCI expanded / SSCI/Scopus Indexed Journal OR peer reviewed Scopus Index Conference with the following details :

Title of the Paper : Diospyros kaki's phytochemical mediated inhibition of GBM by targeting PDK-1: An in-silico docking and machine learning model

Authors Name: Nida E Falak, Harsh Aahra, Asmita Das

Name of Conference : Second IEEE International Conference on “Smart Technologies and System for Next Generation Computing (ICSTSN 2023)

Conference Date and Venue : 21st-22nd April 2023 IFET College of Engineering, Gangarampalaiyam, Tamil Nadu, India

Registration: Done

Status of Paper : Under proceeding

Date of Paper Communication: 20th December 2022

Date of Paper Acceptance: 23 March 2023

Date of Paper Publication:

Date:

Nida E Falak

DEPARTMENT OF BIOTECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi college of Engineering)

Bawana Road, Delhi-110042

CERTIFICATE

I hereby certify that the Project Dissertation titled *Diospyros kaki's* phytochemical mediated inhibition of GBM by targeting PDK-1: A machine learning model, which is submitted by Nida E Falak, 2K21/MSCBIO/61, Delhi Technological University Delhi, in partial fulfilment of the requirement for the award of the degree of Masters in Science, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere.

Place: Delhi

DR. ASMITA DAS

Supervisor

Date:

ACKNOWLEDGEMENT

I want to express sincere appreciation to everyone who assisted me in completing my thesis. I am eternally grateful for my supervisor, **Dr. Asmita Das**, for their mentorship, generosity, and invaluable insights. This endeavour has benefited greatly from their knowledge and constant support. Additionally, I would like to thank **Professor Pravir Kumar, Head of Department** and all the faculty members from **Department of Biotechnology, Delhi Technological University** for their crucial advice and support during my time at the university. My family and friends' steadfast encouragement, sympathy, and enthusiasm are to be greatly appreciated. It has always given me courage to know that they have faith in me.

Place:

Date:

Nida E Falak

ABSTRACT

Glioblastoma Multiforme (GBM), a grade 4 brain tumour, is resistant to standard treatments and exhibits a 100 percent recurrence rate. By obstructing the primary pathways involved in tumour sustenance and progression, GBM can be completely eliminated. Since high levels of p-AKT (Phosphorylated Protein Kinase B) have been detected in recurring GBM, the AKT/mTOR pathway has received the most attention in GBM therapies. AKT can be activated by PDK-1 (3-phosphoinositide-dependent kinase-1) phosphorylation. Targeting and hampering PDK-1 can prevent such activation. Through in-silico analysis, this study seeks to investigate several phytochemicals that have the potential to target PDK-1. The phytochemical profile of *Diospyros kaki* holds immense significance. Carotenoids, tannins, and phenolic compounds impart *Diospyros kaki* excellent anti-oxidant potential making it a promising candidate for anticancer drug compound extraction. After docking 40 selected phytochemicals against PDK-1 protein, the results revealed that Diospyrin and Neodiospyrin show maximum binding energies (-10.6 and -10.4 kcal/mol respectively) comparable to that of the standard inhibitor 2-(5-{{[2R)-2-amino-3-phenylpropyloxy} pyridine-3-yl}8,9dimethoxybenzo[c][2,7]naphthyridine-4-amine (ID-8I1) compound(-10.1 kcal/mol). Validating the docking results by machine learning, multiple analyses were carried out to find the drug likeness, bioactivity and bioavailability of the selected phytochemicals so as to imply them as a therapeutic agent against GBM.

CONTENTS

S. No	Contents	Page no.
	Declaration	II
	Certificate	III
	Acknowledgement	IV
	Abstract	V
	List of Tables	VIII
	List of Figures	IX
	List of Abbreviations	X
1.	Introduction	1-2
2.	Literature review	3-11
2.1	General	3
2.2	Glioblastoma	3
2.3	Mutations involved	5
2.4	AKT/mTOR pathway	6
2.5	PDK-1	7

2.6	Diospyros kaki	8
2.7	Molecular Docking	8
2.8	Standard/Inhibitor used	10
2.9	Machine Learning	10
2.10	MAACS Fingerprints	11
3.	Materials	13-21
3.1	IMMPAT database	13
3.2	PDB	14
3.3	Autodock	15
3.4	PLIP	16
3.5	BioVia Discovery Studio	17
3.6	ChEMBL database	18
3.7	SwissADME	20
3.8	Molinspiration	20
3.9	Lipinski's rule of five	21
4.	Methodology	22-28
6.	Results	29-39
7.	Conclusion	40

LIST OF TABLES

S.No	Table	Page No.
I	Types of Glioblastomas and the mutations involved and their rate of occurrence.	5
II	Integration of Molecular docking with other computational branches resulting in the improved efficiency of the whole process.	9
III	Molecular docking of phytochemicals against target protein (PDK-1) in-silico	30
IV	Lipinski's Rule of Five Analysis (RO5)	32
V	Bioactivity Scores of Phytochemicals retrieved from Molinspiration.	33
VI	List of MAACS fingerprint keys	35-38

LIST OF FIGURES

S.No	Figure legend	Pg. No.
1.	General Workflow of Methodology	2
2.	Depicts the phytochemical mediated inhibition of PDK-1 that affects the P13/AKT/mTor pathway in GBM cells resulting in decreased tumorigenic characteristics and eventually death of GBM tumour cells.	6
3.	Molecular structures of phytochemicals of <i>Diospyros kaki</i>	29
4.	Molecular docking of ligands and target protein (PDK-1) retrieved from PLIP	31
5.	Bioavailability radars of Neodiospyrin, Diospyrin, Vitexin and Flavylum respectively retrieved using swisADME	31
6.	SwisADME analysis reveals phytochemicals Flavylum crosses BBB whereas compounds like Cyanidin and Plumbagin fail to do so.	32
7.	Confusion matrix of Machine Learning accuracy. Depicting True Positive=46, True Negative= 56, False Positive= 16, False Negative=9.	34
8.	Bar graph Depicting different machine learning model and their accuracy	34
9.	Contribution of each key based on their presence in the drug/compound	39

LIST OF ABBREVIATIONS

1. GBM- Glioblastoma Multiforme
2. PDK-1-3-phosphoinositide-dependent kinase-1
3. GSCs- Glioblastoma stem cells
4. RF- RandomForest
5. RO5- Rule of five
6. NRL- Nuclear receptor ligand
7. ICM- Ion Channel Modulator
8. EI- Enzyme Inhibitor
9. KI- Kinase Inhibitor
10. BBB- Blood Brain Barrier
11. TMZ- Temozolomide
12. MS- Mesenchymal
13. CL- Classical
14. PN- Proneural
15. NE-Neural
16. AC - Astrocyte-like
17. OPC - Oligodendrocytic precursor cell-like
18. NPC- Neuronal progenitor cell-like
19. TCGA- The Cancer Genome Atlas
20. Rb- Retinoblastoma
21. RTK- Receptor Tyrosine Kinases
22. PI3K- Phosphoinositide-3-kinase
23. PIP2- Phosphatidylinositol 4,5-bisphosphate
24. PIP3 - Phosphatidylinositol (3,4,5)-trisphosphate
25. mTORC1- mammalian target of rapamycin complex 1
26. HTS- High Throughput Screening
27. ML- Machine Learning
28. MD- Molecular Dynamics
29. AI- Artificial Intelligence
30. PLIP- Protein-Ligand Interaction Profiler

CHAPTER 1: INTRODUCTION

Glioblastoma owes its high chances of recurrence and resistance to therapy to a fraction of GBM cells that become highly active and begin to display characteristics resembling those of multipotent stem cells. These cells are known as GSCs [1], [2]. Several pathways are involved in imparting tumorigenic potential to GSCs [3]. GBM occurrence and resistant to classical drugs like TMZ has been linked to genetic mutated or genetically altered (overexpressed) EGFR [4] or increased phosphorylated AKT levels [2]. Studies reveal significantly increased phosphorylation of AKT protein in GBM tumor cells, that makes the P13/AKT/mTor pathway a potential blockade target. Upon receiving the signal from the RTKs, PI3K converts PIP2 into PIP3 which is further converted to AKT, the most significant player in the pathway. Being highly susceptible to mutations, it cannot be used a drug target site. PDK-1 is a regulatory enzyme which provides the phosphate group and hence phosphorylates AKT. This phosphorylated AKT in turn activates the mTORC1 pathway and induces tumorigenic properties [4].

Numerous proteins, including PDK-1, PDK-2, or the TORC2 complex of the mTOR, phosphorylate AKT. AKT can be stimulated by PDK-1 phosphorylation alone, but PDK-2 or TORC2 facilitate increased phosphorylation [2]. This work intends to target PDK-1 by phytochemicals derived from *Diospyros kaki* since PDK-1 stimulates oncogenesis and tumour sustenance through AKT [5]. Currently, phytochemicals are widely being implied as anticancer drugs due to their vast availability, simple extraction procedures and minimal toxicity [6]. Persimmons (*Diospyros kaki*) belonging to family Ebenaceae, not only excel in taste but also hold great therapeutic value. Being a rich source of polyphenols and antioxidants, their phytochemical configuration suggest that they can be implied as therapeutic agents against cancer cells [7].

40 phytochemicals derived from *Diospyros kaki* were assessed in an in-silico analysis against the PDK-1 protein. The molecular docking between these phytochemicals and PDK-1 was done employing Autodock Vina [6]. Utilizing compound 8I1 as a standard, the phytochemicals with binding affinities comparable to 8i1 were selected. The in-silico screening was performed in several steps, commencing with molecular docking to anticipate the interactions between phytochemicals and PDK-1, and then Machine Learning to verify and validate the docking results. The drug likeness of the selected

phytochemicals has been determined employing various parameters like Lipinski's rule of five, Bioavailability radars and Bioactivity scores

Further we utilize MACCS descriptor analysis, in conjunction with machine learning algorithms, has further enhanced these efforts. By leveraging the power of MACCS descriptors, we have successfully decoded the common chemical functional groups and structural features shared among active drugs, allowing us to make informed decisions regarding compound modification and library design.[8]–[10]

The knowledge obtained from MACCS descriptor analysis can be harnessed to modify phytochemical compounds or design chemical libraries with improved pharmacological properties. Additionally, this information facilitates comparison and similarity searches within drug databases, enabling efficient data mining and identification of potential lead compounds. The integration of MACCS descriptor analysis in drug discovery processes. [11], [12]empowers researchers to make informed decisions regarding compound modification and library design, ultimately accelerating the development of novel and efficacious therapeutic agents.

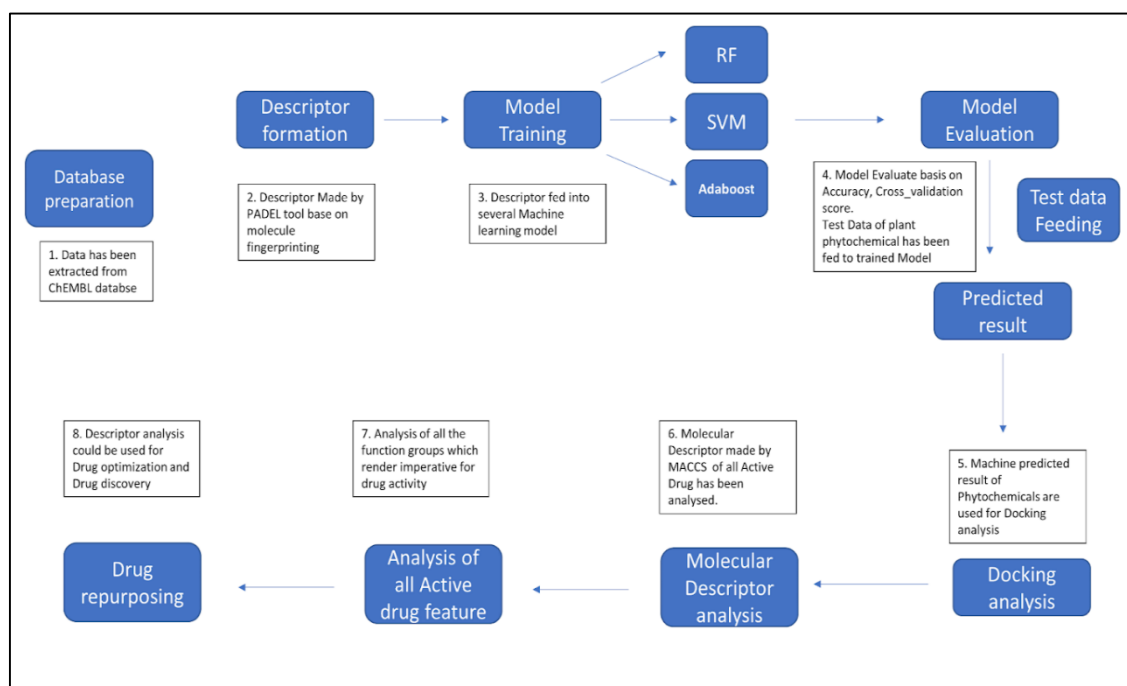


Figure 1: General Workflow of Methodology.

CHAPTER 2: LITERATURE REVIEW

2.1 General

The incidence and mortality of cancer are rising substantially. While conventional approaches to treatment, such as chemotherapy and radiation therapy, can be somewhat beneficial, they frequently lack specificity and can have serious adverse effects. Due to the varied characteristics and complexities of tumours, patients' current treatment choices are inadequate making targeted therapy or personalized medicine necessary. [13] Targeted therapies directly aim at obstructing the molecular pathways that promote tumour sustenance and proliferation. By directly targeting specific cells, the harm to healthy tissues is minimized, in contrast to conventional approaches that might affect both healthy and diseased cells. Advanced genome sequencing techniques have made it easier to pinpoint the molecular as well as genomic aberrations that contribute in cancer emergence. Oncologists can understand more about the unique mutations that can cause cancer in different patients by tumour profiling and provide them targeted therapies accordingly. [13], [14] To assist patients with treatment choices, personalized medicine incorporates biomarker use such as genetic mutations or protein level expression analysis that can be used to anticipate the reaction of a patient to a therapy or identify the patients expected to benefit from a certain drug.

2.2 Glioblastoma

GBM is a grade IV tumour known for its high recurrence rates and resistance to standard treatments. Glioblastoma owes its high chances of recurrence and resistance to therapy to a fraction of GBM cells that become highly active and begin to display characteristics resembling those of multipotent stem cells. These cells are referred to as GSCs [1] Studies reveal that the GSCs are rich in transcription factors like SALL2, SOX2, OLIG2 and POU3F2 and markers like Nestin and CD133 that confers to their neoplastic behaviour. A distinctive characteristic of neural progenitor cells, nestin is an intermediate filament protein linked to neural stem cells. Cell surface glycoprotein CD133, commonly known as Prominin-1, has been linked to stem cell characteristics and tumour development. [15] [16]

Aggressive nature of GBM, existence of self-renewing glioma stem cells, a small concentration of drug reaching the target site due to blood brain barrier, only a few drugs

crossing the BBB, poor prognosis, 100 percent reoccurrence rate are some of the major hurdles in glioblastoma treatment. Expression profiling analysis have revealed four major categories based on the bulk tumor transcription profile: MS, CL, PN and NE [17]. CL and MES tumour types are rich in AC and MES like states while OPC and NPC cellular states are abundant in PN and NE. [18]

Developing targeted therapeutics that can successfully remove GSCs and combat drug resistance requires a thorough understanding of the molecular categories and cellular configurations found in GBM. The heterogeneity of GBM, which includes the existence of GSCs and other molecular subtypes, presents considerable difficulties in creating efficient treatment plans. For enhanced survival for GBM patients, ongoing research is concentrating on finding GSC-specific vulnerabilities and creating medicines that can target and eradicate these cells with precision.[16]

2.3 Mutations involved

Table I: Types of Glioblastomas and the mutations involved and their rate of occurrence.

Tumour type	Characteristics	Mutations involved	Occurrence	References
Primary Tumours (90%)	<ul style="list-style-type: none"> ❖ Extremely Aggressive tumours. ❖ Tend to affect the elderly with greater frequency. ❖ Poor prognosis as they develop without any prior symptoms. ❖ Found in frontal and temporal lobes. 	P13KCA Loss of RB1 gene IDH1/2 mutation PDGFR amplification GLI1 TERT promoter NF1 deletion/mutation MDM2 PTEN mutation/deletion TP53 mutation CDK2A/B deletion MGMT promoter methylation EGFR amplification LOH at 10q	1% 2% 5% 7% 5-22% 10% 11% 7-12% 24-30% 28-31% 31% 36% 22-40% 65%	[19] [20][21], [22] [20]
Secondary Tumours (<10%)	<ul style="list-style-type: none"> ❖ Less prevalent. ❖ Arise from LGGs (Low Grade Glioblastoma) or AAs (Anaplastic Astrocytoma). ❖ Better prognosis. ❖ Most frequently found in the frontal lobe. 	EGFR amplification PDGFR amplification 1p/19q codeletion LOH 19q IDH1/2 mutation TP53 mutation MGMT promoter methylation LOH at 22q	5-7% 7% 15-20% 40-50% 45-50% 65% 75% 70-80%	[20][19] [23] [24]

2.4 AKT/mTOR pathway

The most impacted cellular pathways in glioblastoma, according to TCGA, are RTK/P13K, p53, and Rb pathways [25]. However, some studies also reveal WNT/ β -catenin pathway as one of the significant pathways in GBM, shedding light on a close association involving WNT and P13K/AKT/mTOR pathway. [26]

Lost PTEN function and amplified RTKs like EGFR are two key characteristics of GBM aetiology. AKT/mTOR pathway includes both these attributes as negative regulation of AKT/mTOR pathway is achieved by PTEN, that controls the PIP3 levels and hence regulates the pathway. Loss of PTEN function leaves no surveillance for the pathway and hence it is upregulated. Whereas increased number of receptors (RTKs) and studies showing continuously enhanced EGFR in GBM patients also leads to upregulated AKT/mTOR pathway [22]

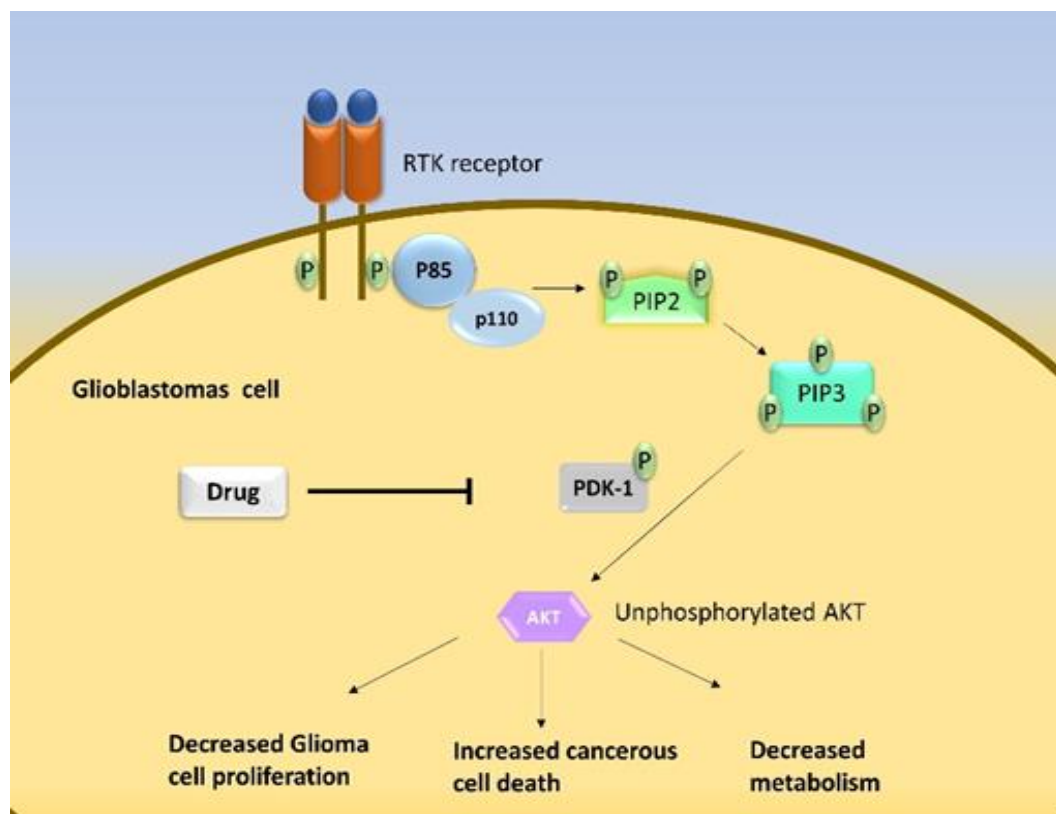


Figure 2: Depicts the phytochemical mediated inhibition of PDK-1 that affects the P13/AKT/mTor pathway in GBM cells resulting in decreased tumorigenic characteristics and eventually death of GBM tumour cells.

The AKT/mTOR pathway is thus, a significant player in imparting tumorigenic properties to the cells like increased glioma cell proliferation, increased cellular metabolism, cellular sustenance and resistance to standard treatment (TMZ). Studies reveal significantly increased phosphorylation of AKT protein in GBM tumor cells, that makes the P13/AKT/mTor pathway a potential blockade target. Upon receiving the signal from the RTKs which are highly amplified in GBM, PI3K converts PIP2 into PIP3 which is further converted to AKT, the most significant player in the pathway. Being highly susceptible to mutations, it cannot be used as a drug target site directly. PDK-1 is a regulatory enzyme which provides the phosphate group and hence phosphorylates AKT. This phosphorylated AKT in turn activates the mTORC1 pathway and induces tumorigenic properties [27]

2.5 PDK-1

PDK-1 is a serine-threonine kinase which acts as a regulatory enzyme and phosphorylates AKT. Studies reveal that PDK-1 and NFkB aids the GBM cells to escape the TMZ drug and develop a resistance against it. GBM cells exhibit the Warburg effect, which is a phenomenon in which cells preferentially produce lactate rather than oxidative phosphorylation regardless of whether there is sufficient oxygen present. Since PDK-1 is necessary for the conversion of pyruvate to lactate, it serves as a checkpoint for regulating the oxidation of glucose [28]. Reversal of the Warburg effect and decreased phosphorylation of the AKT protein, which in turn reduces the tumorigenic traits and tumour resistance to the standard TMZ, are two significant benefits that can be attained by inhibiting PDK-1. Hence, in this study PDK-1 of the AKT/mTOR pathway has been chosen as a possible pharmacological target.

There has been an increase in the number of unwanted effects and toxic reactions brought about by synthetic medications in cancer treatment that has shifted the focus of researchers all over the world on using natural products or phytochemicals. A great demand for investigating undiscovered natural resources to develop drugs against certain diseases has emerged to treat patients who have no other possible medications available. Phytochemicals owe their potential to be used as drugs against cancer or as adjuvant therapy to their non-toxicity, their easy availability and extraction, their efficacy, safety as well as mechanisms of action.

2.6 *Diospyros kaki*

Persimmons (*Diospyros kaki*) belonging to family Ebenaceae, a famous fruit in China and Japan not only excel in taste but also hold great therapeutic value. It has extensively been used in the ancient Chinese medications owing it to its positive health impacts. It has scientifically been proven to act against hypertension, haemorrhages, atherosclerosis, diabetes, throat irritation, insomnia and body temperature maintenance.[7], [29]. The phytochemical configuration of *Diospyros kaki* is rich in polyphenols, carotenoids flavonoids as well as antioxidants which imparts them anti-diabetic, anti-hypercholesterolaemic and even anti-tumour characteristics. [7] [30] Their phytochemical makeup and abundance of vitamins, polyphenols, and antioxidants imply that they can be used as therapeutic agents against cancer cells. Consequently, this study aims to investigate the role of *Diospyros kaki*'s phytochemicals as a GBM inhibitor using in-silico analysis and machine learning.

2.7 Molecular Docking

A molecule's structure and position (collectively called as its pose) when they engage in interaction with a target's binding site, can be examined employing an approach called molecular docking. Molecular docking has revolutionized drug designing and optimizing strategies as simulated testing of compounds lowers overall costs and accelerates the discovery process. Molecular docking forms the basic premise of HTS the de facto approach to screen out the compounds throughout the hit-to-lead optimization process. The structure and orientation-based search algorithms and scoring function are the two fundamental components of the docking process. [31]

The most prevalent softwares utilised for molecular docking include Molegro Virtual Docker, FlexX, DockThor, GOLD, AutoDock and AutoDock Vina. This study employed Autodock Vina and Autodock Perl as they offer rapid as well as precise analysis. The 3D structures of ligands to be screened can be retrieved from databases like PubChem, ZINC and PDB. This study used IMMPAT database for retrieving the phytochemical structures and PDB for PDK-1 protein structure retrieval. [31], [32]

Table II: Integration of Molecular docking with other computational branches resulting in the improved efficiency of the whole process.

	Integrated fields	Pre-docking Screening	Post-docking Screening	References
Molecular Docking	AI and Statistical Approaches	Retrieval of protein structures for screening Optimizing the score	Pose improvement	[31], [32]
	Molecular Dynamics (MD)	Selection of conformations	Phytochemical-target interaction evaluation. Orientation and conformation refinement	[33]
	Ligand-based methods	Retrieval of protein structures for screening	Orientation and conformation selection and scoring	[31], [32]
	Binding Free Energy methods		Orientation and conformation refinement	[32]

Molecular docking can broaden horizons in medical research by identifying potential interactions between the molecular targets and ligands, novel targets for ligands already in use, potential side effects that the drug could cause, repositioning and repurposing a drug that has already received FDA approval. The efficiency of the entire process can be increased by integrating it with several other branches of computational biology including ML, AI, MD, pharmacokinetic and bioavailability analysis.

2.8 Standard/Inhibitor used

To serve as a reference for the molecular docking scoring analysis, a specific inhibitor is retrieved from the PDB database. The comparative analysis of binding affinities of the target-inhibitor complex and the target-ligand complex provides an insight about the extent to which a particular ligand is interacting with the target protein. This study has selected the compound 8,9-dimethoxy-5-(2-aminoalkoxy-pyridin-3-yl)-benzo[c][2,7]naphthyridin-4-ylamine as the standard inhibitor. [34]

2.9 Machine Learning

Python is an extensively employed programming language in the machine learning industry because of its adaptability, sizeable library, and user-friendliness. Using current data from the ChEMBL database can be quite beneficial when it comes to evaluating the outcomes of machine learning models in the context of ligand-drug interactions. ChEMBL is a comprehensive tool that gives users access to a huge library of bioactive compounds, their characteristics, and related biological functions. This database offers data on numerous medications and how they interact with various proteins. The information includes a variety of characteristics, such as ChEMBL IDs, molecular weights, and SMILES sequences. (a compact representation of molecular structure), and IC50 values (a measure of drug potency). The evaluation of in-silico docking findings can be improved by using the ChEMBL data to train machine learning models. With the aid of this dataset, researchers can train their models to determine whether or not the interactions between ligands and target proteins are accurately reflected in their docking results. [35]

The procedure normally entails pulling pertinent information from the ChEMBL data, including molecular weights, SMILES sequences, and Molecule ChEMBL id. The size and mass of the ligands are revealed by their molecular weights, which can reveal whether

or not they have the ability to interact with proteins. SMILES sequences, in contrast, encode the ligands' chemical structures, enabling a more thorough examination of their possible binding abilities. The collected features may then be paired with the associated IC50 values, which function as the ground truth or labels when training a machine learning model. The IC50 values show the amount of a medicine needed to completely block a certain biological process. This knowledge can help the model learn to link certain characteristics to desirable binding affinities throughout the training phase. It is crucial to divide the ChEMBL dataset into training and testing sets that guarantee the machine learning model's dependability and generalizability. The model is trained by utilizing the training set, while its performance on test data is assessed using the testing set. This procedure contributes in evaluating the model's propensity to generalise and deliver precise forecasts for fresh docking outcomes. [36]

The machine learning model may be employed in binding affinity anticipation or IC50 values for brand-new ligand-protein interactions after it has been trained. The accuracy of the docking results may be evaluated by comparing these predictions to experimental data or known IC50 values from the ChEMBL database. This validation stage offers insightful information on the dependability and efficiency of the in-silico docking method.

2.10 MACCS fingerprints

In the field of cheminformatics, one popular technique for producing molecular descriptors or fingerprints is the Molecular ACCESS System (MACCS). The presence or absence of preset substructures or structural properties in a molecule is represented by MACCS keys, which are binary fingerprints. The 166 keys or bits that make up a MACCS fingerprint each stand for a distinct chemical characteristic. Aromatic rings, functional groups, atom kinds, and other structural traits are among the qualities represented in MACCS keys. If the matching characteristic is present in the molecule, each bit is given a value of 1; otherwise, it is given a value of 0 (zero). [37]

Chemical similarity searches, virtual screening, and drug development are just a few of the uses for MACCS fingerprints. Researchers can find related compounds or sift huge chemical datasets to prioritise molecules with desirable structural properties by comparing the MACCS fingerprints of various molecules. A chemical structure is

examined and transformed into a binary fingerprint using preset procedures for each MACCS key to produce MACCS fingerprints. With the use of similarity metrics like the Tanimoto coefficient or Euclidean distance, the generated fingerprint effectively compares to other fingerprints by providing a condensed representation of the structural characteristics of the molecule. Overall, MACCS descriptors offer a method for comparing and capturing molecular structure data in a condensed binary format, which makes computer analysis and prediction jobs in cheminformatics and drug development easier. [38]

Purpose of MACCS analysis

Drug research and development are greatly aided by the examination of effective medications using MACCS fingerprints because it sheds light on the essential characteristics and structural characteristics of effective therapeutic agents. In order to repurpose currently available pharmaceuticals or create new ones with increased efficiency, researchers can find common functional groups, rings, or metals that are commonly present by analysing the fingerprints of known effective drugs.

The capacity of MACCS fingerprints to represent chemical structures in a small binary representation is one of its main features. Large chemical datasets may be stored, compared, and retrieved effectively using this format. Researchers can find recurrent substructures or traits that are connected to therapeutic efficacy by comparing the MACCS fingerprints of known effective medications. These characteristics which might substantially influence the drug's mechanism of action or target interaction, may include certain functional groups, aromatic rings, or even the presence of certain metals.

Researchers learn more about the structural and functional traits that contribute to a drug's efficacy through MACCS analysis for instance, if a certain functional group frequently appears in effective medications, this may indicate that this group is crucial for target binding or particular biological interactions. Using this knowledge, medicinal chemists can concentrate their efforts on adding or changing these essential functional groups to current medications or when developing new drug candidates. [39]

Additionally, MACCS fingerprints serve as a foundation for virtual screening and similarity searches. To find molecules with comparable structural characteristics, researchers can compare the fingerprints of possible drug candidates with those of well-

known effective medications. This strategy aids in the identification of novel compounds that have the desired properties of effective medications or the repurposing of current pharmaceuticals for new applications.

The existence of particular metal ions may be inferred from MACCS fingerprints in addition to functional groups and rings. Metal-containing medications have demonstrated substantial therapeutic benefit, such as platinum-based chemotherapeutic treatments or metalloenzyme inhibitors. Researchers can use the predominance of particular metal characteristics in effective pharmaceuticals to guide the creation of novel metal-based treatments or aid in the optimisation of currently available metal-containing drugs by analysing MACCS fingerprints.[39]

Overall, an organised and effective method for identifying the structural and functional characteristics of effective medications is the MACCS analysis. Researchers can repurpose already available medications, create new ones with better attributes, or concentrate on particular functional groups, rings, or metals that are crucial for pharmacological action by discovering the recurrent characteristics linked with their efficacy. By rationally designing and optimising medication candidates with the use of this information, numerous diseases can eventually be treated more successfully.

CHAPTER 3: MATERIALS

3.1 IMPPAT database

An extensive database that focuses on the medicinal plants found in India is called Indian Medicinal Plants, Phytochemistry and Therapeutics. Researchers, scientists, medical professionals, and anybody else interested in learning about or using the therapeutic characteristics of Indian plants might benefit greatly from IMPPAT. The database includes comprehensive details on a variety of plant species, such as their botanical names, common names, traditional applications, chemical components, pharmacological activity, and pertinent scientific studies. IMPPAT has a significant part in encouraging the discovery and use of Indian medicinal plants in light of the rising interest in natural goods and conventional treatment. It makes it easier to conduct studies on the phytochemical analysis of plant extracts, the discovery of bioactive substances, and the creation of plant-based medicines.[40]

IMPPAT contributes to the in-silico drug development process by offering a comprehensive database of information about the phytochemical makeup of Indian medicinal plants. Information about the chemical components found in various plant species is included in this material. Researchers can anticipate the probable biological functions and characteristics of these chemicals, such as their interactions with certain proteins or enzymes, by computer analysis. Additionally, the database facilitates the computational modelling-based investigation of structure-activity correlations (SAR). Researchers can create prediction models to anticipate the biological activities of comparable substances by analysing the chemical structures of bioactive chemicals and their related biological activities. This makes it possible to find interesting chemicals and synthesise or modify them to improve their qualities and increase their efficacy and decrease their toxicity. [40]

Molecular docking studies and virtual chemical libraries are also included in IMPPAT's in silico capabilities. The database may be used as a springboard for creating digital collections of plant-derived chemical derivatives. These libraries can be utilised for virtual docking experiments where computer algorithms anticipate the binding affinity of a target protein and a chemical complex or for screening against certain therapeutic targets. The discovery of possible medication candidates and the improvement of their binding interactions are both facilitated by such simulations.

3.2 PDB

The protein data bank is an extensive worldwide collection of structural data of molecules such as nucleic acid and protein molecules. It was created in 1971 but initially there was no sufficient data available until 1980 but optimized technologies like crystallographic process and nuclear magnetic resonance (NMR), Cryoelectron microscopy, and theoretical modelling have led to a surge in the number of deposited structures. PDB has an imperative role in structure genomics as well as structural biology and bioinformatics, there are many such as CATH and SCOP which use protein structure deposit in the PDB. PDB contain a large collection of 3D structure of nucleic acid and protein as it stores more than 180000 structure of macromolecules.[41]

It is an invaluable resource for student, researcher and educator worldwide as it offers a pervasive and diverse range of protein structure and enabling researcher to study an

interaction, function, and their folding. The data of structure has been determined and obtain from various techniques such as electron microscopy, magnetic resonance spectroscopy, crystallography etc. Data has been undergone through various rigorous validation to ensure its reliability and precision.

PDB not only contain Protein and nucleic data but also encompasses ligand and small molecules such that play a crucial role in biological process and activity. PDB is very helpful in studying protein-protein interaction, investigate complex molecular interaction and drug discovery, enzyme engineering, and fundamental studies of biological mechanism. Additionally, the PDB provides a variety of resources and tools to aid with data analysis and visualisation. PDB provides an interactive molecular viewer, search functionalities and advance query options. These tools help scientist and researchers to study a structure nuance of nucleic acid and protein molecules and compare molecule structure and gain the knowledge of their functional properties.[41]

3.3 Autodock vina

Protein-ligand docking analysis is used to determine a ligand's affinity to a protein molecule in order to forecast the score and manner of binding. Since its first introduction in 1970, Autodock Vina has stimulated vigorous research activities, assisted in the development of new drugs, and improved existing ones.

A methodology used for exploring and sampling both the positional and structural space is docking programmes, which are based on two key elements: a scoring function and an exploration approach. The free energy of a preprogramed system can be evaluated by the scoring function. A popular piece of software called AutoDock Vina is essential to computational drug development and virtual screening. [42]

The speed and precision with which AutoDock Vina explores the enormous conformational and positional space is one of its key features. It makes use of a hybrid search algorithm that incorporates methods for both global and local optimisation, enabling a thorough investigation of ligand binding poses. AutoDock Vina greatly increases the probability of locating energetically advantageous binding configurations by automatically sampling a variety of ligand orientations and conformations.[33], [43]

Researchers explored a variety of stochastic global optimisation strategies when developing Vina. These techniques included simulated annealing, particle swarm optimisation, genetic algorithms, and others. To speed up the process, additional local optimisation operations and particular optimisation approaches were used. Following substantial research, the Iterated Local Search global optimizer was selected as Vina's strategy. [33], [42]

The programme uses a scoring method to determine how well the ligand and receptor bind using parameters and energy terms that are empirically calculated. The binding energy is calculated using a scoring function that considers intermolecular forces such as hydrogen bonds, van der Waals contacts, and electrostatics. AutoDock Vina can efficiently rank various ligand poses according to their estimated binding affinities by taking into account these criteria.

Additionally, AutoDock Vina provides an intuitive user interface and a large range of adjustable options, enabling researchers to customise the docking procedure to their own requirements. It is compatible with a variety of ligand and receptor configurations because it allows for the integration of different molecular formats. Researchers may discover probable binding sites and comprehend the interactions between ligands and proteins with the help of the software's visualisation tools, which are also provided for analysing and interpreting the docking data.[44]

3.4 PLIP

A potent computational method called PLIP is used to research as well as examine how proteins interact with their small molecule ligands. Understanding these interactions is essential for drug development, molecular biology, and bioinformatics research due to the growth of structural biology and the rising availability of protein-ligand complex structures. Characterising the complex interactions between proteins and ligands is made easier with the help of PLIP, which is both comprehensive and user-friendly. It uses a variety of algorithms and techniques to analyse the binding affinities, locations, and behaviours of ligands to proteins. The three-dimensional nature of these interactions is crucial information gained via the use of a variety of structural analysis tools, and PLIP helps researchers clarify their functional and mechanistic consequences.[45]–[47]

The capacity of PLIP to automatically recognise and categorise ligand-binding sites on protein structures is one of its key characteristics. It examines the surface of the protein, locates putative binding sites, and categorises them into certain ligand classes using advanced algorithms. This capability speeds up the drug development process by enabling researchers to quickly identify possible binding sites and rank them for future examination.

In order to forecast the binding poses of ligands inside protein structures, PLIP also uses molecular docking methods. These docking simulations give important insights into the binding affinities and probable interaction processes by computationally estimating the most advantageous spatial arrangement between the protein and ligand. This knowledge facilitates development and designing of new ligands and the enhancement of current therapeutic candidates[46]

Additionally, PLIP has a rich visualisation feature set that enables scientists to investigate protein-ligand interactions in a very user-friendly way. It creates dynamic 3D protein structure visualisations, emphasising the crucial residues taking part in ligand binding and giving a thorough understanding of the binding modalities. These visual representations make it easier to understand intricate relationships and support the dissemination of research results to a wider scientific community.

A PDB file may be loaded into PLIP, or any structure from the RCSB PDB service can be processed using a four-letter PDB Id or a free text search in the ligand-protein complex. Additionally, the In-silico docking result's output file might be sent to PLIP to examine their covalent and non-covalent interaction.

The ligand-protein complex is typically analysed in the PLIP output. For each ligand-protein interaction, PLIP offers a 2D and 3D interaction graphic as well as a table listing the protein amino acids involved in covalent or hydrogen interactions. You may obtain the outcome file from PLIP in PNG and PyMOL formats as well. Clicking on the overview of the diagram will take you to the specifics of the interaction pattern. [46]

3.5 BioVia Discovery studio

A revolutionary tool called Visual Discovery Studio makes it easier to explore and combine art and science. It serves as a collaborative environment where designers,

scientists, and artists may work together to graphically convey intricate ideas, facts, and thoughts. The studio uses a variety of techniques and technology, including digital design software and conventional art materials, to produce aesthetically attractive and factually correct visualisations.[48]

A potent tool for visualising protein-ligand interactions is visual discovery studio. The ligands and protein are prepared for docking analysis, and the results are visualised. Proteins are prepared by adding a polar hydrogen group, eliminating water from protein molecules, and cutting unneeded ligand or peptide chains. Visual discovery studio gives us a coherent display of binding intricacy of protein and ligand it also provides a 2D diagram of interaction with differentiating the type of bond ligand and protein molecules forming.

3.6 ChEMBL Database

In the realm of medicinal chemistry and drug development, ChEMBL is a useful and commonly used database. It offers a thorough compilation of data on bioactive compounds, their targets, and the biological activities that go along with them. By enabling the research of chemical compounds and their interactions with biological systems, this resource has significantly sped up the medication development process.

ChEMBL is a sizable, publicly accessible drug database that aims to collect information from the pharmaceutical and medical industries as well as from the process of researching and developing medicines. ChEMBL has a sizable collection of drugs and ligands that have been examined on a wide range of proteins and biomolecules. It stores the information about the biological activity of small molecules/drug and data from various medicinal chemical journals. Bioactivity data are share with other database including BindingDB and PubChem Bioassay to allow the scholar to access a comprehensive information[49]

It extracts crucial activity information from research articles published in several journals, including Bioorganic Medicinal Chemistry Letters, Journal of Medicinal Chemistry and Journal of Natural Products. Although these selected journals do not cover every possibility, they have been carefully picked to guarantee the effective use of resources while obtaining a large amount of trustworthy data. The database abstracts for each article include information on the studied substances, tests run, and any relevant target data.

Researchers have access to a wealth of information in ChEMBL, including information on compound structures, binding affinities, and pharmacological profiles. The database contains information from various sources, such as scientific literature, patents, and public databases. With its extensive coverage and continuous updates, ChEMBL offers researchers a robust platform for identifying potential drug targets, investigating structure-activity relationships, and designing new compounds. [49]

The user-friendly interface of ChEMBL, which makes it easy for researchers to search and obtain data, is one of its primary advantages. Users may run sophisticated searches, filter results using certain standards, and get comprehensive annotations for specific substances and targets. Researchers can get insights from the large quantity of information accessible thanks to the database's capabilities for data visualisation and analysis. ChEMBL has made a substantial contribution to the progress of medication research and discovery. It has become a useful resource for scholars throughout the world by combining data from many sources. The database encourages openness and cooperation by granting free access to its contents, enabling researchers to take use of current knowledge and expand on earlier work.[50]

Data access form ChEMBL

The straightforward ChEMBL interface makes it simple to get data. The interface has a search feature that allows users to enter a keyword, protein name, ChEMBL target identifier, or UniPort accession of a target of interest for which a ligand has to be discovered.

Users may quickly retrieve the accompanying bioactivity data using a drop-down menu once they have gotten a target or several targets of interest from the ChEMBL database. They may browse all the available data using this user-friendly functionality, or they can add filters to choose only certain activity kinds. Users can decide, for instance, to only include IC₅₀ and K_i values less than a particular concentration threshold or to concentrate on particular ADMET endpoints. [50]

Each tested drug is fully described in the ensuing bioactivity table, which also includes details on the particular salt form that was employed in the experiment. Additionally, it contains information on the test, a description of the target (including the organism), specifics about the detected activity type, its associated value, and units. Notably, the

table has a link that leads directly to the article from which the data were taken, guaranteeing accessibility and transparency to the original source. To simplify further study and investigation, researchers may quickly export the data from this view as a spreadsheet or text file. By utilizing this feature, users may explore further into the data, carry out their own research, and draw important conclusions from the acquired bioactivity data. [51], [52]

3.7 Swiss ADME

In order to chemically synthesize, develop, test and optimize a drug, one has to access various parameters like biological activity, toxicity and the concentration etc. Pharmacokinetic assessment at the beginning of the discovery phase significantly lowers the likelihood of clinical-phase ADME (Absorption, Distribution, Metabolism, Excretion) -related failures. Swiss ADME is a renowned web tool not only for its reliability and robustness but also for its simplified result analysis that enables effective incorporation to drug discovery via molecular design.

The tool provides a variety of input options like molecular structure and canonical smiles, analysis of a variety of molecules, the ability to save and share results and user-friendly interactive graphs including the boiled egg and the bioavailability radar. The boiled egg aids in evaluating the gastrointestinal absorption and BBB penetration. The white region or the albumin region shows the compounds that are most likely to be passively absorbed by the GIT whereas the yellow yolk region depicts the compounds that possess high BBB permeability. The bioavailability radar on the other hand possesses a pink area which depicts the optimal range of properties like flexibility, saturation, solubility, polarity, molecular weight and lipophilicity. A compound's radar should be limited to the pink region for it to be taken into account as a good drug candidate. [53] By entering the canonical SMILES, the information about the BBB permeability through Boiled-Egg and the Bioavailability radars of potential drug candidates can be obtained.

3.8 Molinspiration

A software platform termed Molinspiration provides a variety of computational tools intended to assist in molecular analysis and alteration. These tools aids in bioactivity anticipation, data visualization and virtual screening. Most important analysis carried out

by Molinspiration in drug development and optimization is the bioactivity score anticipation for drug targets (GPCR, KI, ICM and NM scores).

Molinspiration takes input files in the SMILES or SDfile formats to carry out these analyses. The molecular structure of a substance is represented via SMILES, a string-based syntax, and SDfile (Structure-Data File), a file format that is frequently used to store and transmit chemical structures and related data. Molinspiration implements its computational techniques on the provided input file to calculate the bioactivity scores for the chosen therapeutic targets. These rating systems offer information about the molecule's potential activity or affinity towards the target, assisting in the evaluation of its potential as a therapeutic candidate. The drug development process can be expedited by using this feature to prioritise molecules for additional experimental testing.

3.9 Lipinski's rule of five

The rule of five aids in accessing a compound's drug likeness or efficacy based on two extremely significant factors: a drug's permeability and oral bioavailability. The Lipinski's requirements base themselves on the findings of various studies that prove that easily absorbing orally administered drug attributes fall within these constraints: Mass<500 Da, H-bond donors<5, Hbond acceptors<10, LogP value<5.

The compounds that deviate from these ranges possess a greater probability to show undesirable pharmacokinetic attributes like rapid metabolism, less availability, less permeability and fail to cross the cell membranes. Hence, the rule of five is a gold standard strategy in drug development and optimization techniques that screen out and rank compounds according to their tendency to be orally active.[54]

CHAPTER 4: METHODOLOGY

A. Selection and preparation of ligands/Retrieving phytochemicals and preparation of drug library

Forty bioactive compounds from *Diospyros kaki* were collected from IMPPAT database in pdb format. [40] Using Open Babel GUI, pdbqt formats of the files were obtained.

B. Protein Preparation

The structure of target protein PDK-1 was retrieved from PDB. Along with the target protein structure, the PDK-1 inhibitor's (8I1) structure [34] was also retrieved to be used as a reference.[41]

Using Autodock Vina 1.7.5, Water molecules were removed as they generally do not participate in the binding process. So, in order to simplify the computational calculations and get rid of any potential obstructions and pose distortion in the binding pocket. Polar hydrogens along with Kollman's charges were added as the pdb files lack hydrogens, so in order to attain accurate optimization and calculations, charges and hydrogens were added and pdbqt format of the protein was retrieved.

Open Babel GUI was used to convert inhibitor 8i1 from pdb format to pdbqt format.

C. Molecular docking

- **Protein-ligand docking using Auto dock Vina:** Autodock Vina 1.5.7 employed a computational docking strategy by setting the x, y, and z centre dimensions to -35.04, -26.046, and -1.741, respectively, and the x, y, and z sizes to 58, 48, and 74, respectively, docking analysis was performed. Number of modes and energy ranges were set to 10 and 4 respectively. All phytochemicals are docked against the target protein by using Auto dock Vina and Perl. The standard inhibitor 8i1 was also docked against target protein. [44]
- **Docking/Interaction analysis:** The downloaded outputs from auto dock were analyzed via PLIP and Discovery studio to identify by which amino acid the

ligand binds to the protein. The binding energies of ligands and standard were compared and phytochemicals with high binding energies were selected. [46], [48]

D. Pharmacokinetic and Drug Likeness Screening of selected Phytochemicals: Only phytochemicals that can get past the initial round of binding energy range screening are subjected to further analysis based on their drug likeness and pharmacokinetics.

- **Lipinski's RO5 analysis:** The rule of five assesses drug likeness, or the likelihood that a molecule will be active when taken orally. The selected phytochemicals were subjected to Lipinski's RO5 to analyze their oral activity. [54]
- **In-silico bioavailability analysis:** To comprehend the pharmacokinetics of a drug, it's vital to understand its absorption, distribution, metabolism, and excretion. In order to evaluate the bioavailability radars of the phytochemicals, SwissADME and admetSAR [53] were employed. By providing canonical SMILES of phytochemicals as input, these values can be determined. [53]
- **Bioactivity score:** To determine the druggability characteristics of ligands like NRL, PI, and EI, GPCR, ICM and KI, bioactivity score is required. The scores can be predicted by providing canonical SMILES of phytochemicals as input to Molinspiration.

E. Machine learning Validation

- 1. Data extraction:** Data extraction is the initial stage in the machine learning process. In this instance, the `chembl_webresource_client` Python package was used to extract the data from the ChEMBL database. The protein PDK-1 is the subject of the data extraction, and all ligands that have been thoroughly investigated and analysed in connection to this protein have been developed. The collected dataset contains 1150 ligands. The dataset is then filtered based on the IC50 values when the data extraction is finished. A drug's potency is gauged by its IC50 value, which shows the dose needed to 50% block a certain biological process. The data may be filtered based on IC50 values, which enables the identification of ligands that interact with the protein PDK-1 in a substantial way. Filtering the dataset serves to guarantee that the machine learning model concentrates on ligands that are more likely to be efficient and pertinent in relation to the target protein. Researchers can prioritise ligands that have shown greater

inhibitory capability against the target protein by reducing the dataset based on IC50 values. [50] Depending on the needs and aims of the research, different IC50 filtering criteria may be employed to filter the data. Depending on the desired amount of protein interaction, researchers may decide to include ligands with IC50 values below a particular threshold, suggesting increased potency, or they may concentrate on a specific range of IC50 values. It is significant to remember that the data extraction and filtering processes are essential to guaranteeing the dataset's quality and applicability for subsequent machine learning activities. Researchers can produce a more concentrated and curated dataset that is more suitable for training and validating machine learning models by removing information particularly linked to the protein of interest and filtering based on IC50 values.

2. Data Classification: After extracting the data from the ChEMBL database, the next stage comprises classifying the ligands based on their IC50 values. In this case, a threshold of 100 is used to distinguish between active and inactive medications. If a ligand's IC50 value is less than 100, it is classified as active; otherwise, it is classified as inactive. This categorization process aids in determining which medications are more effective at blocking the target protein. Researchers can learn more about the potential efficacy of these medications by classifying the ligands as either active or inert based on their IC50 values. It's also critical to evaluate any data bias present in the dataset. Any systematic inaccuracy or inconsistency in the dataset that might influence the outcomes or forecasts of the machine learning model is referred to as data bias. To ensure the validity and generalizability of the model, data bias must be addressed. Researchers often do a detailed study of the dataset to look for data bias. They examine various factors such as the distribution of active and inactive drugs, the distribution of IC50 values, and any potential sources of bias in the data collection process.

3. Making descriptors: The PaDEL programme has been used to produce molecular descriptors or fingerprints based on the SMILES sequences of the ligands. PaDEL is a software programme that is frequently used for computational drug design and research. It stands for Prediction and Evaluation of Drug-likeness and Toxicity Liability. It provides a range of molecular descriptors to describe many facets of a molecule's structure and characteristics. In this instance, the structural features of the

ligands generated from their SMILES sequences were used by PaDEL software to generate 308 descriptors. These descriptions offer numerical representations of many molecular characteristics, including size, shape, flexibility, electrostatics, and other structural characteristics. They are significant inputs for machine learning models, allowing them to learn patterns and relationships between molecular structures and their biological activities. In the novel drug identification and development, molecular descriptors are extremely important. In order to anticipate numerous molecular and biological qualities, such as drug-likeness, solubility, toxicity, and activity against target proteins, they let researchers compare and analyse the chemical properties of distinct compounds efficiently. Researchers may train algorithms to understand and recognise patterns in the data by including PaDEL-generated chemical descriptors into the machine learning process. In order to identify possible therapeutic candidates, these models may then be employed to forecast the activity or potency of new ligands based on their chemical structures. [8]

4. Machine Training: The dataset has been split in half throughout the machine learning training phase, with 80% of the data being used to train the model and 20% being used to evaluate its performance. This division makes it possible to assess how effectively the model generalises to new data. The RandomForestClassifier model has been chosen to train the data for this particular case. Popular machine learning method RandomForest uses a group of decision trees to predict outcomes. It is renowned for its capacity to manage intricate connections and deliver reliable outcomes. [55]. The next stage is to assess the model's accuracy using the testing data after it has been trained using the training data. A standard statistic for evaluating the effectiveness of a classification model is accuracy. It calculates the fraction of correctly classified instances out of the total number of instances of the testing dataset. The RandomForestClassifier model's performance on the test data in this instance was found to be 85% accurate. This indicates that, based on the chemical descriptors and other characteristics of the ligands, the model accurately predicted the activity or inactivity of the ligands in 85% of the instances[56]. The model is doing pretty well in forecasting the ligands' activities, as evidenced by its accuracy of 85%. To fully comprehend the performance of the model, it is necessary to conduct a more thorough study and take into account other assessment criteria. Precision, recall, and F1 score are other prominent assessment measures for classification tasks. Out of all cases anticipated as positive, precision is the percentage of accurately predicted

positive instances. The proportion of accurately anticipated positive cases out of the actual positive instances is measured by recall, also known as sensitivity or true positive rate. The accuracy of our Machine learning model has come out to be 80%.

The F1 score, which is a balanced indicator of model performance, is the harmonic mean of accuracy and recall. [Diffuse large B-cell lymphoma outcom] Cross-validation should be used to evaluate the model's stability and generalizability. In cross-validation, the dataset is divided into several subsets, and the model is trained using various combinations of training and testing sets. This lessens the chance of model failure by ensuring that the model performs consistently across various data divisions. In the machine learning process, several models have been employed to train and evaluate the data, including RandomForestClassifier, NuSVC, BaggingClassifier, HistGradientBoostingClassifier, SVC, RidgeClassifier, DecisionTreeClassifier, KNeighborsClassifier, MLPClassifier, AdaBoostClassifier, Lasso, Ridge, and ElasticNet. Among these models, the RandomForestClassifier has provided the best cross-validation result.

5. Phytochemical data preparation and validation: 80 phytochemicals from the Diospyros kaki plant were submitted, and they were found in the IMPPAT database. SMILES sequences have been used to represent the chemical structures of these substances in an Excel file. With the use of the PaDEL programme, molecular fingerprints or descriptors were created in order to anticipate the activity of these phytochemicals and assess their potential as inhibitors. Each compound was given a total of 308 descriptors from this method, each describing a different set of structural and physicochemical characteristics. The resulting descriptors were placed into a RandomForestClassifier to construct a machine learning model for activity prediction. This classifier uses the potent ensemble learning technique of combining different decision trees to provide precise predictions. Due of its capacity to manage complicated interactions, it has been extensively employed in chemoinformatics and drug discovery applications. The activity of the phytochemical substances was then predicted using the trained RandomForestClassifier model based on its descriptors. The model made predictions on whether the chemicals have inhibitory potential by using the learnt patterns and correlations from the training data.

F. Mechanism of MACCS analysis

The ChEMBL database, a useful tool for drug discovery research, provided the data for the MACCS analysis carried out in this case. Purification of the data collected by ChEMBL involved sorting and preparation for additional analysis. The categorization of the data based on the IC50 value was one of the first phases in the study. The half-maximal inhibitory concentration, or IC50, is a measurement of a drug's effectiveness in blocking a particular biological target. The conventional IC50 values used in this instance for categorization varied from 2.5 to 5 ng.

These standards led to the classification of the compounds in the dataset as either active or inactive medications. Drugs were categorised as either active or inactive depending on whether their IC50 values were within the specified range (2.5 to 5 ng) or outside of it. The ligands' molecular structures were created in order to perform the MACCS analysis. The RDKit package, a potent Python cheminformatics toolbox, assisted in this endeavour. The ability to create and modify chemical structures is only one of the many features and tools that RDKit offers for dealing with molecular structures. [57] [[35], [58], <https://www.rdkit.org>"]

The ligands' molecular structures were created and displayed using the RDKit package in a way that was appropriate for MACCS analysis. This process is essential because MACCS fingerprints, which are produced from molecular structures and capture significant structural elements that contribute to the action of the medication, are dependent on them.

The data was analysed using a number of machine learning models, including BaggingClassifier, HistGradientBoostingClassifier, GradientBoostingClassifier, ExtraTreesRegressor, RandomForestClassifier, AdaBoostClassifier, NuSVC, and SVC after the molecular fingerprints had been created. These models were chosen because they were suitable for the supplied dataset and could handle classification tasks.

Cross-validation scores were calculated to gauge each machine learning model's performance. Through the repeated process of training and testing the model on various combinations of the dataset's subsets, cross-validation is a technique that aids in assessing the model's performance. This method offers a more reliable assessment of the model's precision.

The RandomForestClassifier displayed the best efficiency of all the machine learning models evaluated, attaining an accuracy of 80%. An ensemble learning technique called RandomForestClassifier mixes various decision trees to provide predictions. It makes use of the idea of random feature selection and bagging to produce a variety of strong models.

The next step was to extract the key elements that went into the categorization after choosing the most effective model. The RandomForestClassifier's feature_importances method was used to do this. The relative relevance of each feature in the classification process is quantified by this function. The 40 most significant features are displayed in the following graph.

The Machine model performed 89% efficiently after employing these 40 phytochemical features for training, indicating that these features are crucial to the efficacy and potential of drugs. Insights for medication repurposing and development are crucially gained through MACCS analysis of significant RandomForestClassifier characteristics. The ability to pinpoint key chemical descriptors and substructures enables researchers to concentrate on particular traits with significant effects. This targeted approach enhances the selection of existing drugs for repurposing, saving time and costs.

In our analysis, we found that features 79, 86, 93, 95, 96, 109, 111, 144, 135 had a substantial influence on the efficacy and structure of drugs. These characteristics stand for the following groups, in that order: aliphatic carbazole, aromatic pyridine, aromatic phenanthrene, aromatic chrysene, aromatic naphthalene, aromatic acenaphthylene, aromatic benzo[ghi]perylene, and aliphatic acenaphthene. By adding or changing crucial chemical descriptors and substructures, building more effective medication candidates is aided by an understanding of essential properties. Targeting certain illnesses or biological targets, this optimisation technique speeds up the drug development process. Understanding key characteristics is useful for spotting potential adverse responses and off-target consequences. The identification of particular chemical characteristics that cause undesirable side effects or toxicity enables medicinal chemists to enhance safety profiles and optimise drug prospects. Furthermore, understanding important features contributes to the mechanistic understanding of drug activity. Correlating specific molecular descriptors or substructures with classification outcomes sheds light on structure-function relationships. This knowledge guides rational compound design and enhances drug efficacy.

CHAPTER: 6 RESULTS

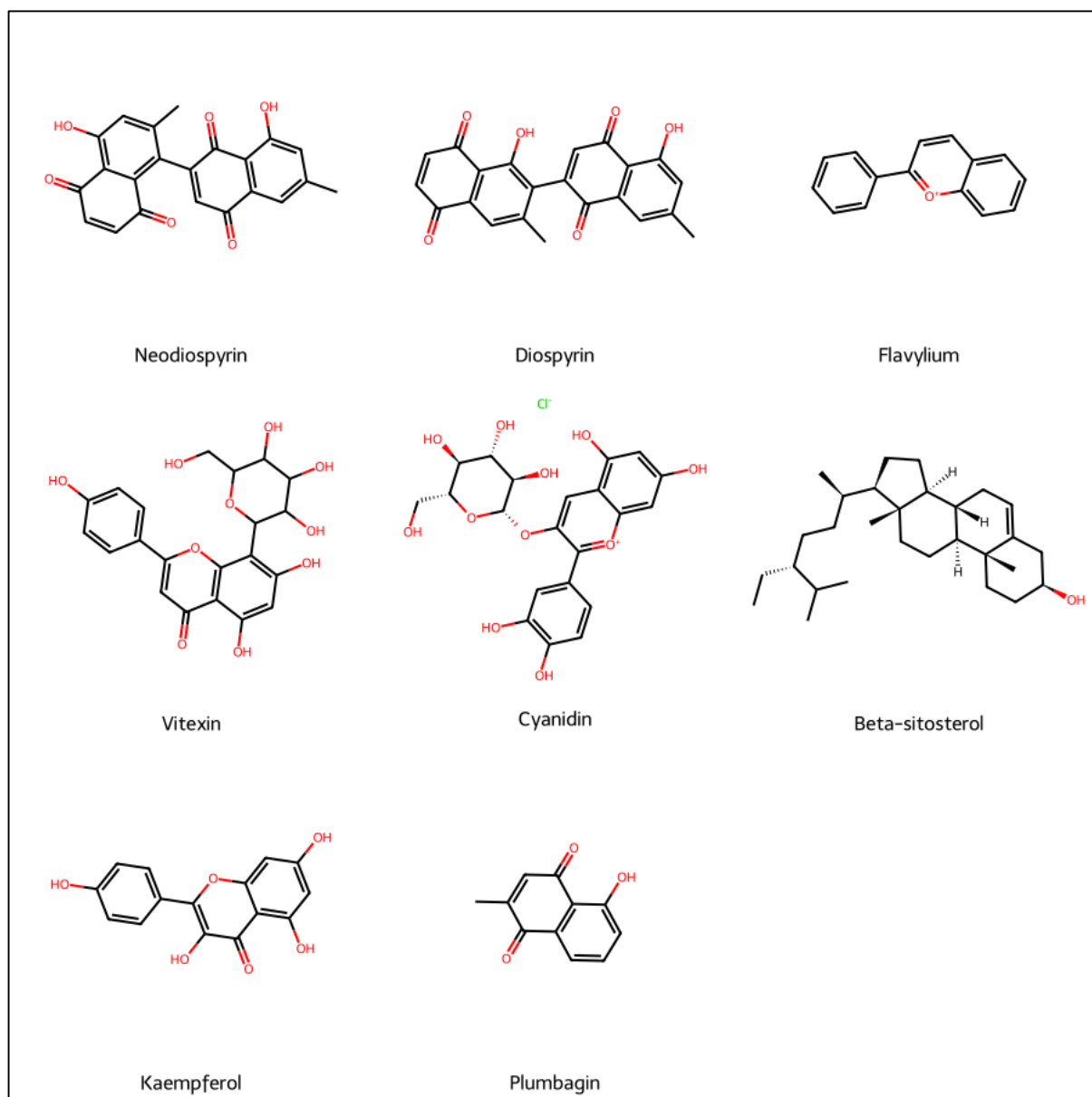


Figure 3: Molecular structures of phytochemicals of *Diospyros kaki*

This study employed a docking analysis to estimate the binding affinity of several phytochemicals, namely Neodiospyrin, Diospyrin, Flavylum, Vitexin, and Beta-sitosterol, to a specific target of interest. The obtained binding energies for these phytochemicals were determined as -10.4, -10.6, -8.4, -9.6, and -8.7, respectively. The obtained binding energies indicate the strength of the interactions between the target protein and each phytochemical. Lower binding energies generally suggest stronger binding affinity and a higher likelihood of forming stable complexes.

Table III: Molecular docking of phytochemicals against target protein (PDK-1) in-silico.

Phytochemical	Binding energy (kcal/mol)	Hydrogen bonds	KI	Amino acid participating
Diospyrin	-10.6	1	16.2 nM	ALA 162
Neodiospyrin	-10.4	2	22.8 nM	SER 160, ALA 162
Compound 8i1 (Standard inhibitor)	-10.1	3	37.8 nM	ALA 162, ASN 210, ASP 223
Vitexin	-9.6	2	88.2 nM	SER 160, ALA 162
Beta-sitosterol	-8.7	1	0.404 μ M	ALA 162
Flavylum	-8.4	0	0.671 μ M	0
Kaempferol	-8.3	2	0.795 μ M	ALA 162, ASP 223
Cyanidin	-8.1	3	1.116 μ M	ALA 162, THR 222, ASP 223
Plumbagin	-7.7	2	2.196 μ M	SER 160, ALA 162

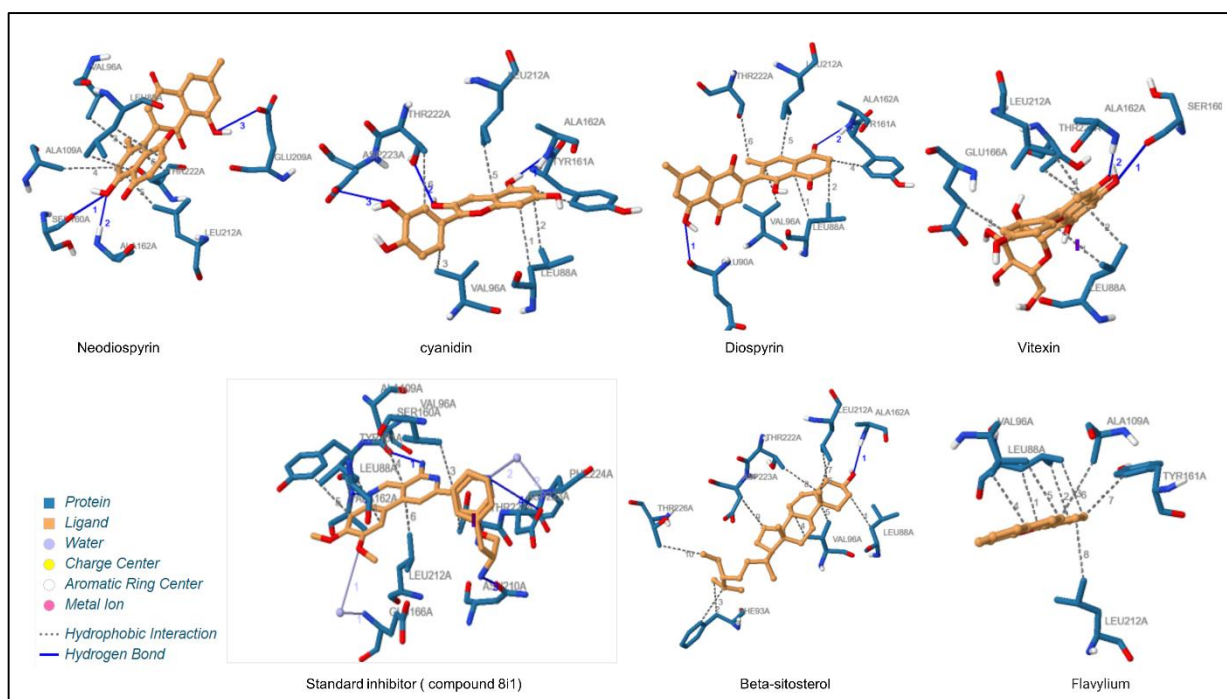


Figure 4: Molecular docking of ligands and target protein (PDK-1) retrieved from PLIP

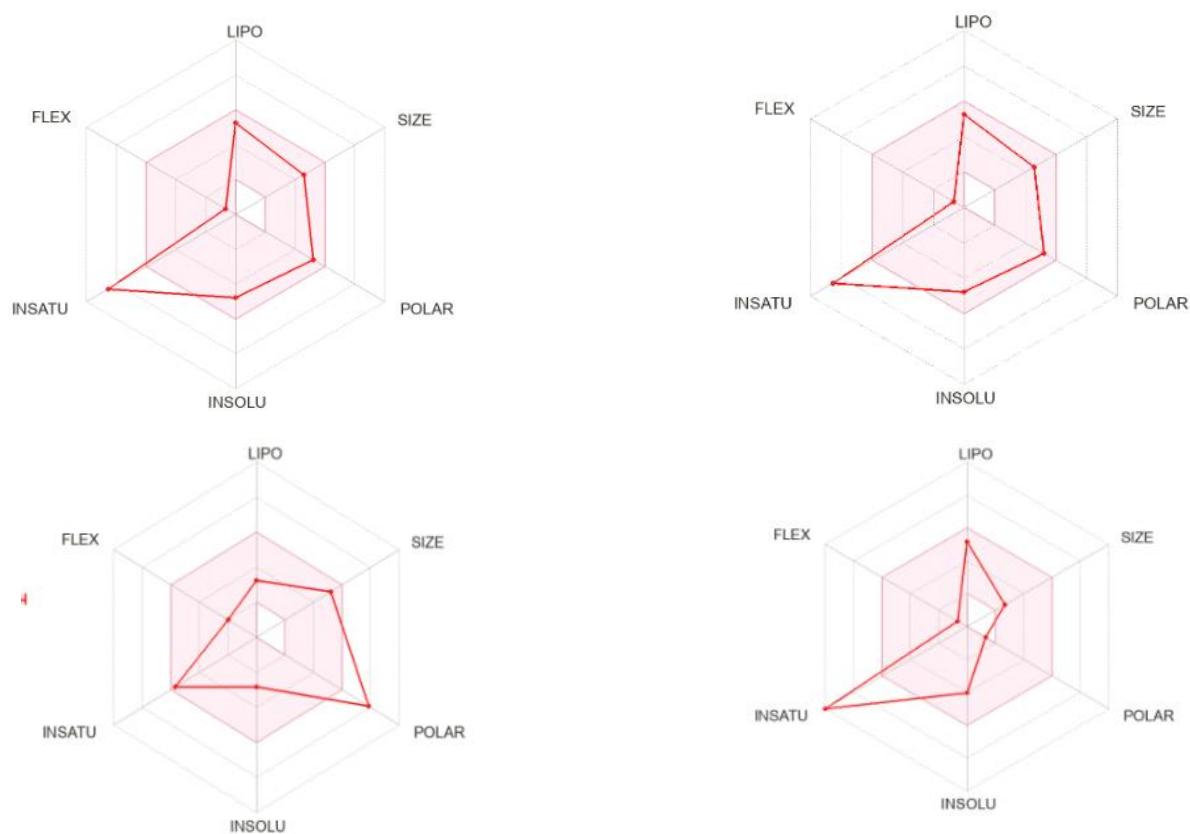


Figure 5: Bioavailability radars of Neodiospyrin, Diospyrin, Vitexin and Flavylum respectively retrieved using swisADME

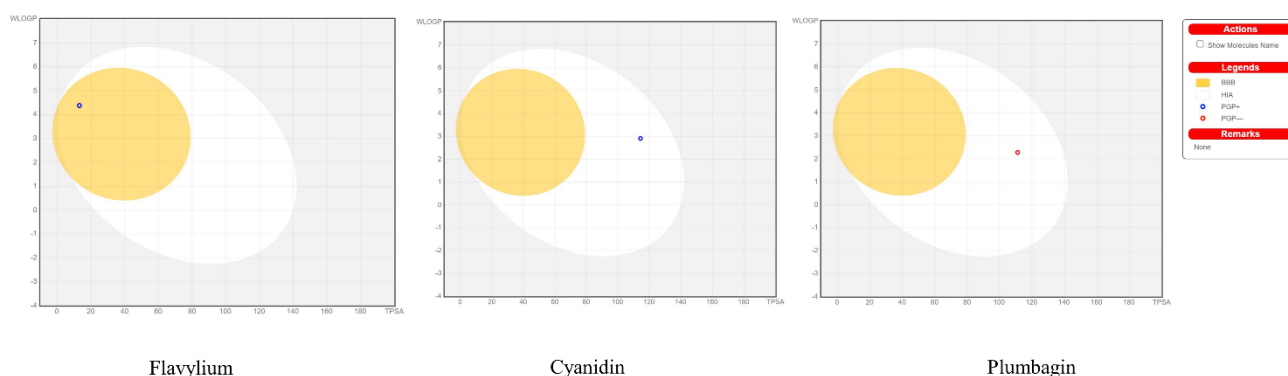


Figure 6: SwisADME analysis reveals phytochemicals Flavylum crosses BBB whereas compounds like Cyanidin and Plumbagin fail to do so.

Following the successful docking analysis, this study conducted further evaluation of drugs using bioinformatics software tools, including swisADME[53] and Molinspiration, to assess their pharmacodynamics and pharmacokinetics properties. Neodiospyrin, Diospyrin, Cyanidin, and Flavylum exhibited no violations, indicating their compliance with the Lipinski rule of five, which evaluates drug-likeness based on physicochemical properties. However, Beta-sitosterol and Vitexin were found to violate the Lipinski rule of five, suggesting potential challenges in their absorption, distribution, metabolism, and excretion.

Table IV: Lipinski's Rule of Five Analysis (RO5) Mass<500, H-bond donors<5, Hbond acceptors<10, LogP value<5

Phytochemicals	Mass	H-bond donor	H-bond acceptor	LogP value	Molar refractivity	No. of violations
Neodiospyrin	374.35	2	6	3.1 1	101.30	0
Cyanidin	287.25	5	5	2.9 1	76.17	0
Diospyrin	374.35	2	6	3.1 1	101.3	0
Vitexin	432.38	7	10	0.0 9	106.61	1
Flavylum	207.25	0	0	4.3 8	66.06	0
Beta-sitosterol	441.72	1	1	8.0 2	133.23	1

Table V: Bioactivity Scores of Phytochemicals retrieved from Molinspiration.

Phytochemical	GPCR	PI	EI	NRL	KI	ICM
Neodiospyrin	-0.10	-0.14	0.23	0.13	0.03	-0.28
Cyanidin	-0.13	-0.30	0.01	0.09	0.02	-0.09
Diospyrin	-0.06	-0.09	0.29	0.09	0.04	-0.18
Vitexin	0.13	0.03	0.46	0.23	0.19	-0.14
Flavylum	-0.61	-0.75	-0.38	-0.65	-0.57	-0.30
Beta-sitosterol	0.14	0.07	0.51	0.73	-0.51	0.04

Additionally, the pharmacokinetics analysis using MOLinspiration software encompassed various features such as GPCR, kinase inhibitor, enzyme inhibitor, NRL and PI. These analyses provide insights into the drugs' potential interactions with specific biological targets and their inhibition activity. The obtained results from these bioinformatics analyses contribute significant information regarding the pharmacokinetics and pharmacodynamics properties of Neodiospyrin, Diospyrin, Cyanidin, Flavylum, Beta-sitosterol, and Vitexin. This comprehensive assessment aids in understanding the potential efficacy and safety profiles of these compounds, guiding further drug development, optimization and research studies.

In order to validate our docking results, a machine learning analysis was conducted on the selected phytochemicals. Among various machine learning models, the RandomForestClassifier was chosen as it exhibited the highest accuracy and F1 value, indicating its effectiveness in anticipating the compound's activity. Using the RandomForestClassifier model, the analysis identified six phytochemicals as potential active drugs for inhibiting the PDK-1 protein. These phytochemicals were Neodiospyrin, Cyanidin, Diospyrin, Vitexin, Flavylum, and Beta-sitosterol. The machine learning model utilized various features and characteristics of the compounds to make predictions on their potential inhibitory activity against the target protein. This machine learning-based approach provides significant information about the potential therapeutic relevance of these phytochemicals in targeting the PDK-1 protein. The identified active drugs could serve as promising candidates for further experimental validation and development as potential inhibitors of PDK-1, a protein of significant interest in various biological processes and disease pathways.

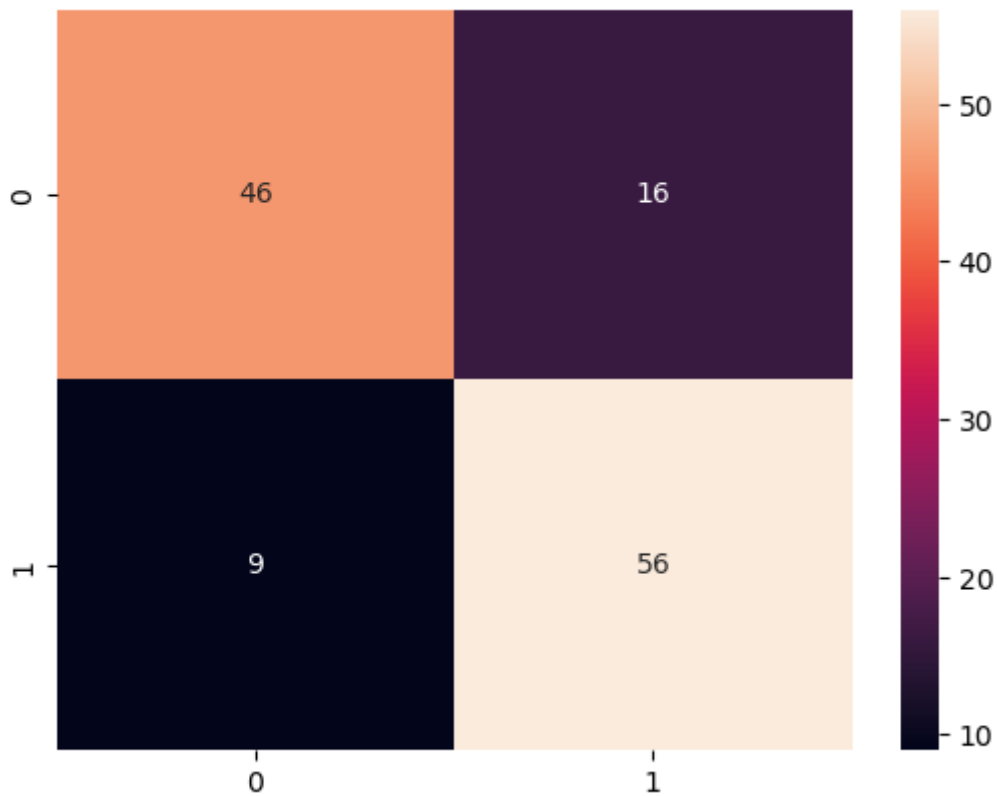


Figure 7: Confusion matrix of Machine Learning accuracy. Depicting True Positive=46, True Negative= 56, False Positive= 16, False Negative=9.

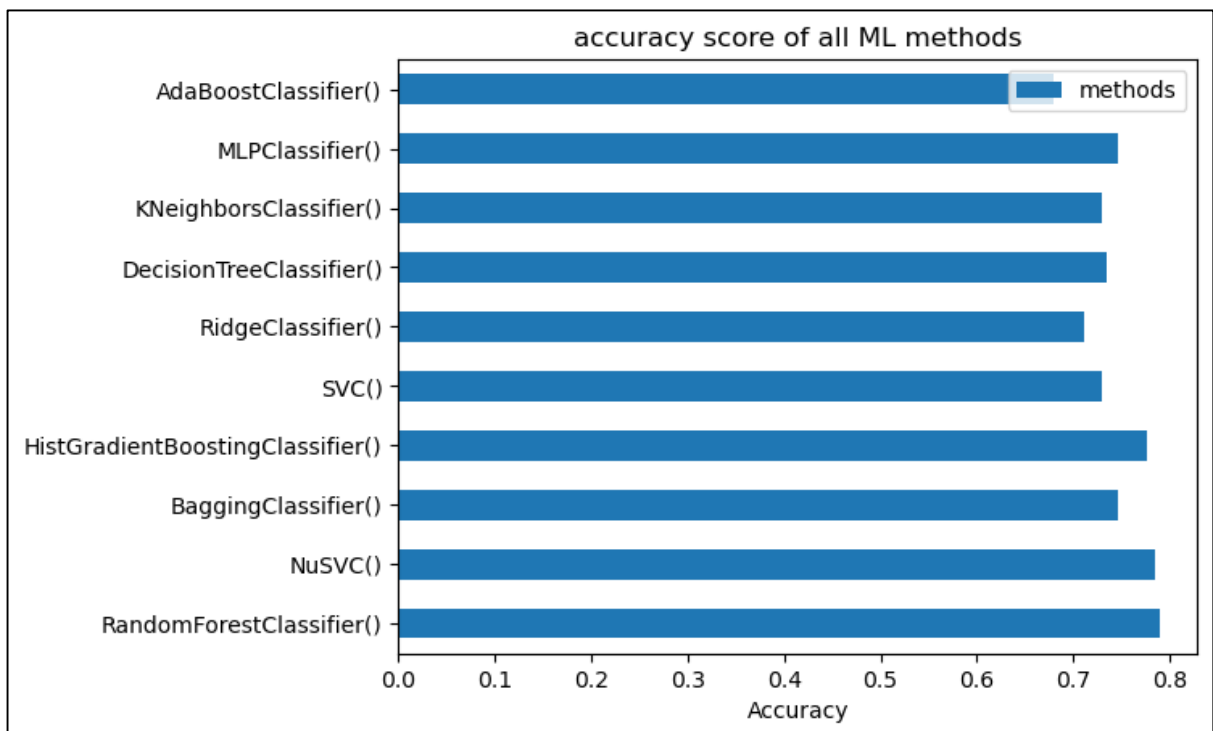


Figure 8: Bar graph Depicting different machine learning model and their accuracy

Table VI: List of MAACS fingerprint keys

MDL MACCS Key	Smart Pattern		
		42	Aliphatic secondary nitrogen
1	Aromatic ring	43	Aromatic primary nitrogen
2	Aliphatic ring	44	Aliphatic primary nitrogen
3	Double bond	45	Aromatic halogen
4	Triple bond	46	Aliphatic halogen
5	Aromatic nitrogen	47	Aromatic ester
6	Aliphatic nitrogen	48	Aliphatic ester
7	Aromatic oxygen	49	Aromatic ether
8	Aliphatic oxygen	50	Aliphatic ether
9	Aromatic sulfur	51	Aromatic amide
10	Aliphatic sulfur	52	Aliphatic amide
11	Aromatic chlorine	53	Aromatic nitrile
12	Aliphatic chlorine	54	Aliphatic nitrile
13	Aromatic bromine	55	Aromatic urea
14	Aliphatic bromine	56	Aliphatic urea
15	Aromatic iodine	57	Aromatic thioether
16	Aliphatic iodine	58	Aliphatic thioether
17	Aromatic hydroxyl	59	Aromatic imide
18	Aliphatic hydroxyl	60	Aliphatic imide
19	Aromatic methoxy	50	Aliphatic ether
20	Aliphatic methoxy	51	Aromatic amide
21	Aromatic amino	52	Aliphatic amide
22	Aliphatic amino	53	Aromatic nitrile
23	Aromatic thiol	54	Aliphatic nitrile
24	Aliphatic thiol	55	Aromatic urea
25	Aromatic nitro	56	Aliphatic urea
26	Aliphatic nitro	57	Aromatic thioether
27	Aromatic carbonyl	58	Aliphatic thioether
28	Aliphatic carbonyl	59	Aromatic imide
29	Aromatic carboxylic acid	60	Aliphatic imide
30	Aliphatic carboxylic acid	56	Aliphatic urea

31	Aromatic sulfonic acid	57	Aromatic thioether
32	Aliphatic sulfonic acid	58	Aliphatic thioether
33	Aromatic phosphonic acid	59	Aromatic imide
34	Aliphatic phosphonic acid	60	Aliphatic imide
35	Aromatic phosphinic acid	61	Aromatic hydrazine
36	Aliphatic phosphinic acid	62	Aliphatic hydrazine
37	Aromatic quaternary nitrogen	63	Aromatic azomethine
38	Aliphatic quaternary nitrogen	64	Aliphatic azomethine
39	Aromatic tertiary nitrogen	65	Aromatic azo
40	Aliphatic tertiary nitrogen	66	Aliphatic azo
41	Aromatic secondary nitrogen	67	Aromatic Schiff base
68	Aliphatic Schiff base	111	Aromatic acenaphthylene
69	Aromatic pyrazole	112	Aliphatic acenaphthylene
70	Aliphatic pyrazole	113	Aromatic acenaphthene
71	Aromatic imidazole	114	Aliphatic acenaphthene
72	Aliphatic imidazole	115	Aromatic fluorene
73	Aromatic thiazole	116	Aliphatic fluorene
74	Aliphatic thiazole	117	Aromatic phenanthrene
75	Aromatic furan	118	Aliphatic phenanthrene
76	Aliphatic furan	119	Aromatic anthracene
77	Aromatic pyrrole		
78	Aliphatic pyrrole	120	Aliphatic anthracene
79	Aromatic pyridine	121	Aromatic pyrene
80	Aliphatic pyridine	122	Aliphatic pyrene
81	Aromatic quinoline	123	Aromatic benz[a]anthracene

82	Aliphatic quinoline	124	Aliphatic benz[a]anthracene
83	Aromatic isoquinoline	125	Aromatic chrysene
84	Aliphatic isoquinoline	126	Aliphatic chrysene
85	Aromatic carbazole	127	Aromatic benzo[a]pyrene
86	Aliphatic carbazole	128	Aliphatic benzo[a]pyrene
87	Aromatic phthalazine	129	Aromatic benzo[b]fluoranthene
88	Aliphatic phthalazine	130	Aliphatic benzo[b]fluoranthene
89	Aromatic phenanthrene	131	Aromatic benzo[k]fluoranthene
90	Aliphatic phenanthrene	132	Aliphatic benzo[k]fluoranthene
91	Aromatic anthracene	133	Aromatic dibenz[a,h]anthracene
92	Aliphatic anthracene	134	Aliphatic dibenz[a,h]anthracene
93	Aromatic phenanthrene	135	Aromatic benzo[ghi]perylene
94	Aliphatic phenanthrene	136	Aliphatic benzo[ghi]perylene
95	Aromatic chrysene	137	Aromatic indeno[1,2,3-cd]pyrene
96	Aliphatic chrysene	138	Aliphatic indeno[1,2,3-cd]pyrene
97	Aromatic benzo[a]pyrene	139	Aromatic naphthalene
98	Aliphatic benzo[a]pyrene	140	Aliphatic naphthalene
99	Aromatic benzo[b]fluoranthene	141	Aromatic acenaphthylene
100	Aliphatic benzo[b]fluoranthene	142	Aliphatic acenaphthylene
101	Aromatic benzo[k]fluoranthene	143	Aromatic acenaphthene

102	Aliphatic benzo[k]fluoranthene	144	Aliphatic acenaphthene
103	Aromatic dibenz[a,h]anthracene	145	Aromatic fluorene
104	Aliphatic dibenz[a,h]anthracene	146	Aliphatic fluorene
105	Aromatic benzo[ghi]perylene	147	Aromatic phenanthrene
106	Aliphatic benzo[ghi]perylene	148	Aliphatic phenanthrene
107	Aromatic indeno[1,2,3-cd] pyrene	149	Aromatic anthracene
108	Aliphatic indeno[1,2,3-cd] pyrene	150	Aliphatic anthracene
109	Aromatic naphthalene	151	Aromatic pyrene
110	Aliphatic naphthalene	152	Aliphatic pyrene
157	Aromatic benzo[a]pyrene	162	Aliphatic acenaphthene
158	Aliphatic naphthalene	163	Aromatic fluorene
159	Aromatic acenaphthylene	164	Aliphatic fluorene
160	Aliphatic acenaphthylene	165	Aromatic phenanthrene
161	Aromatic acenaphthene	166	Aliphatic phenanthrene

We analyzed the MACCS descriptors of active drugs with coherent IC50 values for drug remodeling and optimization. MACCS descriptors are binary fingerprints encoding structural features based on 166 predefined substructural patterns. These patterns capture the presence or absence of specific fragments or functional groups. By analyzing the MACCS descriptors, we identified shared structural characteristics among the active drugs. This information is valuable for modifying and optimizing drugs, as it helps identify key molecular features that can be enhanced. The MACCS descriptors allow us to uncover relationships between molecular structure and drug activity, guiding future modifications and optimization strategies in drug design.

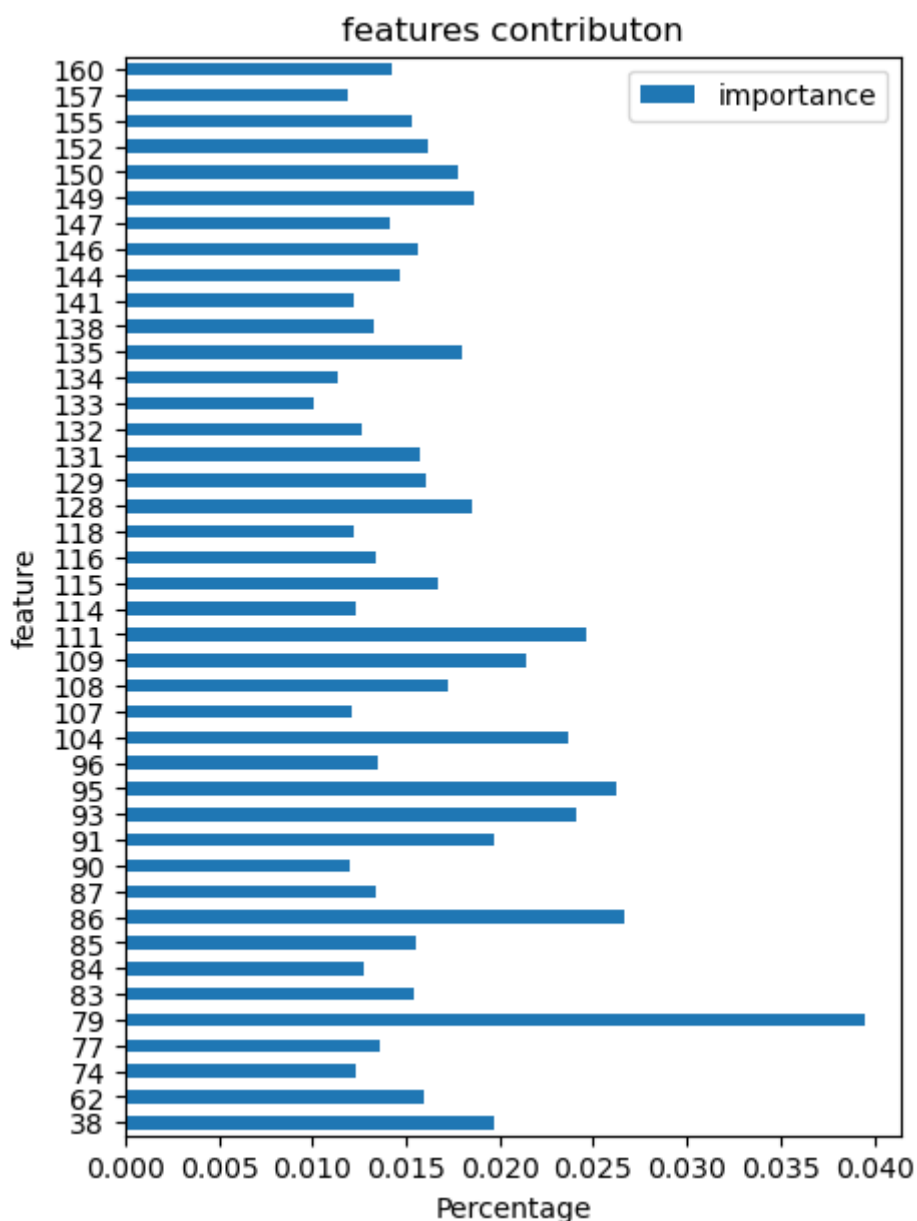


Figure 9: Contribution of each key based on their presence in the drug/compound.

Among the 166 descriptors analysed for our active drugs, the following descriptors were present in each of the active drugs. This observation suggests that these specific functional groups or fragments are consistently found in active drugs, indicating their potential significance in drug activity. The intersection of molecular descriptors present in all the drugs includes descriptors 79, 86, 95, 111, 38, 91, 109, 128, 135, 149, 150, 62, and more. These descriptors indicate the presence of specific chemical structures or functional groups in the active drug molecules. For instance, aromatic pyridine, aliphatic carbazole, aromatic chrysene, aromatic acenaphthylene, aliphatic quaternary nitrogen, aromatic anthracene, aromatic naphthalene, aliphatic benzopyrene, aromatic benzo perylene, aliphatic anthracene, aliphatic hydrazine, and others.

The consistent presence of these descriptors suggests that these structural features, such as aromatic rings or specific functional groups, play a significant role in the activity of the drugs. These findings provide valuable insights into the importance of certain chemical structures or fragments for the activity of our active drugs and can guide further drug design and optimization strategies.

CHAPTER 7: CONCLUSION

The adverse side effects and toxicities associated with synthetic drugs commonly used in cancer treatment have prompted a shift in focus towards exploring alternative therapeutic agents that are both non-toxic and readily available. In this context, phytochemicals derived from natural sources have gained attention as potential anti-cancer agents.

Withania somnifera, a plant known for its various health benefits, possesses a diverse array of bioactive compounds within its phytochemical profile. Therefore, a specific research endeavor was conducted using computational methods (in-silico study) to investigate the potential of utilizing bioactive compounds derived from *Withania somnifera* to suppress the activity of a protein called PDK-1. PDK-1 plays a crucial role in the AKT-mTOR pathway, which is involved in cancer progression.

In addition to targeting PDK-1, the we considered several additional parameters to evaluate the suitability of phytochemicals as potential therapeutic drugs. These parameters encompassed minimal cytotoxicity (toxic effects on cells), maximum bioavailability (ability to reach the target site in the body), bioactivity (capability to induce desired biological effects), and blood-brain barrier (BBB) permeation (ability to cross the protective barrier surrounding the brain). By considering these factors, the we categorized the phytochemicals from *Withania somnifera* based on their potential as therapeutic drugs. Further MACCS analysis of the active compound demonstrated that phytochemicals exhibit drug activity-associated fingerprints. Consequently, slight structural modifications of phytochemicals hold the potential to enhance their efficacy as potent drugs or inhibitors specifically targeting glioblastoma.

The study also suggests that the approach of inhibiting other proteins involved in the AKT-mTOR pathway could be employed, implying a broader application of the research findings. Furthermore, the statement emphasizes the importance of further research in this field to enhance our understanding of the underlying molecular mechanisms responsible for the recurrence and persistence of glioblastoma stem cells (GSCs). By gaining more insights, we can precisely identify the specific proteins involved, potentially leading to the development of more targeted and effective therapeutic strategies for GSC-related cancers.

REFERENCES

- [1] S. Daniele *et al.*, “Dual Inhibition of PDK1 and Aurora Kinase A: An Effective Strategy to Induce Differentiation and Apoptosis of Human Glioblastoma Multiforme Stem Cells,” *ACS Chem Neurosci*, vol. 8, no. 1, pp. 100–114, Jan. 2017, doi: 10.1021/acscchemneuro.6b00251.
- [2] E. Majewska, “AKT/GSK3 β Signaling in Glioblastoma,” 2044, doi: 10.1007/s11064-016-2044-x/Published.
- [3] M. Signore *et al.*, “Combined PDK1 and CHK1 inhibition is required to kill glioblastoma stem-like cells in vitro and in vivo,” *Cell Death Dis*, vol. 5, no. 5, 2014, doi: 10.1038/cddis.2014.188.
- [4] H. Xu *et al.*, “Epidermal growth factor receptor in glioblastoma (Review),” *Oncol Lett*, vol. 14, no. 1, pp. 512–516, 2017, doi: 10.3892/ol.2017.6221.
- [5] H. J. Lin, F. C. Hsieh, H. Song, and J. Lin, “Elevated phosphorylation and activation of PDK-1/AKT pathway in human breast cancer,” *Br J Cancer*, vol. 93, no. 12, pp. 1372–1381, Dec. 2005, doi: 10.1038/sj.bjc.6602862.
- [6] L. E. Mendie and S. Hemalatha, “Molecular Docking of Phytochemicals Targeting GFRs as Therapeutic Sites for Cancer: an In Silico Study,” *Appl Biochem Biotechnol*, vol. 194, no. 1, pp. 215–231, Jan. 2022, doi: 10.1007/s12010-021-03791-7.
- [7] R. Direito, J. Rocha, B. Sepodes, and M. Eduardo-Figueira, “From diospyros kaki L. (persimmon) phytochemical profile and health impact to new product perspectives and waste valorization,” *Nutrients*, vol. 13, no. 9. MDPI, Sep. 01, 2021. doi: 10.3390/nu13093283.
- [8] C. W. Yap, “PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints,” *J Comput Chem*, vol. 32, no. 7, pp. 1466–1474, May 2011, doi: 10.1002/jcc.21707.
- [9] C. W. Yap, “PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints,” *J Comput Chem*, vol. 32, no. 7, pp. 1466–1474, May 2011, doi: 10.1002/jcc.21707.
- [10] H. Moriwaki, Y. S. Tian, N. Kawashita, and T. Takagi, “Mordred: A molecular descriptor calculator,” *J Cheminform*, vol. 10, no. 1, Feb. 2018, doi: 10.1186/s13321-018-0258-y.
- [11] G. Ramos *et al.*, “Fatigue Evaluation through Machine Learning and a Global Fatigue Descriptor,” *J Healthc Eng*, vol. 2020, 2020, doi: 10.1155/2020/6484129.
- [12] S. S. S. J. Ahmed and V. Ramakrishnan, “Systems biological approach of molecular descriptors connectivity: Optimal descriptors for oral bioavailability prediction,” *PLoS One*, vol. 7, no. 7, Jul. 2012, doi: 10.1371/journal.pone.0040654.
- [13] A. C. Tan, D. M. Ashley, G. Y. López, M. Malinzak, H. S. Friedman, and M. Khasraw, “Management of glioblastoma: State of the art and future directions,” *CA Cancer J Clin*, vol. 70, no. 4, pp. 299–312, Jul. 2020, doi: 10.3322/caac.21613.
- [14] A. R. P. Antunes, I. Scheyltjens, J. Duerinck, B. Neyns, K. Movahedi, and J. A. Van Ginderachter, “Understanding the glioblastoma immune microenvironment as basis for

- the development of new immunotherapeutic strategies," *Elife*, vol. 9, Feb. 2020, doi: 10.7554/eLife.52176.
- [15] T. Hara *et al.*, "Interactions between cancer cells and immune cells drive transitions to mesenchymal-like states in glioblastoma," *Cancer Cell*, vol. 39, no. 6, pp. 779-792.e11, Jun. 2021, doi: 10.1016/j.ccell.2021.05.002.
- [16] S. Daisy Precilla, I. Biswas, S. S. Kuduvalli, and T. S. Anitha, "Crosstalk between PI3K/AKT/mTOR and WNT/ β -Catenin signaling in GBM - Could combination therapy checkmate the collusion?," *Cellular Signalling*, vol. 95. Elsevier Inc., Jul. 01, 2022. doi: 10.1016/j.cellsig.2022.110350.
- [17] H. Kim *et al.*, "Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution," *Genome Res*, vol. 25, no. 3, pp. 316-327, Mar. 2015, doi: 10.1101/gr.180612.114.
- [18] C. Neftel *et al.*, "An Integrative Model of Cellular States, Plasticity, and Genetics for Glioblastoma," *Cell*, vol. 178, no. 4, pp. 835-849.e21, Aug. 2019, doi: 10.1016/j.cell.2019.06.024.
- [19] N. Montemurro, "Glioblastoma Multiforme and Genetic Mutations: The Issue Is Not over Yet An Overview of the Current Literature," *Journal of Neurological Surgery, Part A: Central European Neurosurgery*, vol. 81, no. 1. Georg Thieme Verlag, pp. 64-70, 2020. doi: 10.1055/s-0039-1688911.
- [20] R. McLendon *et al.*, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways," *Nature*, vol. 455, no. 7216, pp. 1061-1068, Oct. 2008, doi: 10.1038/nature07385.
- [21] S. Han *et al.*, "IDH mutation in glioma: molecular mechanisms and potential therapeutic targets," *British Journal of Cancer*, vol. 122, no. 11. Springer Nature, pp. 1580-1589, May 26, 2020. doi: 10.1038/s41416-020-0814-x.
- [22] F. Han *et al.*, "PTEN gene mutations correlate to poor prognosis in glioma patients: A meta-analysis," *Onco Targets Ther*, vol. 9, pp. 3485-3492, Jun. 2016, doi: 10.2147/OTT.S99942.
- [23] M. Simon *et al.*, "TERT promoter mutations: A novel independent prognostic factor in primary glioblastomas," *Neuro Oncol*, vol. 17, no. 1, pp. 45-52, Jan. 2015, doi: 10.1093/neuonc/nou158.
- [24] H. Colman, "Toward more informative biomarker-based clinical trials in glioblastoma," *Neuro-Oncology*, vol. 19, no. 7. Oxford University Press, pp. 880-881, Jul. 01, 2017. doi: 10.1093/neuonc/nox063.
- [25] Q. Sun *et al.*, "Up-regulation of MSH6 is associated with temozolomide resistance in human glioblastoma," *Biochem Biophys Res Commun*, vol. 496, no. 4, pp. 1040-1046, Feb. 2018, doi: 10.1016/j.bbrc.2018.01.093.
- [26] A. B. Behrooz, Z. Talaie, F. Jusheghani, M. J. Łos, T. Klonisch, and S. Ghavami, "Wnt and PI3K/Akt/mTOR Survival Pathways as Therapeutic Targets in Glioblastoma," *International Journal of Molecular Sciences*, vol. 23, no. 3. MDPI, Feb. 01, 2022. doi: 10.3390/ijms23031353.

- [27] X. Li *et al.*, "PI3K/Akt/mTOR signaling pathway and targeted therapy for glioblastoma," 2016. [Online]. Available: www.impactjournals.com/oncotarget
- [28] K. K. Umar Velpula and A. J. Tsung, "PDK1: a new therapeutic target for glioblastoma?," *CNS oncology*, vol. 3, no. 3, pp. 177–179, May 01, 2014. doi: 10.2217/cns.14.13.
- [29] G. Hibino, T. Nadamoto, F. Fujisawa, and T. Fushiki, "Regulation of the Peripheral Body Temperature by Foods: A Temperature Decrease Induced by the Japanese Persimmon (kaki, *Diospyros kaki*)," *Biosci Biotechnol Biochem*, vol. 67, no. 1, pp. 23–28, Jan. 2003, doi: 10.1271/bbb.67.23.
- [30] M. S. Butt *et al.*, "Persimmon (*diospyros kaki*) fruit: Hidden phytochemicals and health claims," *EXCLI Journal*, vol. 14. Leibniz Research Centre for Working Environment and Human Factors, pp. 542–561, May 04, 2015. doi: 10.17179/excli2015-159.
- [31] P. H. M. Torres, A. C. R. Sodero, P. Jofily, and F. P. Silva-Jr, "Key topics in molecular docking for drug design," *International Journal of Molecular Sciences*, vol. 20, no. 18. MDPI AG, Sep. 02, 2019. doi: 10.3390/ijms20184574.
- [32] L. Pinzi and G. Rastelli, "Molecular docking: Shifting paradigms in drug discovery," *International Journal of Molecular Sciences*, vol. 20, no. 18. MDPI AG, Sep. 01, 2019. doi: 10.3390/ijms20184331.
- [33] 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*, p. NA-NA, 2009, doi: 10.1002/jcc.21334.
- [34] T. Nittoli *et al.*, "The identification of 8,9-dimethoxy-5-(2-aminoalkoxy-pyridin-3-yl)-benzo[c][2,7]naphthyridin-4-ylamines as potent inhibitors of 3-phosphoinositide-dependent kinase-1 (PDK-1)," *Eur J Med Chem*, vol. 45, no. 4, pp. 1379–1386, Apr. 2010, doi: 10.1016/j.ejmech.2009.12.036.
- [35] A. Korotcov, V. Tkachenko, D. P. Russo, and S. Ekins, "Comparison of Deep Learning with Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets," *Mol Pharm*, vol. 14, no. 12, pp. 4462–4475, Dec. 2017, doi: 10.1021/acs.molpharmaceut.7b00578.
- [36] A. Gaulton *et al.*, "The ChEMBL database in 2017," *Nucleic Acids Res*, vol. 45, no. D1, pp. D945–D954, Jan. 2017, doi: 10.1093/nar/gkw1074.
- [37] M. T. Patrick *et al.*, "Drug Repurposing Prediction for Immune-Mediated Cutaneous Diseases using a Word-Embedding–Based Machine Learning Approach," *Journal of Investigative Dermatology*, vol. 139, no. 3, pp. 683–691, Mar. 2019, doi: 10.1016/j.jid.2018.09.018.
- [38] J. L. Durant, B. A. Leland, D. R. Henry, and J. G. Nourse, "Reoptimization of MDL keys for use in drug discovery," *J Chem Inf Comput Sci*, vol. 42, no. 6, pp. 1273–1280, Nov. 2002, doi: 10.1021/ci010132r.
- [39] N. Cerisier, L. Regad, D. Triki, M. Petitjean, D. Flatters, and A. C. Camproux, "Statistical Profiling of One Promiscuous Protein Binding Site: Illustrated by Urokinase Catalytic Domain," *Mol Inform*, vol. 36, no. 10, Oct. 2017, doi: 10.1002/minf.201700040.

- [40] K. Mohanraj *et al.*, "IMPPAT: A curated database of Indian Medicinal Plants, Phytochemistry and Therapeutics," *Sci Rep*, vol. 8, no. 1, Dec. 2018, doi: 10.1038/s41598-018-22631-z.
- [41] H. M. Berman *et al.*, "The Protein Data Bank," 2000. [Online]. Available: <http://www.rcsb.org/pdb/status.html>
- [42] T. Gaillard, "Evaluation of AutoDock and AutoDock Vina on the CASF-2013 Benchmark," *J Chem Inf Model*, vol. 58, no. 8, pp. 1697–1706, Aug. 2018, doi: 10.1021/acs.jcim.8b00312.
- [43] D. Seeliger and B. L. De Groot, "Ligand docking and binding site analysis with PyMOL and Autodock/Vina," *J Comput Aided Mol Des*, vol. 24, no. 5, pp. 417–422, 2010, doi: 10.1007/s10822-010-9352-6.
- [44] N. T. Nguyen *et al.*, "Autodock Vina Adopts More Accurate Binding Poses but Autodock4 Forms Better Binding Affinity," *J Chem Inf Model*, vol. 60, no. 1, pp. 204–211, Jan. 2020, doi: 10.1021/acs.jcim.9b00778.
- [45] S. Salentin *et al.*, "From malaria to cancer: Computational drug repositioning of amodiaquine using PLIP interaction patterns," *Sci Rep*, vol. 7, no. 1, Dec. 2017, doi: 10.1038/s41598-017-11924-4.
- [46] S. Salentin, S. Schreiber, V. J. Haupt, M. F. Adasme, and M. Schroeder, "PLIP: Fully automated protein-ligand interaction profiler," *Nucleic Acids Res*, vol. 43, no. W1, pp. W443–W447, 2015, doi: 10.1093/nar/gkv315.
- [47] M. F. Adasme *et al.*, "PLIP 2021: Expanding the scope of the protein-ligand interaction profiler to DNA and RNA," *Nucleic Acids Res*, vol. 49, no. W1, pp. W530–W534, Jul. 2021, doi: 10.1093/nar/gkab294.
- [48] "Discovery Studio Life Science Modeling and Simulations."
- [49] D. Mendez *et al.*, "ChEMBL: Towards direct deposition of bioassay data," *Nucleic Acids Res*, vol. 47, no. D1, pp. D930–D940, Jan. 2019, doi: 10.1093/nar/gky1075.
- [50] A. Gaulton *et al.*, "ChEMBL: A large-scale bioactivity database for drug discovery," *Nucleic Acids Res*, vol. 40, no. D1, Jan. 2012, doi: 10.1093/nar/gkr777.
- [51] A. Mayr *et al.*, "Large-scale comparison of machine learning methods for drug target prediction on ChEMBL," *Chem Sci*, vol. 9, no. 24, pp. 5441–5451, 2018, doi: 10.1039/c8sc00148k.
- [52] A. Gaulton *et al.*, "ChEMBL: A large-scale bioactivity database for drug discovery," *Nucleic Acids Res*, vol. 40, no. D1, Jan. 2012, doi: 10.1093/nar/gkr777.
- [53] A. Daina, O. Michielin, and V. Zoete, "SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules," *Sci Rep*, vol. 7, Mar. 2017, doi: 10.1038/srep42717.
- [54] C. A. Lipinski, "Lead- and drug-like compounds: the rule-of-five revolution," *Drug Discov Today Technol*, vol. 1, no. 4, pp. 337–341, Dec. 2004, doi: 10.1016/j.ddtec.2004.11.007.

- [55] C. Sommer and D. W. Gerlich, "Machine learning in cell biology-teaching computers to recognize phenotypes," *J Cell Sci*, vol. 126, no. 24, pp. 5529–5539, Dec. 2013, doi: 10.1242/jcs.123604.
- [56] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Computational and Structural Biotechnology Journal*, vol. 15. Elsevier B.V., pp. 104–116, 2017. doi: 10.1016/j.csbj.2016.12.005.
- [57] R. Rodríguez-Pérez and J. Bajorath, "Interpretation of compound activity predictions from complex machine learning models using local approximations and shapley values," *J Med Chem*, vol. 63, no. 16, pp. 8761–8777, Aug. 2020, doi: 10.1021/acs.jmedchem.9b01101.
- [58] T. R. Lane, D. H. Foil, E. Minerali, F. Urbina, K. M. Zorn, and S. Ekins, "Bioactivity Comparison across Multiple Machine Learning Algorithms Using over 5000 Datasets for Drug Discovery," *Mol Pharm*, vol. 18, no. 1, pp. 403–415, Jan. 2021, doi: 10.1021/acs.molpharmaceut.0c01013.

PAPER NAME

FINAL D.DRAFT(N).docx

WORD COUNT

13286 Words

CHARACTER COUNT

79708 Characters

PAGE COUNT

56 Pages

FILE SIZE

1.6MB

SUBMISSION DATE

May 27, 2023 12:15 PM GMT+5:30

REPORT DATE

May 27, 2023 12:16 PM GMT+5:30

● 8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 5% Internet database
- 2% Publications database
- Crossref database
- Crossref Posted Content database
- 6% Submitted Works database

● Excluded from Similarity Report

- Bibliographic material

8% Overall Similarity

Top sources found in the following databases:

- 5% Internet database
- Crossref database
- 6% Submitted Works database
- 2% Publications database
- Crossref Posted Content database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	dspace.dtu.ac.in:8080 Internet	3%
2	Technical University of Cluj-Napoca on 2022-08-31 Submitted works	<1%
3	icstsn2023.ifet.ac.in Internet	<1%
4	Colorado School of Mines on 2022-04-08 Submitted works	<1%
5	dataset.drugtargetcommons.org Internet	<1%
6	Patrick K. Jjemba, Brian K. Kinkle, Jodi R. Shann. "In-situ enumeration ... Crossref	<1%
7	nature.com Internet	<1%
8	University of Leeds on 2023-05-18 Submitted works	<1%

9	University of Hertfordshire on 2023-05-23	<1%
	Submitted works	
10	Asia Pacific University College of Technology and Innovation (UCTI) on...	<1%
	Submitted works	
11	Roehampton University on 2023-05-01	<1%
	Submitted works	
12	University of Melbourne on 2018-05-30	<1%
	Submitted works	
13	Queen Mary and Westfield College on 2023-05-08	<1%
	Submitted works	
14	BPP College of Professional Studies Limited on 2023-05-02	<1%
	Submitted works	
15	swissmodel.expasy.org	<1%
	Internet	
16	Staffordshire University on 2020-11-26	<1%
	Submitted works	
17	The British College on 2023-05-23	<1%
	Submitted works	
18	theses.whiterose.ac.uk	<1%
	Internet	
19	deepblue.lib.umich.edu	<1%
	Internet	
20	Ipswich Girls' Grammar School on 2013-08-30	<1%
	Submitted works	

21	Jayze daCunhaXavier, Francisco Wagner de Queiroz Almeida-Neto, Pri... Crossref	<1%
22	Pugazhenthan Thangaraju, Gopinathan Narasimhan, Vijayakumar Aru... Crossref	<1%
23	University of Bedfordshire on 2021-05-23 Submitted works	<1%
24	academic.oup.com Internet	<1%
25	vital.seals.ac.za:8080 Internet	<1%
26	biorxiv.org Internet	<1%
27	Leiden University on 2020-02-20 Submitted works	<1%
28	Leiden University on 2023-04-19 Submitted works	<1%
29	Martin Weisel, Hans-Marcus Bitter, François Diederich, W. Venus So, R... Crossref	<1%
30	Universiti Teknologi Malaysia on 2020-05-19 Submitted works	<1%
31	"Handbook of Chemoinformatics", Wiley, 2003 Crossref	<1%
32	Adamson University on 2018-08-08 Submitted works	<1%

33	Cho, June-Haeng, Mi-Young Lee, Irshad Ahmed Baig, Na-Reum Ha, Jou...	<1%
	Crossref	
34	Jenny Balfer, Jürgen Bajorath. "Introduction of a Methodology for Visu...	<1%
	Crossref	
35	University of Edinburgh on 2019-08-16	<1%
	Submitted works	
36	University of Greenwich on 2022-08-30	<1%
	Submitted works	
37	Xin Yang, Yifei Wang, Ryan Byrne, Gisbert Schneider, Shengyong Yang. ...	<1%
	Crossref	
38	downloads.hindawi.com	<1%
	Internet	
39	go.gale.com	<1%
	Internet	
40	link.springer.com	<1%
	Internet	
41	mts.intechopen.com	<1%
	Internet	
42	pubs.rsc.org	<1%
	Internet	
43	revues.imist.ma	<1%
	Internet	
44	hindawi.com	<1%
	Internet	

45	jcheminf.com Internet	<1%
46	science.gov Internet	<1%
47	Mauro S. Nogueira, Oliver Koch. "The Development of Target-Specific ... Crossref	<1%
48	Nozawa, H.. "Phosphorylation of ribosomal p70 S6 kinase and rapamy... Crossref	<1%
49	Sarah Naomi Bolz, Melissa F. Adasme, Michael Schroeder. "Toward an ... Crossref	<1%
50	University of Sydney on 2022-01-04 Submitted works	<1%
51	Delhi Technological University on 2019-05-29 Submitted works	<1%
52	University of Ulster on 2023-04-19 Submitted works	<1%



Your paper, ID:ICSTSN 194, has been ACCEPTED

1 message

ICSTSN 2023 <icstsn2023@ifet.ac.in>

Thu, 23 Mar 2023 at 10:02 pm

To: nidaefalak380@gmail.com, ashrahansh@gmail.com, asmitadas1710@dcce.ac.in

Dear Author,

Congratulations!!!

The review and selection process for your paper ID ICSTSN-194 entitled "Diospyros kakis phytochemical mediated inhibition of GBM by targeting PDK-1: An in-silico docking and machine learning model." has been completed. Based on the recommendations from the reviewers assigned for your paper, I am pleased to inform you that your paper has been **ACCEPTED** by the **Technical Program Committee (TPC) for ORAL PRESENTATION** which is organized by IFET College of Engineering, Villupuram, Tamil Nadu, India during 21st - 22nd, April 2023. I am also glad to inform you that the proceedings of ICSTSN 2023 will be submitted for inclusion in IEEE Xplore.

Note : Conference will be held in both OFFLINE and ONLINE MODE.

Registration

- You are further requested to do the following

- You are requested to kindly register at the earliest (after paying the Conference Registration Fees) using the Registration Link.
<https://forms.gle/Xxe8VAm3YC1hoepT8>
- Registration will be closed on **27th March 2023**.
- IEEE members can avail the membership benefits on registration fees. Please attach the scanned copy of your IEEE membership card in the Google form.

Final submission Checklist

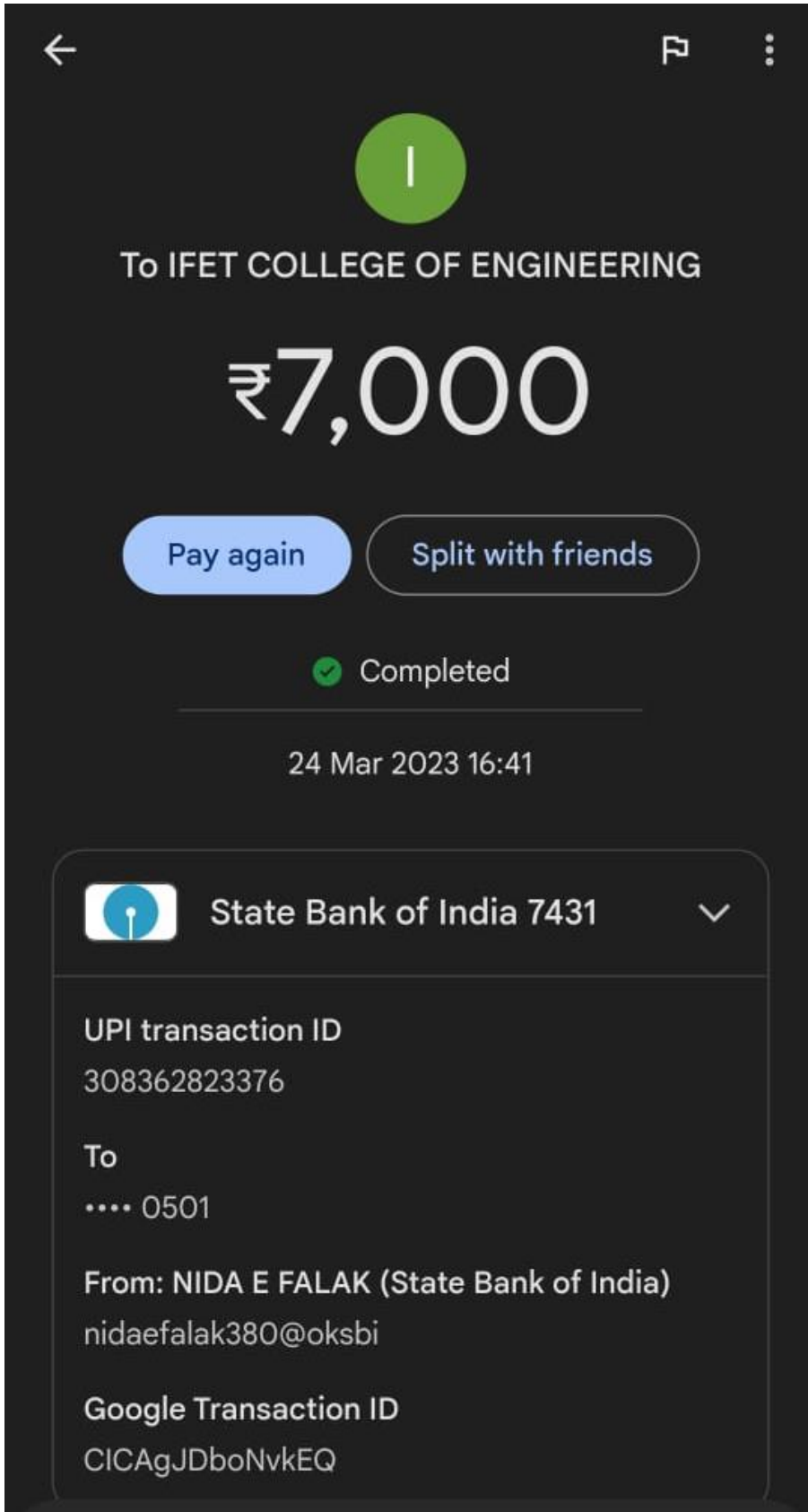
- The following documents have to be submitted along with the camera-ready paper on or before **27.03.2023**.

1. Camera ready paper in IEEE double column format (in Microsoft office word file) should be uploaded in the CMT portal.
2. Filled in Google form.
3. Proof of registration fee paid.

(The final paper should not exceed 6 pages, IEEE Xplore does not support pages more than 6. **Rs.500 will be charged for every extra page)**

With Regards,

Conference Coordinator
ICSTSN-2023





IFET COLLEGE OF ENGINEERING

AN AUTONOMOUS INSTITUTION



Approved by AICTE, Permanently Affiliated to Anna University,

Recognized under section 2(f) & 12(B) of UGC Act-1956. Recognized Research Centre of Anna University (ECE)

IFET Nagar, Gangarampalayam, Villupuram - 605 108. Website: www.ifet.ac.in



BEST PAPER AWARD

PRESENTED TO

MS. NIDA E FALAK

PG STUDENT, DELHI TECHNOLOGICAL UNIVERSITY

for the paper entitled

**DIOSPYROS KAKI'S PHYTOCHEMICAL MEDIATED INHIBITION OF
GBM BY INHIBITING PDK-1: AN IN-SILICO
AND MACHINE LEARNING MODEL**

in

Second IEEE International Conference on "Smart Technologies and Systems for Next Generation Computing (ICSTSN 2023)" held at IFET College of Engineering on 21st and 22nd April 2023

Matsyale. S.

CONFERENCE CHAIR

[Signature]

PRINCIPAL

[Signature]

CHAIRMAN

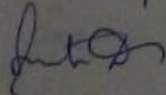
DEPARTMENT OF BIOTECHNOLOGY
DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi college of Engineering)
Bawana Road, Delhi-110042

CERTIFICATE

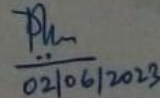
I hereby certify that the Project Dissertation titled Diospyros kaki's phytochemical mediated inhibition of GBM by targeting PDK-1: A machine learning model, which is submitted by Nida E Falak, 2K21/MSCBIO/61, Delhi Technological University Delhi, in partial fulfilment of the requirement for the award of the degree of Masters in Science, is a record of the project work carried out by the students under my supervision. To the best of my knowledge this work has not been submitted in part or full for any Degree or Diploma to this University or elsewhere

Place: Delhi:

Date: 02/06/23



DR. ASMITA DAS
SUPERVISOR
Department of biotechnology
Delhi technological university



02/06/2023

PROF. PRAVIR KUMAR
HEAD OF DEPARTMENT
Department of biotechnology
Delhi technological university

iii

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi college of Engineering)

Bawana Road, Delhi-110042

DECLARATION

I, Nida E Falak, 2K21/MSCBIO/61 of MSc. Biotechnology, hereby declare that the project Dissertation titled ***Diospyros kaki's phytochemical mediated inhibition of GBM by targeting PDK-1: A machine leaning model***, which is submitted by me to the Department of Biotechnology, Delhi Technological University, Delhi in partial fulfilment of the requirement for the award of the degree of Master of Science Biotechnology and submitted to the Department of Biotechnology, Delhi Technological University, Delhi, is an authentic record of my own carried out during the period from January-May 2023 under the supervision of **Dr. Asmita Das**.

The matter presented in this report has not been submitted by me for the award for any other degree of this or any other institute/University. The work has been accepted in SCI/SCI expanded / SSCI/Scopus Indexed Journal OR peer reviewed Scopus Index Conference with the following details :

Title of the Paper : Diospyros kaki's phytochemical mediated inhibition of GBM by targeting PDK-1: An in-silico docking and machine learning model

Authors Name: Nida E Falak, Harsh Aahra, Asmita Das

Name of Conference : Second IEEE International Conference on "Smart Technologies and System for Next Generation Computing (ICSTSN 2023)

Conference Date and Venue : 21st-22nd April 2023 IFET College of Engineering, Gangarampalaiyam, Tamil Nadu, India

Registration: Done

Status of Paper : Under proceeding

Date of Paper Communication: 20th December 2022

Date of Paper Acceptance: 23 March 2023

Date of Paper Publication:

Date: 02 | 06 | 23

Nida-e-falak
Nida E Falak