# INTERPRETABLE ENSEMBLE LEARNING PREDICTS GLIOBLASTOMA SENSITIVITY TO NATURAL COMPOUNDS

**A Thesis Submitted**
**In Partial Fulfillment of the Requirements**
**for the Degree of**

## MASTER OF TECHNOLOGY
**in**
## BIOINFORMATICS

**by**

**SOMYA PARASHAR**
**(Enrollment No. 23/BIO/01)**

**Under the Supervision of**
**Prof. PRAVIR KUMAR**
**Delhi Technological University**


**Department of Biotechnology**

**DELHI TECHNOLOGICAL UNIVERSITY**
**(Formerly Delhi College of Engineering)**
**Shahbad Daulatpur, Main Bawana Road, Delhi-110042. India**

**May, 2025**

# ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my project guide and mentor, Prof. Pravir Kumar, for accepting me in his lab and allowing me to carry out my final semester project under his mentorship. His knowledgeable insight and support throughout the duration of the project has helped me develop a better understanding of the nature of my research topic and gain an appreciation for academics. He has been a great source of reliance by engaging me with new ideas and demanding high quality and precision in all of my endeavours. I wholeheartedly thank him for his support, for advising me and giving me the intellectual freedom to work.

I would also like to take this opportunity to thank the Ph.D. Scholars in the lab Ms. Shefali and Ms. Shruti and Dr. Rahul for their availability and guidance during this project. Their insights and friendly suggestions have been instrumental towards this work. I am also very grateful to the faculty of Department of Biotechnology, Delhi Technological University whose teachings have equipped me with necessary skills to work as individual as well as helping me grow as a person.

Finally, I would like to thank my family who has always believed in me and urged me to excel in my endeavours and my friends and batchmates Ms. Rishi Mrinal and Mr. Akshay Hatwal for their emotional support through my journey.

# DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

## CANDIDATE'S DECLARATION

I, **SOMYA PARASHAR**, hereby certify that the work which is being presented in the thesis entitled **"Interpretable Ensemble Learning Predicts Glioblastoma Sensitivity to Natural Compounds"** in partial fulfillment of the requirements for the award of the Degree of Master of Technology, submitted in the Department of Biotechnology, Delhi Technological University is an authentic record of my own work carried out during the period from January 2025 to May 2025 under the supervision of **Prof. Pravir Kumar**.

The matter presented in the thesis has not been submitted by me for the award of any other degree of this or any other Institute.

**Candidate's Signature**

This it to certify that the student has incorporated all corrections suggested by the examiner in the thesis and the statement made by the candidate is correct to the best of our knowledge.

**Signature of Supervisor**

# DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

## CERTIFICATE BY THE SUPERVISOR

Certified that **Somya Parashar** (Roll No. 23/BIO/01) has carried out their search work presented in this thesis entitled **"Interpretable Ensemble Learning Predicts Glioblastoma Sensitivity to Natural Compounds"** for the award of **Master of Technology** from Department of Biotechnology, Delhi Technological University, Delhi, under my supervision. The thesis embodies results of original work, and studies are carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

20/05/2025

**Prof. Pravir Kumar (Supervisor)**
Dean, International Affairs
Department of Biotechnology
Delhi Technological University

30.05.25

**Prof. Yasha Hasija**
Head of Department
Department of Biotechnology
Delhi Technological University

Head of the Department
Department of Biotechnology
Delhi Technological University
(Formerly Delhi College of Engg.)
Bawana Road, Delhi-110042

Date: 20.05.2025

# ABSTRACT

Glioblastoma is a belligerent heterogeneous type of brain tumor, inherently difficult to treat with a dismal prognosis. Conventional therapies are rendered ineffectual by various limitations associated with the blood-brain barrier (BBB), tumor microenvironment, and adaptability manifested by the tumors. This study developed an interpretable machine learning framework to predict drug sensitivity in GBM and identify potential new therapeutic candidates. An XGBoost regression model was trained on a curated dataset integrating drug response data (from GDSC with baseline transcriptomic profiles Drug features included 1024-bit Morgan fingerprints and 9 key physicochemical/ADME properties, while cell line features comprised 100 gene expression markers selected via Recursive Feature Elimination. The final model demonstrated strong predictive performance, achieving a mean $R^2 \sim 0.833$ and a mean RMSE ~ 1.060 across five repeated train-test splits. SHAP analysis provided crucial insights into model predictions, identifying key drivers of model learning and the expression levels of GBM-relevant genes like ZEB2 and ABCB6. The model was subsequently used to screen a filtered subset of the COCONUT natural product database, identifying several compounds predicted to exhibit high potency. CNP0152293.3 was the top identified compound.

# LIST OF PUBLICATIONS

1. Journal publication in Ageing Research Reviews “E2 conjugating enzymes: A silent but crucial player in ubiquitin biology.”
2. Journal publication in Ageing Research Reviews “Ubiquitin E3 ligases assisted technologies in protein degradation: Sharing pathways in neurodegenerative disorders and cancer”
3. Poster Presentation in SNCI (Society for Neurochemistry, India) 2025, Delhi Chapter at Jamia Hamdard University on “Identifying Blood-Brain Barrier-Permeable Drug Candidates for PARP1 Targeting in Glioblastoma”

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# LIST OF SYMBOLS, ABBREVIATIONS AND NOMENCLATURE

| | |
|---|---|
| 2-HG | 2-hydorxyglutarate |
| ADME | Absorption, Distribution, Metabolism, Excretion |
| AI | Artificial Intelligence |
| BBB | Blood Brain Barrier |
| CHI3L1 | Chitinase-3-like Protein 1 |
| COCONUT | Collection of Open Natural Products |
| COSMIC | Catalogue of Somatic Mutations in Cancer |
| EGFR | Epidermal Growth Factor Receptor |
| ERK | Extracellular signal Regulated Kinase |
| GABRA1 | Gamma-aminobutyric acid type A receptor |
| GBM | Glioblastoma Multiforme |
| GBM | Gradient Boosting Machine |
| G-CIMP | Glioma CpG Island Methylator |
| GDSC | Genomics of Drug Sensitivity in Cancer |
| GSC | Glioblastoma Stem Cell |
| HGNC | HUGO Gene Nomenclature Committee |
| IDH1 | Isocitrate Dehydrogenase 1 |
| MAE | Mean Absolute Error |
| MAPK | Mitogen-Activated Protein Kinases |
| MGMT | O6-methyguanine-DNA methyltransferase |
| ML | Machine Learning |
| mTOR | Mechanistic Target of Rapamycin |
| MW | Molecular Weight |

| | |
|---|---|
| NF1 | Neurofibromin 1 |
| PANTHER | Protein Analysis Through Evolutionary Relationships |
| PDGFRA | Platelet Derived Growth Factor Alpha |
| PI3K | Phosphoinositidase 3-kinase |
| PTEN | Phosphatase and Tensin homolog deleted on chromosome 10 |
| RB | Retinoblastoma |
| RFECV | Recursive Feature Elimination with Cross Validation |
| RMSE | Root Mean Square Deviation |
| RTK | Receptor Tyrosine Kinase |
| SHAP | Shapley Additive Explanations |
| TCGA | The Cancer Genome Atlas |
| TERT | Telomerase Reverse Transcriptase |
| TMZ | Temozolomide |
| TNF | Tumor Necrosis Factor |
| TPSA | Topological Polar Surface Area |
| XAI | Explainable Artificial Intelligence |
| XGBoost | Extreme Gradient Boosting |

# CHAPTER 1

# INTRODUCTION

## 1.1 Background

Glioblastoma Multiforme (GBM) or Glioblastoma is a belligerent brain neoplasm that exhibits malignant nature [1]. It originates from astrocytes, which are star-like glial cells of the central nervous system. [2]. WHO categorizes Glioblastoma as Grade IV, reflecting its tendency to progress swiftly and invasively in brain or spinal cord [3]. Glioblastoma primarily develop in cerebral hemispheres, that are responsible for multiple functions including reasoning, sensory processing and motor functions [4]. Occurrence of tumor in brain stem or spinal cord is relatively rare.

This malignancy can occur in anyone regardless of age but it is commonly reported in adults of age 45-75 years, with the median age of 64. It exhibits a marginally higher prevalence among men [5]. Glioblastoma can also occur in children; however, the pediatric cases present with different biological characteristics. Primary glioblastomas originate independently in ageing adults, secondary glioblastomas stem from lower-grade diffuse astrocytoma that have undergone malignant progression [6]. Glioblastoma is heterogenous and aggressive, with an inherent tendency of resistance to treatment and recur [7]. Standard treatments include surgical resection, radiotherapy, and chemotherapy, with average survival period about 12-15 months

post-diagnosis [8]. One-year survival rate is about 40% but this dwindles down to less than 10% as five-year survival rate. The blood brain barrier (BBB), the tumor microenvironment and intrinsic resistance mechanisms pose limits and impede effective therapy with conventional treatments [9]. Tumor heterogeneity is a likely a malefactor to developing chemo-resistance, particularly after extended exposure to conventional chemotherapeutics. Variability in cellular subpopulations, genetic mutations and signaling pathways also contribute to the differential drug responses, rendering the standard care of treatments ineffective over time. The dynamic nature of Glioblastoma requires development of adaptable and scalable approaches. This complexity necessitates the use of data-driven strategies such as in-silico modeling and machine learning to accelerate drug discovery and prioritize compounds for laboratory screening.

Computational frameworks offer the unique ability to integrate data from various sources, including high-throughput drug screening, transcriptome profiles of patient-derived or established cell lines and molecular descriptors such as chemical fingerprints and pharmacokinetic properties. Machine Learning (ML) algorithms excel at recognizing non-linear associations that drive the drug response across cancer cells [10]. ML models have been shown to enable identification of novel compounds with high therapeutic potential, including those that have not yet been tested in specific disease conditions [11]. They have also been shown to identify existing drugs as repurposing candidates [12]. ML-based approaches are in the unique position to screen the enormous chemical space and reduce the dependency on empirical methods. Moreover, incorporation of interpretability allows for predictions that are biologically meaningful, such as determine and rank significant pharmacophore scaffolds, or molecular descriptors that contribute to the likelihood of model prediction [13]. Such insights facilitate a deeper understanding of drug-disease association, by highlighting pathways and mechanisms of resistance. Biologically interpretable results inspire confidence in drug development efforts and streamline the preclinical validation.

## 1.2 Literature Review

Glioblastoma is aggressive and highly heterogeneous, with multiples mechanisms contributing to its malignancy and resistance to treatment [7]. Molecularly, Glioblastoma has a complex landscape of genetic and epigenetic alterations. Based on such information, The Cancer Genome Atlas (TCGA) categorizes glioblastoma in 4 molecular subgroups [14]. Classical glioblastoma has chromosome 7 amplified with loss of chromosome 10; *EGFR* amplification or *EGFRvIII* mutation and exclusive disruption of RB pathway via deletion of the tumor suppressor gene *CDKN2A* while strong expression of Notch and Sonic hedgehog signaling [14]. Mesenchymal subtype correlates with the deletion of *NF1* tumor suppressor located on chromosome 17, expression of *CHI3L1* and *MET*, and enrichment of genes of TNF superfamily and NF-kB signaling like *TRADD, RELB,* and *TNFRS1A* [14]. Genes typically expressed in neurons *NEFL, GABRA1, SYT1* and *SLC12A5* are strongly upregulated in Neuronal glioblastoma, the least understood with limited studies [14]. Modifications in *PDGFRA* and mutations in *IDH1*, *TP53* are commonly observed in the Proneural subtype, and interestingly, mutations in the *PIK3CA* have also been reported mutually exclusive of *PDGFRA* abnormalities [14]. Among the identified subtypes – mesenchymal is the most aggressive and invasive and has been found to exhibit resistance to multiple therapies, leading to worse prognosis [15].

### 1.2.1 Genetic and Epigenetic Disruptions in Glioblastoma

Disruptions in several signaling pathways through various combinations of mutations, copy number variations and gene fusions drive glioblastoma development and progression. Key genetic disruptions include *EGFR* amplification, reported in 40-50% of primary glioblastoma, often accompanied by *EGFRvIII* mutation, *PDGFRA* amplification (13%) and *MET* receptor (4%) [16]. Mutations or deletions in the *PTEN* tumor suppressor gene (30-40%) result in unchecked activation of the PI3K-AKT cascade [17]. Mutated *TP53* is roughly reported in one-third glioblastoma cases,

prevalent in *IDH*-mutant or proneural subtypes [18]. Aberrations like +7/-10, related with the classical subtype is a characteristic of primary glioblastoma [19]. Point mutations present upstream of *TERT* occur in nearly 83% of primary glioblastoma cases, enabling the expression of catalytic subunit of telomerase [20]. Alteration of the RB pathway occurs directly by mutations, deletions, or epigenetic changes at the *RB* locus, and gene amplifications of *CDK4*, *CDK6* and *CCND2* have been reported [21]. Multi-faceted dysregulation of major pathways drives glioblastoma pathogenesis.

Epigenetic dysregulation also contributes to disease pathogenesis. *IDH1* mutation results in intracellular buildup of the oncometabolite 2-HG, causing a global hypermethylation phenotype of G-CIMP (Glioma CpG Island Methylator) that silences several genes like *RBP1* and *G0S2* [22]. MGMT promoter methylation occurs in ∼45% of adult patients, suppressing DNA repair enzyme O6-methylgaunine-DNA methyltransferase and leads to increased sensitivity to Temozolomide [23]. In quiescent Glioblastoma Stem Cells (GSCs), H3K4me3 and H3K27me3 keep the genes in a transcriptionally poised state to rapidly reactivate proliferation and differentiation pathways upon receiving right signals [24]. *KDM5B* and *KDM6A/B* are overexpressed in quiescent GSCs, regulating the transition from quiescent to proliferative state by demethylation of H3K4 and H2K27, respectively [25]. *KDM4C*, demethylase for H3K9, induces c-Myc expression and inhibits apoptogenic tendencies of p53 [26]. *UBE2V2* that regulates HK16 acetylation is also highly expressed in GBM marks similarity between GSCs and embryonic progenitor cells [27]. BPTF recognizes H3K4me3 and H6K16ac marks, supports GSC maintenance and self-renewal [28]. Chromatin remodeling enables GSCs to adopt a therapy-resistant phenotype by dynamic modification of histone marks. Tumor suppressor miRNAs like miR-7, miR-34a, miR-128 and OncomiRs miR-10B, miR-21, miR-93 regulate cell survival, proliferation and migration, apoptosis, invasion, angiogenesis, stemness, radioresistance and chemoresistance in glioblastoma [29]. lncRNAs CASC7/9, AGAP2-AS1, NEAT1, LINC1426, LINC01446, PART1, MNX1-AS1, DCST1-AS1, AC016045.3, HOTAIRM1, lnc-TALC, MALAT1 promote tumorigenesis, proliferation, invasion, angiogenesis and TMZ resistance in glioblastoma cell lines and

tissues [30]. The genetic alterations drive tumor growth while the epigenetic changes reinforce malignant gene expression.

### 1.2.2 Key Signaling Pathways in Glioblastoma Progression

Despite the complex nature of mutations in glioblastoma, it has been noted that three signaling pathways are dysregulated in majority of glioblastoma cases. Amplifications in *EGFR, PDGFRA* and *MET* constitutively activate RTK/PI3K/AKT/mTOR cascade, ensuing cellular proliferation, growth and survival [31]. Downstream of RTKs, mutational events in *PI3KC* or loss of *PTEN* activity further establish that the PI3K pathway is highly active in glioblastoma, enhancing cell motility and invasiveness [31]. The p53 tumor suppressor pathway is silent in glioblastoma. Mutations and deletions in the *TP53* gene, and amplified *MDM2*, *MDM4* genes that code of p53 degrading proteins ensure that DNA damage response, apoptosis and cellular senescence is compromised in glioblastoma [32]. Deletion of *CDKN2A* encoding MDM2 inhibitor, ARF, also blunts the p53 mediated cell-cycle arrest [33]. The Rb tumor suppressor pathway regulates the G1-S checkpoint in cell cycle [34]. In glioblastoma, *CDKN2A* deletion allows for overexpression of the CDK4 and CDK6 [35]. Disease pathogenesis requires a combination of proliferative signaling and loss of two tumor suppressor genes. Other pathways also contribute to glioblastoma development. Aberrant activation of developmental pathways like Notch, Wnt/β-catenin and Hedgehog occurs in glioblastoma cells and GSCs [36]. *VEGF* is also commonly upregulated and angiogenesis pathways contribute to tumor development [37]. The mesenchymal transformation correlates with the activation of NF-kB and STAT3 signaling, arising from *NF1* or *PTEN* loss [38]. Multiple oncogenic activations with epigenetic dysregulation synergistically contribute to disease progression through dynamic reprogramming of gene expression.

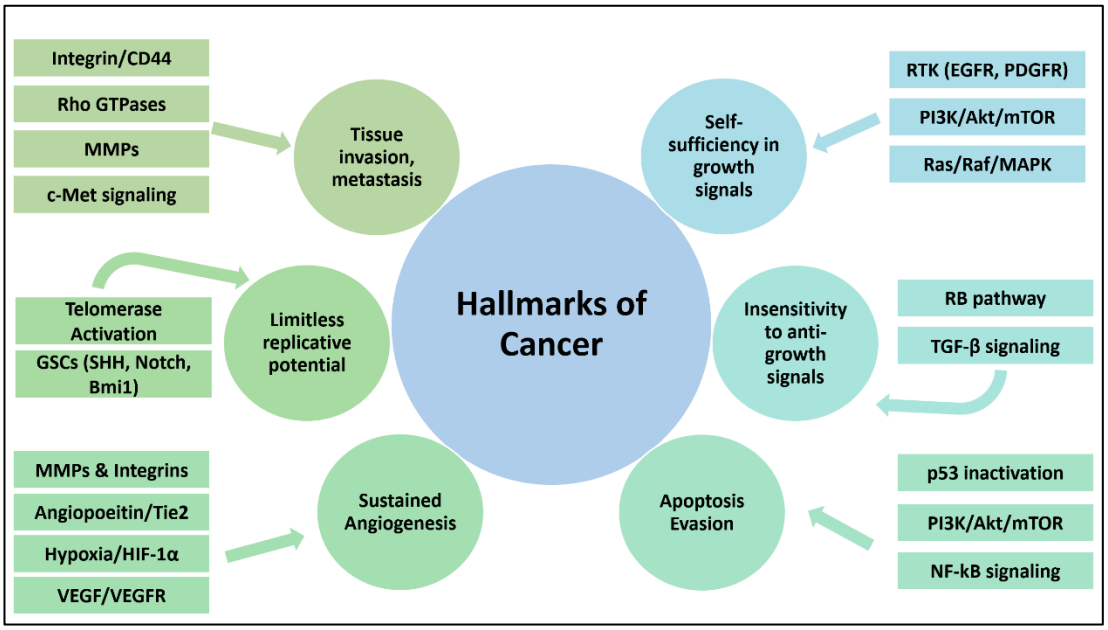


**Fig 1.1 Dysregulated pathways in glioblastoma contributing to cancer hallmarks**.

### 1.2.3 Glioblastoma Cell Lines and Their Characteristics

Preclinical research on glioblastoma relies heavily on in vitro cell line models. A number of human glioblastoma cell lines have been established from patient tumors, propagated long-term in culture to recapitulate specific genetic, epigenetic and phenotypic tumor characteristics. Most frequently used cell lines for *in vitro* studies include U87, U251, LN-229, A172 and T98G. A literature survey revealed that three most used cell lines were U-87 (60%), U-251(41%), followed by T98G (26%) [39]. These cell lines are a popular choice for investigative studies due to their robust proliferative ability and the extensive baseline information available for them. The U87 (Uppsala 87 Malignant Glioma), established in 1968 in University of Uppsala, Sweden, forms large vascularized tumors [40], expressing mutant *PTEN*, *PI3K* and *Akt* and deletions in *CDKN2A* and *ARF* [41]. U251 cell line, established in 1973 at the same university displays high infiltrative and invasive growth pattern, mutant PTEN, upregulated PI3K and Akt, non-functional p53 and aberrantly expressed cell-cycle control proteins, retaining astrocytic lineage markers such as GFAP, S100β [40], [41]. T98G is the polyploid variant of T98, is anchorage-independent tumor line on account of the highly expressed *ACTA2* gene which contributes to cell motility, mutant *TP53*

and *PTEN* [42], [43]. The LN-229 cells possess a methylated MGMT promoter, mutated TP53, homozygous deletions in the p16 and p14ARF, and undergo apoptosis induced by the Fas ligand [44].

Use of cell lines offers a convenient and reproducible platform for in-vitro analyses. Indefinite culturing under controlled conditions allows for high-throughput screening of drug candidates and genetic alterations. Due to human origin of these cell lines, findings can be often translated to human tumor biology. Cell lines have been instrumental in developing standard treatments, for e.g. TMZ [45]. However, cell lines do not fully express the heterogeneity and microenvironment of disease in vivo. Cell culture also lacks the three-dimensional architecture, hypoxic gradients, and interactions with immune cells, vasculature and stroma present in tumors. This may lead to discrepancies in studying drug response because the tumor microenvironment is known to induce protective stress response, or prevents the drug from reaching tumor cells [46]. Researchers often employ patient-derived xenografts and GSC neurospheres that likely preserve tumor phenotype [47]. Nevertheless, traditional cell lines like U87 and U251 remain standard in glioblastoma research — often used as stepping stone for generating hypothesis before moving into complex models.

### 1.2.4 Standard Care of Treatment and Therapeutic Challenges

Currently, standard care regimens particularly for the IDH-wildtype Grade IV tumors follows the Stupp protocol, which consists of extensive surgical removal of the tumor, radiation therapy and concurrent TMZ administration [48]. Surgical resection aims to reduce tumor burden and is facilitated by techniques such as 5-ALA fluorescence guidance and intraoperative mapping, which helps preserve neurological function [49]. Radiotherapy typically involves fractionated dosing of 60 Gy over six weeks, while chemotherapy with TMZ is administered daily during radiotherapy and followed by six adjuvant cycles [48]. TMZ efficacy is closely associated with the methylation of the *MGMT* promoter, with methylated tumors demonstrating better therapeutics responses [50]. Additional therapies, namely Tumor Treating Fields (TTFs), have shown modest improvements in survival when combines with standard therapy but

remain limited by high cost and logistical challenges [51]. Despite the aggressive nature of treatment regimens, the prognosis of glioblastoma remains dismal, with average survival around 15 months and 7% rate of survival past 5 years. This depressing outlook is driven by several challenges — the high invasive nature of tumor, intra-tumor heterogeneity, and the presence of therapy-resistant GSCs.

Furthermore, BBB constitutes significant challenge to effective drug delivery, severely restricting therapy agents. BBB is a specialized endothelial barrier in cerebral capillaries that strictly regulates entry of molecules into the brain parenchyma [52]. Tight junctions between endothelial cells, along with pericytes and astrocytes, exclude nearly 98% of small molecule drugs from the brain under normal conditions [52]. In disease state, the BBB is partially disrupted; glioblastoma induced angiogenesis produces leaky, abnormal vessels and expression levels of transport and junction proteins are altered tumor vasculature [53]. BBB disruption in glioblastoma is highly heterogenous and often incomplete. Studies report that all glioblastoma patients harbor significant regions where the BBB remains intact [54]. BBB thus contributes to treatment resistance by limiting drug delivery. Most chemotherapeutic agents and targeted kinase inhibitors have poor BBB penetration [55]. TMZ, which is BBB-permeable, may not reach all tumor cells uniformly due to regional blood flow differences. Methods of circumventing BBB are currently being explored —osmotic or ultrasound-mediated BBB disruption can transiently open tight junctions, and design of BBB-penetrant drug analogues and nanoparticles [56], [57], [58]. Some early-phase trials such as focused ultrasound to open the BBB for chemotherapy have shown that repeated safe BBB modulation is feasible [59]. BBB's role in glioblastoma is two-fold. It contributes to glioblastoma progression by fostering a protective niche for tumor cells, and promoting selection of invasive cells that migrate into healthy brain, as well as acting as a central factor in therapy resistance by preventing uniform drug delivery.

Additionally, the tumor microenvironment also presents another therapeutic hindrance due to its complex and immunosuppressive nature, comprising of GSCs, microglia, macrophages, neutrophils, lymphocytes, and neuronal cells, interacting dynamically to promote tumor growth, progression, and resistance to therapy [60]. GSCs contribute

to therapeutic resistance by secreting chemokines and factors promoting angiogenesis, that facilitate endothelial cell growth and attract macrophages, leading to immunosuppressive milieu [61]. These immunosuppressive characteristics hinder the efficacy of immune-based treatments, as glioblastoma's unique brain profiles and cellular heterogeneity limit the benefit of such therapies [62]. Interactions between the glioblastoma cells and tumor microenvironment can induce resistance to both chemotherapy and radiotherapy [63]. The astrocytes within the microenvironment assist in cell survival, promoting drug resistance and forming physical barriers that prevent therapeutic agents from reaching tumor cells [64]. Recurrence is nearly universal, and treatment options in the recurrent setting remain palliative, with no universally accepted standard of care. These barriers necessitate the urgent need for more targeted, penetrant, and adaptive therapeutic strategies that can address both the molecular complexity of glioblastoma and its protective microenvironment.

### 1.2.5 Machine Learning in Cancer Research

Computational models and machine learning are increasingly being investigated as powerful tools to decode tumor complexities, predict therapeutic response, and accelerate the discovery of more effective interventions, in oncology and in glioblastoma. This shift in paradigm of research comes from availability of biomedical data such as genomics, transcriptome profiles, methylation status, imaging of tumor sizes etc. from patient cohorts, as well as *in vitro* preclinical studies.

Artificial Intelligence (AI) is mainly an avenue devoted to studying and development of algorithms and systems that enforce machines to execute tasks that necessitate human-like intelligence [65]. AI is a computer science discipline rooted in mathematical theorems and statistical techniques to make predictions after learning from training dataset and evaluating on test data set. AI is used in automation of functions such as reasoning, problem-solving, decision-making, perception such as learning from data [66]. Machine Learning (ML) is an AI subset that concentrates on developing models and devising algorithms to explore and scrutinize data without being explicitly programmed [67].

Supervised Learning uses class labels to learn and predict those outcomes for new data and is broadly utilized for classification and regression endeavors in biomedical research such as oncology [68]. A supervised ML model might be trained to classify tumor biopsy samples as benign or malignant based on labeled training images, or to predict a patient's survival time from past patient data. Methods such as support vector machines and neural networks have accurately mapped inputs to outputs in tasks like tumor type classification and gene expression-based drug sensitivity prediction [69]. The success of success of supervised learning in cancer is evident in studies where classifiers learned to distinguish cancer subtypes of leukemia or predict treatment response with high accuracy [70]. Supervised methods require datasets with class labels to facilitate model learning [71]. Unsupervised Learning pursues patterns in data without class labels. Clustering and dimensionality reduction techniques can discover hidden patterns without pre-defined labels. In oncology, unsupervised learning approaches have been employed to describe novel patient subgroups and tumor subtypes [14]. Principal Component Analysis (PCA) and hierarchical clustering have also helped visualize highly dimensional genomic data and uncover features that differentiate patient subpopulations [72]. These methods generate hypotheses and insights that supervised methods might miss, though linking clusters back to clinical outcomes requires rigorous analysis.

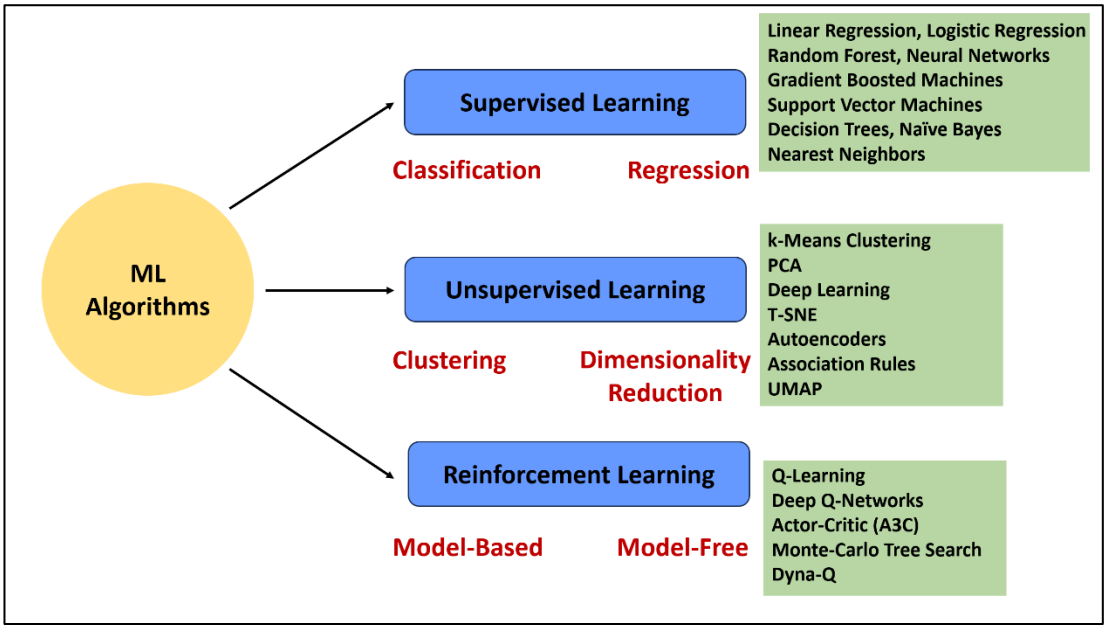


**Fig 1.2 Types of Machine learning algorithms**

Supervised ML uses many algorithms, as per the problem requirement. These include linear regression, logistic regression, decision trees, SVMs, k-nearest neighbors, naïve Bayes, random forest [71].

Linear regression is used to predict a continuous dependent variable outcome based on one or more input features that are independent [73]. A linear relationship is presumed between response and predictor variables, and the outcome is modeled as a weighted sum of the input features and a bias term. The optimal coefficients are determined by minimizing the cost function, which is typically the mean squared error (MSE) — so that there is best fit for the observed data. It is widely used in biomedical research to model dose-response relationships and prognostic score estimation [74], [75].

Logistic regression is a classification algorithm that models the probabilities of a binary or multi-class categorical outcome as a function of input features. Sigmoid function is applied to map prediction to [0,1], representing class probabilities, and maximum likelihood function is used to train the model [76]. It finds use in clinical outcome prediction, biomarker-based disease classification, and epidemiological modeling due to its interpretability and statistical robustness [77], [78].

SVMs served tasks of both classification and regression by identifying an optimal hyperplane that maximizes the margins between different classes in a high-dimensional space [79]. Kernel functions such as polynomial, radial, are used to model complex non-linear relationships due to their ability to handle sparse data and robustness to overfitting [80]. They are particularly convenient for high-dimensional datasets like gene expression profiles [81].

The k-nearest neighbor algorithm is a non-parametric, instance-based learning method that classifies or regresses a data point based on the label of its k closest neighbor [82]. Euclidean distance is the metric for measuring distance and this algorithm does not require a training phase — decisions are made at the time of inference, which can be computationally intensive. It has bene used in cancer subtype classification and phenotype clustering, but it is sensitive to data scaling and irrelevant features [83], [84].

Naïve Bayes is a probabilistic classifier based on Bayes' Theorem and it assumes independence between input features [85]. It estimates the posterior ability of each class and selects the one with the highest probability. It is speedy, simple and effective in multi-class classification problems such as tumor-type predictions and microbiome-based disease classification [86], [87].

A decision tree is a non-parametric supervised learning algorithm that partitions the feature space into a hierarchy of decision rules based on input features [88]. Each internal node represents a decision based on a feature threshold, and each leaf node represents an output label or value [89]. Tree construction occurs through recursive binary splitting to minimize impurity measures *e.g.,* Gini index for classification and MSE for regression. Decision Trees offer the advantage of being interpretable, can suitably model non-linear associations, making them valuable for transparent decision making required in clinical settings [90].

Ensemble learning methods combine multiple models to produce a robust predictor. Outputs of diverse learners are aggregated to achieve higher accuracy and generalization [91]. Ensemble models have been applied to tasks like gene expression-bases prognosis, prediction and radiomic image analysis, frequently outperforming large models. An ensemble allows for integration of predictions from separate models, such analyzing imaging and genomics to improve the overall accuracy in predicting a patient's outcome [92], [93]. Common methods include random forests and gradient boosting machines. Random Forest is an ensemble learning algorithm that constructs a large number of decision tress during training and outputs the mode or mean for classification and regression, respectively [94]. Each tree is trained on a bootstrapped subgroup of the features when nodes are split, that enhances model diversity and reduce overfitting [94]. Random forests are highly robust and interpretable to an extent owing to feature importance, and are widely used in biomarker discovery, clinical prognosis, and treatment response modeling [95], [96], [97]. In a recent review of ML models for glioblastoma survival, random forest was the single most popular algorithm among researchers [98]. Gradient Boosting builds a strong predictor by combining multiple weak learners, typically decision tress, in a sequential manner [99]. Unlike bagging techniques used in random forest, which trains trees independently on

bootstrapped datasets, gradient boosting fits each new model to the residual errors of the combined ensemble so far [100]. At each iteration, a specific loss function is minimized using gradient descent approach. This iterative correction of residuals leads to high model accuracy and flexibility, making gradient boosting well suited for genomics and clinical prediction tasks. Gradient boosting supports custom loss functions and regularization techniques, for e.g., learning rate, tree depth limits, shrinkage, which enhance its robustness and prevent overfitting. It can be implemented using multiple algorithms. Gradient Boosting Machine (GBM) is the classical implementation, introduced by Friedman, where each tree is added in a stage-wise manner to correct the errors of the ensemble so far [101]. LightGBM is developed by Microsoft and improves training efficiency and memory usage by utilizing histogram-based binning and leaf-wise growth strategy, allowing handling of large and highly dimensional datasets [102]. CatBoost, developed by Yandex, handles categorical features natively without preprocessing and uses symmetric trees for better generalization [103]. XGBoost or Extreme Gradient Boosting is actually an optimized version of GBM that includes regularization, sparsity awareness, and parallelized tree construction [104]. It is widely used in biomedical ML tasks due to its speed, scalability and superior performance.

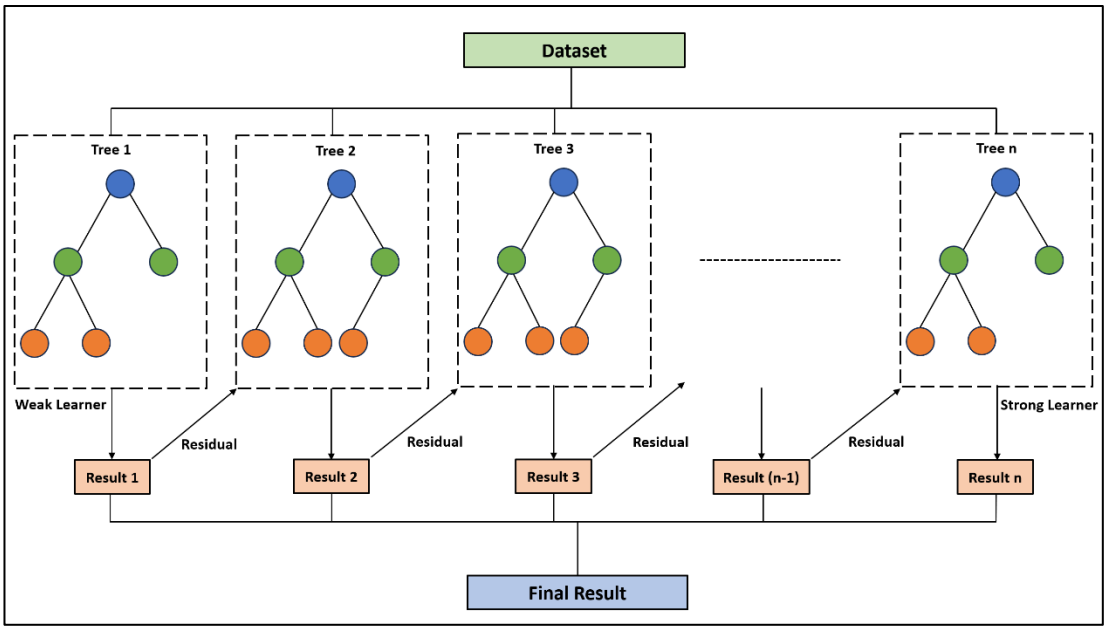


**Fig 1.3 Structure of XGBoost algorithm**

Reinforcement Learning revolves around rewarding and penalizing actions in an environment and adjusting behavior accordingly, to improve decision making over time and maximize cumulative reward. In oncology, it has been applied to treatment pathway optimization [105], adaptive dosing of anti-cancer drugs [106], and adaptive radiotherapy for better clinical decision support [107].

As the clinical space grows exponentially, application of ML is poised to become a regular feature before undertaking any preclinical or animal model studies. For trust and confidence in predictions made by the ML models, it is necessary that these models inspire confidence in their outcomes and are interpretable. An interpretable ML model is one whose predictions are readily understood by humans, the simplest example of which is the decision trees, with their clear if-then rule-based logic. Unlike black-box neural networks, interpretable models provide transparency, which is important for acceptance in clinical settings due to safety concerns. Absence of explainability can introduce safety concerns and erode trust in ML recommendations.

To address this, researchers either design inherently interpretable models or apply explainability techniques to complex models. Explainable AI (XAI) studies such approaches that enable model working to human comprehension and reasoning [108]. It contributes to meaningful deployment of framework that base their decisions on relevant and justifiable features. SHAP, abbreviated for Shapley Additive exPlanations, attributes model predictions to individual features by computing Shapley values, that represent the average impact on model predictability [109]. Permutation-based feature importance assesses the feature significance by computing performance loss upon feature randomization [110]. LIME, or Local Interpretable Model-agnostic Explanations, give a local justification by considering a subset of data when approximating explanations for model inputs and outputs [111]. Grad-CAM is a local method use for explain convolutional neural networks commonly used for image classification, by producing heatmaps highlighting the image regions most important for model prediction [112]. The interest in XAI methods in oncology is on the rise. A recent study predicting glioblastoma patient outcomes chose a simple classification tree over a more complex model specifically for its high interpretability, noting that tree's decision path could be readily understood and validated [113]. This is a growing

area of research where interpretable or explainable ML allows for data-driven insights to be translated into practice with transparency and accountability.

### 1.2.6 ML Approaches in Glioblastoma

ML has become a focal apparatus in oncology, enabling the discovery of hidden pattern in biomedical data to improve cancer diagnosis, prognosis and treatment selection. Methods like ANNs and decision trees were applied to cancer diagnosis as early as in the 1980s [114]. Over the past decades, increasing computing power and the explosion of genomic and imaging data has propelled ML into widespread use. Now ML techniques are used across all facets of oncology, from risk assessment and early detection to outcome prediction and therapy planning [115]. Well-designed ML models can outperform traditional approaches; one AI system achieved 99% accuracy in recognizing breast cancer metastases in pathology images in contrast to 81% by pathologists [116]. Such examples illustrate the potential of ML to assist in clinical decision making with improved speed and accuracy.

Medical image analysis is a very prominent field where ML models are deployed. Deep Learning, although a black-box model, can analyze radiological images and digitized histopathological slides to detect tumors and classify malignancies with great accuracy [117], [118]. CNNs have been shown to outperform multiple human pathologists in classifying tumor histology, highlighting the value of ML in diagnostics [119]. A landmark study by the TCGA employed unsupervised learning to cluster glioblastoma tumors into molecular subtypes using gene expression data [14]. This discovery has deepened biological understanding and suggested tailored treatment strategies. ML models have also been used for predictive oncology, by integrating multi-omics data to predict the likelihood of recurrence or survival, helping to stratify patients by risk in breast cancer.

In glioblastoma specifically, researchers are exploring ML models in multiple ways to manage the disease. In a study using the deep-feedforward ANNs, survival in glioblastoma was predicted with 90% accuracy using multimodal neuroimaging data [120]. Multivariate decision tree models were built to develop image-based

biomarkers, owing to regional genetic diversity in tumor segments leading to associations between copy number variations (CNV) and localized imaging features [121]. Many models have been built to predict *IDH1* mutations using MR-based radiomics data through algorithms like random forest [122] and gradient tree boosting [123]. SVM-based model have been applied to delineate regions of necrotic tissue in patients being treated with radiation and chemotherapy [124]. *PTEN* mutation status could be predicted using deep learning radiomics [125] and SVM [126]. The tool GBMDriver is built using three different algorithms Adaboost, SVM and XGBoost for classifying glioblastoma mutations as disease-driving or neutral [127]. An integrated model, developed using the tumor infiltrating lncRNAs, identified patients that would benefit the most from immunotherapy [128]. Multiple predictive models for glioblastoma patients survival exist [129], [130], [131], [132], [133]. Predictions of *MGMT* methylation through genetic algorithms [134] and deep learning-based approaches [135] have elaborated the much-needed spotlight on epigenetic features governing disease progression. However, in the scope of novel compounds or drug discovery, fewer ML models exist that either target singular proteins like FAK [136] or target single cellular type like C6, U251 and U87 without interpretability [137].

While these studies are promising, glioblastoma has relatively fewer cell lines and patient-derived cultures. This leads to overfitting or low confidence in the model. Transfer learning and multi-cancer datasets have been used to mitigate this issue but the assumption that other cancer data can inform glioblastoma predictions persists [138]. Multiple models with high accuracies are black-boxes and hence do not readily explain the reason behind a prediction made. This clashes with the clinical context where biological rationale is imperative to consider the prediction worthwhile. The integration of SHAP in ML models for oncology is limited to risk assessment [139], diagnosis [140], survival prediction [141], treatment recommendation [142], and to some extent classification [143].

Currently no models exist that integrate drug discovery efforts with explainable modules of ML. There remains a notable lack of ML models that integrate chemical features of drugs with molecular profiles of glioblastoma cell lines in an interpretable framework. There are no published studies that provide explainable drug sensitivity

prediction models that allow researchers to trace back predictions to actionable biological or chemical features. Moreover, explainable ML techniques such as SHAP or LIME have not been systematically applied to jointly analyze drug descriptors and genomic features in the context of glioblastoma therapeutic efforts. This gap severely limits the translation of *in silico* predictions into experimental designs, hindering efforts to prioritize candidate compounds. Addressing this shortcoming could provide a scalable, cost-effective framework for rational drug discovery using natural or understudied compounds.

### 1.3 Objective

To develop an interpretable ML framework capable of predicting drug sensitivity in glioblastoma cell lines, measured as the half-maximal inhibitory concentration $IC_{50}$, validate the predictive performance using statistical metrics, propose putative targets and map the predicted targets to appropriate pathways using pathway enrichment analysis.

# CHAPTER 2

# METHODS

## 2.1 Dataset Selection

A comprehensive drug sensitivity dataset was curated by integrating pharmacological response data from both the GDSC1 and GDSC2 releases of the Genomics of Drug Sensitivity in Cancer (GDSC) database employing a 29 glioblastoma cell line panel [144]. GDSC1 was initially considered for its broader compound coverage. However, to enhance data quality and precision, overlapping drug-cell line measurements from GDSC2 were preferentially used to replace corresponding entries from GDSC1. GDSC2 implements improved experimental protocols, including acoustic compound dispensing (Echo555) for greater accuracy and CellTiter-Glo luminescence-based viability assays for enhanced sensitivity.

Both GDSC1 and GSDC2 datasets utilize a consistent computational pipeline for IC50 calculation. It involves fitting a sigmoidal dose-response model via non-linear regression using the gdscIC50 R package and subsequently log-transformation of IC50 values in μM [145]. This standardized downstream processing allows for direct replacement of GDSC1 IC50 values with more precise and reproducible value for identical drug-cell line pairs, allowing for a harmonized dataset.

Transcriptome profiles were generated by processing the raw Affymetric Human Genome U219 array data, ArrayExpress accession E-MTAB-3610 for these cell lines using the Robust Multi-array Average (RMA) normalization procedure using the

Bioconductor R package. Cell line identity between pharmacological and genomic features was confirmed using cell line metadata from COSMIC. These integrated features were used to develop the ML matrices for drug-sensitivity model.

## 2.2 Feature Generation and Selection

To prepare the data for ML, distinct feature sets were generated for both drugs and the cell lines. Subsequently, a feature selection strategy was employed to identify the most predictive molecular features.

### 2.2.1 Drug Feature Generation

Canonical SMILES strings for each compound were programmatically retrieved using their names via the PubChem database (using the pubchempy Python library) [146]. Compounds which were returned with no SMILES were manually checked in other databases like LINCS data portal [147] and Therapeutic Target Database [148]. Compounds for which SMILES could not be obtained were excluded from further analysis. A total of 435 unique compounds were used for further investigation.

For each compound, Morgan fingerprints were generated using the RDKit cheminformatics toolkit [149]. These fingerprints were calculated with a radius of 2 and hashed into a 1024-bit vector. Each bit in this vector reflects the inclusion or absence of a particular circular chemical substructures within the molecule.

The SwissADME web server (http://www.swissadme.ch/index.php) was utilized to compute a total of 9 physicochemical and ADME (Absorption, Distribution, Metabolism, Elimination) related properties were calculated [150]. These included Molecular Weight (MW), number of rotatable bonds, number of H-bond donors, number of H-bond acceptors, Topological Polar Surface Area (TPSA), Consensus Log P *i.e.* lipophilicity, ESOL Log S *i.e.* predicted aqueous solubility, BBB-permeability predicted as Yes/No, encoded as 1/0, and the number of Lipinski rule violations [151].

### 2.2.2 Cell Line Feature Generation

Baseline gene expression data for 29 glioblastoma cell lines, derived from RMA-normalized Affymetrix Human Genome U219 arrays was used as cell line features.

Probeset IDs from the microarray data were systematically converted to official HUGO Gene Nomenclature Committee (HGNC) gene symbols using the hgu219.db Bioconductor annotation package in R. In instances where multiple probes were mapped to singular gene symbol, the median expression value was taken as the representative expression level for that gene in each cell line. This resulted in an initial expression matrix of approximately 19,434 unique genes.

### 2.3 Assembly of Paired Feature Matrix (X) and Target Vector (Y)

The drug features and cell line gene expression features were combined with the log-transformed IC50 values to create the final dataset. Each row in the dataset corresponded to a specific drug-cell line experiment. The feature vector (X) for each row was constructed by concatenating the drug's feature vector (Morgan fingerprints with physicochemical and ADME properties) with the complete baseline gene expression profile of the corresponding cell line. The target variable in each row was the experimental $\log(IC_{50})$ value for the given drug-cell line pair. This assembly created a dataset of samples, each distributed by features.

### 2.4 Gene Feature Selection using RFECV

Given the high dimensionality of the transcriptome, Recursive Feature Elimination with Cross Validation (RFECV) was utilized to select an optimal subset of the most predictive gene features, while retaining aa drug-derived features. An XGBoost Regressor model was used as the estimator within the RFECV framework, configured with GPU acceleration. Key parameter for this estimator included n_eestimator=100, learning_rate=0.1, and max_depth=6. RFECV iteratively assessed the feature importances, and removed the least important gene features in steps of 1000 genes. This process was guided by 5-fold cross validation, optimizing for the negative MSE. An optimal subset of 100 gene expression features that maximized cross-validated predictive performance was identified. These 100 selected genes, along with all drug features, were used for training the final predictive model.

### 2.5 Model Training and Evaluation

A definitive predictive model was trained utilizing the optimized feature set. The full dataset, comprising 11,588 drug-cell line pair samples was utilized. The feature vector for each sample consisted for the 33 drug-derived features combined with the 100 gene expression features selected by RFECV, resulting in a total of 1133 features per sample. An XGBoost regressor algorithm was employed for this final model, configured for GPU acceleration with key hyperparameters.

To robustly assess performance and generalizability, a repeated train-test split strategy was implemented. The data was randomly partitioned into training (80%) and testing (20%) set five independent times, each with a different random seed. The model was trained on each training fold and its predictions evaluated against the corresponding unseen fold test. Performance was quantified using Root Mean Squared Error (RMSE), R-squared ($R^2$), Pearson correlation coefficient, Spearman rank correlation, and Mean Absolute Error (MAE), with the final reported metrics being the mean and standard deviation across these five repeats.

### 2.6 Model Interpretation using SHAP Analysis

To obtain insights of the trained XGBoost model and identify key feature contributions, SHAP analysis was conducted. The shap.TreeExplainer, specifically designed for tree-based ensemble models, was applied to the final XGBoost model trained in the last repeat of the evaluation phase. SHAP values were computed for a representative subset of samples from the corresponding test set to determine the effect of each feature on individual predictions. Global feature importance was assessed by calculating the mean SHAP value for every feature across all explained samples, visualized using summary bar plots and beeswarm plots. The beeswarm plots additionally illustrated the distribution and direction of feature effects relative to their actual values. To further understand the association between specific feature values and their influence on the predicted log($IC_{50}$), SHAP dependence plots were generated for the top-ranking features, also revealing potential interaction effects between

features by coloring points based on a second, automatically selected interacting feature.

### 2.7 Drug Group Analysis for Key Fingerprint Bits

To elucidate the potential chemical or mechanism-based significance of the most influential drug fingerprint bits identified through SHAP analysis, a group-based pathway analysis was performed. For each high-impact fingerprint, all drugs from the initial training set of 435 compounds were identified. These drug lists were then cross-referenced with their own primary targets and targeted pathway available in GDSC1 and GDSC2 datasets. The frequency of target pathways within each drug group was then tabulated to identify commonly targeted pathways, thereby allowing inference of the likely chemical class represented by the important fingerprint bit.

### 2.8 *In Silico* Screening of COCONUT Compound Library

To identify potentially novel anti-glioblastoma compounds, the trained XGBoost model was used to screen the COCONUT (Collection of Open Natural Products) database [152].

Compounds were extracted from the COCONUT database. For each compound, the canonical SMILES string was obtained. A multi-step filtering process was applied to refine this initial library. Only those compounds annotated at level 5 were considered. Other filters on drug features were applied after studying their feature contributions to the model learning through SHAP analysis, such as MW < 800, TPSA between 100 and 200 $Å^2$, H-bond donors 3-5, H-bond acceptors < 10, Rotatable bonds < 10. This filtering resulted in 290 compounds whose pharmacokinetic properties were obtained through the Swiss-ADME web browser. Morgan fingerprints of the selected compounds were generated using RDKit. A representative glioblastoma gene expression profile was created by averaging the RMA-normalized gene expression of selected 100 genes across all 29 cell lines in the training dataset. A single vector representing the average glioblastoma transcriptomic profile was thus generated.

For each COCONUT compound, the feature vector was concatenated with the fixed average gene expression vector, and this combined feature vector was fed into the trained XGBoost model to obtain a predicted log($IC_{50}$) value. The screened compounds were subsequently ranked based on their predicted values, with the lower values indicating higher predicted potency. Putative targets of the top 10 compounds were predicted through the SuperPred web server [153]. The 14 unique putative target genes were then subject to enrichment analysis through the ShinyGO v0.82 PANTHER as the pathway database and FDR cutoff of 0.05 using Uniprot IDs [154], [155].

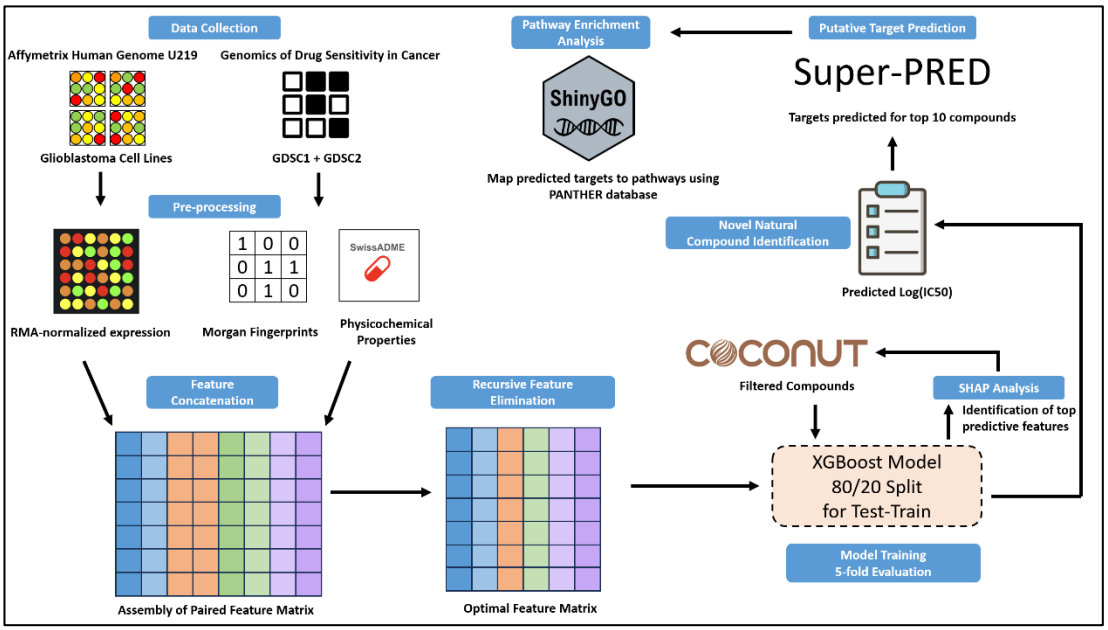


**Fig 2.1. Methodology**

# CHAPTER 3

# RESULTS

## 3.1 Predictive model based on XGBoost demonstrates robust performance in glioblastoma drug sensitivity prediction

An XGBoost regression model was developed to predict drug sensitivity *i.e.* log($IC_{50}$) in glioblastoma cell lines using integrated drug chemical features and cell line transcriptome data. RFECV identified an optimal subset of 100 gene expression features that when combined with all 1033 drug features, yielded the best performance.

The model performance was rigorously evaluated with 5 repeats of an 80/20 split. Across these repeats, the model demonstrated strong predictive accuracy and robustness, as noted in Table 1.

**Table 1. Model Evaluation Metrics**

| Evaluation Metric | Value |
|---|---|
| RMSE | 1.0600 ± 0.0225 |
| $R^2$ | 0.8332 ± 0.0083 |
| Pearson Correlation Coefficient | 0.9133 ± 0.0047 |
| Spearman Rank Correlation | 0.8766 ± 0.0063 |
| MAE | 0.8101 ± 0.0155 |

RMSE indicates that the model's predicted log($IC_{50}$) typically deviates by a factor of 1.0600 ± 0.0225 from the experimental values. The variance in drug sensitivity is around 0.8332 ± 0.0083. Furthermore, a high degree of correlation was observed between predicted and actual sensitivities, as shown by the Pearson correlation coefficient of 0.9133 ± 0.0047 0063 ($p < 0.001$ for all repeats) and Spearman rank correlation of 0.8766 ± 0.0063 ($p < 0.001$ for all repeats). The Mean Absolute Error was 0.8101 ± 0.0155. These metrics indicate that the model can accurately predict log($IC_{50}$) and explain over 83% of the variance in drug sensitivity.

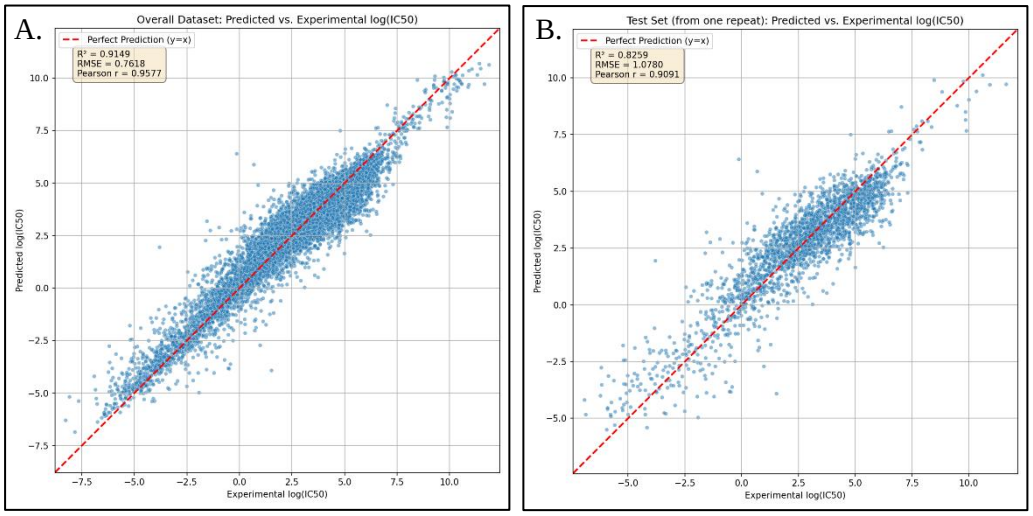


**Fig 3.1 Scatter Plot for predicted vs experimental log($IC_{50}$) values for (A) overall dataset. (B) test set from single repeat**

Scatter plots were generated to visualize the predictive performance. Evaluation of the model's fit for the entire dataset used for developing the train/test split, comprising the 11,588 drug-cell lines, is seen in Fig 3.1(A). The plot demonstrates an excellent fit overall, with data points tightly distributed along the y=x line. The performance metrics calculated were $R^2$ of 0.9149, RMSE of 0.7168 and Pearson's correlation coefficient of 0.9577. These metrics are likely influenced by the inclusion of training data points, and they confirm the model's capacity to learn the underlying relationships. The model's ability to generalize unseen data was visualized using a representative test set, comprising 20% of the data randomly selected from one of the five evaluation repeats in Fig 3.1(B). The data points cluster closely around the y=x

line and the model achieved an $R^2$ value of 0.8259, RMSE od 1.0780 and Pearson's correlation coefficient of 0.9091, further delineating the strong linear agreement between the predicted and observed sensitivities on unseen data. These values are also close to the final metrics of the 5-fold validated model.

Combined, these scatter plots illustrate the model's strong predictive capabilities, both in terms of fit to the overall data and its ability to generalize effectively to test samples.

### 3.2 SHAP analysis unveils key determinants of predicted drug sensitivity

Global SHAP analysis revealed that drug physicochemical properties, distinct chemical substructures and specific gene expression levels were all significant contributors.

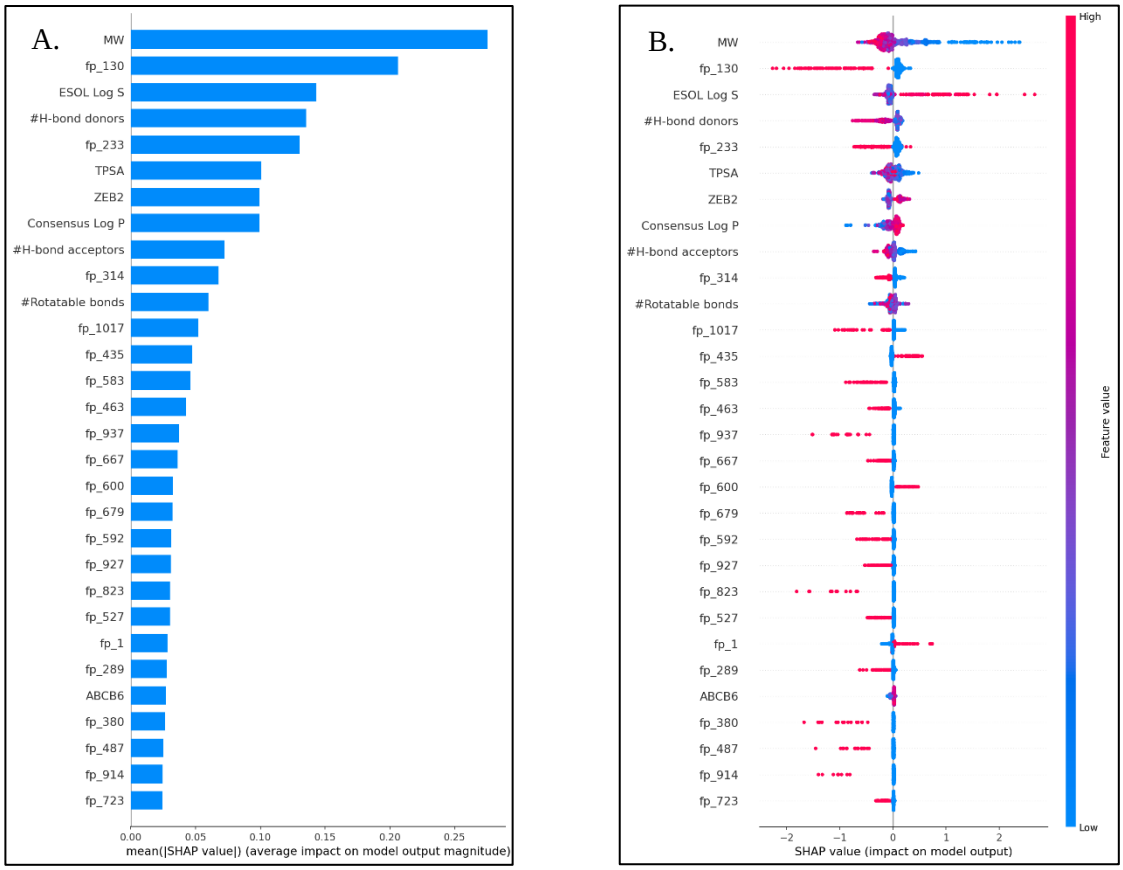


**Fig 3.2(A). SHAP summary plot of top 30 features contributing to model. (B) SHAP beeswarm plot of top 30 features.**

Molecular weight is the most impactful feature, followed by the drug fingerprint fp_130, ESOL Log S (solubility), H-bond donors, drug fingerprint fp_233, TPSA, gene ZEB2, Consensus Log P (lipophilicity), H-bond acceptors and Rotatable Bonds. SHAP dependence plots elucidated the manner in influence of these top features.

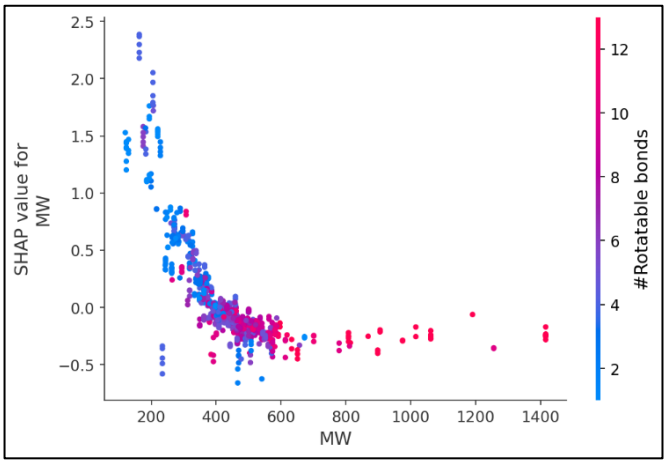


**Fig 3.3 SHAP dependence plot for Molecular Weight**

For MW, there is a clear downward trend in SHAP values as molecular weight increases, particularly up to 700 Da. Low MW, less than 400 Da is associated with positive SHAP values, strongly pushing the model to predict higher log($IC_{50}$). These molecules tend to be less flexible. Medium MW (400-600 Da) cluster around zero and show less influence on model predictability. Higher MW is associated with negative SHAP values, pushing the model to predict lower log($IC_{50}$). These molecules are predominantly the ones with higher number of rotatable bonds. The model likely associates higher MW with better intracellular engagement or target interaction properties in glioblastoma cell lines. Low MW are likely to non-specific or susceptible to efflux.

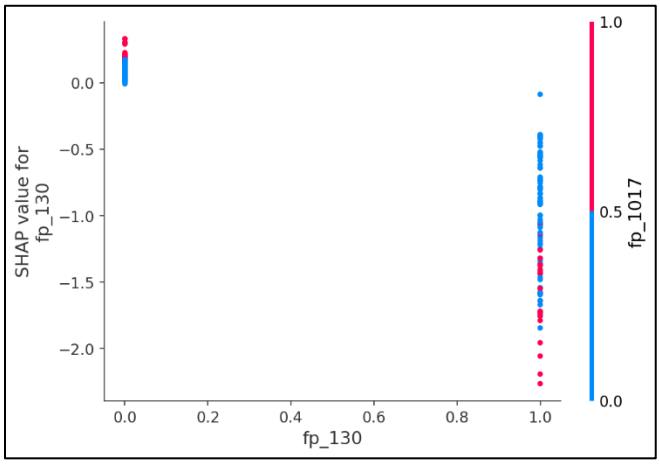


**Fig 3.4(A). SHAP dependence plot for fingerprint fp_130**

SHAP analysis highlighted fingerprint fp_130 as a potent predictor of drug sensitivity. Its presence (fp_130 =1) in Fig 3.4(A) is associated with strongly negative SHAP values, guiding the model to predict low log($IC_{50}$) values or sensitivity. This effect is further amplified when fingerprint fp_1017 is also present, indicating a synergistic effect between chemical features.

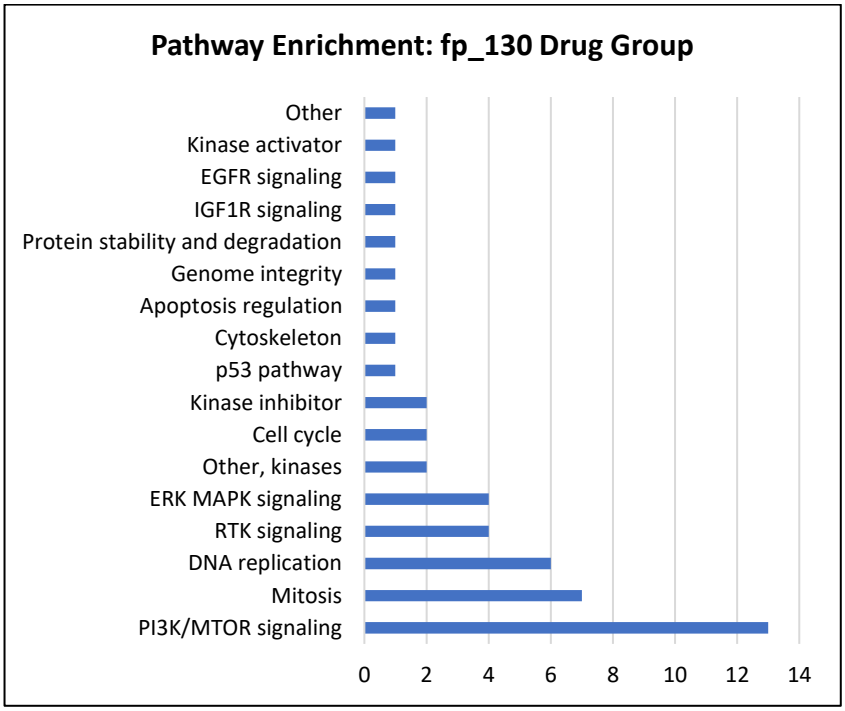


**Fig 3.4(B). Pathway Distribution for drugs containing fingerprint bit fp_130**

A group analysis was conducted to identify the drugs containing this fingerprint. 49 drugs and their targeted pathways (Fig 3.4(B)) reveals a striking enrichment for compounds targeting the PI3K/mTOR signaling pathway. Other represented pathways include Mitosis, DNA replication, RTK signaling, ERK/MAPK signaling. The model has strongly learned that drugs possessing this feature are highly effective likely due to their impact on PI3K/mTOR survival pathway or cell division process.

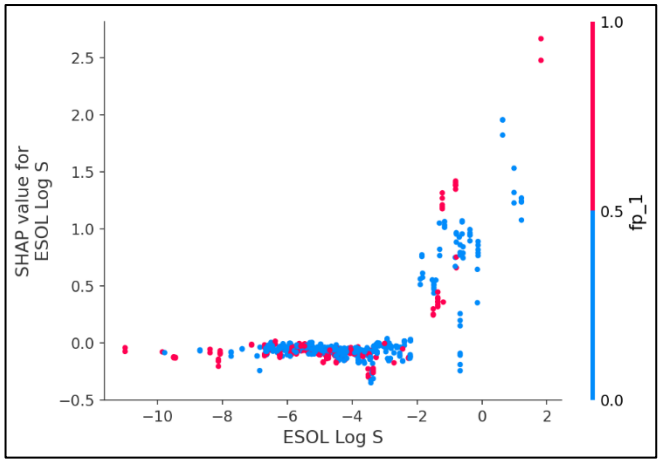


**Fig 3.5. SHAP dependence plot for ESOL Log S (solubility)**

ESOL Log S stands for Estimated Solubility on logarithmic scale. Low solubility *i.e.* ESOL Log S < -3, the SHAP values are aggregated around or slightly below zero indicating poor aqueous solubility has minimal effect on predicted log($IC_{50}$). Moderate to high solubility shows an upward trend toward SHAP values, reaching +2, indicating that higher solubility is associated with higher prediction of log($IC_{50}$). The model associates higher predicted aqueous solubility with increased predicted resistance. This effect is more pronounced in compounds lacking the fp_1 substructure. Poorly soluble drugs show neutral or slightly-sensitive associated SHAP contributions. It may be so that high solubility compounds may be more prone to efflux-mediated clearance, leading to apparent resistance, due to high transporter expression.

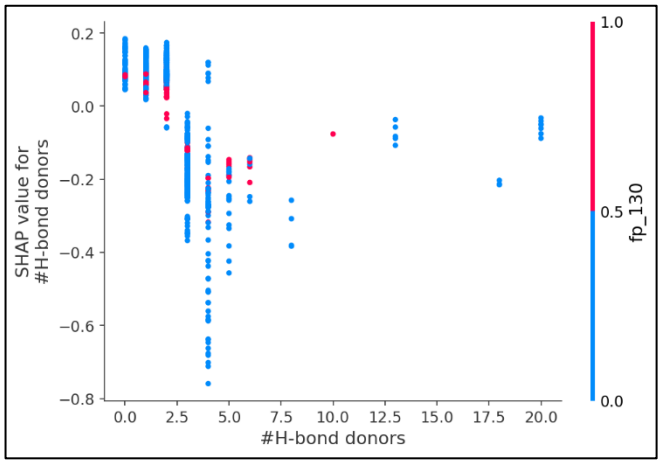


**Fig 3.6 SHAP dependence plot for H-bond donors**

There is a non-linear trend for H-bond donors. Low donors (0-2) are associated with SHAP values clustering around zero or slightly positive, generally predicting a slight resistance or minimal impact. Moderate donors (3-5_ show most strongly negative SHAP values down to -0.8. But as the number of donors increases beyond 5, the SHAP values tend to move back towards zero or less negative. While fp_130 itself is a strong predictor of sensitivity, its interaction with H-bond donors is most pronounced when fp_130 is absent. It could possibly mean that there is a need for an optimal number of H-bond donors is more necessary in drugs where this substructure is not present. The model has learned a specific optimal range for H-bond donors (4-5) that predicts drug sensitivity. This is in line with Lipinski's rule of 5.

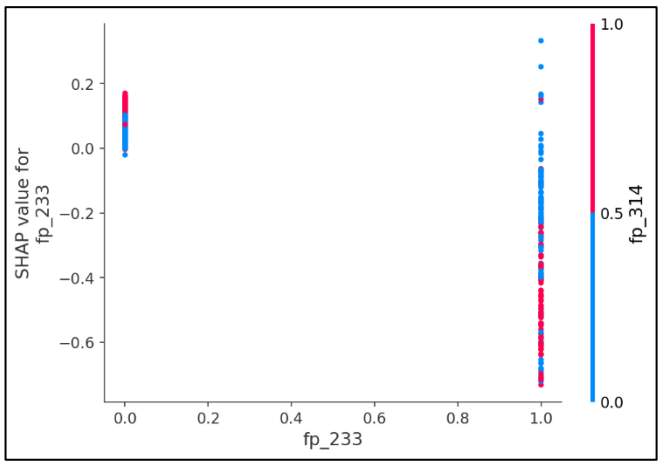


**Fig 3.7(A). SHAP dependence plot for fingerprint fp_233**

SHAP analysis in Fig. 3.7(A) highlighted fingerprint fp_233 as another strong predictor of drug sensitivity. Its presence (fp_233 =1) is associated with strongly negative SHAP values, guiding the model to predict low log($IC_{50}$) values or sensitivity. This effect might be amplified when is further amplified when fingerprint fp_314 is also present, but on its own fp_233 remains a strong drug sensitivity predictor.

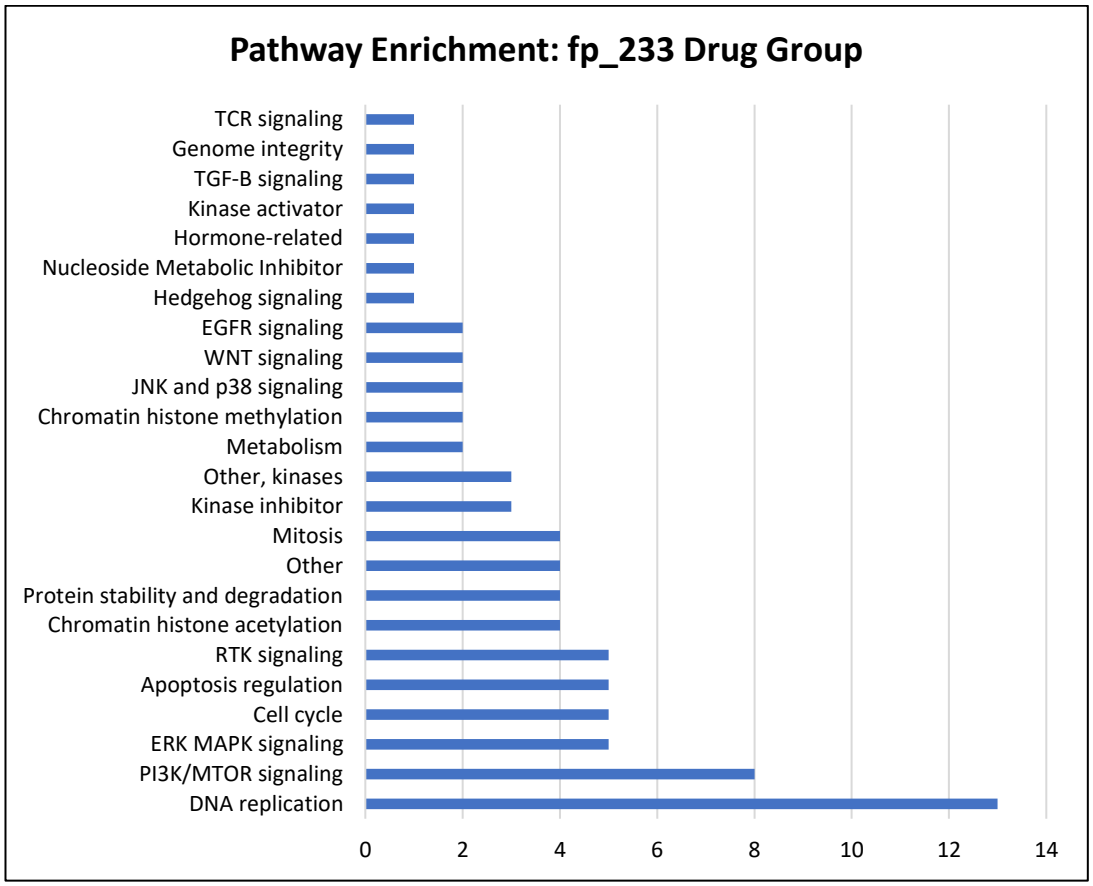


**Fig 3.7(B). Pathway Distribution for drugs containing fingerprint bit fp_233**

A group analysis was conducted to identify the drugs containing this fingerprint. 80 drugs and their targeted pathways in Fig. 3.7(B) reveals a striking enrichment for compounds targeting the DNA replication, PI3K/mTOR signaling pathway, ERK/MAPK signaling, cell cycle, apoptosis regulation and RTK signaling. Other represented pathways include chromatin histone acetylation and methylation, mitosis, metabolism, WNT signaling, JNK and p38 signaling, and EGFR signaling. The model

has strongly learned that drugs possessing this feature are highly effective likely due to their impact on DNA replication or PI3K/mTOR pathway.

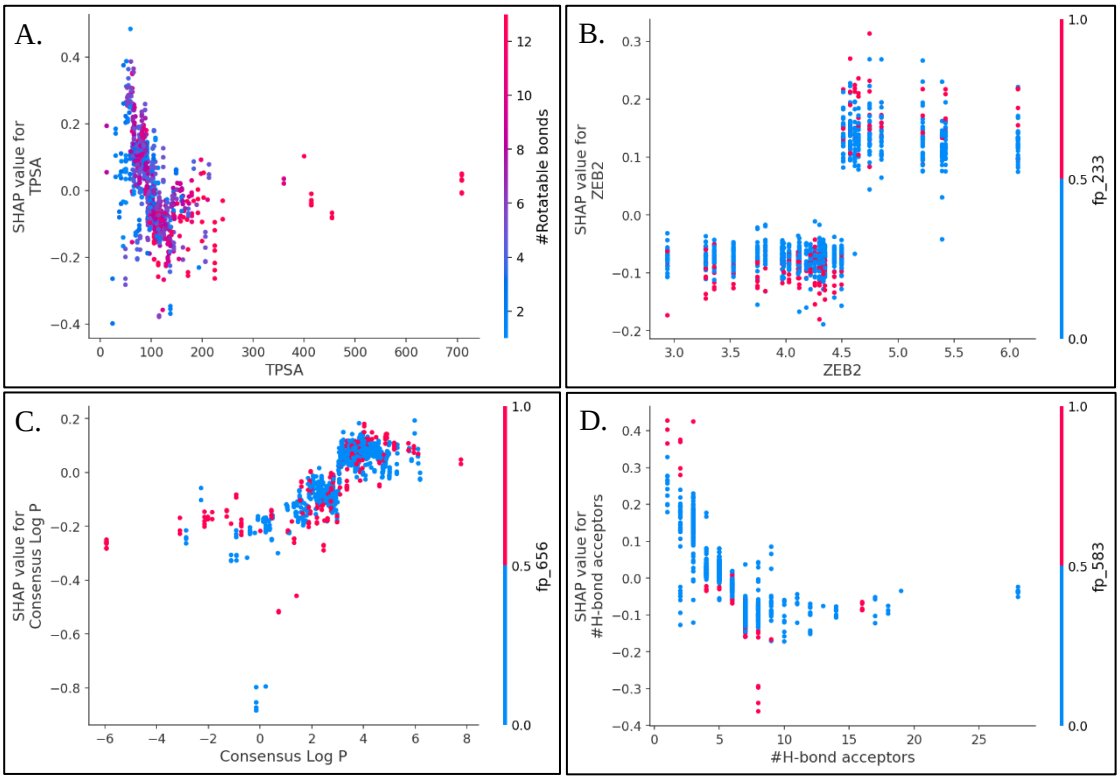


**Fig 3.8. SHAP dependence plots for (A) TPSA (B) *ZEB2* (C) Consensus Log P (D) H-bond acceptor**

TPSA is another strong feature contributing to model learning and prediction capabilities as shown in Fig 3.8(A). Higher TPSA, greater 100 $Å^2$ is a significant predictor of drug sensitivity *i.e.* lower log($IC_{50}$), potentiated by the presence of molecular flexibility as shown through interaction with higher number of rotatable H-bonds. The plot of *ZEB2* in Fig 3.8(B) shows that lower gene expression is contributes to predicted sensitivity, while a higher gene expression contributes to predicted resistance. *ZEB2* is a transcriptional repressor involved in epithelial-mesenchymal transition and stemness [156]. So, it aligns with the notion that high expression predicts for resistance. Lipophilicity in Fig 3.8(C) shows that high hydrophilicity (Log P < 0) is strongly associated with predicted sensitivity, potentially modulated by the presence

of the drug fingerprint fp_656. There is a strong inverse relationship between the H-bond acceptors and the SHAP values as illustrated in Fig 3.8(D). Moderate number of H-bond acceptors (5-10) exhibit negative SHAP values, pointing to higher sensitivity. This effect is partially influenced by the presence of the bit fp_583.

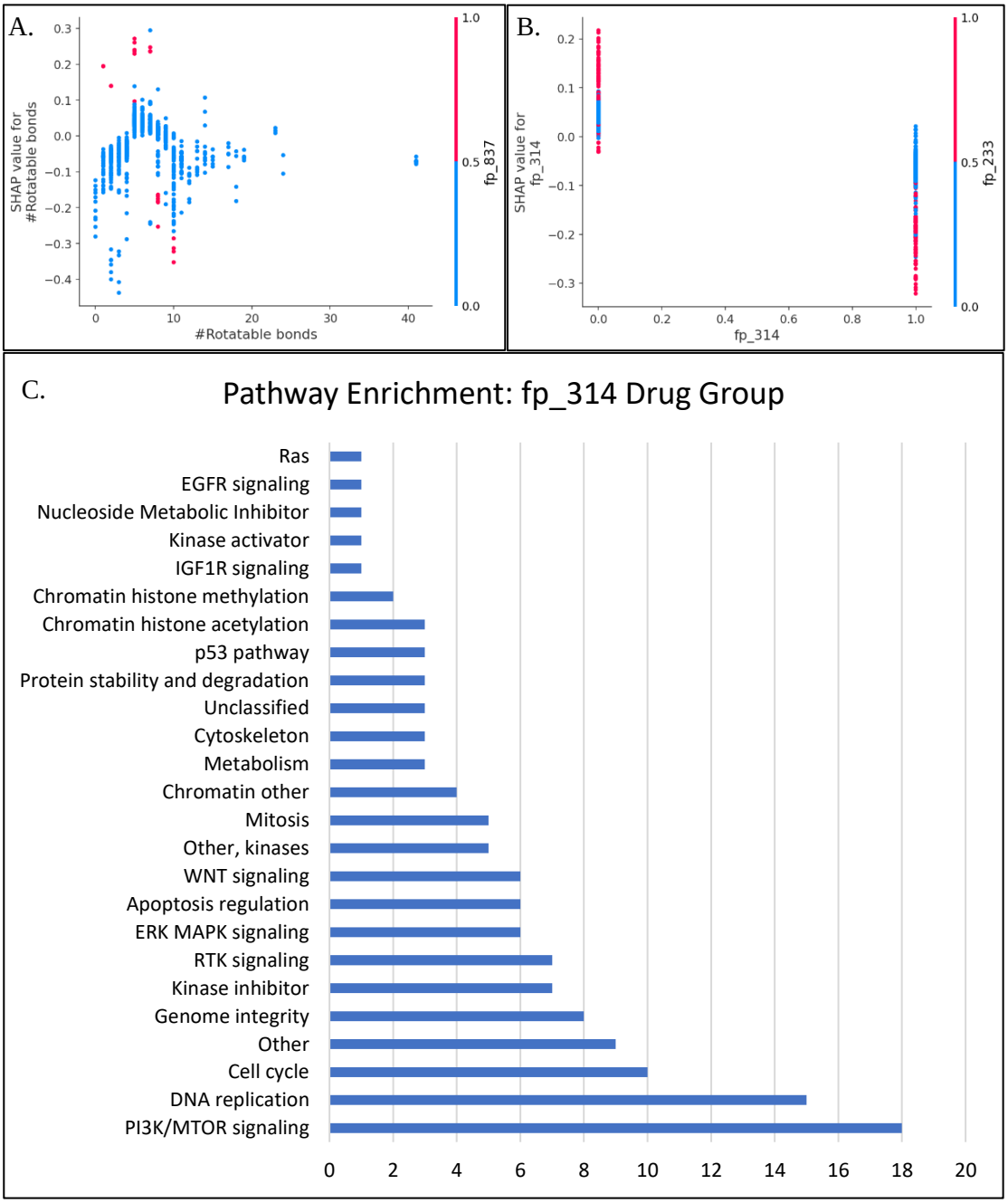


**Fig 3.9. SHAP dependence plots for (A) Rotatable Bonds (B) drug fingerprint fp_314 (C) Pathway Distribution for drugs containing fingerprint bit fp_314.**

The contribution of Rotatable Bonds to model prediction is significant as illustrated in Fig 3.9(A). Less than 10 rotatable bonds point to drug sensitivity, or lower log($IC_{50}$) values. Presence of fp_817 does not have a huge impact on this contribution. Presence of the substructure fp_314 strongly contributes to negative SHAP values, and drug sensitivity as shown in Fig 3.9(B). This effect works synergistically with the fingerprint fp_233 discussed earlier in the text. Drug group analysis identified 131 drugs with this fingerprint in Fig 3.9(C), enriched for targeting pathways like PI3K/mTOR signaling, cell cycle, genome integrity, RTK signaling, apoptosis regulation among many other.

### 3.3 In silico screening of COCONUT database identifies novel natural product candidates with high predicted potency

The XGBoost model was used to perform an *in-silico* screen against the COCONUT natural product database. The library was filtered for compounds with the highest annotation level 5. Additional filters such as MW < 800, TPSA between 100 and 200 $Å^2$, H-bond donors 3-5, H-bond acceptors < 10, Rotatable bonds < 10, Log P < 2 were applied to further reduce the chemical space for search. A total of 290 compounds were finally filtered. The solubility and BBB-permeability were retrieved through the Swiss ADME web browser, since these parameters are not provided in the COCONUT database. The predicted log($IC_{50}$) values of the top 10 compounds from the trained model are noted in Table 2, along with their highest probability( > 90%) predicted targets from SuperPred.

**Table 2. Top 10 COCONUT compounds screened using trained model and their predicted sensitivity**

| Compound | Predicted Log($IC_{50}$) | Predicted $IC_{50}$ (in nM) | Predicted Targets |
|---|---|---|---|
| CNP0152293.3 | -4.019322395 | 17.96513407 | CTSD, COX-1, TIF1-α, MAOA |
| CNP0347714.1 | -3.898481131 | 20.27267963 | CTSD, COX-1, TIF-α, MAOA, APE1, ADAM10 |
| CNP0347714.2 | -3.898481131 | 20.27267963 | CTSD, COX-1, TIF-α, MAOA, APE1, ADAM10 |
| CNP0353870.1 | -3.831660509 | 21.67359653 | APE1, CTSD, NTRK3, TIF1-α |
| CNP0196054.2 | -3.698587179 | 24.75848107 | TDP1, COX-1, CTSD, MAOA, TIF1- α, APE1 |
| CNP0324164.1 | -3.432158709 | 32.31710221 | HSP90β, STAT3 |
| CNP0142637.2 | -3.103508234 | 44.89143615 | CTSD, COX-1, TDP1, APE1, TIF1- α |
| CNP0343540.1 | -3.031371355 | 48.24942564 | APE1, HIF1-α, CTSD, LSD1, PDGFRA |
| CNP0091821.1 | -2.884468317 | 55.88449447 | CTSD, APE1, ADAM10, CK2 |
| CNP0223869.2 | -2.832907915 | 58.84149856 | ERK2, APE1, HIF1- α, TDP1, CTSD, TIF1- α |

The pathway enrichment analysis of the 14 unique putative targets through ShinyGO with the PANTHER database revealed significant enrichment for pathways highly relevant to glioblastoma in Fig. 3.10.

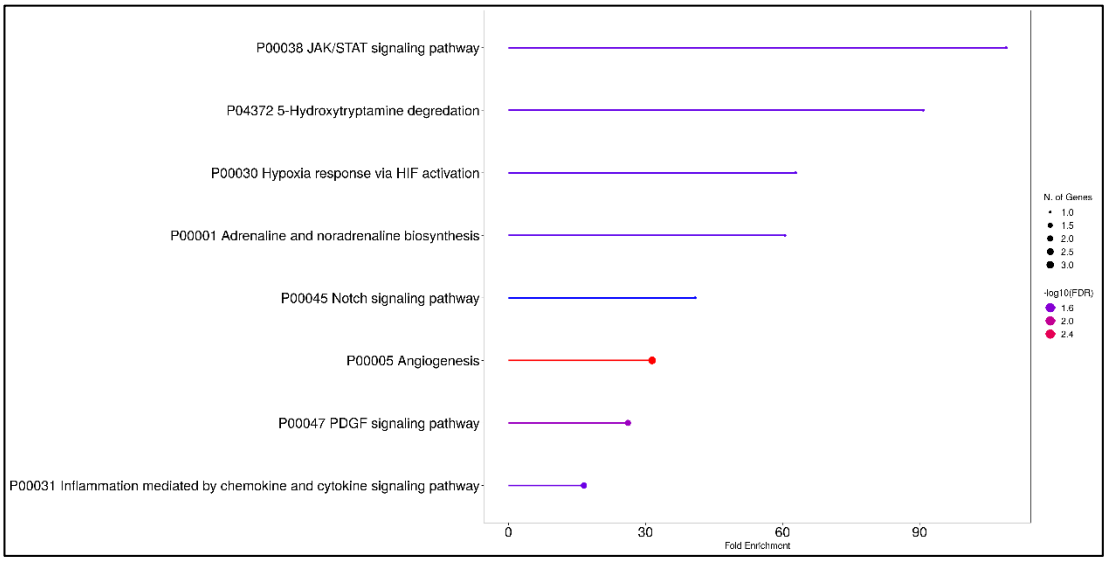


**Fig 3.10 Pathway enrichment plot for predicted targets of the top 10 COCONUT compounds**

Angiogenesis was the top hit with a fold enrichment of 31.4 and -log10(FDR) of 2.4. The presence of 3 genes from predicted targets highlights the importance of these findings. Angiogenesis is essential for glioblastoma growth and vascularity, so it is likely that several of the top predicted compounds may exert their anti-cancer activity by inhibiting formation of new blood vessels. The PDGF signaling pathway also showed significant fold enrichment 26.1 and statistical significance ( $-\log_{10}$(FDR) of nearly 2.0-2.2. The involvement of 2 targets could be a novel avenue of potentially new anti-glioblastoma drugs through the disruption of the PDGFR signaling axis, that is well established driver of glioblastoma proliferation. The enrichment of the inflammation mediated by chemokine and cytokine signaling pathway points towards the model identifying compounds that may modulate tumor microenvironment or inflammatory responses generated by microglia and macrophages. 16-fold enrichment and possibly two target genes allude towards a possible therapeutic avenue. Pathways

such as JAK/STAT signaling, Hypoxia response via HIF activation, 5-Hydroxytryptamine degradation, Adrenaline and noradrenaline biosynthesis also demonstrated high fold enrichment and statistically significant FDR vales. It is possible that the top drug candidates may also affect cellular signaling required for proliferation and survival, stress responses common in the tumor milieu and neurotransmitter pathways that may influence glioblastoma cells. Notch signaling pathway had a lower statistical significance, but a high fold enrichment of 40.8, and is highly evidenced in glioblastoma stemness and resistance. This analysis suggests that the top 10 compounds are likely to interfere with key biological processes essential for glioblastoma progression.

**TABLE 3. Pathway Enrichment Analysis of Predicted Targets**

| Pathway | Pathway Genes | Fold Enrichment | FDR | Genes | |
|---|---|---|---|---|---|
| JAK/STAT signaling pathway | 15 | 108.9571 | 0.032881 | 1 | STAT3 |
| 5-Hydroxytryptamine degredation | 18 | 90.79762 | 0.032881 | 1 | MAOA |
| Hypoxia response via HIF activation | 26 | 62.85989 | 0.03514 | 1 | HIF1A |
| Adrenaline and noradrenaline biosynthesis | 27 | 60.53175 | 0.03514 | 1 | MAOA |
| Notch signaling pathway | 40 | 40.85893 | 0.045385 | 1 | ADAM10 |
| Angiogenesis | 156 | 31.42995 | 0.001606 | 3 | STAT3, HIF1A, PDGFRA |
| PDGF signaling pathway | 125 | 26.14971 | 0.019357 | 2 | STAT3, PDGFRA |
| Inflammation mediated by chemokine and cytokine signaling pathway | 198 | 16.50866 | 0.031657 | 2 | STAT3, PTGS1 |

# CHAPTER 4

# DISCUSSION

The efforts to propose novel compounds as potential drug candidates, or repurpose existing drugs to altogether different indications are often narrowed to single targets and fail to consider the gene expression context. Many of them also restrict themselves to classification tasks such as predicting whether a compound will be sensitive or resistant, accompanied by a lack of interpretability when the pipeline assumes a machine learning framework, be it neural networks or deep learning strategies. Feature selection plays an essential and deterministic role for model performance.

In this work, I have incorporated gene expression features, along with drug Morgan fingerprints, and drug features like MW, lipophilicity, H-bond donors, H-bond acceptors, rotatable bonds, solubility, and BBB-permeability along with Lipinski's violations. The model trained on XGBoost regressor had a strong performance, rigorously evaluated using 5-fold validation, with RMSE of $1.0600 \pm 0.0225$ and $R^2$ of $0.8332 \pm 0.0083$, with a strong correlation between experimental and predicted values of $\log(IC_{50})$. SHAP analysis revealed that MW had the most significant contribution to model learning, followed by the drug fingerprint fp_130, solubility, H-bond donor, TPSA and so on. Multiple drug fingerprints that contributed to model associations were enriched in drug targeting pathways like PI3K/mTOR, DNA replication, cell cycle, MAPK signaling, apoptosis regulation and WNT signaling majorly. Few drugs targeted the chromatin histone acetylation and methylation,

genome integrity, EGFR signaling, IGF1R signaling, metabolism and even the nucleoside metabolic inhibitors.

Two genes of interest were also highlighted in the SHAP summary plot. *ZEB2* is Zinc finger E-box Binding homeobox 2, overexpressed in glioblastoma, and knockout of which inhibits cell and tumorigenesis [157]. Suppression of ZEB2 induces apoptosis in glioblastoma cell lines, indicating that overexpression of this gene has oncogenic role [158]. The other gene of interest, though not as strongly contributing to the model learning is *ABCB6*, ATP binding cassette subfamily B member 6 that functions as transport protein for porphyrin transport, drug resistance and protection against stress [159]. ABCB6 expression is also strongly correlated with histological tumor grade, with significant upregulation in glioblastoma cell lines [160]. This may contribute to drug efflux leading to higher resistance.

Out of the proposed novel compounds, CNP0152293.3 or 10-Dehydrobaccatin V ranked the highest. It belongs to the diterpenoids super class and its predicted targets include CTSD, COX-1, TIF1-α, MAOA. The top 5 compounds belong to the super class diterpenoids and alkaloids (CNP0353870.1). Diterpenoids like Triptolide, Crocetin and Phytol is well established, from inducing senescence phenotype in cancer cells to inhibition of cancer survival genes [161]. Alkaloids like Vincristine, Vinblastine, Campothecin, Paclitaxel and Docetaxel have a long history of investigation as anti-cancer agents in multiple malignancies [162]. The COCONUT compounds have been predicted to target multiple proteins. Cathepsin D (*CTSD*) is closely associated with clinical malignancy and is overexpressed in radioresistant cells [163]. Cyclooxygenase-1 (*COX-1*) also known as *PTGS1*, is constitutively expressed in brain tissue and overexpressed in glioblastoma cells, inhibition of which abrogates tumor cell migration [164]. Transcription intermediary factor 1-α (*TIF1-α*), also known as *TRIM24*, is an oncogenic coactivator of *STAT3* in glioblastoma and enhances EGFR-driven tumorigenesis [165]. Monoamine oxidase A (*MAOA*) oxidizes monoamine neurotransmitter, leading to reactive oxygen species that drive cancer [166]. Inhibition of these in TMZ-resistant cells can reduce tumor progression. *APE1* or *APEX1* is the major apurinic/apyrimidinic endonuclease of the base excision repair

pathway to mitigate DNA damage, that guides drug tolerance in glioblastoma, if suppressed [167]. It is also correlated with glioblastoma recurrence and increased immunosuppressive tumor microenvironment [168]. The metalloproteinase *ADAM10* is highly expressed in GSCs, and inhibition can lead to phenotypes that are more amenable to therapy [169]. *NTRK3* fusions have been reported in glioblastomas that potentially drive the tumor, along with *NTRK1/2* fusions [170]. Tyrosyl-DNA phosphodiesterase 1 (*TDP1*) is a potential biomarker along with topoisomerase 1 in glioblastoma for irinotecan treatment [171]. STAT3 , part of major JAK-STAT signaling pathway, is overexpressed in glioblastoma tissues and is required for proliferation and potency of GSCs [172]. HSP90 is upregulated, influences stemness, drives up glucose consumption in glioblastoma cells [173]. ERK2 is a protein kinase of the MAPK pathway. It has been reported that ERK suppression is correlated with autophagy activation and tumor suppression [174]. Lysine specific demethylase 1 (LSD1) can cooperate with histone deacetylase inhibitors to regulate cell death in glioblastoma cell lines [175]. Inhibition of LSD1 also induces senescence in glioblastoma cells [176]. Hypoxia-inducible factor-1α (HIF-1α) mediates maintenance of GSCs under hypoxic conditions through Notch signaling[177]. Platelet-derived growth factor receptor alpha (PDGFRA) amplification is a poor prognostic marker for IDH wild-type glioblastoma [177]. The enrichment of targets within pathways like angiogenesis and PDGF signaling aligns with known glioblastoma progression mechanisms and provides further rationale for the potential efficacy of these compounds. This analysis of predicted targets for the top computationally screened compounds further supports the biological relevance of the patterns learned by the model.

It should be noted that this model provides a stepping stone, a preliminary step towards robust drug discovery paradigms. Further *in silico* studies such as molecular docking, simulations followed by *in vitro* assays to elucidate their mechanism of action would propel greater impact on the community researching glioblastoma.

# CHAPTER 5

# CONCLUSIONS, FUTURE SCOPE AND SOCIAL IMPACT

## 5.1 Conclusions

This study successfully developed and validated a robust ML framework, centered on an XGBoost model, for predicting drug sensitivity in a panel of 29 glioblastoma cell lines. By integrating log transformed IC50 values for 435 unique compounds with transcriptomic profiles derived from RMA-normalized ArrayExpress E-MTAB-3610 data, a high-performance model was trained. This model utilized a carefully selected feature set, comprising 100 key gene expression markers identified through RFECV, alongside drug specific chemical and physicochemical features. 5-fold evaluation demonstrated stable predictive power, achieving average R2 of approximately 0.833 and an RMSE of approximately 1.060, indicating that the model effectively captured key determinants of drug response in glioblastoma context.

A significant contribution of this work is the deep model interpretability achieved through SHAP analysis, that enable identification of specific genes (*ZEB2, ABCB6*), drug physicochemical properties and distinct chemical substructures most strongly associated with predicted sensitivity or resistance. High ZEB2 expression consistently predicted resistance, aligning with its role in epithelial-mesenchymal transition, while specific fingerprint bits were linked through group pathway analysis to drugs targeting critical oncogenic pathways like PI3K/mTOR signaling and DNA replication. This

also lent a biological relevance of the learned feature-response relationship The utility of this work was further illustrated by screening a filtered subset of the COCONUT natural product database, identifying several novel compounds with highly potent predicted IC50 values against an average glioblastoma cell line, warranting further investigation as potential anti-cancer therapeutics.

### 5.2 Future Scope

In the future, it may become necessary to incorporate additional omic layers such as somatic mutation profiles, copy number variations, DNA methylation patterns, or even proteomics data to provide a more comprehensive portrait of molecular mechanisms in glioblastoma. There are many glioblastoma cohort datasets available on the TCGA and GDC Data Portal hosted on the National Cancer Institute. Availability of cohort histopathological data and imaging could lend a new angle to the study. Of course, integration of more sophisticated drug representations through graph neural networks or convolutional neural networks may be explored to understand chemical determinants of drugs and make the pipeline multi-layered. Molecular docking, simulation and experimental approaches form the next rung of the ladder in combatting glioblastoma. Experimental validation of the top-ranked natural products would involve *in vitro* testing in glioblastoma cell lines, to investigate their anti-cancer activity and determine experimental IC50 values.

### 5.3 Social Impact

Glioblastoma remains the most devastating brain cancer with multiple barriers to successful treatments and a grim prognosis. The immense cost and high attrition rates accompanying traditional drug development necessitate innovative approaches. The ML framework developed in this study offers a computationally efficient and scalable strategy to navigate the vast chemical space of natural products and prioritize candidate compounds for glioblastoma. These models can accelerate the early stages of discovering novel therapeutics.

Natural products have historically been a rich source of anti-cancer compounds, yet their systematic exploration is often hampered by their structural complexity and the challenge of isolating and testing them at scale. This work demonstrates a path to rationally identify promising natural product candidates, potentially uncovering novel scaffolds and mechanism against glioblastoma. The emphasis on model interpretability through SHAP analysis promotes transparency and allows for deeper biological understanding of predicted drug responses. Explainability is crucial for building trust in AI-driven decision-making within biomedical research and can guide experimental work for new potential leads. This research commits to the broader field of precision oncology, aiming to tailor therapeutic strategies to specific molecular characteristics. While focused on preclinical models, the principals and potential lead compounds identified here could contribute to the development of new therapies that ultimately improve clinical outcomes and enhance the overall well-being of individuals diagnosed with this disease.

# REFERENCES

[1] H.-G. Wirsching, E. Galanis, and M. Weller, “Glioblastoma,” *Handb. Clin. Neurol.*, vol. 134, pp. 381–397, 2016, doi: 10.1016/B978-0-12-802997-8.00023-2.
[2] M. Yao *et al.*, “Cellular origin of glioblastoma and its implication in precision therapy,” *Cell. Mol. Immunol.*, vol. 15, no. 8, pp. 737–739, Aug. 2018, doi: 10.1038/cmi.2017.159.
[3] L. M. Wang, Z. K. Englander, M. L. Miller, and J. N. Bruce, “Malignant Glioma,” *Adv. Exp. Med. Biol.*, vol. 1405, pp. 1–30, 2023, doi: 10.1007/978-3-031-23705-8_1.
[4] F. Seker-Polat, N. Pinarbasi Degirmenci, I. Solaroglu, and T. Bagci-Onder, “Tumor Cell Infiltration into the Brain in Glioblastoma: From Mechanisms to Clinical Perspectives,” *Cancers*, vol. 14, no. 2, p. 443, Jan. 2022, doi: 10.3390/cancers14020443.
[5] A. Colopi *et al.*, “Impact of age and gender on glioblastoma onset, progression, and management,” *Mech. Ageing Dev.*, vol. 211, p. 111801, Apr. 2023, doi: 10.1016/j.mad.2023.111801.
[6] H. Ohgaki and P. Kleihues, “The definition of primary and secondary glioblastoma,” *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.*, vol. 19, no. 4, pp. 764–772, Feb. 2013, doi: 10.1158/1078-0432.CCR-12-3002.
[7] S. S. K. Yalamarty *et al.*, “Mechanisms of Resistance and Current Treatment Options for Glioblastoma Multiforme (GBM),” *Cancers*, vol. 15, no. 7, p. 2116, Apr. 2023, doi: 10.3390/cancers15072116.
[8] A. Rodríguez-Camacho *et al.*, “Glioblastoma Treatment: State-of-the-Art and Future Perspectives,” *Int. J. Mol. Sci.*, vol. 23, no. 13, p. 7207, Jun. 2022, doi: 10.3390/ijms23137207.
[9] A. Ou, W. K. A. Yung, and N. Majd, “Molecular Mechanisms of Treatment Resistance in Glioblastoma,” *Int. J. Mol. Sci.*, vol. 22, no. 1, p. 351, Dec. 2020, doi: 10.3390/ijms22010351.
[10] D. Bertsimas and H. Wiberg, “Machine Learning in Oncology: Methods, Applications, and Challenges,” *JCO Clin. Cancer Inform.*, no. 4, pp. 885–894, Oct. 2020, doi: 10.1200/CCI.20.00072.
[11] A. N. Lima, Philot ,Eric Allison, Trossini ,Gustavo Henrique Goulart, Scott ,Luis Paulo Barbour, Maltarollo ,Vinícius Gonçalves, and K. M. and Honorio, “Use of machine learning approaches for novel drug discovery,” *Expert Opin. Drug Discov.*, vol. 11, no. 3, pp. 225–239, Mar. 2016, doi: 10.1517/17460441.2016.1146250.
[12] Z. Tanoli, Vähä-Koskela ,Markus, and T. and Aittokallio, “Artificial intelligence, machine learning, and drug repurposing in cancer,” *Expert Opin. Drug Discov.*, vol. 16, no. 9, pp. 977–989, Sep. 2021, doi: 10.1080/17460441.2021.1883585.
[13] M. R. Karim *et al.*, “Explainable AI for Bioinformatics: Methods, Tools and Applications,” *Brief. Bioinform.*, vol. 24, no. 5, p. bbad236, Sep. 2023, doi: 10.1093/bib/bbad236.
[14] R. G. W. Verhaak *et al.*, “Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1,” *Cancer Cell*, vol. 17, no. 1, pp. 98–110, Jan. 2010, doi:

10.1016/j.ccr.2009.12.020.
[15] Y. Kim *et al.*, "Perspective of mesenchymal transformation in glioblastoma," *Acta Neuropathol. Commun.*, vol. 9, no. 1, p. 50, Mar. 2021, doi: 10.1186/s40478-021-01151-4.
[16] 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.
[17] J. A. Benitez *et al.*, "PTEN regulates glioblastoma oncogenesis through chromatin-associated complexes of DAXX and histone H3.3," *Nat. Commun.*, vol. 8, no. 1, p. 15223, May 2017, doi: 10.1038/ncomms15223.
[18] Y. Zhang *et al.*, "The p53 Pathway in Glioblastoma," *Cancers*, vol. 10, no. 9, p. 297, Sep. 2018, doi: 10.3390/cancers10090297.
[19] I. Crespo *et al.*, "Detailed Characterization of Alterations of Chromosomes 7, 9, and 10 in Glioblastomas as Assessed by Single-Nucleotide Polymorphism Arrays," *J. Mol. Diagn.*, vol. 13, no. 6, pp. 634–647, Nov. 2011, doi: 10.1016/j.jmoldx.2011.06.003.
[20] S. R. Bollam, M. E. Berens, and H. D. Dhruv, "When the Ends Are Really the Beginnings: Targeting Telomerase for Treatment of GBM," *Curr. Neurol. Neurosci. Rep.*, vol. 18, no. 4, p. 15, Mar. 2018, doi: 10.1007/s11910-018-0825-7.
[21] I. Crespo *et al.*, "Molecular and Genomic Alterations in Glioblastoma Multiforme," *Am. J. Pathol.*, vol. 185, no. 7, pp. 1820–1833, Jul. 2015, doi: 10.1016/j.ajpath.2015.02.023.
[22] H. Noushmehr *et al.*, "Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma," *Cancer Cell*, vol. 17, no. 5, pp. 510–522, May 2010, doi: 10.1016/j.ccr.2010.03.017.
[23] D. Sturm *et al.*, "Paediatric and adult glioblastoma: multiform (epi)genomic culprits emerge," *Nat. Rev. Cancer*, vol. 14, no. 2, pp. 92–107, Feb. 2014, doi: 10.1038/nrc3655.
[24] H.-M. Chen, A. Nikolic, D. Singhal, and M. Gallo, "Roles of Chromatin Remodelling and Molecular Heterogeneity in Therapy Resistance in Glioblastoma," *Cancers*, vol. 14, no. 19, p. 4942, Oct. 2022, doi: 10.3390/cancers14194942.
[25] B. B. Liau *et al.*, "Adaptive chromatin remodeling drives glioblastoma stem cell plasticity and drug tolerance," *Cell Stem Cell*, vol. 20, no. 2, pp. 233-246.e7, Feb. 2017, doi: 10.1016/j.stem.2016.11.003.
[26] D. H. Lee *et al.*, "Histone demethylase KDM4C controls tumorigenesis of glioblastoma by epigenetically regulating p53 and c-Myc," *Cell Death Dis.*, vol. 12, no. 1, p. 89, Jan. 2021, doi: 10.1038/s41419-020-03380-2.
[27] A. A. Hamed *et al.*, "A brain precursor atlas reveals the acquisition of developmental-like states in adult cerebral tumours," *Nat. Commun.*, vol. 13, no. 1, p. 4178, Jul. 2022, doi: 10.1038/s41467-022-31408-y.
[28] A. L. Green *et al.*, "BPTF regulates growth of adult and pediatric high-grade glioma through the MYC pathway," *Oncogene*, vol. 39, no. 11, pp. 2305–2327, Mar. 2020, doi: 10.1038/s41388-019-1125-7.
[29] M. Chen, Z. Medarova, and A. Moore, "Role of microRNAs in glioblastoma," *Oncotarget*, vol. 12, no. 17, pp. 1707–1723, Aug. 2021, doi: 10.18632/oncotarget.28039.
[30] C. T. Stackhouse, G. Y. Gillespie, and C. D. Willey, "Exploring the Roles of

lncRNAs in GBM Pathophysiology and Their Therapeutic Potential," *Cells*, vol. 9, no. 11, p. 2369, Oct. 2020, doi: 10.3390/cells9112369.
[31] M. Colardo, M. Segatto, and S. Di Bartolomeo, "Targeting RTK-PI3K-mTOR Axis in Gliomas: An Update," *Int. J. Mol. Sci.*, vol. 22, no. 9, Art. no. 9, Jan. 2021, doi: 10.3390/ijms22094899.
[32] J. Gao *et al.*, "Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal," *Sci. Signal.*, vol. 6, no. 269, p. pl1, Apr. 2013, doi: 10.1126/scisignal.2004088.
[33] L. E. Huang, "Impact of CDKN2A/B Homozygous Deletion on the Prognosis and Biology of IDH-Mutant Glioma," *Biomedicines*, vol. 10, no. 2, p. 246, Jan. 2022, doi: 10.3390/biomedicines10020246.
[34] C. Giacinti and A. Giordano, "RB and cell cycle progression," *Oncogene*, vol. 25, no. 38, pp. 5220–5227, Aug. 2006, doi: 10.1038/sj.onc.1209615.
[35] V. Juric and B. Murphy, "Cyclin-dependent kinase inhibitors in brain cancer: current state and future directions," *Cancer Drug Resist.*, vol. 3, no. 1, pp. 48–62, Mar. 2020, doi: 10.20517/cdr.2019.105.
[36] V. Kumar *et al.*, "The Role of Notch, Hedgehog, and Wnt Signaling Pathways in the Resistance of Tumors to Anticancer Therapies," *Front. Cell Dev. Biol.*, vol. 9, p. 650772, Apr. 2021, doi: 10.3389/fcell.2021.650772.
[37] D. A. Reardon *et al.*, "A Review of VEGF/VEGFR-Targeted Therapeutics for Recurrent Glioblastoma," *J. Natl. Compr. Canc. Netw.*, vol. 9, no. 4, pp. 414–427, Apr. 2011, doi: 10.6004/jnccn.2011.0038.
[38] B. Yamini, "NF-κB, Mesenchymal Differentiation and Glioblastoma," *Cells*, vol. 7, no. 9, p. 125, Aug. 2018, doi: 10.3390/cells7090125.
[39] M. T. C. Poon, M. Bruce, J. E. Simpson, C. J. Hannan, and P. M. Brennan, "Temozolomide sensitivity of malignant glioma cell lines – a systematic review assessing consistencies between in vitro studies," *BMC Cancer*, vol. 21, no. 1, p. 1240, Nov. 2021, doi: 10.1186/s12885-021-08972-5.
[40] E. Radaelli *et al.*, "Immunohistopathological and neuroimaging characterization of murine orthotopic xenograft models of glioblastoma multiforme recapitulating the most salient features of human disease," *Histol. Histopathol.*, vol. 24, no. 7, pp. 879–891, Jul. 2009, doi: 10.14670/HH-24.879.
[41] N. Ishii *et al.*, "Frequent Co-Alterations of TP53, p16/CDKN2A, p14ARF, PTEN Tumor Suppressor Genes in Human Glioma Cell Lines.," *Brain Pathol.*, vol. 9, no. 3, pp. 469–479, 1999, doi: 10.1111/j.1750-3639.1999.tb00536.x.
[42] X. Wang, J. Chen, Y. Liu, C. You, and Q. Mao, "Mutant TP53 enhances the resistance of glioblastoma cells to temozolomide by up-regulating O6-methylguanine DNA-methyltransferase," *Neurol. Sci.*, vol. 34, no. 8, pp. 1421–1428, Aug. 2013, doi: 10.1007/s10072-012-1257-9.
[43] F. B. Furnari, H. Lin, H.-J. S. Huang, and W. K. Cavenee, "Growth suppression of glioma cells by PTEN requires a functional phosphatase catalytic domain," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 94, no. 23, pp. 12479–12484, Nov. 1997, doi: 10.1073/pnas.94.23.12479.
[44] C. Gratas *et al.*, "Fas Ligand Expression in Glioblastoma Cell Lines and Primary Astrocytic Brain Tumors," *Brain Pathol.*, vol. 7, no. 3, pp. 863–869, Jul. 1997, doi: 10.1111/j.1750-3639.1997.tb00889.x.
[45] S. Y. Lee, "Temozolomide resistance in glioblastoma multiforme," *Genes Dis.*,

vol. 3, no. 3, pp. 198–210, May 2016, doi: 10.1016/j.gendis.2016.04.007.
[46] Y. Xiao and D. Yu, “Tumor microenvironment as a therapeutic target in cancer,” *Pharmacol. Ther.*, vol. 221, p. 107753, May 2021, doi: 10.1016/j.pharmthera.2020.107753.
[47] K. Kristoffersen, Villingshøj ,Mette, Poulsen ,Hans Skovgaard, and M.-T. and Stockhausen, “Level of Notch activation determines the effect on growth and stem cell-like features in glioblastoma multiforme neurosphere cultures,” *Cancer Biol. Ther.*, vol. 14, no. 7, pp. 625–637, Jul. 2013, doi: 10.4161/cbt.24595.
[48] 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.
[49] S. S. Chelliah, E. A. L. Paul, M. N. A. Kamarudin, and I. Parhar, “Challenges and Perspectives of Standard Therapy and Drug Development in High-Grade Gliomas,” *Mol. Basel Switz.*, vol. 26, no. 4, p. 1169, Feb. 2021, doi: 10.3390/molecules26041169.
[50] A. Karachi, F. Dastmalchi, D. A. Mitchell, and M. Rahman, “Temozolomide for immunomodulation in the treatment of glioblastoma,” *Neuro-Oncol.*, vol. 20, no. 12, pp. 1566–1572, Nov. 2018, doi: 10.1093/neuonc/noy072.
[51] R. Stupp *et al.*, “Effect of Tumor-Treating Fields Plus Maintenance Temozolomide vs Maintenance Temozolomide Alone on Survival in Patients With Glioblastoma: A Randomized Clinical Trial,” *JAMA*, vol. 318, no. 23, pp. 2306–2316, Dec. 2017, doi: 10.1001/jama.2017.18718.
[52] N. Rabah, F.-E. Ait Mohand, and N. Kravchenko-Balasha, “Understanding Glioblastoma Signaling, Heterogeneity, Invasiveness, and Drug Delivery Barriers,” *Int. J. Mol. Sci.*, vol. 24, no. 18, Art. no. 18, Jan. 2023, doi: 10.3390/ijms241814256.
[53] L. Cheng *et al.*, “Glioblastoma Stem Cells Generate Vascular Pericytes to Support Vessel Function and Tumor Growth,” *Cell*, vol. 153, no. 1, pp. 139–152, Mar. 2013, doi: 10.1016/j.cell.2013.02.021.
[54] J. N. Sarkaria *et al.*, “Is the blood-brain barrier really disrupted in all glioblastomas? A critical assessment of existing clinical data,” *Neuro-Oncol.*, vol. 20, no. 2, pp. 184–191, Jan. 2018, doi: 10.1093/neuonc/nox175.
[55] H. K. Brar, J. Jose, Z. Wu, and M. Sharma, “Tyrosine Kinase Inhibitors for Glioblastoma Multiforme: Challenges and Opportunities for Drug Delivery,” *Pharmaceutics*, vol. 15, no. 1, p. 59, Dec. 2022, doi: 10.3390/pharmaceutics15010059.
[56] A. Shergalis, A. Bankhead, U. Luesakul, N. Muangsin, and N. Neamati, “Current Challenges and Opportunities in Treating Glioblastoma,” *Pharmacol. Rev.*, vol. 70, no. 3, pp. 412–445, Jul. 2018, doi: 10.1124/pr.117.014944.
[57] T. Sun *et al.*, “Ultrasound-mediated delivery of flexibility-tunable polymer drug conjugates for treating glioblastoma,” *Bioeng. Transl. Med.*, vol. 8, no. 2, p. e10408, 2023, doi: 10.1002/btm2.10408.
[58] L. Tang, Y. Feng, S. Gao, Q. Mu, and C. Liu, “Nanotherapeutics Overcoming the Blood-Brain Barrier for Glioblastoma Treatment,” *Front. Pharmacol.*, vol. 12, Nov. 2021, doi: 10.3389/fphar.2021.786700.
[59] A. M. Sonabend *et al.*, “Repeated blood–brain barrier opening with an implantable ultrasound device for delivery of albumin-bound paclitaxel in patients with recurrent glioblastoma: a phase 1 trial,” *Lancet Oncol.*, vol. 24, no. 5, pp. 509–522, May 2023, doi: 10.1016/S1470-2045(23)00112-2.

[60] L. G. Tataranu *et al.*, “Glioblastoma Tumor Microenvironment: An Important Modulator for Tumoral Progression and Therapy Resistance,” *Curr. Issues Mol. Biol.*, vol. 46, no. 9, Art. no. 9, Sep. 2024, doi: 10.3390/cimb46090588.
[61] H. Lin, C. Liu, A. Hu, D. Zhang, H. Yang, and Y. Mao, “Understanding the immunosuppressive microenvironment of glioma: mechanistic insights and clinical perspectives,” *J. Hematol. Oncol.J Hematol Oncol*, vol. 17, no. 1, p. 31, May 2024, doi: 10.1186/s13045-024-01544-7.
[62] Y. Liu, F. Zhou, H. Ali, J. D. Lathia, and P. Chen, “Immunotherapy for glioblastoma: current state, challenges, and future perspectives,” *Cell. Mol. Immunol.*, vol. 21, no. 12, pp. 1354–1375, Dec. 2024, doi: 10.1038/s41423-024-01226-x.
[63] J. White, M. P. J. White, A. Wickremesekera, L. Peng, and C. Gray, “The tumour microenvironment, treatment resistance and recurrence in glioblastoma,” *J. Transl. Med.*, vol. 22, no. 1, p. 540, Jun. 2024, doi: 10.1186/s12967-024-05301-9.
[64] D. K. Tripathy, L. P. Panda, S. Biswal, and K. Barhwal, “Insights into the glioblastoma tumor microenvironment: current and emerging therapeutic approaches,” *Front. Pharmacol.*, vol. 15, p. 1355242, Mar. 2024, doi: 10.3389/fphar.2024.1355242.
[65] C. Zhang and Y. Lu, “Study on artificial intelligence: The state of the art and future prospects,” *J. Ind. Inf. Integr.*, vol. 23, p. 100224, Sep. 2021, doi: 10.1016/j.jii.2021.100224.
[66] I. H. Sarker, “AI-Based Modeling: Techniques, Applications and Research Issues Towards Automation, Intelligent and Smart Systems,” *SN Comput. Sci.*, vol. 3, no. 2, p. 158, Feb. 2022, doi: 10.1007/s42979-022-01043-x.
[67] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Comput. Sci.*, vol. 2, no. 3, p. 160, Mar. 2021, doi: 10.1007/s42979-021-00592-x.
[68] D. Bertsimas and H. Wiberg, “Machine Learning in Oncology: Methods, Applications, and Challenges,” *JCO Clin. Cancer Inform.*, vol. 4, p. CCI.20.00072, Oct. 2020, doi: 10.1200/CCI.20.00072.
[69] G. Riddick *et al.*, “Predicting in vitro drug sensitivity using Random Forests,” *Bioinformatics*, vol. 27, no. 2, pp. 220–224, Jan. 2011, doi: 10.1093/bioinformatics/btq628.
[70] V. Alcazer *et al.*, “Evaluation of a machine-learning model based on laboratory parameters for the prediction of acute leukaemia subtypes: a multicentre model development and validation study in France,” *Lancet Digit. Health*, vol. 6, no. 5, pp. e323–e333, May 2024, doi: 10.1016/S2589-7500(24)00044-X.
[71] J. A. Richards, “Supervised Classification Techniques,” in *Remote Sensing Digital Image Analysis*, J. A. Richards, Ed., Cham: Springer International Publishing, 2022, pp. 263–367. doi: 10.1007/978-3-030-82327-6_8.
[72] H. Wang *et al.*, “Integrative single-cell transcriptome analysis reveals a subpopulation of fibroblasts associated with favorable prognosis of liver cancer patients,” *Transl. Oncol.*, vol. 14, no. 1, p. 100981, Jan. 2021, doi: 10.1016/j.tranon.2020.100981.
[73] X. Su, X. Yan, and C.-L. Tsai, “Linear regression,” *WIREs Comput. Stat.*, vol. 4, no. 3, pp. 275–294, 2012, doi: 10.1002/wics.1198.
[74] N. Orsini, R. Li, A. Wolk, P. Khudyakov, and D. Spiegelman, “Meta-Analysis for Linear and Nonlinear Dose-Response Relations: Examples, an Evaluation of Approximations, and Software,” *Am. J. Epidemiol.*, vol. 175, no. 1, pp. 66–73, Jan.

2012, doi: 10.1093/aje/kwr265.
[75] O. Trédan *et al.*, "Validation of prognostic scores for survival in cancer patients beyond first-line therapy," *BMC Cancer*, vol. 11, no. 1, p. 95, Mar. 2011, doi: 10.1186/1471-2407-11-95.
[76] T. G. Nick and K. M. Campbell, "Logistic Regression," in *Topics in Biostatistics*, W. T. Ambrosius, Ed., Totowa, NJ: Humana Press, 2007, pp. 273–301. doi: 10.1007/978-1-59745-530-5_14.
[77] E. C. Zabor, C. A. Reddy, R. D. Tendulkar, and S. Patil, "Logistic Regression in Clinical Studies," *Int. J. Radiat. Oncol.*, vol. 112, no. 2, pp. 271–277, Feb. 2022, doi: 10.1016/j.ijrobp.2021.08.007.
[78] J. G. Liao and K.-V. Chin, "Logistic regression for disease classification using microarray data: model selection in a large p and small n case," *Bioinforma. Oxf. Engl.*, vol. 23, no. 15, pp. 1945–1951, Aug. 2007, doi: 10.1093/bioinformatics/btm287.
[79] S. Suthaharan, "Support Vector Machine," in *Machine Learning Models and Algorithms for Big Data Classification: Thinking with Examples for Effective Learning*, S. Suthaharan, Ed., Boston, MA: Springer US, 2016, pp. 207–235. doi: 10.1007/978-1-4899-7641-3_9.
[80] A. Tharwat, "Parameter investigation of support vector machine classifier with kernel functions," *Knowl. Inf. Syst.*, vol. 61, no. 3, pp. 1269–1302, Dec. 2019, doi: 10.1007/s10115-019-01335-4.
[81] Y. Lee and C.-K. Lee, "Classification of multiple cancer types by multicategory support vector machines using gene expression data," *Bioinforma. Oxf. Engl.*, vol. 19, no. 9, pp. 1132–1139, Jun. 2003, doi: 10.1093/bioinformatics/btg102.
[82] E. Ozturk Kiyak, B. Ghasemkhani, and D. Birant, "High-Level K-Nearest Neighbors (HLKNN): A Supervised Machine Learning Model for Classification Analysis," *Electronics*, vol. 12, no. 18, p. 3828, Sep. 2023, doi: 10.3390/electronics12183828.
[83] T. J. Loftus *et al.*, "Phenotype clustering in health care: A narrative review for clinicians," *Front. Artif. Intell.*, vol. 5, Aug. 2022, doi: 10.3389/frai.2022.842306.
[84] C. Wang *et al.*, "Exploratory study on classification of lung cancer subtypes through a combined K-nearest neighbor classifier in breathomics," *Sci. Rep.*, vol. 10, no. 1, p. 5880, Apr. 2020, doi: 10.1038/s41598-020-62803-4.
[85] P. Sarang, "Naive Bayes," in *Thinking Data Science: A Data Science Practitioner's Guide*, P. Sarang, Ed., Cham: Springer International Publishing, 2023, pp. 143–152. doi: 10.1007/978-3-031-02363-7_7.
[86] B. H. Do, C. Langlotz, and C. F. Beaulieu, "Bone Tumor Diagnosis Using a Naïve Bayesian Model of Demographic and Radiographic Features," *J. Digit. Imaging*, vol. 30, no. 5, pp. 640–647, Oct. 2017, doi: 10.1007/s10278-017-0001-7.
[87] M. Langarizadeh and F. Moghbeli, "Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review," *Acta Inform. Medica*, vol. 24, no. 5, pp. 364–369, Oct. 2016, doi: 10.5455/aim.2016.24.364-369.
[88] P. Geurts, A. Irrthum, and L. Wehenkel, "Supervised learning with decision tree-based methods in computational and systems biology," *Mol. Biosyst.*, vol. 5, no. 12, pp. 1593–1605, Nov. 2009, doi: 10.1039/B907946G.
[89] P. Sarang, "Decision Tree," in *Thinking Data Science: A Data Science Practitioner's Guide*, P. Sarang, Ed., Cham: Springer International Publishing, 2023, pp. 75–96. doi: 10.1007/978-3-031-02363-7_4.

[90] M. P. Hendriks *et al.*, “Clinical decision trees support systematic evaluation of multidisciplinary team recommendations,” *Breast Cancer Res. Treat.*, vol. 183, no. 2, pp. 355–363, Sep. 2020, doi: 10.1007/s10549-020-05769-1.
[91] R. Polikar, “Ensemble Learning,” in *Ensemble Machine Learning: Methods and Applications*, C. Zhang and Y. Ma, Eds., New York, NY: Springer, 2012, pp. 1–34. doi: 10.1007/978-1-4419-9326-7_1.
[92] L. Brunese, F. Mercaldo, A. Reginelli, and A. Santone, “An ensemble learning approach for brain cancer detection exploiting radiomic features,” *Comput. Methods Programs Biomed.*, vol. 185, p. 105134, Mar. 2020, doi: 10.1016/j.cmpb.2019.105134.
[93] R. Banerjee, B. Marathi, and M. Singh, “Efficient genomic selection using ensemble learning and ensemble feature reduction,” *J. Crop Sci. Biotechnol.*, vol. 23, no. 4, pp. 311–323, Sep. 2020, doi: 10.1007/s12892-020-00039-4.
[94] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.
[95] J. Li *et al.*, “A multicenter random forest model for effective prognosis prediction in collaborative clinical research network,” *Artif. Intell. Med.*, vol. 103, p. 101814, Mar. 2020, doi: 10.1016/j.artmed.2020.101814.
[96] A. Acharjee, J. Larkman, Y. Xu, V. R. Cardoso, and G. V. Gkoutos, “A random forest based biomarker discovery and power analysis framework for diagnostics research,” *BMC Med. Genomics*, vol. 13, no. 1, p. 178, Nov. 2020, doi: 10.1186/s12920-020-00826-6.
[97] X. Su, A. T. Peña, L. Liu, and R. A. Levine, “Random forests of interaction trees for estimating individualized treatment effects in randomized trials,” *Stat. Med.*, vol. 37, no. 17, pp. 2547–2560, 2018, doi: 10.1002/sim.7660.
[98] R. Poursaeed, M. Mohammadzadeh, and A. A. Safaei, “Survival prediction of glioblastoma patients using machine learning and deep learning: a systematic review,” *BMC Cancer*, vol. 24, p. 1581, Dec. 2024, doi: 10.1186/s12885-024-13320-4.
[99] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.
[100] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artif. Intell. Rev.*, vol. 54, no. 3, pp. 1937–1967, Mar. 2021, doi: 10.1007/s10462-020-09896-5.
[101] J. H. Friedman, “Greedy Function Approximation: A Gradient Boosting Machine,” *Ann. Stat.*, vol. 29, no. 5, pp. 1189–1232, 2001.
[102] G. Ke *et al.*, “LightGBM: A Highly Efficient Gradient Boosting Decision Tree,” in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds., Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf
[103] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, “CatBoost: unbiased boosting with categorical features,” in *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Curran Associates, Inc., 2018. [Online]. Available:

https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf
[104] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[105] F. Ekundayo, "Reinforcement learning in treatment pathway optimization: A case study in oncology," *Int. J. Sci. Res. Arch.*, vol. 13, no. 2, Art. no. 2, 2024, doi: 10.30574/ijsra.2024.13.2.2450.
[106] J.-N. Eckardt, K. Wendt, M. Bornhäuser, and J. M. Middeke, "Reinforcement Learning for Precision Oncology," *Cancers*, vol. 13, no. 18, p. 4624, Sep. 2021, doi: 10.3390/cancers13184624.
[107] D. Niraula, J. Jamaluddin, M. M. Matuszak, R. K. T. Haken, and I. E. Naqa, "Quantum deep reinforcement learning for clinical decision support in oncology: application to adaptive radiotherapy," *Sci. Rep.*, vol. 11, no. 1, p. 23545, Dec. 2021, doi: 10.1038/s41598-021-02910-y.
[108] P. P. Angelov, E. A. Soares, R. Jiang, N. I. Arnold, and P. M. Atkinson, "Explainable artificial intelligence: an analytical review," *WIREs Data Min. Knowl. Discov.*, vol. 11, no. 5, p. e1424, 2021, doi: 10.1002/widm.1424.
[109] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of Machine Learning Models Using Improved Shapley Additive Explanation," in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, Niagara Falls NY USA: ACM, Sep. 2019, pp. 546–546. doi: 10.1145/3307339.3343255.
[110] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer, "Permutation importance: a corrected feature importance measure," *Bioinformatics*, vol. 26, no. 10, pp. 1340–1347, May 2010, doi: 10.1093/bioinformatics/btq134.
[111] N. Ullah, M. Hassan, J. A. Khan, M. S. Anwar, and K. Aurangzeb, "Enhancing explainability in brain tumor detection: A novel DeepEBTDNet model with LIME on MRI images," *Int. J. Imaging Syst. Technol.*, vol. 34, no. 1, p. e23012, Jan. 2024, doi: 10.1002/ima.23012.
[112] M. M. M, M. T. R, V. K. V, and S. Guluwadi, "Enhancing brain tumor detection in MRI images through explainable AI using Grad-CAM with Resnet 50," *BMC Med. Imaging*, vol. 24, no. 1, p. 107, May 2024, doi: 10.1186/s12880-024-01292-7.
[113] T. Geldof, N. Van Damme, I. Huys, and W. Van Dyck, "Patient-Level Effectiveness Prediction Modeling for Glioblastoma Using Classification Trees," *Front. Pharmacol.*, vol. 10, p. 1665, 2019, doi: 10.3389/fphar.2019.01665.
[114] J. A. Cruz and D. S. Wishart, "Applications of Machine Learning in Cancer Prediction and Prognosis," *Cancer Inform.*, vol. 2, pp. 59–77, Feb. 2007.
[115] R. Cuocolo, M. Caruso, T. Perillo, L. Ugga, and M. Petretta, "Machine Learning in oncology: A clinical appraisal," *Cancer Lett.*, vol. 481, pp. 55–62, Jul. 2020, doi: 10.1016/j.canlet.2020.03.032.
[116] I. Z. Shen and L. Zhang, "Digital and Artificial Intelligence-based Pathology: Not for Every Laboratory – A Mini-review on the Benefits and Pitfalls of Its Implementation," *J. Clin. Transl. Pathol.*, vol. 000, no. 000, pp. 000–000, Apr. 2025, doi: 10.14218/JCTP.2025.00007.

[117] A. A. Ahmed, M. Abouzid, and E. Kaczmarek, “Deep Learning Approaches in Histopathology,” *Cancers*, vol. 14, no. 21, Art. no. 21, Jan. 2022, doi: 10.3390/cancers14215264.
[118] X. Jiang, Z. Hu, S. Wang, and Y. Zhang, “Deep Learning for Medical Image-Based Cancer Diagnosis,” *Cancers*, vol. 15, no. 14, Art. no. 14, Jan. 2023, doi: 10.3390/cancers15143608.
[119] S. Dabeer, M. M. Khan, and S. Islam, “Cancer diagnosis in histopathological image: CNN based approach,” *Inform. Med. Unlocked*, vol. 16, p. 100231, Jan. 2019, doi: 10.1016/j.imu.2019.100231.
[120] P. H. Luckett *et al.*, “Predicting survival in glioblastoma with multimodal neuroimaging and machine learning,” *J. Neurooncol.*, vol. 164, no. 2, pp. 309–320, Sep. 2023, doi: 10.1007/s11060-023-04439-8.
[121] L. S. Hu *et al.*, “Radiogenomics to characterize regional genetic heterogeneity in glioblastoma,” *Neuro-Oncol.*, vol. 19, no. 1, pp. 128–137, Jan. 2017, doi: 10.1093/neuonc/now135.
[122] Y. Choi *et al.*, “IDH1 mutation prediction using MR-based radiomics in glioblastoma: comparison between manual and fully automated deep learning-based approach of tumor segmentation,” *Eur. J. Radiol.*, vol. 128, Jul. 2020, doi: 10.1016/j.ejrad.2020.109031.
[123] Y. Sakai *et al.*, “MRI Radiomic Features to Predict IDH1 Mutation Status in Gliomas: A Machine Learning Approach using Gradient Tree Boosting,” *Int. J. Mol. Sci.*, vol. 21, no. 21, p. 8004, Oct. 2020, doi: 10.3390/ijms21218004.
[124] X. Hu, K. K. Wong, G. S. Young, L. Guo, and S. T. Wong, “Support Vector Machine (SVM) Multi-parametric MRI Identification of Pseudoprogression from Tumor Recurrence in Patients with Resected Glioblastoma,” *J. Magn. Reson. Imaging*, vol. 33, no. 2, pp. 296–305, Feb. 2011, doi: 10.1002/jmri.22432.
[125] H. Chen *et al.*, “Deep Learning Radiomics to Predict PTEN Mutation Status From Magnetic Resonance Imaging in Patients With Glioma,” *Front. Oncol.*, vol. 11, p. 734433, 2021, doi: 10.3389/fonc.2021.734433.
[126] Y. Li *et al.*, “Radiogenomic analysis of PTEN mutation in glioblastoma using preoperative multi-parametric magnetic resonance imaging,” *Neuroradiology*, vol. 61, no. 11, pp. 1229–1237, Nov. 2019, doi: 10.1007/s00234-019-02244-7.
[127] M. Pandey, P. Anoosha, D. Yesudhas, and M. M. Gromiha, “Identification of potential driver mutations in glioblastoma using machine learning,” *Brief. Bioinform.*, vol. 23, no. 6, p. bbac451, Nov. 2022, doi: 10.1093/bib/bbac451.
[128] H. Zhang *et al.*, “Machine learning-based tumor-infiltrating immune cell-associated lncRNAs for predicting prognosis and immunotherapy response in patients with glioblastoma,” *Brief. Bioinform.*, vol. 23, no. 6, p. bbac386, Nov. 2022, doi: 10.1093/bib/bbac386.
[129] Y. Kim, K. H. Kim, J. Park, H. I. Yoon, and W. Sung, “Prognosis prediction for glioblastoma multiforme patients using machine learning approaches: Development of the clinically applicable model,” *Radiother. Oncol.*, vol. 183, p. 109617, Jun. 2023, doi: 10.1016/j.radonc.2023.109617.
[130] E. Audureau *et al.*, “Prognostic factors for survival in adult patients with recurrent glioblastoma: a decision-tree-based model,” *J. Neurooncol.*, vol. 136, no. 3, pp. 565–576, Feb. 2018, doi: 10.1007/s11060-017-2685-4.
[131] H. Moradmand, S. M. R. Aghamiri, R. Ghaderi, and H. Emami, “The role of

deep learning-based survival model in improving survival prediction of patients with glioblastoma," *Cancer Med.*, vol. 10, no. 20, pp. 7048–7059, 2021, doi: 10.1002/cam4.4230.
[132] J. C. Peeken, J. Hesse, B. Haller, K. A. Kessel, F. Nüsslin, and S. E. Combs, "Semantic imaging features predict disease progression and survival in glioblastoma multiforme patients," *Strahlenther. Onkol.*, vol. 194, no. 6, pp. 580–590, Jun. 2018, doi: 10.1007/s00066-018-1276-4.
[133] H. G. Yoon *et al.*, "Multi-Parametric Deep Learning Model for Prediction of Overall Survival after Postoperative Concurrent Chemoradiotherapy in Glioblastoma Patients," *Cancers*, vol. 12, no. 8, Art. no. 8, Aug. 2020, doi: 10.3390/cancers12082284.
[134] D. T. Do, M.-R. Yang, L. H. T. Lam, N. Q. K. Le, and Y.-W. Wu, "Improving MGMT methylation status prediction of glioblastoma through optimizing radiomics features using genetic algorithm-based machine learning approach," *Sci. Rep.*, vol. 12, no. 1, p. 13412, Aug. 2022, doi: 10.1038/s41598-022-17707-w.
[135] S. Faghani, B. Khosravi, M. Moassefi, G. M. Conte, and B. J. Erickson, "A Comparison of Three Different Deep Learning-Based Models to Predict the MGMT Promoter Methylation Status in Glioblastoma Using Brain MRI," *J. Digit. Imaging*, vol. 36, no. 3, pp. 837–846, Jun. 2023, doi: 10.1007/s10278-022-00757-x.
[136] Y. Zhao, X. He, and Q. Wan, "Combined machine learning models, docking analysis, ADMET studies and molecular dynamics simulations for the design of novel FAK inhibitors against glioblastoma," *BMC Chem.*, vol. 18, no. 1, p. 203, Oct. 2024, doi: 10.1186/s13065-024-01316-x.
[137] B. J. Neves *et al.*, "Efficient identification of novel anti-glioma lead compounds by machine learning models," *Eur. J. Med. Chem.*, vol. 189, p. 111981, Mar. 2020, doi: 10.1016/j.ejmech.2019.111981.
[138] J. Ju *et al.*, "Two-Step Transfer Learning Improves Deep Learning-Based Drug Response Prediction in Small Datasets: A Case Study of Glioblastoma," *Bioinforma. Biol. Insights*, vol. 19, p. 11779322241301507, 2025, doi: 10.1177/11779322241301507.
[139] A. Ghasemi, S. Hashtarkhani, D. L. Schwartz, and A. Shaban-Nejad, "Explainable artificial intelligence in breast cancer detection and risk prediction: A systematic scoping review," *Cancer Innov.*, vol. 3, no. 5, p. e136, Oct. 2024, doi: 10.1002/cai2.136.
[140] O. O. Oladimeji, H. Ayaz, I. McLoughlin, and S. Unnikrishnan, "Mutual information-based radiomic feature selection with SHAP explainability for breast cancer diagnosis," *Results Eng.*, vol. 24, p. 103071, Dec. 2024, doi: 10.1016/j.rineng.2024.103071.
[141] R. O. Alabi, M. Elmusrati, I. Leivo, A. Almangush, and A. A. Mäkitie, "Machine learning explainability in nasopharyngeal cancer survival using LIME and SHAP," *Sci. Rep.*, vol. 13, no. 1, p. 8984, Jun. 2023, doi: 10.1038/s41598-023-35795-0.
[142] S. Tang *et al.*, "Prostate cancer treatment recommendation study based on machine learning and SHAP interpreter," *Cancer Sci.*, vol. 115, no. 11, pp. 3755–3766, Nov. 2024, doi: 10.1111/cas.16327.
[143] N. Abuzinadah *et al.*, "Improved Prediction of Ovarian Cancer Using Ensemble Classifier and Shaply Explainable AI," *Cancers*, vol. 15, no. 24, Art. no. 24,

Jan. 2023, doi: 10.3390/cancers15245793.
[144] W. Yang *et al.*, "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells," *Nucleic Acids Res.*, vol. 41, no. D1, pp. D955–D961, Jan. 2013, doi: 10.1093/nar/gks1111.
[145] D. J. Vis, L. Bombardelli, H. Lightfoot, F. Iorio, M. J. Garnett, and L. F. Wessels, "Multilevel models improve precision and speed of IC50 estimates," *Pharmacogenomics*, vol. 17, no. 7, pp. 691–700, May 2016, doi: 10.2217/pgs.16.15.
[146] S. Kim *et al.*, "PubChem 2025 update," *Nucleic Acids Res.*, vol. 53, no. D1, pp. D1516–D1525, Jan. 2025, doi: 10.1093/nar/gkae1059.
[147] V. Stathias *et al.*, "LINCS Data Portal 2.0: next generation access point for perturbation-response signatures," *Nucleic Acids Res.*, vol. 48, no. D1, pp. D431–D439, Jan. 2020, doi: 10.1093/nar/gkz1023.
[148] X. Chen, Z. L. Ji, and Y. Z. Chen, "TTD: Therapeutic Target Database," *Nucleic Acids Res.*, vol. 30, no. 1, pp. 412–415, Jan. 2002, doi: 10.1093/nar/30.1.412.
[149] H. Zhou and J. Skolnick, "Utility of the Morgan Fingerprint in Structure-Based Virtual Ligand Screening," *J. Phys. Chem. B*, vol. 128, no. 22, pp. 5363–5370, Jun. 2024, doi: 10.1021/acs.jpcb.4c01875.
[150] 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, no. 1, p. 42717, Mar. 2017, doi: 10.1038/srep42717.
[151] P. Agarwal, J. Huckle, J. Newman, and D. L. Reid, "Trends in small molecule drug properties: A developability molecule assessment perspective," *Drug Discov. Today*, vol. 27, no. 12, p. 103366, Dec. 2022, doi: 10.1016/j.drudis.2022.103366.
[152] M. Sorokina, P. Merseburger, K. Rajan, M. A. Yirik, and C. Steinbeck, "COCONUT online: Collection of Open Natural Products database," *J. Cheminformatics*, vol. 13, no. 1, p. 2, Jan. 2021, doi: 10.1186/s13321-020-00478-9.
[153] J. Nickel *et al.*, "SuperPred: update on drug classification and target prediction," *Nucleic Acids Res.*, vol. 42, no. W1, pp. W26–W31, Jul. 2014, doi: 10.1093/nar/gku477.
[154] The UniProt Consortium, "UniProt: the Universal Protein Knowledgebase in 2025," *Nucleic Acids Res.*, vol. 53, no. D1, pp. D609–D617, Jan. 2025, doi: 10.1093/nar/gkae1010.
[155] S. X. Ge, D. Jung, and R. Yao, "ShinyGO: a graphical gene-set enrichment tool for animals and plants," *Bioinformatics*, vol. 36, no. 8, pp. 2628–2629, Apr. 2020, doi: 10.1093/bioinformatics/btz931.
[156] T. Liao and M. Yang, "Revisiting epithelial-mesenchymal transition in cancer metastasis: the connection between epithelial plasticity and stemness," *Mol. Oncol.*, vol. 11, no. 7, pp. 792–804, Jul. 2017, doi: 10.1002/1878-0261.12096.
[157] S. Qi *et al.*, "ZEB2 Mediates Multiple Pathways Regulating Cell Proliferation, Migration, Invasion, and Apoptosis in Glioma," *PLoS ONE*, vol. 7, no. 6, p. e38842, Jun. 2012, doi: 10.1371/journal.pone.0038842.
[158] S. Safaee *et al.*, "Silencing ZEB2 Induces Apoptosis and Reduces Viability in Glioblastoma Cell Lines," *Molecules*, vol. 26, no. 4, Art. no. 4, Jan. 2021, doi: 10.3390/molecules26040901.
[159] G. Song *et al.*, "Molecular insights into the human ABCB6 transporter," *Cell Discov.*, vol. 7, no. 1, pp. 1–11, Jul. 2021, doi: 10.1038/s41421-021-00284-z.

[160] S.-G. Zhao *et al.*, “Increased Expression of ABCB6 Enhances Protoporphyrin IX Accumulation and Photodynamic Effect in Human Glioma,” *Ann. Surg. Oncol.*, vol. 20, no. 13, pp. 4379–4388, Dec. 2013, doi: 10.1245/s10434-011-2201-6.
[161] S. Kamran, A. Sinniah, M. A. M. Abdulghani, and M. A. Alshawsh, “Therapeutic Potential of Certain Terpenoids as Anticancer Agents: A Scoping Review,” *Cancers*, vol. 14, no. 5, p. 1100, Feb. 2022, doi: 10.3390/cancers14051100.
[162] F. Ballout, Z. Habli, A. Monzer, O. N. Rahal, M. Fatfat, and H. Gali-Muhtasib, “Anticancer Alkaloids: Molecular Mechanisms and Clinical Manifestations,” in *Bioactive Natural Products for the Management of Cancer: from Bench to Bedside*, A. K. Sharma, Ed., Singapore: Springer Singapore, 2019, pp. 1–35. doi: 10.1007/978-981-13-7607-8_1.
[163] W. Zheng *et al.*, “Inhibition of Cathepsin D (CTSD) enhances radiosensitivity of glioblastoma cells by attenuating autophagy,” *Mol. Carcinog.*, vol. 59, no. 6, pp. 651–660, Jun. 2020, doi: 10.1002/mc.23194.
[164] M. T. Ferreira *et al.*, “Cyclooxygenase Inhibition Alters Proliferative, Migratory, and Invasive Properties of Human Glioblastoma Cells In Vitro,” *Int. J. Mol. Sci.*, vol. 22, no. 9, p. 4297, Apr. 2021, doi: 10.3390/ijms22094297.
[165] D. Lv *et al.*, “TRIM24 is an oncogenic transcriptional co-activator of STAT3 in glioblastoma,” *Nat. Commun.*, vol. 8, no. 1, p. 1454, Nov. 2017, doi: 10.1038/s41467-017-01731-w.
[166] S. Kushal *et al.*, “Monoamine oxidase A (MAO A) inhibitors decrease glioma progression,” *Oncotarget*, vol. 7, no. 12, pp. 13842–13853, Feb. 2016, doi: 10.18632/oncotarget.7283.
[167] T. Ströbel *et al.*, “Ape1 guides DNA repair pathway choice that is associated with drug tolerance in glioblastoma,” *Sci. Rep.*, vol. 7, no. 1, p. 9674, Aug. 2017, doi: 10.1038/s41598-017-10013-w.
[168] A. L. Hudson *et al.*, “Glioblastoma Recurrence Correlates With Increased APE1 and Polarization Toward an Immuno-Suppressive Microenvironment,” *Front. Oncol.*, vol. 8, p. 314, Aug. 2018, doi: 10.3389/fonc.2018.00314.
[169] E. J. Siney, A. Holden, E. Casselden, H. Bulstrode, G. J. Thomas, and S. Willaime-Morawek, “Metalloproteinases ADAM10 and ADAM17 Mediate Migration and Differentiation in Glioblastoma Sphere-Forming Cells,” *Mol. Neurobiol.*, vol. 54, no. 5, pp. 3893–3905, Jul. 2017, doi: 10.1007/s12035-016-0053-6.
[170] Y. Wang, P. Long, Y. Wang, and W. Ma, “NTRK Fusions and TRK Inhibitors: Potential Targeted Therapies for Adult Glioblastoma,” *Front. Oncol.*, vol. 10, p. 593578, Nov. 2020, doi: 10.3389/fonc.2020.593578.
[171] W. Wang *et al.*, “Tyrosyl-DNA Phosphodiesterase 1 and Topoisomerase I Activities as Predictive Indicators for Glioblastoma Susceptibility to Genotoxic Agents,” *Cancers*, vol. 11, no. 10, Art. no. 10, Oct. 2019, doi: 10.3390/cancers11101416.
[172] W. Fu, X. Hou, L. Dong, and W. Hou, “Roles of STAT3 in the pathogenesis and treatment of glioblastoma,” *Front. Cell Dev. Biol.*, vol. 11, Feb. 2023, doi: 10.3389/fcell.2023.1098482.
[173] X. Kang, J. Chen, and J. Hou, “HSP90 facilitates stemness and enhances glycolysis in glioma cells,” *BMC Neurol.*, vol. 22, p. 420, Nov. 2022, doi: 10.1186/s12883-022-02924-7.
[174] K. Yang, L. Luan, X. Li, X. Sun, and J. Yin, “ERK inhibition in glioblastoma

is associated with autophagy activation and tumorigenesis suppression," *J. Neurooncol.*, vol. 156, no. 1, pp. 123–137, Jan. 2022, doi: 10.1007/s11060-021-03896-3.
[175] M. M. Singh *et al.*, "Inhibition of LSD1 sensitizes glioblastoma cells to histone deacetylase inhibitors," *Neuro-Oncol.*, vol. 13, no. 8, pp. 894–903, Aug. 2011, doi: 10.1093/neuonc/nor049.
[176] C. D. Saccà
*et al.*, "Inhibition of lysine-specific demethylase LSD1 induces senescence in Glioblastoma cells through a HIF-1α-dependent pathway," *Biochim. Biophys. Acta BBA - Gene Regul. Mech.*, vol. 1862, no. 5, pp. 535–546, May 2019, doi: 10.1016/j.bbagrm.2019.03.004.
[177] L. Qiang *et al.*, "HIF-1α is critical for hypoxia-mediated maintenance of glioblastoma stem cells by activating Notch signaling pathway," *Cell Death Differ.*, vol. 19, no. 2, pp. 284–294, Feb. 2012, doi: 10.1038/cdd.2011.95.

# LIST OF PUBLICATIONS

1. Journal publication in Ageing Research Reviews "E2 conjugating enzymes: A silent but crucial player in ubiquitin biology."

Ageing Research Reviews 108 (2025) 102740

Contents lists available at ScienceDirect

Ageing Research Reviews

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

## E2 conjugating enzymes: A silent but crucial player in ubiquitin biology

Somya Parashar [a], Aastha Kaushik [a], Rashmi K. Ambasta [b], Pravir Kumar [a,*,1,2,3]

[a] Molecular Neuroscience and Functional Genomics Laboratory, Department of Biotechnology, Delhi Technological University (Formerly Delhi College of Engineering), Shahbad Daulatpur, Bawana Road, Delhi 110042, India
[b] Department of Medicine, Vanderbilt University Medical Center (VUMC), Nashville, TN, USA

ARTICLE INFO

*Keywords:*
E2 conjugating enzymes
Structural determinants
PTMs
Deubiquitinase
NDDs

ABSTRACT

E2 conjugating enzymes serve as the linchpin of the Ubiquitin-Proteasome System (UPS), facilitating ubiquitin (Ub) transfer to substrate proteins and regulating diverse processes critical to cellular homeostasis. The interaction of E2s with E1 activating enzymes and E3 ligases singularly positions them as middlemen of the ubiquitin machinery that guides protein turnover. Structural determinants of E2 enzymes play a pivotal role in these interactions, enabling precise ubiquitin transfer and substrate specificity. Regulation of E2 enzymes is tightly controlled through mechanisms such as post-translational modifications (PTMs), allosteric control, and gene expression modulation. Specific residues that undergo PTMs highlight their impact on E2 function and their role in ubiquitin dynamics. E2 enzymes also cooperate with deubiquitinases (DUBs) to maintain proteostasis. Design of small molecule inhibitors to modulate E2 activity is emerging as promising avenue to restrict ubiquitination as a potential therapeutic intervention. Additionally, E2 enzymes have been implicated in the pathogenesis and

*Abbreviations:* Aβ, Amyloid beta; ACBD 3/5, Acyl-CoA binding domain containing protein 3/5; AD, Alzheimer's Disease; ALS, Amyotrophic Lateral Sclerosis; AMPAR, AMPA Receptor; APC/C, Anaphase Promoting Complex/Cyclosome; APPBP1, Amyloid Beta Precursor, Protein Binding Protein 1; AR, Androgen Receptor; ATG8, Autophagy-related protein; β-TrCP1, beta Transducin repeat Containing Protein 1; B4GALT1, beta-1.4-galactosyltransferase 1; BIRC 6/7, Baculoviral IAP Repeat Containing 6/7; B-Myb, MYB-related protein B; BRCA1, Breast Cancer Type 1 susceptibility protein; BARD1, BRCA1 Associated RING Domain protein 1; BST2, Bone Marrow Stromal Antigen 2; CDK, Cyclin Dependent Kinase; Cdt1, Chromatin licensing and DNA replication Factor 1; CGAS, cyclic GMP-AMP synthase; CHIP, C-terminus Hsc70-Interacting Protein; CNOT, CCR4-NOT transcription complex; CRL, Cullin RING E3 Ligases; CTLH, C-terminal to LiSH; DDR, DNA damage response; Drp1, Dynamin-related protein 1; DUB, Deubiquitinase; EBV, Epstein Barr Virus; ERAD, Endoplasmic Reticulum Associated Degradation; FACET-IP, Fractionated ACETylation IP; FAK, Focal Adhesion Kinase; FLT3, Fms-like Tyrosine Kinase 3; FANCD2/I, Fanconi Anemia Complementation group D2/I; G2BR, UBE2G2 Binding Region; GID, Glucose-induced degradation-deficient E3; Gp78, glycoprotein 78; HD, Huntington's Disease; HDAC, Histone Deacetylase; HECT, Homologous to E6-AP Carboxyl Terminus; HERP, Homocysteine ER-induced Protein; Hmmr, Hyaluronan-Mediated Motility Receptor; HOIP, HOIL-1 Interacting Protein; HRD1, HMG-CoA reductase degradation protein 1; HRMS, High-resolution MS; HURP, Hepatoma Up-Regulated Protein; IKZF1/3, Ikaros family Zinc Finger 1/3; ISG15, Interferon Stimulated Gene 15; LPP, Lipoma Preferred Partner; LUBAC, Linear Ubiquitin Chain Assembly Complex; March1/7, Membrane Associated RING CH-1/7; MHtt, mutant Huntingtin; NAE1, NEDD8 activating enzyme subunit 1; NDD, Neurodegenerative diseases; NEDD, Neural precursor cell Expressed Developmentally Down-regulated; NEMO, NF-kB Essential Modulator; NF-kB, Nuclear Factor kappa-light-chain enhancer of B cells; NMDA, N-methyl-D-aspartate; NPRL2, Nuclear Protein Localization Regulator 2; NTD, N-terminal Domain; NuSAP, Nuleolar and Spindle Associated Protein; ORC1, Origin Recognition Complex subunit 1; OspG/I, Outer Surface Protein G/I; OTUB1, OTU deubiquitinase; PD, Parkinson's Disease; PIAS1, Protein inhibitor of activated STAT1; PML-RARα, Promyelocytic leukemia-Retinoic Acid Receptor alpha; PolyQ, polyglutamine; PCNA, Proliferating Cell Nuclear Antigen; PROTOMAP, Protein Topography and Migration Analysis Platform; PTM, Post translational modifications; RBR, RING between RING; RIG-I, Retinoic Acid Inducible Gene I; RING, Really Interesting New Gene; RNF, RING Finger protein; RPB8, RNA Polymerase II subunit B8; RWD, RING finger and WD containing domain; SAE, SUMO activating enzyme; SCA, Spinocerebellar Ataxia; SALL1, Sal-like protein 1; SCF, Skp1-Cullin1-F-box protein; SEPTM, Serial enrichment of PTMs; SILAC, Stable isotope labeling by/with amino acids in cell culture; SLC7A11, Solute carrier family 7 member 11; SMURF1/2, Smad Ubiquitin Regulatory Factor 1/2; SNUSP, serial NEDD8-ubiquitin substrate profiling; SQSTM1, Sequestosome 1; STAT1, Signal Transducer and Activator of Transcritption 1; SUMO, Small Ubiquitin-like Modifier; SYMD3, SET and MYD domain containing protein 3; TDP-43, TAR DNA, binding protein-43; TMEM135, Transmembrane protein 135; TRAF6, TNF-receptor associated Factor 6; TRIM, Tripartite Motif Containing; TSSC5, Tumor-suppressing subchromosomal transferable fragment cDNA; Ub, Ubiquitin; UBA, Ubiquitin activating; UBC, Ubiquitin conjugating; Ubl, Ubiquitin-like; UFD, Ubiquitin Fold Domain; UFC1, Ubiquitin-Fold Modifier Conjugating enzyme 1; UFM1, Ubiquitin-Fold Modifier 1; UPS, Ubiquitin Proteasome System; ZNF, Zinc finger protein.

* Correspondence to: Molecular Neuroscience and Functional Genomics Laboratory, Delhi Technological University, Delhi, India.
*E-mail addresses:* kpravir@gmail.com, pravirkumar@dtu.ac.in (P. Kumar).
[1] ORCID ID: 0000–0001-7444–2344
[2] Room# FW4TF3, Mechanical Engineering Building, Shahbad Daulatpur, Bawana Road, Delhi 110042
[3] https://www.ncbi.nlm.nih.gov/myncbi/1DW465XG9bp5p/bibliography/public/

https://doi.org/10.1016/j.arr.2025.102740
Received 1 February 2025; Received in revised form 14 March 2025; Accepted 19 March 2025
Available online 5 April 2025

2. Journal publication in Ageing Research Reviews “Ubiquitin E3 ligases assisted technologies in protein degradation: Sharing pathways in neurodegenerative disorders and cancer”

Ageing Research Reviews 96 (2024) 102279

Contents lists available at ScienceDirect

Ageing Research Reviews

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

Review article

# Ubiquitin E3 ligases assisted technologies in protein degradation: Sharing pathways in neurodegenerative disorders and cancer

Aastha Kaushik [a,1], Somya Parashar [a,1], Rashmi K. Ambasta [b], Pravir Kumar [a,*,2,3]

[a] Molecular Neuroscience and Functional Genomics Laboratory, Department of Biotechnology, Delhi Technological University (Formerly DCE), Delhi 110042, India
[b] Department of Biotechnology and Microbiology, SRM University-Sonepat, Haryana, India

*Abbreviations:* AR, Androgen Receptor; ARIH1, Ariadne-1 homolog; ABCB1, ATP-binding cassette sub-family B member 1; AUTAC, Autophagy Targeting Chimera; APP, Aβ precursor protein; BMI1, B lymphoma Mo-MLV insertion region 1 homolog; BIRC7, Baculoviral IAP Repeat-Containing protein 7; BCL-XL, B-cell lymphoma-extra-large; BAG5, Bcl-2-associated athanogene 5; BECN1, Beclin-1; β-TrCP, Beta-Transducin repeats-Containing Proteins; BRI3, Brain Protein I3; BARD1, BRCA1-Associated RING Domain protein 1; BRD4, Bromodomain-containing protein 4; Cdh1, Cadherin-1; CREB, cAMP Response Element-Binding protein; CDC20, Cell Division Cycle 20; CPP, Cell Penetrating Peptide; CIAP, cellular Inhibitor of Apoptosis; CRBN, Cereblon; CDT1, Chromatin licensing and DNA replication factor 1; CLL, Chronic lymphocytic leukemia; CMML, Chronic Myelomonocytic Leukaemia; CHIP, C-terminus of Hsc70 Interacting Protein; Cul1, Cullin 1; CRL, Cullin-RING Ubiquitin E3 Ligase; CNrasGEF, Cyclic Nucleotide ras Guanine-nucleotide-Exchange Factor; CCNF, Cyclin F; CDK, Cyclin-dependent Kinase; DAXX, Death domain-associated protein; DUBs, Deubiquitinating enzymes; DLBCL, Diffuse large B cell lymphoma; DDR, DNA-damage response; E6-AP, E6-associated protein; EMI1, Early Mitotic Inhibitor 1; 4EHP, eIF4E-Homologous Protein; EBPβ, Enhancer-Binding Protein β; ErbB4, Erb-B2 Receptor Tyrosine Kinase 4; EBV, Epstein Barr Virus; ER, Estrogen Receptor; ERα, Estrogen receptor α; ELK1, ETS Like-1 protein; ERK1/2, Extracellular signal-regulated kinase 1/2; FANC, FA Complementation; Fbxo7, F-box domain-containing protein; FBXW7, F-box/WD repeat-containing protein 7; FGFR1, Fibroblast Growth Factor Receptor 1; Fzr1, Fizzy-related protein; FLIP, FLICE-like inhibitory protein; FL, Follicular lymphoma; FOXO3, Forkhead box O3; GID complex, Glucose-Induced Degradation Deficient complex; HOIL1, Haem-oxidised IRP2 ubiquitin ligase-1; HSPA2, Heat Shock Protein Family A Hsp70 Member 2 protein; HSF1, Heat Shock Transcription Factor 1; HACE1, HECT domain and Ankyrin repeat Containing E3 ubiquitin protein ligase 1; HECW1/2, HECT C2 and WW domain containing E3 ubiquitin protein ligase 1/2; HOIP, HOIL-1-interacting protein; HoxA10, Homeobox A10; HECT, Homologous to the E6AP-Carboxy terminus; HspBP1, HSPA Hsp70-Binding Protein 1; HER3, Human Epidermal growth factor Receptor 3; HTT, Huntingtin; HD, Huntington's Disease; HyT, Hydrophobic Tagging; IKZF1/3, Ikaros family Zinc Finger protein 1/3; IMiDs, Immune Modulatory Drugs; CLIPTAC, In-cell click-formed proteolysis targeting chimeras; IPSCs, Induced pluripotent stem cells; IDOL, Inducible Degrader of the low-density Lipoprotein receptor; ITM2b, integral membrane protein; JNK, Jun N-terminal Kinase; LATS1, Large tumor suppressor kinase 1; LRRK2, leucine-rich repeat kinase 2; LUBAC, Linear Ubiquitin Assembly Complex; LAPTM5, Lysosomal-associated transmembrane protein 5; MCL, Mantle cell lymphoma; MZL, Marginal zone lymphoma; MARCH5, Membrane-Associated ring finger C3HC4 5; MCRPC, Metastatic castration-resistant prostate cancer; LC3, Microtubule-associated proteins 1A/1B light chain 3B; MAM, mitochondria-associated membrane; MITOL, Mitochondrial ubiquitin Ligase; MAPK, Mitogen-activated protein kinases; SMAD2, Mothers against decapentaplegic homolog 2; MDM2, Mouse Double Minute 2 homolog; MDMX, Murine Double Minute X; MuRF1, Muscle RING-finger protein-1; MHtt, mutant Huntingtin; N4BP, NEDD4-binding protein; NEDD4, Neural precursor cell Expressed Developmentally Down-regulated protein 4; NDD, Neurodegenerative Diseases; NFT, Neurofibrillary tangles; NOTCH1, Neurogenic locus notch homolog protein 1; NLRP3, NLR family pyrin domain containing 3 protein; NMDA receptor, N-methyl-D-aspartate receptor; NRF2, Nuclear Factor Erythroid 2; NMR, Nuclear Magnetic Resonance; NRBP1, Nuclear Receptor Binding Protein 1; OPA1, Optic Atrophy 1; OPTN, Optineurin; PK, Parkin; PD, Parkinson Disease; Peli1, Pellino E3 ubiquitin ligase; PUB, Peptide N-glycanase/UBA or UBX-containing proteins; PTEN, Phosphatase and Tensin homolog; PI3K, phosphatidylinositol 3-Kinases; Plk1, Polo-like Kinase 1; PARP, Poly ADP-ribose polymerases; PolyQ, Polyglutamate; Prp19, Pre-mRNA-processing factor 19; PCNSL, Primary CNS lymphoma; PDL-1, Programmed death-ligand 1; PML-RARα, Promyelocytic Leukemia/Retinoic Acid Receptor α; PIAS1, Protein Inhibitor of Activated STAT1; PKCζ, Protein Kinase C zeta; PERK, Protein Kinase RNA-Like ER Kinase; PROTAC, Proteolytic Targeting Chimera; PINK-1, PTEN-induced putative kinase-1; RBCK1, RanBP-type and C3HC4-type zinc finger-containing protein 1; RASSF5, Ras association domain-containing protein 5; RING, Really Interesting New Genes; Rock2, Rho protein kinase 2; RBR, RING Between RING; RNF4, Ring Finger 4; RNAP II, RNA Polymerase II; SAG, Sensitive to Apoptosis Gene; SQSTM1, Sequestosome1; SIAH1, Seven in Absentia Homolog 1; STAT1, Signal Transducer and Activator of Transcription 1; SNON protein, Ski novel protein; SCF, Skp1-Cullin-F-box protein; SMURF1/2, Small ubiquitination regulatory factor 1/2; SLL, Small Lymphocytic Lymphoma; SNIPER, Specific and Nongenetic IAP-based Protein Erasers; SKP1/2, S-phase Kinase associated Protein 1/2; STING, Stimulator of Interferon Genes; SOD1, Superoxide Dismutase-1; SYVN1, Synoviolin 1; TDP43, TAR DNA Protein-43; TPR, Tetratricopeptide repeat; TF-PROTACs, Transcription Factor PROTACs; TGF-β, Transforming Growth Factor β; TRAP, Translocon protein; TIR-1, Transport Inhibitor Response-1; TRIM, Tripartite Motif; TNBC, Triple Negative Breast Cancer; TRAID1, Two RING fingers and DRIL1; UPS, Ubiquitin Proteasome System; Ubc, Ubiquitin-conjugating complex; UBR5, Ubiquitin-protein ligase N-recognin 5; USP7, Ubiquitin-specific processing protease 7; VCP, Valosin-Containing Protein; VHL, Von Hippel-Landau; WM, Waldenström's Macroglobulinemia; XBP1, X-box Binding Protein 1; XPC, Xeroderma Pigmentosum complementation group C; XRC, X-Ray Crystallography; ZNF179, Zinc Finger protein179; AMPA receptor, α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid receptor; BACE1, β-site APP-cleaving enzyme 1.

* Correspondence to: Department of Biotechnology, International Affairs, Molecular Neuroscience and Functional Genomics Laboratory, Delhi Technological University (Formerly Delhi College of Engineering), Room# FW4TF3, Mechanical Engineering Building, Shahbad Daulatpur, Bawana Road, Delhi 110042, India.
*E-mail address:* pravirkumar@dtu.ac.in (P. Kumar).
[1] Both authors contributed equally to this work.
[2] **ORCID ID:** 0000-0001-7444-2344
[3] **Scopus ID:** 14831447800

https://doi.org/10.1016/j.arr.2024.102279

Available online 21 March 2024

3. Poster Presentation in SNCI (Society for Neurochemistry, India) 2025, Delhi Chapter at Jamia Hamdard University on “Identifying Blood-Brain Barrier-Permeable Drug Candidates for PARP1 Targeting in Glioblastoma”

Society for Neurochemistry, India (SNCI) Delhi Local Chapter
&
Department of Toxicology, School of Chemical and Life Sciences
Jamia Hamdard, New Delhi

Certificate of Appreciation

This is to certify that Prof./Dr./Ms./Mr. Ms. Somya Parashar
of Delhi Technological University has
Participate as **Delegate** in the Two Days National Symposium on **“Neurochemistry and Emerging Therapeutics: Challenges and Opportunities in Neuroscience”** held from 16th April 2025 to 17th April 2025 at Convention Centre, Jamia Hamdard, New Delhi. He/She has also presented in Young Investigator/Poster Session.

Prof. Mohammad Akram
Organizing Co-Chairperson

Prof. Suhel Parvez
Organizing Chairperson

Prof. Prakash Babu Phanithi
Secretary General (HQ), SNCI

# DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-42

## PLAGIARISM VERIFICATION

Title of the Thesis **"Interpretable Ensemble Learning Predicts Glioblastoma Sensitivity to Natural Compounds"** Total Pages **75** Name of the Scholar **Somya Parashar**

Supervisor

**Prof. Pravir Kumar**

**Department of Biotechnology**

This is to report that the above thesis was scanned for similarity detection. Process and outcome is given below:

Software used: **Turnitin**, Similarity Index: **2%**, Total Word Count: **15,012**

Date: **20/05/2025**

Somya

**Candidate's Signature**

20/05/2025

**Signature of Supervisor**

# 2% Overall Similarity

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

## Filtered from the Report

- Bibliography
- Quoted Text
- Cited Text
- Small Matches (less than 8 words)

## Match Groups

**21 Not Cited or Quoted 2%**
Matches with neither in-text citation nor quotation marks

**0 Missing Quotations 0%**
Matches that are still very similar to source material

**0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation

**0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

## Top Sources

1% Internet sources
1% Publications
1% Submitted works (Student Papers)

## Integrity Flags

**1 Integrity Flag for Review**

**Replaced Characters**
**56 suspect characters on 18 pages**
Letters are swapped with similar characters from another alphabet.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.