Machine Learning Models to Probe the CYP3A4 Mediated Drug Metabolic Profiles

By

Varda Mian

MS BI-5 00000359057

Supervised By:

Dr. Yusra Sajid Kiani

School of Interdisciplinary Engineering and Sciences (SINES)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan.

October 2022

Machine Learning Models to Probe the CYP3A4 Mediated Drug Metabolic Profiles

A thesis submitted in partial fulfilment of the requirement for the degree of Master's in Bioinformatics

By

Varda Mian

MS BI-5 00000359057

Supervised By:

Dr. Yusra Sajid Kiani

School of Interdisciplinary Engineering and Sciences (SINES)

National University of Sciences and Technology (NUST)

Islamabad, Pakistan.

October 2022

Annex A to NUST Letter No. 0972/102/Exams/Thesis-Cert dated 23 Dec 16.

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Ms. Varda Mian Registration No. 00000359057 of **SINES** has been vetted by undersigned, found complete in all aspects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature with stamp: _____

Name of Supervisor: Dr. Yusra Sajid Kiani

Date: _____

Signature of HoD with stamp: _____

Date: _____

Countersign by

Signature (Dean/Principal): _____

Date: _____

Dedication

I dedicate this work to my parents, for their endless love, support, and encouragement. I also dedicate this thesis to my beloved husband, Dr. Hashim Naeem, without whom this would not have been possible.

Certificate of Originality

I hereby declare that the research work presented in this thesis has been generated by me as a result of my own research work. Moreover, none of its contents are plagiarized or submitted for any kind of assessment or higher degree. I have acknowledged and referenced all the main sources of help in this work.

_____

Varda Mian

NUST00000359057-MSBI-5

ACKNOWLEDGMENT

First and foremost, I would like to praise and thank Allah the Almighty, the Most Gracious, and the Most Merciful for giving me the strength and courage to complete this thesis.

I would like to express my gratitude and sincere thanks to Dr. Yusra Sajid Kiani and Dr. Ishrat Jabeen, this research would not have been possible without their constant support and guidance, and for that I am extremely grateful.

I would like to acknowledge my sister, Hadiya Mian, for her readily available help whenever I needed it, and my husband, Dr. Hashim Naeem, for his support throughout this whole endeavor. A huge thanks is owed to my friends, especially Fatima Ahmed, for their moral support, encouragement, and help in the completion of this degree.

Last but not the least, I would like to acknowledge School of Interdisciplinary Engineering and Sciences for providing the infrastructure that made this research possible.

Table of Contents

List of Abbreviations

ADME	Absorption, Distribution, Metabolism, Excretion
ADR	Adverse Drug Reaction
ANN	Artificial Neural Network
AUC	Area Under the Curve
CCR	Corrected Classification Rate
CNS	Central Nervous System
CYP	Cytochrome P450
DDI	Drug-drug Interactions
DNN	Deep Neural Network
DT	Decision Tree
FCFD	Functional-Class Fingerprint Descriptors
FN	False Negative
FP	False Positive
GALAS	Global, Adjusted Locally According to Similarity
GOLD	Genetic Optimization for Ligand Docking
GST	Glutathione S-Transferase
IC_{50}	Inhibitory Potency
IG	Information Gain
kNN	K-Nearest Neighbor
MCC	Mathew's Correlation Coefficient
MD	Molecular Dynamic
MLP	Multilayer Perceptron
MSA	Multiple Sequence Alignment
NB	Naïve Bayesian
NDA	New Drug Application
NSCLC	Non-Small Cell Lung Cancer
QSAR	Quantitative Structure and Activity Relationship
RF	Random Forest
ROC	Receiver Operating Characteristic

RP	Recursive Partitioning
SAR	Structure and Activity Relationship
SE	Sensitivity
SP	Specificity
SULT	Sulfotransferases
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
UGT	Glucuronosyltransferase

List of Figures

List of Tables

Abstract

Cytochrome P450s (CYP) are a diverse group of Heme-containing proteins found in all kingdoms of life, that participate in vital life processes including oxidization of endogenous and exogenous compounds. Of the 57 CYP isoforms, CYP3A4 is the most abundant isoform in humans. CYP3A4 is highly promiscuous in substrate specificity and allows the accommodation of compounds diverse in size and structure, which leads to CYP3A4-mediated metabolism of up to 50% of all marketed drugs. However, the ability of CYP3A4 to adjust two or more similar or different molecules may also lead to adverse drug-drug interactions (DDIs), as the inhibition or induction of CYP3A4 by one drug can lead to adverse effects in the *in vivo* metabolism of other drugs. Pharmacokinetic issues due to the inhibition or induction of CYP isozymes are accredited for the failure of nearly 80% of drugs during development. Therefore, it is important to analyze cytochrome interactions before preclinical trials to ensure the success during the drug development process. The current study aims to utilize supervised machine learning techniques and molecular modeling strategies on publicly available CYP3A4 inhibition data to predict CYP inhibition through the development of a predictive model and the identification of 3D features responsible for CYP3A4 inhibition. Five models were built to predict CYP3A4 Inhibition on two refined different datasets of CYP3A4 inhibitors: Support Vector Machine, Logistic Regression, Decision Tree, Random Forest, and Multilayer Perceptron. The Support Vector Machine and Logistic Regression models built on the more refined dataset outperformed all others, with accuracies of 98% and 96% indicating superior performance. Therefore, these two models built on the chosen hyperparameters are suitable for the prediction of CYP3A4 inhibition in new chemical entities and can assist in the drug developmental process. Additionally, all models in the more refined dataset resulted in accuracies over 80% indicating the stabilities of the models on the data used and highlighting the importance of the refined features and data refining in general over the use of noisy raw data. The results draw attention to the importance of increased lipophilicity, vander waals surface area on pharmacophoric points, number of aromatic and rotatable bonds, percentage of Nitrogen atoms, topological distances between Nitrogen and Oxygen, and Nitrogen and Sulfur, and overall negative charge on a molecule in CYP3A4 inhibition. Thus, this study assists in understanding the key CYP3A4 interactions, prediction of CYP3A4 inhibition and the optimization of the toxicological profiles of new chemical entities.

CHAPTER 1
INTRODUCTION

1: Introduction

1.1Drug Discovery

Drug discovery is an interdisciplinary process which combines biology, chemistry, and pharmacology to identify potential new medicines[1]. The need for drug discovery and development begins with disease prevalence and medical necessity. The design and discovery of new drugs is becoming increasingly essential as new diseases are discovered. The COVID19 pandemic being a stark reminder of the possibility that new diseases can still arise, and the disastrous consequences of not having treatment readily available[2]. Aside from this, drug resistance is also a major cause for the need of new drugs, the improper use and over prescription of antibiotics can result in the development of drug resistance in bacteria, resulting in staggering death tolls and rendering the need for new antibiotics to counteract the drug resistance clear. In a similar vein, new drugs are also needed to contend with drug resistance as a result of viruses and parasites becoming immune to currently available treatments, the multiple drug resistance of the Plasmodium parasites resulting in reduced efficacy of currently available malaria treatment being a notable example[3]. Research and understanding of disease pathways and metabolism are increasing year by year, as a result we are now equipped with the tools needed to treat and potentially even cure diseases that were previously thought to be untreatable or unpreventable. The capacity to treat or improve the treatment of such conditions is where the future of drug discovery has the biggest potential.

Drug discovery begins with the identification of a potential biological target involved in a biological pathway that is believed to be behaving dysfunctional in people with a specific disease, and after years of tests and trials, ultimately results in the introduction of a new drug into the market. On average it takes between 10-15 years for new drug compounds to be approved, costing nearly $1-2 billion. Therefore, it is a huge achievement for a drug to pass clinical trials, however 90% of drug candidates fail during clinical studies with an even higher failure rate if preclinical failures are included[4]. Of the limited number of compounds that enter clinical trials, only 1 in 10 will even reach the market[5]. With such a high rate of failure, it is safe to assume that most candidate compounds will fail before reaching the market, and for those that do succeed, the duration and costs are extremely high making the process of drug design and discovery and extremely process.

The drug development process, summarized in Figure 1, is a 5-stage process consisting of 1. Drug Discovery, 2. Preclinical Development, 3. Clinical Phases, 4. FDA Review and 5. Post

Marketing Monitoring. During the drug discovery phase of drug design, the target is identified which is usually a nucleic acid sequence or protein that is involved in gene regulation or intracellular signaling of a pathway that plays a significant role in a disease. This is done through intensive mining of the available biomedical data. The target is then validated to ensure that it indeed involved in the disease mechanism and elicits a biological response upon testing *in vivo* and in vitro that can be measured and regulated[6]. This is arguable the most critical part of drug design, as many of the reasons why drugs fail in the later stages of drug design, including lack of safety and efficacy, can ultimately be traced back to poor target validation[7]. Once the target is validated, a series of processes are undertaken in order to identify a naturally occurring or synthetic small molecule, known as a lead, which interacts with the previously selected target. The lead compound undergoes multiple tests, screening procedures, and optimization to ensure satisfactory absorption, distribution, metabolism, and excretion (ADME) results before it is ready for animal testing[8].

The next phase of drug development is the preclinical phase which determines if a drug is safe for human trials by first performing extensive testing on animal models. Preclinical trials test the drug's efficacy, toxicity and pharmacokinetic *in vivo* and in vitro with unrestricted dosages. The drug then enters clinical development which take place using human subjects and consists of the following phases:

1. Phase 1: Safety, tolerability, pharmacodynamic and pharmacokinetic effects of the drug are tests on up to 100 healthy volunteers. Safe dosage is also determined during this step.
2. Phase 2: Effectiveness of the drug on up to 500 patients with the target disease is tested.
3. Phase 3: Hypothesis of efficacy and adverse effects determined in large scale testing on up to 5,000 patients with the target disease.
4. Phase 4: Studies conducted after FDA approval to monitor the adverse effects of the drug post marketing[5,9].

After the clinical trials have been completed and delivered positive outcomes, the data is compiled in a New Drug Application (NDA) and submitted to the FDA for review. The NDA must prove the safety and efficacy of the drug, and then may be accepted or rejected by the FDA. If a drug is rejected, the applicant is given a reason as to why, as well as what information

would be required for the application to be accepted. Once an application is approved, it is ready for marketing[10].

The strict regulation of the drug development process has ensured therapeutic safety in populations taking the drugs, but this also means that very few of the compounds that enter the trials reach FDA approval. In fact, of 5,000-10,000 compounds that enter preclinical studies, only 5 reach clinical trials and only 1 reaches approval for marketing[9,11]. This begs the question of why the drug failure rate is so low and is an area of interest for those in the drug development and pharmaceutical industries.

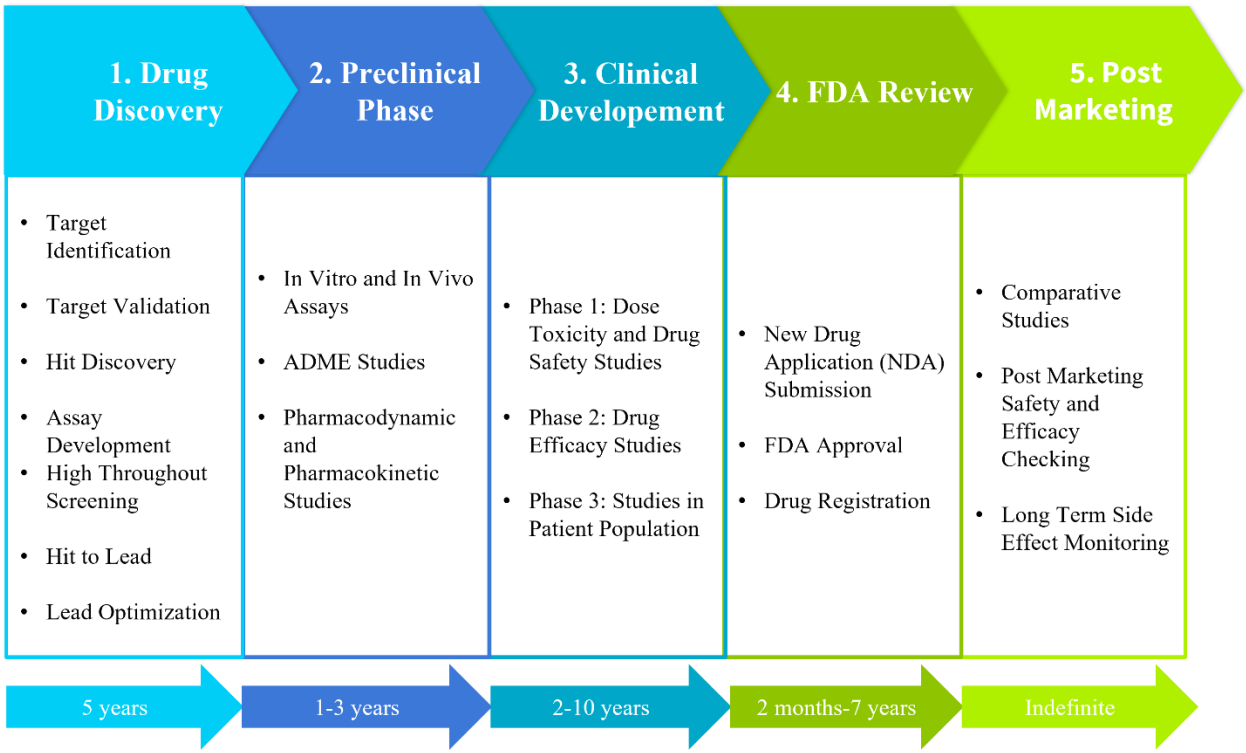

Figure 1. Drug Development Process Overview, adapted from Moein et al.[12] and Reda et al.[13]

1.1 Why Drugs Fail

Given the high expenditure of research and development in drug development, it is important to understand the reasons for drug failure so as to avoid failures in the later stages. Premature failure of drugs is preferable over late-stage drug failure as the costs of development increase exponentially as the drug moves further along the development cycle. Lack of efficacy is the main reason for drug failure in preliminary stages. While this implies that efficacy should be well established by the time a drug reaches clinical trials, research has shown that lack thereof is still a major contributing factor for late-stage failures[14]. The staggeringly high attrition rates in preclinical and clinical development are contributed to lack of safety, efficacy, and poor pharmacokinetics with pharmacokinetics being a major contributing factor resulting in 39% of failures in drug design[15]. Identifying these factors early on could aid in reducing both the cost and time required to get a new drug into the market. Since ADME properties play a role in both toxicity and efficacy, they are crucial in differentiating successful drugs from those likely to fail[16]. In fact, recent studies by Palmer et al. have shown that the incorporation of strong pharmacokinetic and pharmacodynamic principles leads to higher successes in antimicrobial drug development[17].

1.2 Drug Metabolism

Improvements can be made in the drug development process by focusing on methods that will reduce the number of failures during the preclinical and clinical trials. The best way to do this is to direct more effort towards determining early on if a drug candidate has all the required properties and characteristics of a drug. Applying more focus on metabolites studies during the initial stages of drug design is key in determining the success of a drug as it plays a role in tethering together the various disciplines of drug development including drug discovery, drug safety, clinical development, and pharmaceutical development, as well as project management and regulatory affairs[18]. Moreover, dose regimen designs also depend on metabolic studies. Defined as “the biotransformation of exogenous compounds by living organisms, usually through specialized enzymatic systems”[19], drug metabolism is essentially the modification of a drug substrate to make it more polar and thus assist its removal from the body, thus avoiding the undesirable side-effects associated with the accumulation of drugs within the body. The metabolism of majority of xenobiotics including orally administered drugs takes in two stages. During stage 1,

functionalization occurs, during which a functional moiety is attached to the drug substrate, resulting in a more water-soluble substrate. in the enhance water-solubility of the xenobiotic. Next is stage 2 of metabolism, during which conjugation reactions occur, involving the addition of a molecule, usually sulfates or glucuronic acid molecules, to the drug substrate leading to the formation of an intermediate. Conjugation results in the increased solubility of the drug molecules leading to their enhanced removal from the body[20].

In most cases, metabolism leads to the inactivation of a drug, however on some occasions it leads to the formation of an active metabolite which is the major circulating active agent solely or partially responsible for pharmacological response[21,22]. On other occasions the metabolism of a drug may result in a toxic metabolite that interacts with cells by covalently binding to cellular constituents, stimulating peroxidation, and decomposing cellular lipids[23] leading to the initiation or aggravation of a variety of toxicities including hepatotoxicity, nephrotoxicity, and pulmonary toxicities[24]. Another aspect to take into consideration are the major liabilities associated with drug-drug interactions (DDIs) caused by enzyme inhibition or induction of major metabolizing enzymes[22]. Hence, having a good understanding of the metabolism of new chemical entities is needed during the development of new drugs.

Studies have shown that increased effort in applying pharmacokinetic principles during the drug design process has decreased attrition from 40% in 1990 to 10% in 2000[18]. Extensive research has shown that phase 1 metabolizing enzymes, especially Cytochrome P450s (CYPs) mediate majority of drug activation, while phase 2 metabolic enzymes like glucuronosyltransferase (UGTs), glutathione S-transferase (GSTs) and sulfotransferases (SULT) play a smaller role[24], as summarized in Figure 2.

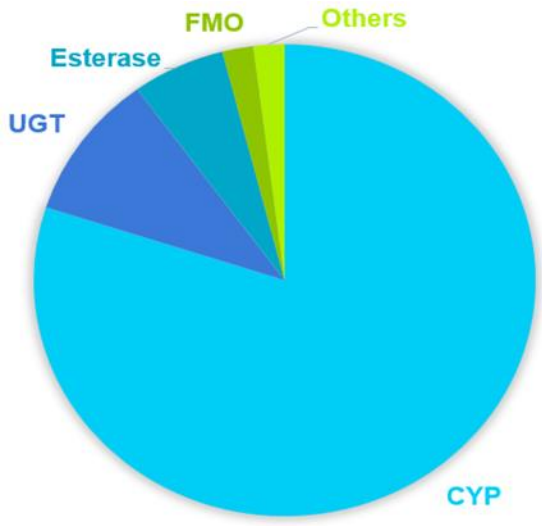

Figure 2. Contribution of Enzymes to Metabolism of Marketed Drugs, adapted from Williams et al.[25]

1.3 Cytochrome P450 in Drug Metabolism

The CYPs are a superfamily of heme containing enzymes embedded mainly in the lipid bilayer of the endoplasmic reticulum, mitochondria of hepatocytes and in the intestines[26]. Found in almost all organisms, CYPs are in involved in the metabolism of a wide variety of drugs and xenobiotics including drugs, chemicals, pollutants, steroids, bile and fatty acids, vitamin D and other natural products. These phase 1 metabolic enzymes catalyze a range of oxidative and reductive biotransformation including carbon hydroxylation, heteroatom oxidation, bond oxidation, hydrocarbon desaturation, and halocarbon dehalogenation[27]. The CYP monooxygenases reactions typically involve the insertion an oxygen atom from dioxygen into a C–H bond to give the corresponding alcohol, while the remaining oxygen atom is reduced to water, depicted as follows:

$$NADPH + H^+ + O2 + RH \rightarrow NADP^+ + H2O + ROH$$

Here, RH represents an oxidizable drug substrate and ROH is the hydroxylated metabolite, and CYP catalyzes the overall reaction. The two electrons required in the process are provided by NADH or NADPH cofactors[28].

In mammals, 57 CYP genes are present that are divided into 18 families and 42 subfamilies based on amino acid sequence homology[29]. Seven of these 57 isoforms including CYP1A2, CYP2C9, CYP2C18, CYP2C19, CYP2D6, CYP2E1 and CYP3A4 are responsible for the metabolism of more than 90% of all currently used clinical use drugs[30]. Table 1 displays the role played of each of these isoforms. The requirement to be considered a member of the CYP family is at least 40% homology to another member of the family, while >55% homology is required for members in the same subfamily[31]. CYPs are classified into 2 categories based on their intracellular location; 1. microsomal cytochrome P450, which are present mainly in the microsomes of liver cells and represents about 14% of the microsomal fraction of liver cells, and 2. mitochondrial cytochrome P450, which are present in mitochondria of many tissues but are particularly abundant in the liver and other tissues such as adrenal cortex, testis, ovary, placenta, and kidney. CYPs have a helix-rich secondary structure that encloses the active site, the heme cofactor is located in the bottom area of the active site with the iron tethered to a cysteine thiolate[32]. The tertiary structure of CYP enzymes is highly conserved with a characteristic protein fold shared among the superfamily, however the only amino acid residue that is completely conserved is the proximal cysteine ligand of the haem group[33]. However, these isoenzymes differ in pharmacogenetics, substrate specificities, inducibility, and susceptibility to inhibition by competing drugs[34].

Since so many medications are metabolized by such a small group of enzymes, the risk of DDIs, adverse drug reactions (ADRs) and decreased drug efficacy due to CYP inhibition or induction remain high. DDIs may occur when the intake of one drug inhibits the CYP-mediated metabolism of another drug, usually in a dose dependent manner, resulting in the decrease in metabolism and possible accumulation of the other drug, which may lead to toxicity[34]. Therefore, proper screening of drugs against CYP isoforms is essential in drug design, and lack thereof can result in serious ADRs and DDIs. Therefore, it is of vital importance for the for the pharmaceutical industry and professionals to identify the likely effect of a drug candidate in terms of interactions with any CYP isoform well before time to avoid the losses associated with drug developmental failures at the later stages. This aspect is crucial to determine whether the drug will be worth the time, money, and resources it takes to develop it, and alludes to the lives that can be saved when proper CYP screening protocols are applied.

Table 1. Summary of CYP Isoforms

CYP Isoform	Role in Metabolism of Drugs (percentage)	Location
CYP1A2	2%	Liver
CYP2E1	4%	Liver, Brain, etc.
CYP2B6	4%	Liver
CYP2C9	10%	Liver
CYP2C19	10%	Liver
CYP2D6	28%	Liver, Central Nervous System (CNS)
CYP3A4	47%	Liver, Small Intestine, etc.

1.4 Cytochrome P450 3A4

As far as the CYP450 isoforms are concerned the CYP3A4 enzyme is the most widely expressed and is typically found in the liver and gastrointestinal tract. The human CYP3A4 protein is encoded by the CYP3A4 gene, which is part of the Cytochrome P450 genes positioned at chromosome 7q22[35]. Its structure contains a heme group in the active site and consists of a small β-strand N terminal and a larger α-helical C-terminal domain and adopts the typical fold of the cytochrome P450 superfamily[36]. An important feature of CYP3A4 is its promiscuous substrate specificity which allows it to accommodate and oxidize a wide array of substrates with many different structural features[29]. This characteristic of CYP3A4 can be attributed to its large substrate binding cavity, which is studied to be open, flexible, and able to accommodate volumes ranging from 1173 to 2862 cubic angstroms ($Å^3$), and accounts for its importance in the metabolism of approximately 50% of all clinically used drugs.An interesting feature of the CYP3A4 crystal structure is the cluster of seven phenylalanine residues above the active site which form forming a hydrophobic core, reducing the active site to 520 $Å^3$. This is surprising given the generous size of some CYP3A4 inhibitors; however, mutagenesis studies show that these residues play a role in substrate metabolism and can be relocated or displaced to accommodate the metabolism of larger substrates[36,37].

This, as well as its ability to accommodate one large substrate or multiple smaller substrates of the same or different types makes CYP3A4 of practical interest in the pharmaceutical industry as well as in the conventional drug design process. The wide array of drugs metabolized by CYP3A4 include benzodiazepines, calcium channel blockers, cyclosporine, macrolide antibiotics, opioids, several statins, and it also contributes to the metabolism of steroid hormones[38,39], making CYP3A4 a key player in drug metabolism as it is sensitive to changes in CYP3A4 activity and expression level.

In addition, CYP3A4 is also inhibited by a wide range of xenobiotics, such as erythromycin, grapefruit juice, ketconazole, and HIV protease inhibitors, all of which have been proven to react with other medications and result is serious adverse effects[3940]. Among inhibitors like these, there are a few high energy binding interactions that tend to occur, for example high affinity binders tend to have one or more aromatic rings capable of forming numerous pi–pi stacking interactions with active phenylalanine residues, be lipophilic, and make polar interactions with D76, R106, S119 and R372[38]. However, despite this knowledge, predicting DDIs due to CYP3A4 is still relatively difficult due to the flexible nature of its binding pocket and the diversity of its ligand interactions.

Since many pharmaceuticals have narrow therapeutic indices and are CYP3A4 substrates, their dose needs to be adjusted for optimal therapy; preferably their concentrations are monitored continuously to prevent adverse drug reactions[41]. Interactions between simultaneously administered drugs may affect drug clearance and the outcome of drug therapies and are therefore also of major importance for the pharmaceutical industry.

1.5 Drug-Drug Interactions

Drug-drug interactions are becoming an increasing issue in healthcare. Around 20-40% of DDIs are due to polypharmacy, as older adults tend to consume multiple drugs to treat a greater number of comorbidities. However, this use of multiple drugs and the resultant DDIs are responsible for the induction of adverse drug reactions and decreased efficacy of therapeutics. Changes in physiology, pharmacokinetics and pharmacodynamics in elderly patients also make them prone to altered drug responses including prolonged drug half-lives ADRs and DDIs which escalates the prevalence and severity with an increased number of medications[42]. Moreover, ADRs

and DDIs have been found to be associated with prolonged hospital stay, increased morbidity and mortality[43]. DDIs can be pharmacokinetic-related, where drug absorption, distribution, metabolism, or excretion are affected or inhibited due to interaction with another drug, or pharmacodynamic related, where a drug modifies the responsiveness of tissues towards another drug by either having an agnostic or antagonistic effect. Of these, pharmacokinetic DDIs are the most common, and mostly involve impaired drug removal because of interference with hepatic metabolism, renal excretion, or transcellular transport[44]. Despite standard screening procedures, DDIs are still present in 16-41% of oncology patients, suggesting the need for better screening procedures in not only oncology but other medicinal fields as well[45].

Furthermore, it is notable that majority of pharmacokinetic DDIs interactions arise due to the effect of previously administered drugs on the hepatic Cytochrome P450s (CYPs). Whereby, the mechanism of CYP450 inhibition and induction predominantly result in the occurrence of the most clinically significant DDIs[46]. DDIs often involve isozymes of CYP and result in alterations in drug bioavailability that can lead to serious adverse events or decreased drug efficacy[47]. It is common knowledge that many drugs are metabolized by CYP3A4 mediated reactions, however there are also some drugs which are activated by the enzyme. Substances such as grapefruit juice and some drugs, interfere with the action of CYP3A4 resulting in either amplified or weakened action of other drugs modified by CYP3A4, causing them to accumulate to toxic levels and create adverse side effects. CYP3A4 can be inhibited in one of three ways; competitively, non-competitively, or mechanism based.

A well-known example of CYP3A4 inhibition is that of grape juice, where researchers discovered bergamottin, a furanocoumarin found in grape juice, increased the bioavailability of drugs in patients who had consumed the liquid[48]. An example of the severe DDIs that can result from CYP3A4 inhibition includes the antihypertensive and anti-anginal drug Mibefradil (Posicor), which was was withdrawn from the market in 1998 based on its tendency to inhibit the CYP3A4-mediated metabolism of drugs treating cardiovascular diseases. The resultant DDIs were the cause of higher mortality in patients with congestive heart failure who were also taking Mibefradil[49]. Similarly, antihistamines terfenadine (Seldane), astemizole (Hismanal) and cisapride (Propulsid) are all drugs that were withdrawn from the U.S. market due to metabolic inhibition by other drugs that led to life-threatening arrhythmias[50]. Therefore, understanding a drug's potential to inhibit

CYP3A4 is an essential step in drug development. Current studies in drug development attempt to do this by applying traditional and novel machine learning algorithms to chemical datasets in an attempt to discover non-linear patterns within the data.

1.6 Problem Statement and Proposed Solution

Often drug failure is not detected until the late stages of drug discovery and development as the interactions and properties of the drug candidates that lead to failure, such as DDIs, metabolism, and toxicity are only exposed during clinical studies. CYP3A4 is an important compound in the metabolism of drugs but is susceptible to inhibition and induction by a diverse range of compounds resulting in undesirable DDIs and toxic side effects that can ultimately lead to drug failure. Therefore, the development of models that can predict the likelihood of a drug candidate to inhibit CYP3A4 can assist in saving time, money and resources in the drug discovery and development process. The proposed solution is to develop an in-silico profiler using a variety of machine learning techniques including Support Vector Machine, Logistic Regression, Decision Tree, Random Forest, and Multilayer Perceptron to predict likelihood of chemical entities to inhibit CYP3A4. These machine learning models can also provide insight into what features of inhibitors are important to look out for in new drug candidate compounds.

1.7 Objectives

- To elucidate the impact of 2D structural descriptors on CYP3A4 inhibitory profiles.
- Developing machine learning models to probe the CYP3A4 inhibition profiles of new chemical entities.

CHAPTER 2

LITERATURE REVIEW

2: Literature Review

2.1 Computational Techniques

Drug mediated CYP3A4 inhibition can lead to ADRs associated with accumulation of the drug toxic levels or reduced efficacy of other administered drugs. Therefore, it is of interest to conduct screening protocols for potential new drugs to avoid or attenuate potential DDIs. Currently, hepatocytes are used to determine the risk of DDIs associated with CYP3A4 metabolism however these in vitro assays are time-consuming and provide limited information about the structure activity relationship of CYP3A4 with the compound causing its inhibition. Computational techniques, on the other hand, have been used to provide insight into CYP3A4 inhibition. The ability of *in silico* approaches to evaluate a large number of compounds with a relatively low cost and reduce the need for a number of experimental studies makes it a more attractive approach for determining CYP3A4 inhibition of drug candidates, and thus improve success rate.

2.1.1 Structure Based Approaches

2.1.1.1 Molecular Docking

Molecular docking is becoming an increasingly important tool in drug discovery, and has been commonly used for such since the 1980's. It allows us to model the atomic level interaction between a small molecule and a protein which in turn gives us information about favored binding orientation of the ligand within the protein and is a valuable tool for investigating metabolic stereoselectivity[51]. Docking works by first predicting the ligand pose within the binding pocket and then providing scores to assess the binding affinity of each of the ligand poses[52]. Molecular docking studies have been performed to determine the interaction of molecules with CYPs, especially CYP3A4.

Molecular docking studies have often been used to determine the binding mode of anticancer drugs as well as carcinogens with CYP3A4, as many cancer drugs are metabolized by this CYP isoform. For example, A study conducted by Maréchal et al. to predict the ability of 33 commonly used cancer drugs to inhibit CYP3A4 determined that the residues in the phenylalanine cluster (Phe108, Phe213, Phe219, Phe220, Phe241, Phe304 and Phe215), as well as Arg212 and

Glu37 are important residues in the isoforms active site[53]. The rigid and flexible docking approaches used by Panneerselvam et al. on the cancer drugs Cytarabine, daunorubicin, doxorubicin and vincristine observed that S119, R212 and R372 are the major drug-binding residues in CYP3A4[54]. Using the GOLD (Genetic Optimization for Ligand Docking) software version 3.3.1, in combination with the Chemscore scoring function to dock the potent natural carcinogen Aflatoxin B_1(AFB_1) into the CYP3A4 binding site, the importance of the amino-acid residues Leu210, Leu211, and Phe304 were confirmed as indispensable for the positive homotropic cooperativity of both AFB_1 oxidations[55]. Chemotherapeutic agents for the treatment of Non-Small Cell Lung Cancer (NSCLC) such as gemcitabine, cisplatin, carbo platin, docetaxel, and paclitaxel were docked by induced fit against CYP3A4 using Schrödinger suite 2014 by Subhani et al. during which ARG105 was identified as a key residue involved in drug binding, along with Pro107, Ser119 and Arg212[56].

Zhou et al. docked the compound Miltirone in major CYP isoforms and discovered the importance of Vander Waals interactions with Phe57, Arg212, Phe215, Thr309, Ala370, Arg372 and Met37, and polarity with Gly481 and Phe213 residues[57]. A docking study performed by Ashour M used the C-Docker protocol of Discovery Studio to dock 38 compounds found in active plant extracts against CYP3A4 and elucidated the importance of hydrogen or ionic bonding with Arg 212 and Glu374, π bonding with Arg105, and hydrogen bonding with Arg375, Asn441, Cys442, Gly48, Ile443 and Pro434 in the binding mode of flavonoids[58].

Thus, multiple docking studies elucidate the role and importance of the following amino acid residues in CYP3A4 binding; Met37, Phe57, Phe108, Phe213, Phe215, Phe219, Phe220, Phe241, Phe304, Phe215, Pro107, Thr209, Ala 370, Arg105, Arg212, Arg372, Glu37, Ser119, Leu210, Leu211 and Gly481.

2.1.1.2 Molecular Dynamics Simulations

Molecular Dynamics (MD) Simulations are a structure based computational method that use approximations based on Newtonian physics to simulate atomic motions over time based on a model of the interatomic interactions. This technique plays a practical role in drug design and development as it assists in probing molecular properties that are either difficult or impossible to determine in a wet lab. They are also useful in generating hypothesis' that can be used as a baseline for new experimental work to substantiate structural and functional properties of a potential drug[59].

MD simulations have proven to be useful in confirming important residues within the CYP3A4 binding pocket was well as critical interactions between the enzyme and its inhibitors and substrates. MD Simulations were used to examine the differences in structural and dynamic properties between CYP3A4 complexes with the substrate progesterone and the inhibitor metyrapone by Hwangseo et al. using AMBER force fields. The study intimated that the flexible loop containing amino acid residues Asp214, Phe215 and Leu216 was responsible for the diverse substrate specificity of CYP3A4. Hydrogen bonding between the Ser119 sidechain and a carbonyl group was found to be a stabilizing binding force for inhibitors and substrates in the active site of CYP3A4, and the interaction between a structural water molecule and the heme group was seen to be a significant stabilizing force[60].

A study by Han et al. built MD simulations using the LEaP module of the AMBER16 package and applied AMBER force field ff14SB and General AMBER Force Fields to study the mechanisms causing the effects of CYP3A4 variants on the differential kinetic profiles of acalabrutinib and ibrutinib. A 30-ns molecular dynamic simulation was conducted through which it was determined that the clearance rate of the compounds was mediated by distance between the redox site and the heme iron atom of CYP3A4[61].

A combination of molecular dynamics simulations and free-energy calculations was used by Bren et al. to determine the origin of the positive homotropic cooperativity in the binding of ketoconazo to CYP3A4, and thus provided insight into the mechanism of CYP3A4 inhibition. The 10-ns simulation determined that the main driving force for ketocanzao binding was shape complementarity through Vander Waals forces, which coincides with what we know about CYP3A4 and its binding promiscuity. The presence of nonpolar residues and flexibility of CYP3A4's binding site also contribute towards the high interaction energy of ketocanzao and the enzyme[62].

Teixiera et al. studied the binding modes of ligands into CYP3A4 conformations by building MD simulations through GROMACS version 3.2.1 package and the GROMOS96 43a1 force field to model both the protein and ligands alone and in a complex, for five 10-ns MD simulations. The study confirmed the importance of the residues Arg105, Arg212, Glu374, Ser119, Thr309, Phe213, Phe215, and Phe304 in stabilization, accommodating space and orienting ligands for enzymatic action by CYP3A4[63]. Analysis of results of MD simulations and docking protocols

performed by Kiani et al. indicate similar results while also confirming the importance of Arg106 and Arg372 in the stabilization of CYP3A4-inhibitor complexes[64].

2.1.2 Ligand Based Approaches

2.1.2.1 QSAR

Quantitative Structure and Activity Relationship (QSAR) is a ligand-based computational method for identifying the relationships between the structural properties of chemical compounds and their biological activities. QSAR models employ data from the molecular structure of ligands and examines physiochemical properties, therapeutic activities, and pharmacokinetic parameters to obtain a reliable statistical model for prediction of the activities of new chemical entities and thus predict the best molecules for a target. While previously QSARs only modeled linear relationships using Hansch analysis, it is now used to generate multiple linear regression models using 3D grid-based approaches. The fundamental underlying principle being that 'similar structures behave similarly'. QSARs have applications all throughout the preclinical phase of drug development as it helps in the prediction of ADMET properties early in the drug discovery process, and ultimately result in experimentally testable hypotheses'[65]. Recent advancements in QSAR modelling involve the prediction of CYP mediated metabolism based on previously known inhibitor data and provide important insight into what structural properties are essential in CYP3A4 based inhibition. For example, Didziapetris et al. used datasets from PubChem and NCBI to develop a Structure Activity Relationship (SAR) model using the GALAS (Global, Adjusted Locally According to Similarity) approach to predict the CYP3A4 inhibition and determine the properties of interest.

Their model determined that the presence of a strong basic group or an acidic group in a compound reduces its probability to inhibit CYP3A4, that higher inhibition was associated with increase in molecule size of the molecule as well as association with hydrophobic aliphatic or aromatic, and highlighted the importance of molecular weight and lipophilicity in determining a compounds tendency and level of CYP3A4 inhibition[66]. A model by Roy et al. showed the importance of a U shape conformation of the substrate for optimal inhibitory activity and showed that the logp descriptor had the greatest effect on inhibition of CYP3A4[67]. The generalized model developed by Riley et al. established the importance of lipophilicity and heme interactions on the inhibition potency[68].

This is corroborated through work by Lewis et al., whose QSAR model on 3 series of CYP3A4 inhibitors concluded that lipophilicity is a major contributing factor for inhibition, regardless of the structural class of the inhibitor. They also determined that hydrogen bonding between the inhibitors and at least one active site donor or acceptor amino acid as common interactions in CYP3A4 inhibitors[69].

2.1.2.2 Pharmacophore Modeling

A pharmacophore is a computational technique implemented in rational drug design, defined as "a molecular framework that carries (phoros) the essential features responsible for a drug's (pharmacon) biological activity". Ligand based pharmacophore approaches involve superposing a set of 3D structures that are representative of essential interactions between the ligand and target, to extract similarities in their chemical features. Such approaches have become key in facilitating the drug discovery process, specifically virtual screening of chemical compounds, de novo design and lead optimization[70]. Pharmacophore modelling are often applied in combination with other computational methods such as QSAR, machine learning, and MD simulations.

Kaur et al. used rationally designed compounds to test their pharmacophore model of CYP3A4. Through this, they were able to confirm important features in compounds for inhibition, including a flexible backbone, H-bond donor/acceptor moieties, and aromatic side group analogous. Their model also reaffirmed the essential role of hydrophobic interactions near the enzyme heme group and phenylalanine cluster in the ligand binding process[71]. This further substantiates the contribution of hydrophobic interactions and hydrogen bonding in a compounds affinity to binding to CYP3A4 that was previously verified in previous pharmacophoric studies[72,73]. However, these studies also indicate that building an accurate pharmacophore-based model for predicting CYP3A4 inhibition proves difficult, mostly in part to the enzymes promiscuous binding site that allows substrates with a diverse range of features to bind[72].

2.2 Machine Learning Techniques

While computational methods can provide important insights into CYP3A4 interactions and inhibition, they tend to be computationally expensive and tend to pose problems when incorporating them into software tools. Machine learning provides a solution to this problem as they are less computationally expensive, and provide accurate results in a timely manner, making them ideal for usage in web applications and software tools. Another advantage of machine learning is that it allows the in-silico investigation of CYP interactions early in drug development so that investigators are allowed to select compounds that are less likely to fail due to undesirable pharmacokinetics later stages of the drug design process[38]. Machine learning techniques are frequently used for substrate specificity prediction as they can model complex nonlinear relationships from large datasets of enzyme-substrate interaction data[74]. Machine learning approaches for CYP3A4 inhibition prediction has gained traction in recent years. The models generated for CYP inhibitors, including decision trees, support vector machines (SVMs), k-nearest neighbors (kNN), random forests, artificial neural networks (ANN), and deep learning to predict the likelihood of CYP inhibition of a compound[75,76,77].

2.2.1 Support Vector Machine

Support vector machine (SVM) is one of the most widely used deep learning algorithms which allows the classification of linear as well as non – linear data into separate groups with the help of hyperplanes. A hyperplane can be defined as a decision boundary that differentiates between two or more classes in SVM. Any datapoint that falls on either side of the hyperplane can be categorized into separate classes. However, finding the optimal hyperplane that maximizes the distance between those datapoints is the key to a good learning model. SVM falls under the category of supervised machine learning algorithms because this classification provides a learning basis for future data processing. It primarily uses labeled input data to perform training, and after numerous training examples, can be used to perform training on unlabeled data. Needless to say, this algorithm has proven its potential for structure – activity relationship analysis in the drug development process[78]. Table 2 displays research carried out in the past on the implementation of SVMs in CYP3A4 inhibition prediction.

2.2.2 Decision Trees

Decision Tree (DT) also belongs to the family of supervised machine learning algorithms with the difference being that it follows a hierarchical, top – down approach and has a tree – like structure consisting of one root node, branches, internal and leaf nodes. The root node has no incoming branches, but it does have outgoing branches which lead to the internal nodes – also known as the decision nodes. These decision nodes help in further segregation of the data into classes or groups which are denoted by the leaf nodes and these leaf nodes are a representation of all the possible outcomes within a dataset. All of these node types aid in classification by conducting evaluations on the basis of pre – set features. The most crucial step in this algorithm is deciding the hierarchy of the features based on their level of importance for classification. This is mainly done by using the Information Gain (IG) criteria, according to which the feature with the highest IG is placed at the root node. In terms of applicability, Decision Trees are extensively being used in the pharmaceutical domain with noteworthy prediction accuracy[79]. This algorithm type has also been used in the prediction of CYP3A4 inhibition that is being shown in Table.

2.2.3 Random Forest

The Random Forest (RF) classification is an extension of the Decision Tree algorithm performed by using an ensemble approach known as bootstrap aggregation, or bagging. This method works by constructing multiple independent decision trees, each using a sub – sample of the original dataset. This is done by sampling with replacement, meaning that duplicates of the observations can be used and that samples created will be independent of each other with respect to their observations. Each of these samples is used to construct a decision tree and the consensus among all of these trees aids in predicting the final outcome. Decision made by majority vote will be more accurate and statistically significant than the decision predicted by each of those trees individually. Additionally, this approach is not as sensitive to overfitting as other algorithms. Several researchers are of the view that a consensus – based approach would be more effective in predicting drug – target activity than each machine learning model individually[80]. Table shows research that has already been carried out on the application of Random Forest in CYP3A4 inhibition prediction.

2.2.4 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a simple and easy – to – understand supervised machine learning algorithm which is based on the assumption that similar entities are supposed to exist in proximity to each other. A model trained by using KNN classification will predict the values of new datapoints by calculating how close those points are to the points in the training set. In order to identify which of the points are closest to a certain datapoint, distance of the type Euclidean, Manhattan, Minkowski or other is calculated between them. The points closest to that query datapoint will be grouped together and the rest will fall into another group. 'K' in KNN is an essential parameter that defines the number of nearest neighbors to be considered for classification. This method is easy to implement, adapts effectively and uses only two hyperparameters which are distance metric and k – value. As for its applicability, KNN is widely used in QSAR analysis during the process of drug development[81]. This algorithm type has also been used in the prediction of CYP3A4 inhibition that is being displayed in Table.

2.2.5 Logistic Regression

Logistic Regression is one of the most popular supervised machine learning techniques being used today. It is used to predict a categorical dependent variable by the help of a given set of independent variables. This type of classification does not provide the researcher with a discrete outcome of the variable, rather it gives a probabilistic value between 0 and 1 which divides the date into two separate groups. It is used when our dependent variable is dichotomous in nature. This simply means that the variable will have only two possible outcomes, for example, whether a person has a disease or not. Logistic regression has become a valuable tool in the field of machine learning today, especially because of the ease of training data as compared to other techniques. This methodology can be used for differentiating between cancerous and non – cancerous cells during drug development[82].

Table 2. Assessment of Machine Learning Models for CYP3A4 Inhibitors in Literature

Sr.	Model Type	Data	Descriptors	Results
1.	SVM	DRUGDEX system Drug Information Handbook, Flockart CYP interaction Table, Literature	Constitutional, Geometrical, Topological, RDF, Molecular walk counts, 3D MoRSE, BCUT, WHIM, Galvez topological Charge indices, GETAWAY, 2D autocorrelations, Functional groups, Charge, Atom-centered, Aromaticity indices, Empiricals, Randic Molecular profiles, Molecular properties	Sensitivity (SE) = 92% Specificity (SP) = 97.3% Concordance = 96% MCC = 0.893[83]
2.	SVM	Experimental QSAR dataset	Physical properties, Surface areas, Atom and Bond counts, Pharmacophore feature descriptors, topological, electrostatic potential, hydrogen bond behavior, shape molecular polarizability	$Q^2 = 0.4 - 0.51 \pm 0.01$ RMSE = 0.36-0.45[84]
3.	SVM	PubChem AID 1851	Classical topological Atom type	AUC = 0.87 Accuracy = 81.1%[85]
4.	SVM	BindingDB ChEMBL	Molecular Descriptors: AlogP, ES_Count_aas C,ES_Count_dS	Sensitivity = 1.000 Specificity = 1.000

			ES_Count_sOH ES_Count_ssNH2 ES_Count_ssS ES_Count_sssN ES_Count_sssNH ES_Sum_aaN ES_Su-m_aaS ES_Sum_aasC ES_Sum_dNH, ES_Sum_dO, ES_Sum_sCl ES_Sum_sF Num_AromaticRings CHI_V_3_C Kappa_2_AM SC_3_CH Wiener	Prediction Accuracy Substrates (Q+) = 0.006 Prediction Accuracy Inhibitors (Q-) = 1.00 MCC = 0.502[86]
5.	SVM	Curated from FDA approved drug list and DRUGDEX system	Charge Analysis Descriptors Topological Descriptors Constitutional Descriptors	Corrected Classification Rate (CCR) = 66.4[87]
6.	K Nearest Neighbor (KNN)	BindingDB ChEMBL	DS 2D Descriptors	Sensitivity = 1.000 Specificity = 1.000 Prediction Accuracy Substrates (Q+) = 0.004 Prediction Accuracy

				Inhibitors (Q-) = 1.00 MCC = 0.502[86]
7.	KNN	Curated from FDA approved drug list and DRUGDEX system	Charge Analysis Descriptors Topological Descriptors Constitutional Descriptors	(CCR) = 64.2[87]
8.	Recursive Partitioning (RP)	BindingDB ChEMBL	DS 2D Descriptors	SE = 0.796 SP = 0.801 Q+ = 0.844 Q- = 0.802 MCC = 0.886[86]
9.	RP	BindingDB ChEMBL	DS 2D Descriptors ECFP-6 Fingerprints	SE = 0.827 SP = 0.832 Q+ = 0.917 Q- = 0.880 MCC = 0.832[86]
10.	RP	Commercially available data	Augmented Atom Descriptors	r^2 = 0.82[88]
11.	Naïve Bayesian (NB)	BindingDB ChEMBL	DS 2D Descriptors	SE = 0.852 SP = 0.809 Q+ = 0.211 Q- = 0.809 MCC = 0.842[86]
12.	NB	BindingDB ChEMBL	DS 2D Descriptors ECFP-6 Fingerprints	SE = 0.902 SP = 0.894 Q+ = 0.954 Q- = 0.886 MCC = 894[86]
13.	Random Forest (RF)	PubChem AID 1851 PubChem AID 844 ChEMBL	Morgan Fingerprints	MCC = 0.68 AUC = 0.94 TPR = 0.74

		ADME		TNR = 0.92 PPV = 0.87 BA = 0.83[51]
14.	RF	Curated from FDA approved drug list and DRUGDEX system	Charge Analysis Descriptors Topological Descriptors Constitutional Descriptors	CCR = 65.8[87]
15.	Artificial Neural Network (ANN)	Curated from FDA approved drug list and DRUGDEX system	Charge Analysis Descriptors Topological Descriptors Constitutional Descriptors	CCR = 62.8[87]
16.	Decision Tree	PubChem AID 1851	1D, 2D, 3D Descriptors Atom Type Electro Topological State Descriptors Crippen's logP Molar refractivity extended topochemical atom, McGowan volume, molecular linear free energy relation,ring counts, count of chemical substructures identified by Laggner	Accuracy = 72.3% Sensitivity = 76.3% Specificity = 72.0% G-mean = 74.1%[89]
17.	Decision Tree	Commercially available Fujitsu database	Constitutional Electrostatic Geometric	Accuracy =72.58% Sensitivity = 82.64% Specificity = 53.85% Kappa = 0.38 MCC = 0.379[90]

18.	Laplacian-modified naïve Bayesian method	PubChem AID 1851	VolSurf+ functional-class fingerprint descriptors (FCFPs)	AUC = 0.9 MCC = 0.61[91]
19.	GXBoost	PubChem AID 1851 PubChem AID 884	PubFP Descriptors PaDEL Descriptors	Accuracy = 89.4%[92]
20.	Multilayer Perceptron (MLP)	Inhouse Experimental Assay Dataset	Molecular Descriptors Atom Pair Descriptors Pharmacophoric Donor-Accept or Pair Descriptors	r^2 = 0.79[93]
21.	Deep Neural Network (DNN)	AID 884	PaDEL-1D and 2D Descriptors PubChem Fingerprints	ACC = 0.884 SP = 0.652 SE = 0.951 MCC = 0.649 AUC = 0.929[94]

CHAPTER 3

MATERIALS AND METHODOLOGY

3. Materials and Methodology

To predict and classify compounds as either inhibitors or noninhibitors of CYP3A4, a large data set of CYP3A4 inhibitors with varying inhibition potencies was curated. The dataset was curated and refined after which class labels were assigned. After which, descriptors were generated and refined to build machine learning models and the best model was determined using model validation techniques.

3.1 Dataset Curation

Data collection and curation is the first and arguably the most important step for the development of machine learning models. The dataset containing inhibitors with inhibitory potency (IC_{50}) values against CYP3A4 that was used in this study was curated from two publicly available datasets. The first was a large dataset downloaded through ChEMBL, while the second was a smaller dataset obtained through PubChem.

3.1.1 ChEMBL CYP3A4 Dataset

The ChEMBL dataset initially consisted of 11,460 entries of CYP3A4 inhibitors with IC_{50} values ranging from 0.053 – 77,624,711.66 nM. The data within the dataset was computed from a variety of sources, including the BindigDB database, DrugMatrix, Drugs for Neglected Diseases Initiative, Patent Bioactivity Data, as well as from scientific literature. However, the dataset also contains many duplicate and blank entries that require refining before it can be used for classification purposes.

3.1.2 PubChem CYP3A4 Dataset

PubChem AID 686941 was also included into the classification dataset. It consists of 31 compounds tested for inhibition against CYP3A4 as well as CYP3A5. To determine inhibitory activity of the compounds, the protocol included incubating 10 nM of the enzyme with substrate concentrations equal to their respective Michaelis constant (Km) concentrations, with 0.1 M potassium phosphate buffer, pH 7.4, and at 37 C. Analysis was performed using an API4000 mass spectrometer, and inhibitory activity was calculated after achieving chromatographic separation. The dataset consists of 12 columns including IC_{50} in µM, Compounds ID (CID), Substrate ID (SID) and Panel name.

In order to make the PubChem data compatible with the ChEMBL data, all the entries were removed for CYP3A5 inhibition for each compound. The SMILES code for the remaining entries relating to CYP3A4 were compared with those in the ChEMBL dataset, and entries with novel SMILES codes were merged with the ChEMBL dataset data. The IC_{50} values were converted from μM to nM. Following this, full dataset refinement was achieved.

3.1.3 Dataset Refining

The ChEMBL dataset contains many duplicate, inconsistent, and blank entries. As the aim of this study is to perform classification based on IC_{50} values, any entries with no inhibition data were removed. The dataset was further refined by removing entries with inconsistent inhibitory potency values, such as non-absolute values. After the final refinement a dataset with 4.314 CYP3A4 inhibitors was used for further study.

3.1.4 Class Label Application

Additionally, further preprocessing was performed by applying class labels to the data. In this study, an activity threshold of 1,000 nM for highly actives, 1,000-1,500 nM for actives/efficient, and <1,500 for least actives or inactives was used. The intermediate class was later on removed, and actives were given the class label ‘1’ while inactives were given the class label ‘0’, indicating a compound to be an inhibitor or non-inhibitor respectively. Figure 3 shows the distribution of the data, with a total of 3,051 inactives and 1,059 actives.

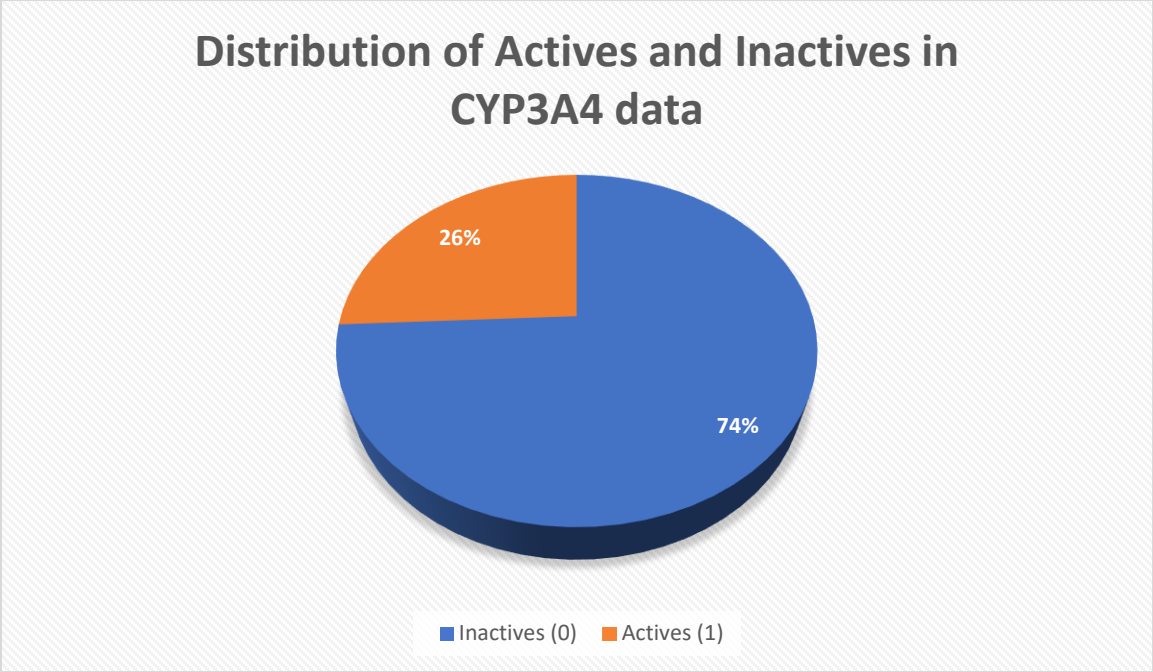

Figure 3. Distribution of Actives and Inactives in CYP3A4 Data

3.2 Descriptor Generation

Initially, 5,669 2D descriptors were calculated using the alvaDesc tool version 2.0.8[95]. The type of descriptors can be divided into 33 categories including Constitutional indices, Ring descriptors, Topological indices, Walk and path counts, Connectivity indices, Information indices, 2D matrix-based descriptors, 2D autocorrelations, Burden eigenvalues, P_VSA-like descriptors, ETA indices, Edge adjacency indices, Geometrical descriptors, 3D matrix-based descriptors, 3D autocorrelations, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, Randic molecular profiles, Functional group counts, Atom-centred fragments, Atom-type E-state indices, Pharmacophore descriptors, 2D Atom Pairs, 3D Atom Pairs, Charge descriptors, Molecular properties, Drug-like indices, CATS 3D descriptors, WHALES descriptors, MDE descriptors, and Chirality descriptors.

3.2.1 Descriptor Refining

The descriptor data set generated required refinement due to the presence of ‘na’ values, multiple 0 value columns, and columns with little to no variance, refining of the descriptor data was needed. In order to remove the ‘na’ values, they were converted to ‘0’ through excel. All the 0 values were then removed together through python. As a result, 2,522 empty descriptors were

removed. Descriptors with low variance were removed by calculating the variance for each column, then removing descriptors with variance below 0.5. Overall, 1,181 useful descriptors remained after refining.

3.3 Feature Engineering

Feature Engineering is the process of converting raw data into useful features that help us to understand our model better and increase its predictive power. This was done through the MinMaxScaler() function in the SciKitLearn Library. This aims to transform the data in all the columns such that each value is proportionally within the range 0 – 1. Normalizing data in this way serves to improve the performance and training reliability of the model by transforming the data to be on a similar scale while also preserving the shape of the data. Figure 4 and Figure 5 display a snippet of the data before and after scaling.

IC50	MW	AMW	Sv	Se	Sp	GD	0T	nSK	0A	...	s3_numSharedNeighbors	s2_numRotBonds	s3_numRotBonds
2900.0	387.44	7.596863	30.8789	52.4359	31.5344	0.082621	51	27	6	...	0.0	0.0	0.0
13000.0	444.60	7.939286	36.9703	55.9742	39.1935	0.072581	56	32	22	...	0.0	1.0	1.0
16500.0	555.71	7.311974	46.6421	76.7229	48.6938	0.056410	76	40	17	...	0.0	0.0	1.0
8700.0	421.92	8.438400	33.5785	50.5561	34.8866	0.075862	50	30	18	...	0.0	0.0	0.0
6200.0	378.43	8.226739	30.4336	46.8943	30.9661	0.082011	46	28	12	...	0.0	0.0	0.0

Figure 4. Data Before Scaling

IC50	MW	AMW	Sv	Se	Sp	GD	0T	nSK	0A	...	s3_numSharedNeighbors	s2_numRotBonds
0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	...	0.0	0.000000
0.001235	0.000443	0.000427	0.000427	0.000433	0.000443	0.003891	0.008065	0.015873	0.034483	...	0.0	0.030303
0.002469	0.000887	0.000855	0.000854	0.000865	0.000885	0.007782	0.016129	0.031746	0.068966	...	0.0	0.000000
0.003704	0.001330	0.001282	0.001280	0.001298	0.001328	0.011673	0.024194	0.047619	0.103448	...	0.0	0.000000
0.004938	0.001773	0.001709	0.001707	0.001731	0.001771	0.015564	0.032258	0.063492	0.137931	...	0.0	0.000000

Figure 5. Data After Scaling

3.4 Feature Elimination

Feature elimination is the process of selecting only those features or descriptors within a dataset that are more relevant in predicting the target variable, which is the Class label in this case. The most relevant features were determined by using the feature_importances() function, which assigns a value to each feature or descriptor based on the role it plays in predicting the Class label. The higher the feature importance value, the more of a role it plays. The top 20 important

descriptors were selected to create a descriptor subset. All the models were run on the two Descriptor sets; 1. 1179 Descriptor Set, and 2. 20 Descriptor set. The model performances were evaluated and compared for all models on both descriptor sets.

3.6 Splitting Data

Both inhibitor datasets were randomly split into separate training and testing sets using the train_test_split() function in the sklearn.model_selection library. For all models, the data was split so that 30% of the data was used for testing while the remaining 20% was used for training. However, for the model built using the Random Forest Classifier, 33% of the data was used for testing as opposed to 30%, as this provided better results.

After splitting the data, the test set contained the descriptor data for 327 inhibitors and 906 noninhibitors, while the training set contained the descriptor data for 732 inhibitors and 2,145 noninhibitors. This data is displayed in Figure 6.

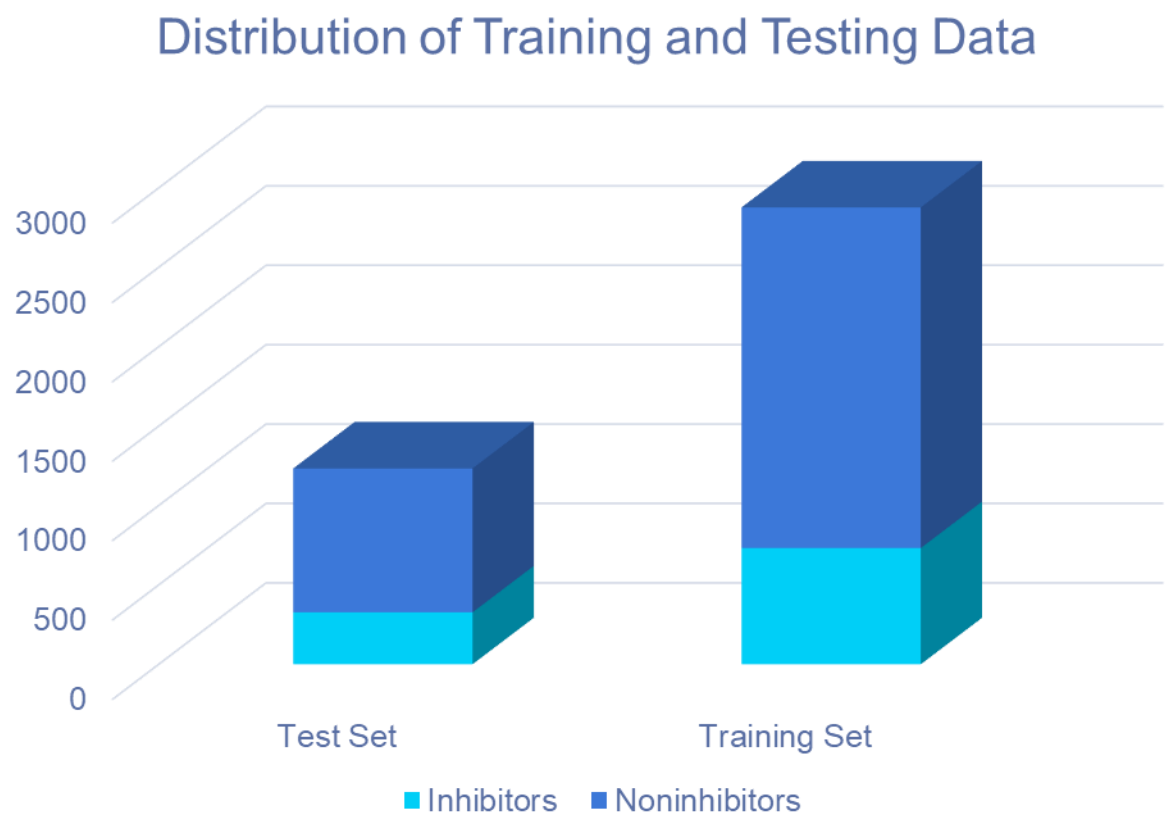

Figure 6. Distribution of Training and Testing Data

3.7 Machine Learning Models

Machine Learning is a modern technique that identifies the data patterns used to classify the data into desired classes. Several types of machine learning models can be applied to data depending upon the data type of attributes and classification labels. Here in this research, the aim is to perform binary classification in order to predict and differentiate between inhibitors and non-inhibitors of human CYP3A4. All machine learning models were built on the Jupyter Notebook platform in Python.

3.7.1 Logistic Regression

The LogisticRegression classifier was implemented using Python and Scikit-Learn. The model built a classifier that predicts whether a compound in our dataset is an inhibitor or not. Hyperparameter Optimization was used to tune and select the parameters that provided the best accuracy. After hyperparameter optimization, the selected parameters for the model included: C=100, penalty = 'l1', solver='liblinear', random_state=0, max_iter=50. Where C is a regularization parameter that determines the model's reliance on the training data and can be set to values such as 1, 10, 100 or 1000. the higher the C value, the more the model relies on the training data. The penalty function also performs regularization by reducing the coefficients of variables that are less contributive towards predicting the Class label. L1 regularization penalizes the sum of absolute values of the weights.

The 'liblinear' solver was used, which performs L1 regularization and is optimal for high dimension datasets such as the CYP3A4 dataset used in this study. Random state was set to 0, and the default settings were used for the remaining parameters. The value of the C hyperparameter was adjusted for best results. C = 1.0 gave the best results and was thus selected as the hyperparameter for the model. The solver parameter determines the algorithm used for optimization. The Liblinear solver, or Library for Large Classification solver, uses a Coordinate Descent algorithm which optimizes by achieving fairly accurate minimization along coordinate directions. The max_iter = 50 parameter ensures that the solver does not iterate more than 50 times, helping to prevent overfitting. Lastly, the random state parameter simply controls randomness in the machine learning model so that it is reproduceable. Whereas for all other models the default parameters were used.

3.7.2 Decision Tree

The DecisionTree Classifier was built using the SciKitLearn Library in Python. It aims to produce a flowchart-like tree where the branches in the tree represent decision rules and the leaf nodes represent outcomes. Hyperparameter Optimization was used to tune and select the parameters that provided the best accuracy. After hyperparameter optimization, the selected parameters for the model included: criterion='gini', max_depth=3, and random_state=0. The criterion parameter determines the criteria with which to split the branches in the tree. The gini index criteria establishes branching by calculating the frequency of mislabeling a compound when it is randomly selected. The Gini Index is calculated using the following formula[96]:

$$\text{GiniIndex} = 1 - \sum_{j} p_j^2$$

Where p_j is the probability of class j.

The random state once again controls the randomness of the model, and the max_depth parameter controls the complexity of the model by determining how deep the tree will be. All other parameters are kept at their default.

3.7.3 Random Forest

The RandomForest Classifier was imported from the ScikitLearn library in python to build the model. It builds multiple decision trees on samples of the data and then uses majority vote to classify CYP3A4 inhibitors and noninhibitors. The parameters that provided the best results included: n_estimators=100, random_state=0. Where the n_estimators = 100 ensures that 100 decision trees will be made before the model takes majority consensus, and random state determines the randomness of the model[97].

3.7.4 Support Vector Machine (SVM)

A support vector machine model was built by importing the SVM classifier from the ScikitLearn library. It aims to map the data points onto a grid so that a hyperplane can be constructed that separates inhibitor and oninhibitor data as best as possible, linearly, or nonlinearly. Hyperparameter Optimization was used to tune and select the parameters that provided the best accuracy. After hyperparameter optimization, the selected parameters for the model based on all

1179 descriptors included: kernel = rbf, C=1000, max_iter=100. The best parameters selected for the model created for 20 Descriptors included: kernel = linear, C=1000, max_iter=100. The kernel parameter is the mathematical function that decides how to manipulate the data. A linear kernel separates the data by creating a hyperplane that is a single line.

The Radial Basis Function, or rbf, kernel separates the data by creating a nonlinear hyperplane. The C parameter helps in lower misclassification of the data by setting margins, the higher the C value, the smaller the margin, as shown in Figure 7. A C value of 1000 indicates high reliance on the training data and smaller margins in the hyperplane. The max_iter parameter prevents overfitting by ensuring that the kernel does not iterate more than 100 times[98].

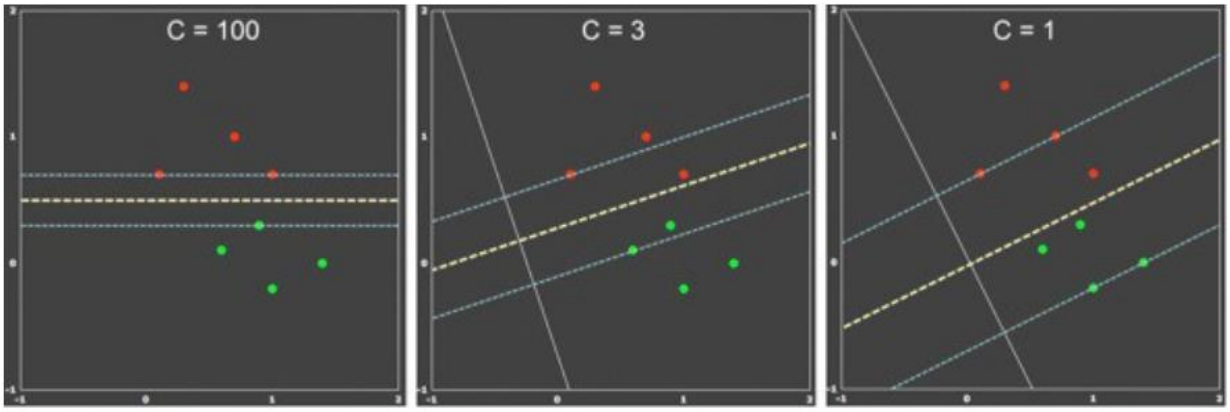

Figure 7. Effect of Varying C parameter on SVM Margins

3.7.5 Multilayer Perceptron (MLP)

A multilayer perceptron model was generated using the MLPClassifier in the SciKitLearn library. It is a type of feedforward artificial neural network that consists of weighted input and hidden layers that are used to send a signal to the output layer which in turn makes a decision or prediction about whether a data point is to be classified as an inhibitor of noninhibitor of CYP3A4. Figure 8 displays a simplified version of a multilayer perceptron.

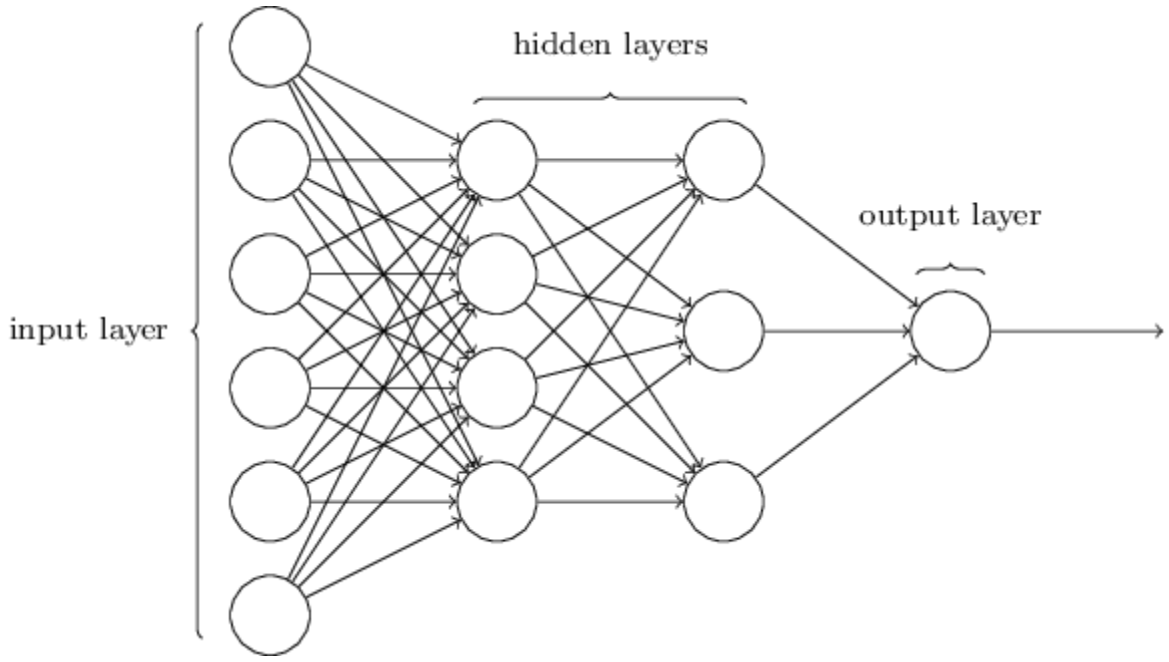

Figure 8. Example Multilayer Perceptron

Hyperparameter Optimization was used to tune and select the parameters that provided the best accuracy. After hyperparameter optimization, the selected parameters for the model generated using all 1179 descriptors included: hidden_layer_sizes=(256,128,64,32), activation="relu", random_state=1, max_iter=100. On the other hand, the model generated using only 20 descriptors had the following parameters: hidden_layer_sizes=(100, 50, 30),alpha = 0.05, learning_rate = 'constant', activation="relu", random_state=1. The hidden_layer_sizes parameter determines the number of hidden layers used in the model, as well as the number of nodes used in each layer. The 1179 descriptor model generates 4 hidden layers with decreasing nodes in each layer, while the 20-descriptor model works best with 3 hidden layers, with 100, 50, and 30 nodes in the respective layers. The activation parameter determines which activation function is used to linearly maps the weighted inputs to the output of each neuron. The Rectified Linear Unit function, or relu function, returns 0 if it receives any negative input, and returns the value itself for any positive value x. The random state parameter again controls the randomness of the model.

The alpha parameter is a penalty term and helps to prevent overfitting by constraining the weights. The learning rate parameter controls how much to change the weights in the model based on the estimated error. A constant learning rate keeps the learning rate the same for each step while

other learning rate values may cause it to vary at each step. All other parameters were kept at their default values[99].

3.7 Model Performance Evaluation

The performances of the models were calculated and compared using Accuracy, Classification Error, Sensitivity, Specificity, Precision, True Positive Rate (TP rate), False Positive Rate (FP rate) Area under the receiver operating characteristic curve (AUC ROC), and Mathews Correlation Coefficient (MCC).

3.7.1 Classification Accuracy

Classification Accuracy is defined as the ratio of correct predictions to the total number of input samples. It is calculated through the following formula[100].

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

Where TP indicates the number of true positives, TN indicates the number of true negatives, FP indicates the number of false positives and FN indicates the number of false negatives.

3.7.2 Classification Error

Classification error is the ratio of incorrect predictions to the total number of input samples. It is calculated through the following formula[101].

$$\text{Classification error} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

3.7.3 Specificity

Specificity is the defined as the proportion of true negatives that are correctly identified by the model, and as calculated by[100]:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

3.7.4 Sensitivity

Sensitivity is defined as the measure of how well a model can detect positive instances, also known as the true positive rate. It is be calculated by[100]:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

3.7.5 Precision

Precision is the ratio of correct positive prediction to the total number of positive predictions. It can be calculated by[101]:

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

3.7.6 False Positive Rate

The FP rate is the ratio of incorrect positive predictions to the total number of actual negative predictions. It can be calculated by[101]:

$$\text{FP rate} = \frac{\text{FP}}{\text{FP} + \text{TN}}$$

3.7.7 AUC ROC

AUC ROC is the measure of the 2D space under a ROC graph. The ROC curve provides a measure of performance across all possible classification thresholds by plotting the rate of true positives with respect to the rate of false positives. AUC helps to determine the ability of the model to distinguish between classes by determining the probability that the model will rank a randomly chosen positive entry higher than that of a randomly chosen negative entry. The higher the AUC value, the better the model can distinguish between inhibitors and noninhibitors[101].

3.7.8 Mathew's Correlation Coefficient (MCC)

MCC determines how good a model is at differentiating classes in classification by measuring the difference between predicited values and actual values. MCC values range between +1 and -1, where +1 indicates a completely correct classifier and -1 indicates a completely incorrect classifier. It is calculated by[100]:

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

3.8 K-Fold Cross Validation

K-fold cross validation was used to calculate the accuracy of all models to better estimate the skill of the models. K-fold cross validation i works by randomly and evenly splitting the dataset into 'k' parts. The model is then trained on all subsets of the data except 1 (k-1 parts). The excluded, or holdout set, is then used to test the accuracy model. This is repeated k number of times, until every subset of the data has been used once as the holdout set for testing. The final accuracy is then calculated as the average accuracy of all k sub models[102]. Figure 9 is an example of how k-fold cross validation works[103].

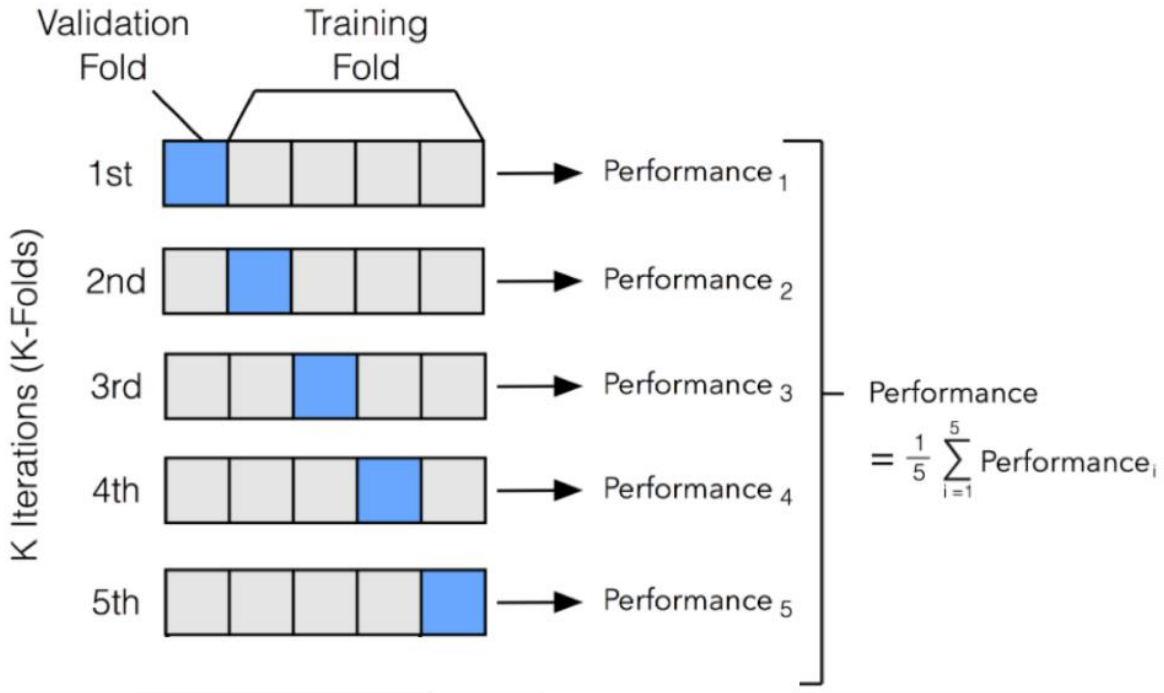

Figure 9. K Fold Cross Validation Visualization

CHAPTER 4

RESULTS

4: Results

The sim of the study is to build machine learning models to predict CYP3A4 inhibition and determine which features are important for decision making. Thus, this section presents the results of the performance evaluation matrices for each of the machine learning models in order to identify the best one. Additionally, through assessment of the model performances, and visualizing the decision trees, descriptors that are important for CYP3A4 inhibition can be defined.

4.1 Feature Importance

As previously mentioned, two descriptor datasets were generated, and models were built separately for both. Initially, the Alvadesc software generated 5,226 descriptors using the refined CYP3A4 inhibitor dataset. After removal of columns with all or mostly '0' values, missing values, and columns with variance less than 0.5, we were left with the first descriptor dataset, consisting of 1179 descriptors belonging to 33 different classes or descriptor types. The second descriptor dataset was generated by further refining through the calculation of feature importance. Feature importance is a function in Python that is ranks features in a dataset based on their importance by calculating the model performance error whenever a certain feature is adjusted. Highly important features results in higher performance error when adjusted, and less important features result in little to no change in prediction error when adjusted[104].

After calculating feature importance, the descriptors were ranked and ordered from most to least important. The top 20 most important descriptors were selected and used to create a second dataset of refined descriptors to compare model performances. Table 3 displays and describes the selected descriptors and presents them along with their relative importances. The important descriptor types include Chirality, P_VSA-like, 2D Atom Pairs, Pharmacophore, IC_{50}, Atom-centered fragments, Atom Pairs, Charge, and Edge adjacency indices Descriptors, the frequencies of which are displayed in Table 4. Figure 10 helps to visualize the importance of each descriptor.

Table 3. Descriptors used in Refined Descriptor Subset

Sr.	Descriptor	Descriptor Type	Description	Score
1.	CATS2D_06_DL	Pharmacophore	CATS2D Donor-Lipophilic at lag 06	0.00861321
2.	H-049	Atom-centered fragments	H attached to C3(sp3)/C2(sp2)/C3(sp2)/C3(sp)	0.006592495
3.	P_VSA_charge_1	P_VSA-like	P_VSA-like on partial charges, bin 1	0.003446934
4.	CATS2D_07_DL	Pharmacophore	CATS2D Donor-Lipophilic at lag 07	0.005586357
5.	T(O..S)	Atom Pairs	sum of topological distances between O..S	0.005313494
6.	P_VSA_MR_7	P_VSA-like	P_VSA-like on Molar Refractivity, bin 7	0.005094096
7.	s2_size	Chirality	number of heavy atoms of the substituent 2 normalized by the atoms shared	0.00506147
8.	s3_numRotBonds	Chirality	number of rotatable bonds of the substituent 3	0.004985541
9.	P_VSA_ppp_con	P_VSA-like	P_VSA-like on potential pharmacophore points, con – conjugated atoms	0.004748966
10.	s2_numAroBonds	Chirality	number of aromatic bonds of the substituent 2	0.004584241

11.	s4_numAroBonds	Chirality	number of aromatic bonds of the substituent 4	0.004557113
12.	qnmax	Charge	maximum negative charge	0.004550546
13.	Eig01_EA(dm)	Edge adjacency indices	eigenvalue n. 1 from edge adjacency mat. weighted by dipole moment	0.004194529
14.	P_VSA_ppp_ar	P_VSA-like	P_VSA-like on potential pharmacophore points, ar – aromatic atoms	0.004016411
15.	nLevel9	Chirality	number of neighbouring atoms of the chiral centre (level 9)	0.004010727
16.	F10[C-O]	2D Atom Pairs	Frequency of C – O at topological distance 10	0.003991062
17.	nLevel8	Chirality	number of neighbouring atoms of the chiral centre (level 8)	0.003945131
18.	T(N..O)	2D Atom Pairs	sum of topological distances between N..O	0.003644445
19.	s3_numAroBonds	Chirality	number of aromatic bonds of the substituent 3	0.00362814
20.	T(N..N)	2D Atom Pairs	sum of topological distances between N..N	0.003456152

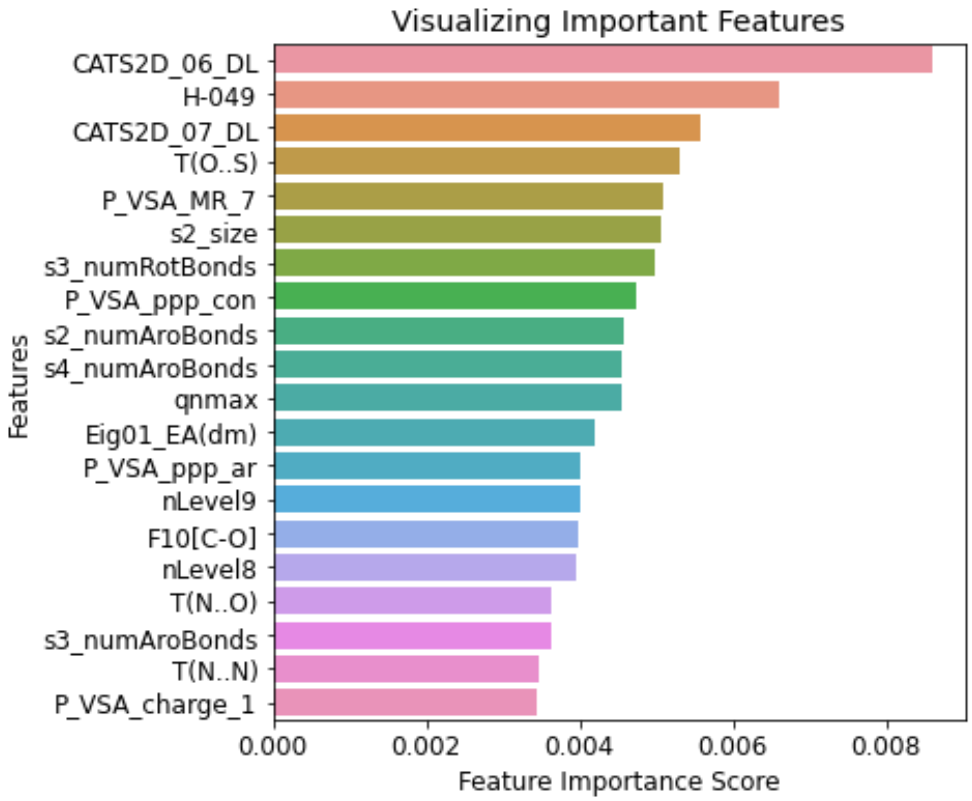

Figure 10. Descriptor Importance

Table 4. Descriptor Type and Count

Descriptor Type	Count
Chirality	7
P_VSA-like	4
2D Atom Pairs	3
Pharmacophore	2
Atom-centered fragments	1
Atom Pairs	1
Charge	1
Edge adjacency indices	1
Grand Total	**20**

4.2 Machine Learning Models

4.2.1 Logistic Regression

4.2.1.1 Model Performance for All Descriptors

The logistic regression model was built on the training dataset consisting of all 1179 descriptors as mentioned in the previous chapter. This model was tested on the test set and model performance was checked with the help of parameters like accuracy, precision, specificity, and sensitivity by taking inhibitors as the positive and noninhibitors as a negative class. Table 5 shows compares the cross validated accuracies for the training and test sets and displays the other performance evaluators for the test set, like classification error, specificity, sensitivity, precision, TP rate, FP rate, AUC and MCC. The model shows high prediction accuracy for both training and testing data with good specificity, and high sensitivity and precision. A low FP rate indicates low false positive predictions, or low chance of classifying noninhibitors as inhibitors. This can also be seen in the confusion matrix in Figure 11. The ROC curve can be seen in Figure 12, and the high AUC value of 0.769 indicates that the model is a good model whose predictions are not random. The MCC of 0.538 indicates a high correlation between true and predicted values.

The descriptors with the highest influence on classification were P_VSA_charge_1, ChiA_Dz(p), StsC, O-059, chiralMoment, s1_size, ATSC8i, SpDiam_Dt, D/Dtr04, T(O..Br), MaxDD, ATSC4m, ChiA_Dz(m), ChiA_Dz(Z), F09C-C, SsCH3, Yindex, ATSC8e, nROR, and SpDiam_B(m), their respective coefficients and the y intercept of the logistic regression are seen in Table 6 .

Table 5. Logistic Regression Model Evaluation: 1179 Descriptor Set

Data	Accuracy	Specificity	Sensitivity	Precision	FP Rate	AUC	Classification Error	MCC
Training	0.923	0.881	0.935	0.963	0.119	0.883	0.077	0.791
Testing	0.818	0.663	0.876	0.876	0.337	0.769	0.182	0.538

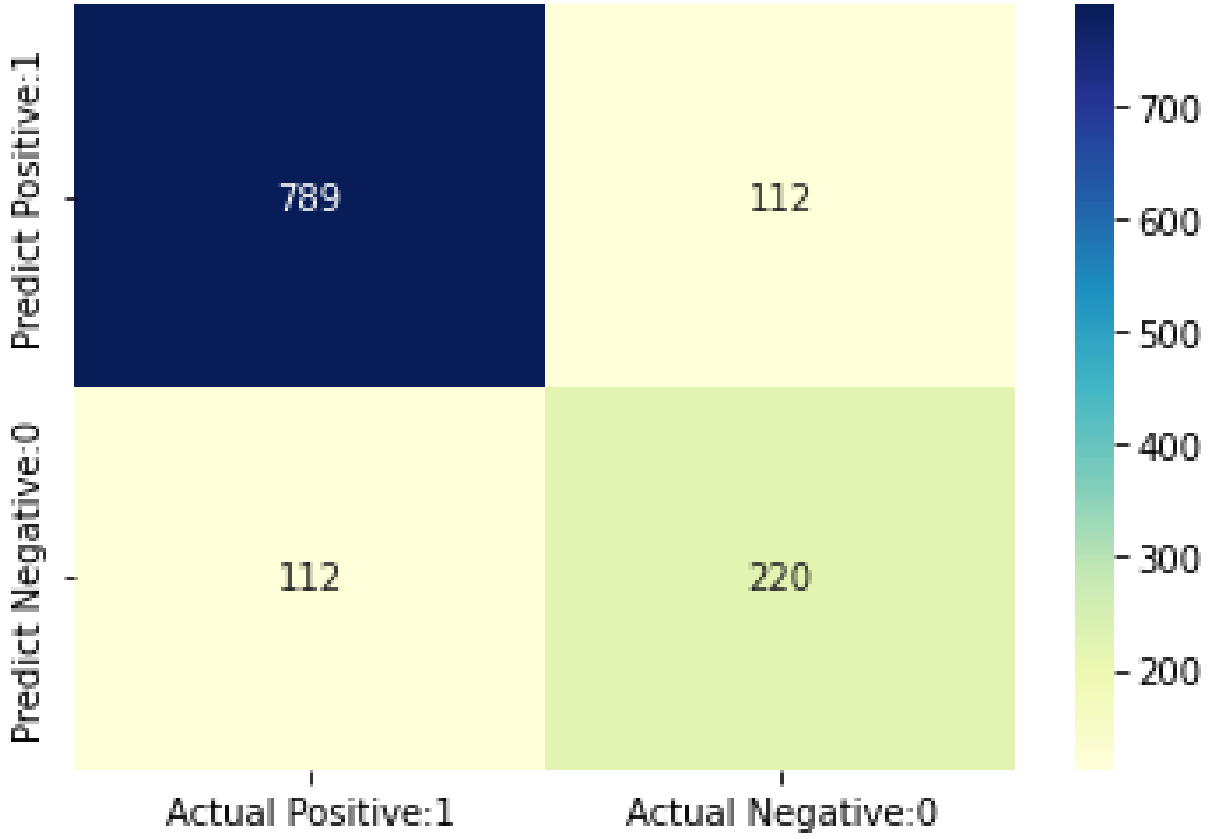

Figure 11. Logistic Regression Confusion Matrix: 1179 Descriptor Set

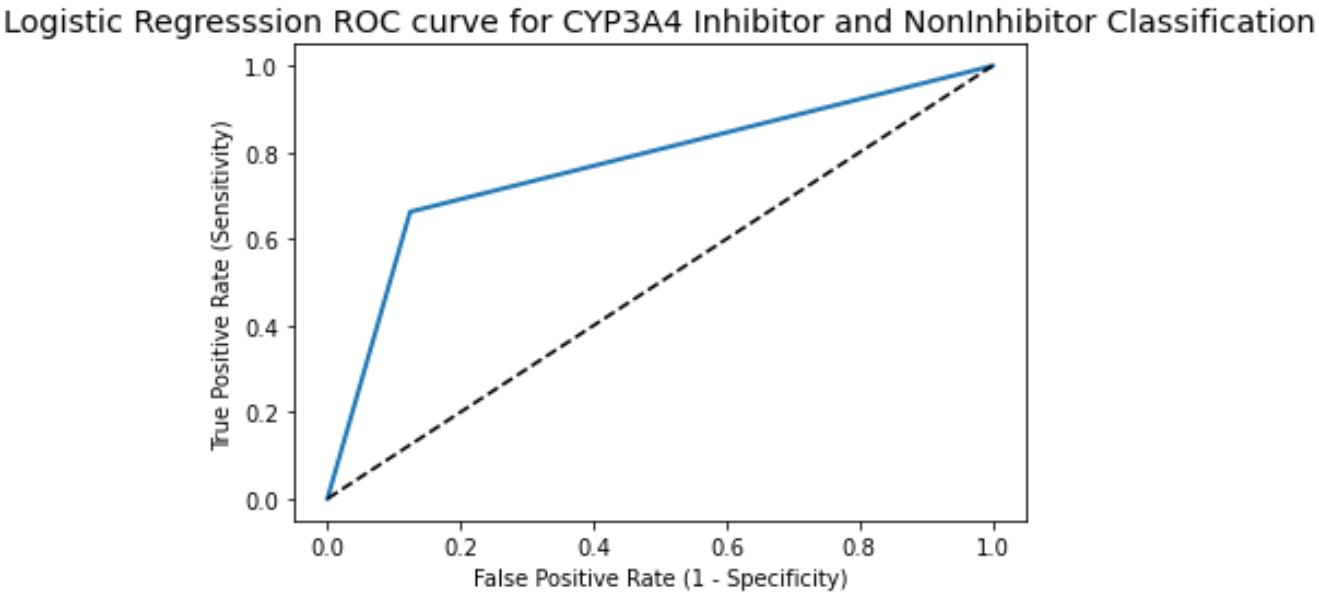

Figure 12. Logistic Regression ROC: 1179 Descriptor Set

Table 6. Logistic Regression Coefficients 1179 Descriptor Set

Sr.	Descriptors	Coefficients
0	Intercept	-1.06262893
1	SpDiam_B(m)	50.02708668
2	ATSC7m	39.45904162
3	nROR	38.14392325
4	J_Dz(m)	37.69750303
5	J_Dz(Z)	34.94809515
6	meanDistFromCC	34.85094625
7	SpMAD_B(m)	31.49826766
8	ChiA_D/Dt	30.86856477
9	P_VSA_ppp_A	-31.67862306
10	P_VSA_m_5	-32.82728164
11	nCIR	-33.02395796
12	chiralMoment	-33.15895957
13	J_Dz(p)	-34.30527496

14	O-059	-37.76760288
15	s1_size	-37.84041623
16	ChiA_Dz(p)	-39.4024263
17	IDET	-39.62337482
18	Wi_Dz(p)	-61.25856468
19	StsC	-67.26435048
20	SpDiam_Dt	-73.27013629

4.2.1.2 Model Performance for 20 Descriptors

The logistic regression model built with the second set of 20 descriptors yielded even higher accuracies than the model built on all descriptors, as seen in Table 7. The specificity, sensitivity, and precision rates are also seen to be high, while the FP rate is very low, indicating an even better model. The confusion matrix in Figure 13 shows the high number of correctly predicted compounds and the sparse number of incorrectly predicted compounds. The ROC curve in Figure 14 is nearly at 45° with a high AUC value of 0.968, indicative of a nearly perfect model. This is also demonstrated in the high MCC value of 0.923. The model performances for the two descriptor sets can be compared in Table 8. The model coefficients and intercept are displayed in Table 7.

Table 7. Logistic Regression Model Evaluation: 20 Descriptor Set

Data	Accuracy	Specificity	Sensitivity	Precision	FP Rate	AUC	Classification Error	MCC
Training	0.979	0.938	0.994	0.978	0.062	0.980	0.021	0.946
Testing	0.962	0.922	0.988	0.970	0.078	0.968	0.031	0.923

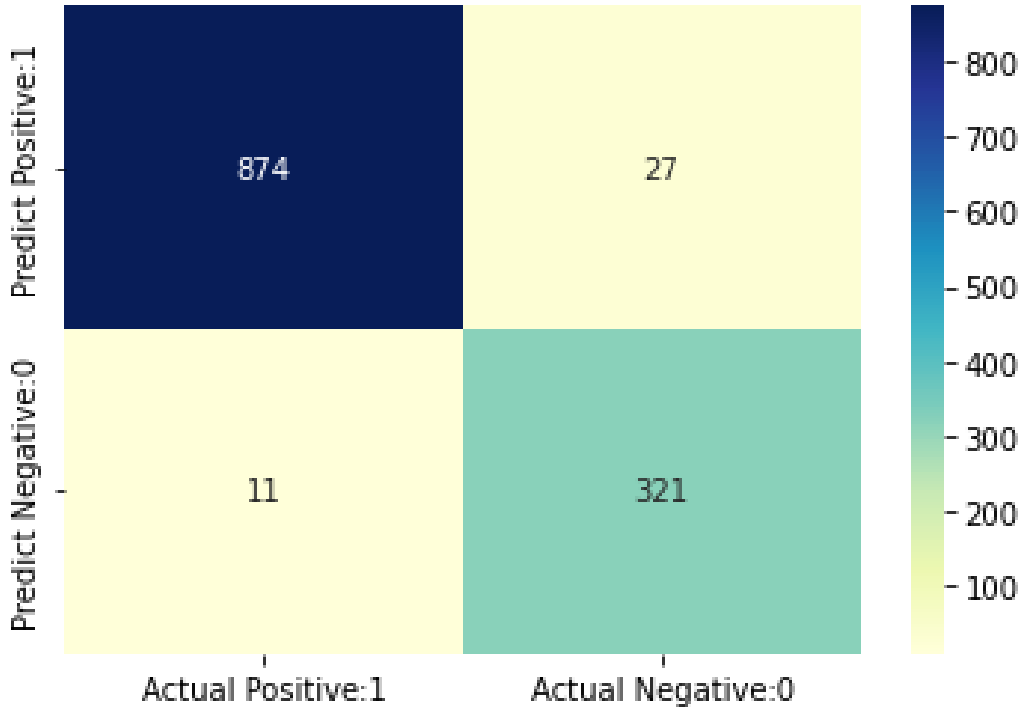

Figure 13. Logistic Regression Confusion Matrix for 20 Descriptors

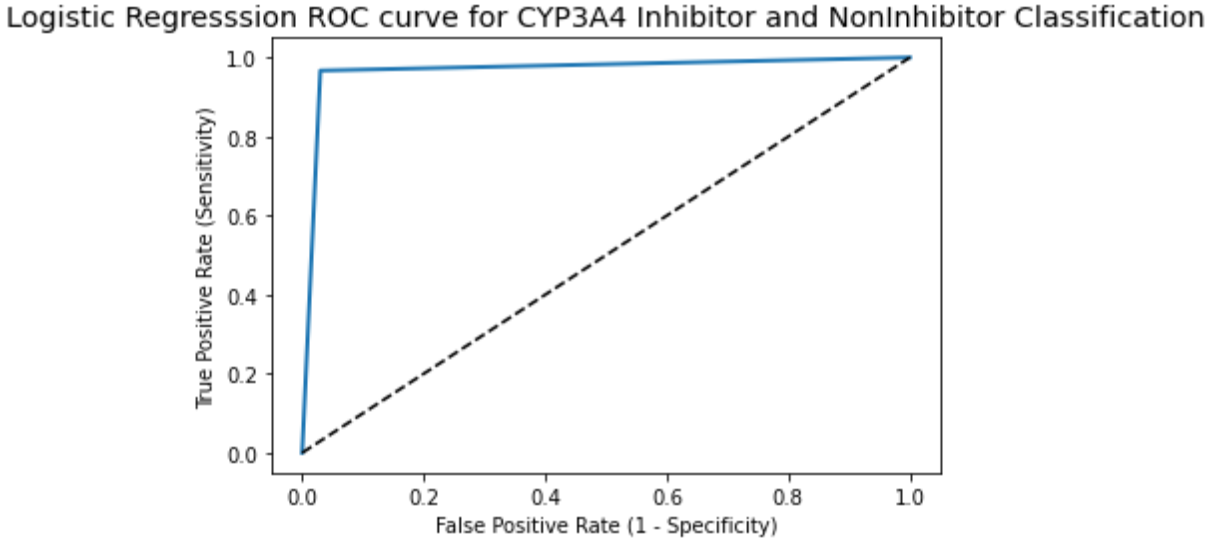

Figure 14. Logistic Regression ROC: 20 Descriptor Set

Table 8. Logistic Regression Coefficients: 20 Descriptor Set

Sr.	Descriptor	Coefficient
0	Intercept	0.915
1	CATS2D_06_DL	4.401
2	T(O..S)	3.715
3	nLevel9	2.083
4	P_VSA_ppp_ar	1.932
5	H-049	1.842
6	nLevel8	0.675
7	s2_numAroBonds	0.614
8	P_VSA_MR_7	0.531
9	P_VSA_charge_1	0.524
10	s3_numAroBonds	0.330
11	CATS2D_07_DL	0.248
12	Eig01_EA(dm)	0.112
13	s2_size	-0.398
14	s4_numAroBonds	-1.077
15	s3_numRotBonds	-1.240
16	F[10C-O]	-1.4970
17	qnmax	-1.658
18	P_VSA_ppp_con	-1.827
19	T(N..N)	-2.209
20	T(N..O)	-2.444

4.2.2 SVM

4.2.2.1 Model Performance for All Descriptors

For the SVM built on all 1179 descriptors, Table 9 displays a comparison of the performance metrics. We can see that this model results in extremely good accuracy in both training and testing, with values above 80%. This is also seen in the confusion matric in Figure 16

and the ROC curve in Figure. 15The MCC value above 50% tells us that the predictions are likely not random and the AUC of above 80% indicates extremely good predictive performance. Overall, based on these results, we can conclude that the model is can accurately predict whether a molecule is an inhibitor or noninhibitor of CYP3A4.

Table 9. SVM Model Evaluation: 1179 Descriptor Set

Data	Accuracy	Specificity	Sensitivity	Precision	FP Rate	AUC	Classification Error	MCC
Training	0.998	0.999	0.998	1.000	0.001	0.997	0.002	0.995
Testing	0.848	0.718	0.895	0.897	0.282	0.805	0.153	0.612

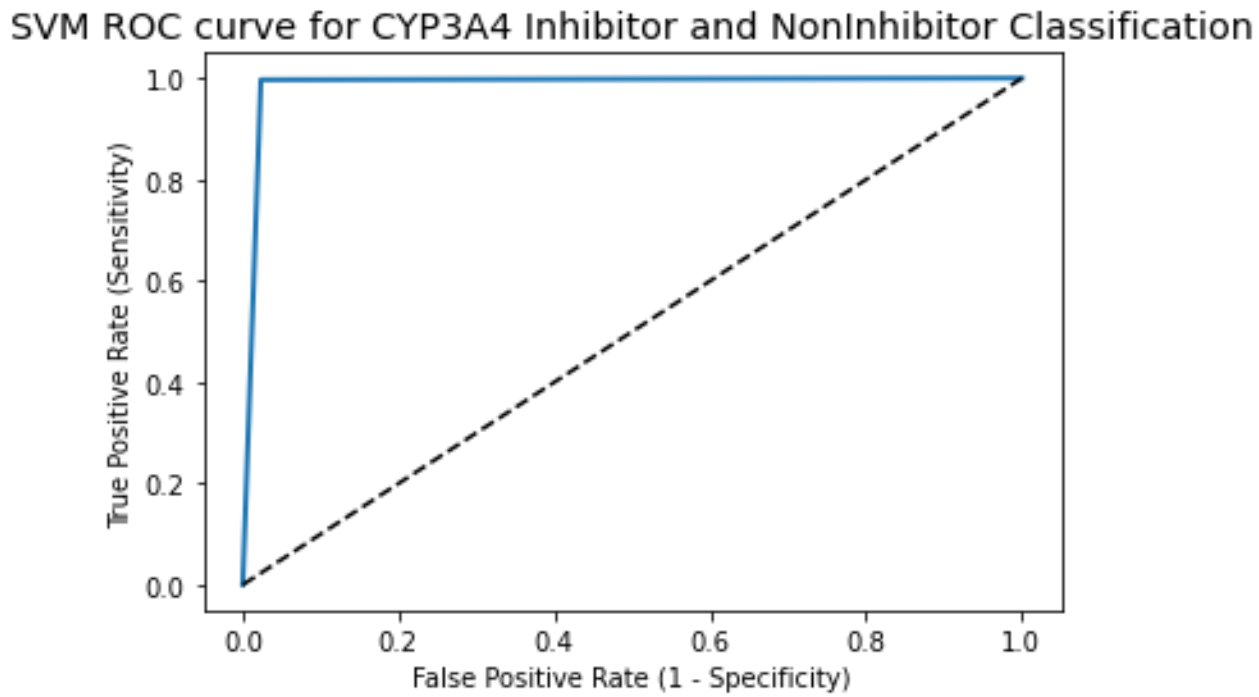

Figure 15. SVM ROC: 1179 Descriptor Set

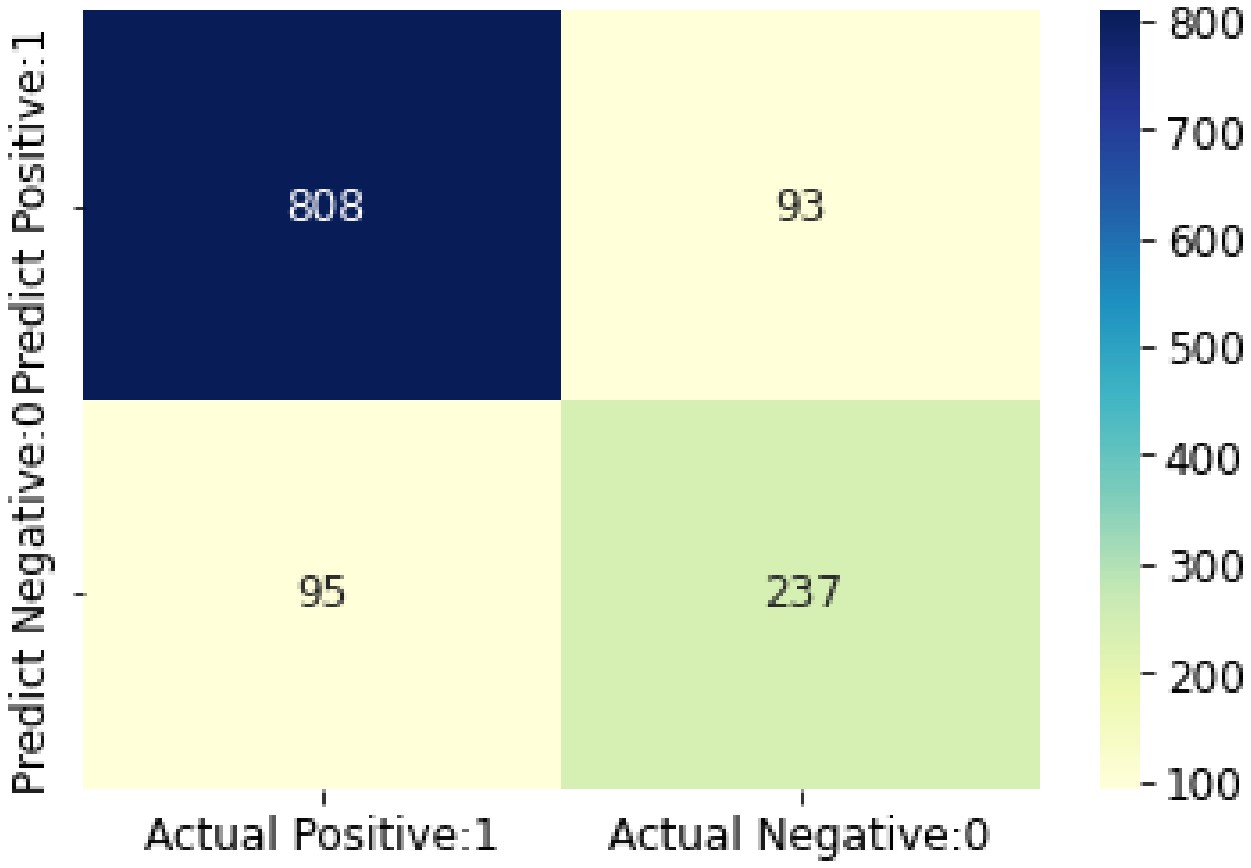

Figure 16. SVM Confusion Matrix: 1179 Descriptor Set

4.2.2.2 Model Performance for 20 Descriptors

From the performance metrics in Table 10, we can see that the 20-descriptor set once against yields higher accuracies overall as compared to the 1179 descriptor set. The confusion matrix in Figure 17 shows the high number of correctly predicted compounds and the low number of incorrectly predicted compounds. The ROC curve in Figure 18 is nearly at 45° with a high AUC value of 0.987, indicative of a nearly perfect model. This is also demonstrated in the high MCC value of 0.956. The model performances for the two descriptor sets can be compared in Table 17.

Table 10. SVM Model Evaluation: 20 Descriptor Set

Data	Accuracy	Specificity	Sensitivity	Precision	FP Rate	AUC	Classification Error	MCC
Training	0.989	0.959	1.000	0.986	0.041	0.993	0.011	0.972
Testing	0.982	0.940	0.999	0.977	0.060	0.987	0.018	0.956

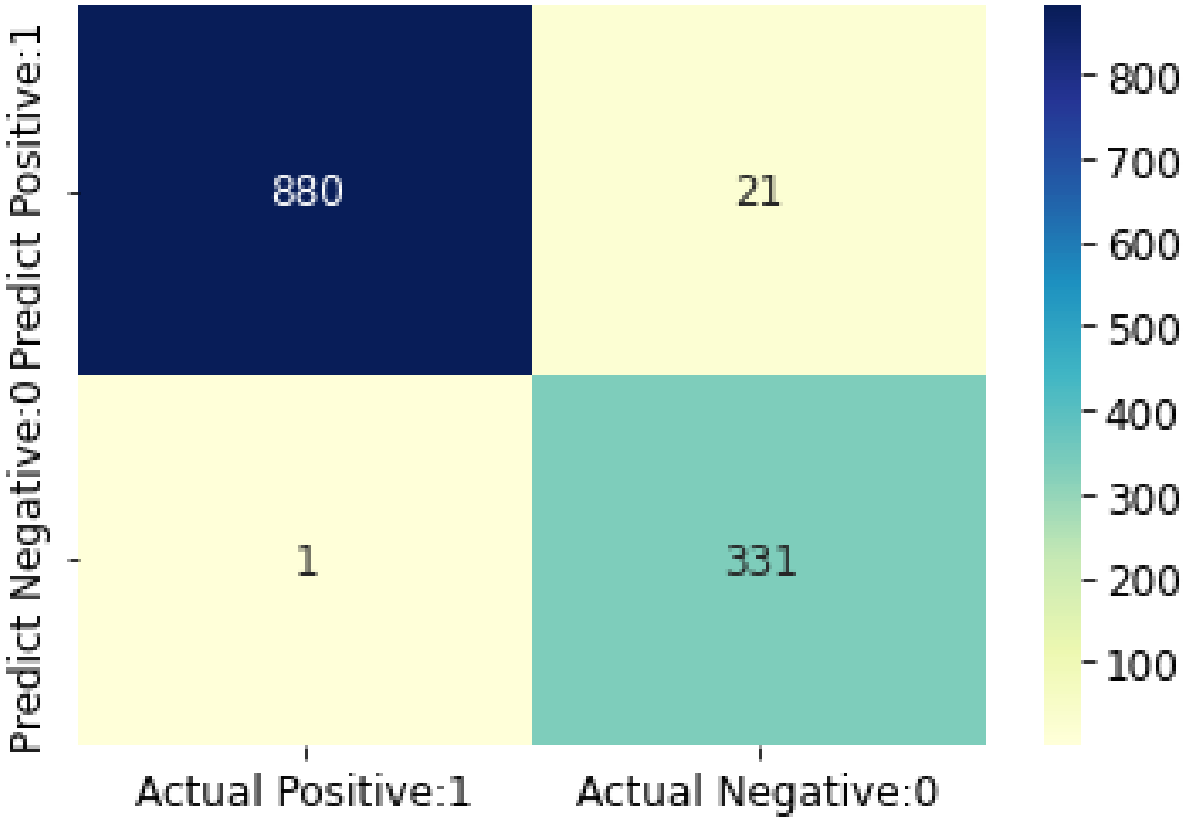

Figure 17. SVM Confusion Matrix: 20 Descriptor Set

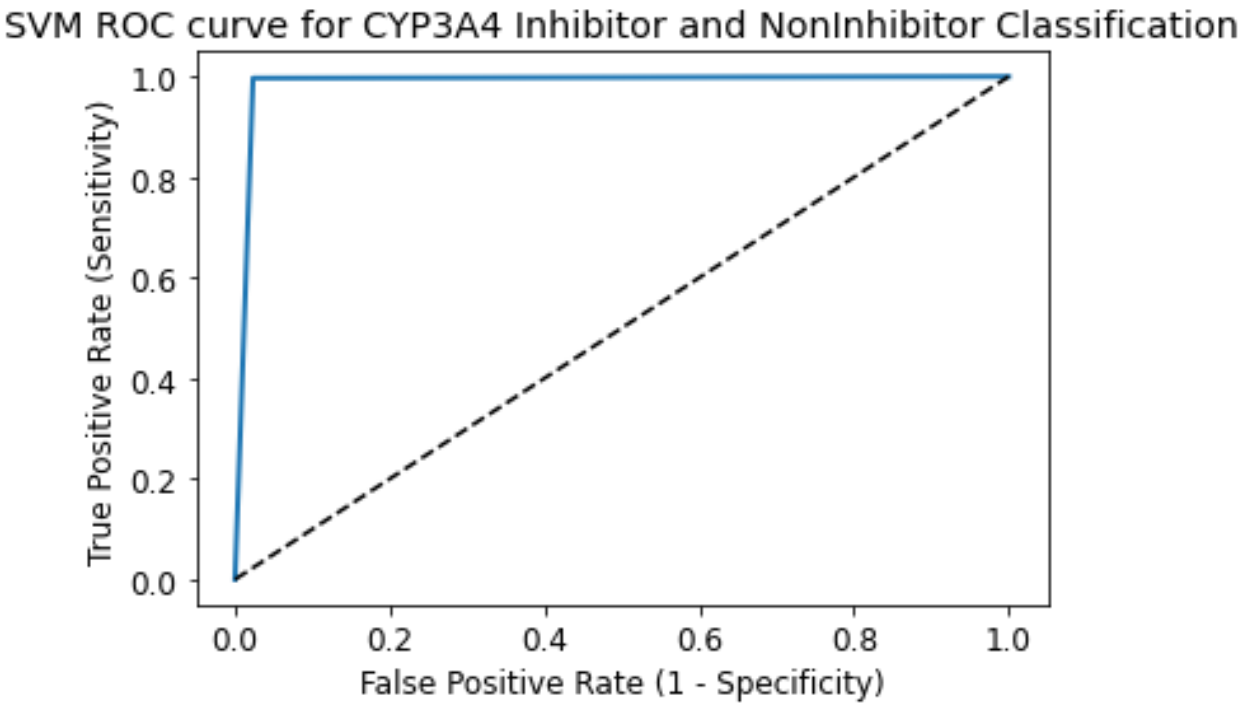

Figure 18. SVM ROC: 20 Descriptor Set

4.2.3 Decision Tree

4.2.3.1 Model Performance for All Descriptors

The Decision Tree in Figure 19 shows that, for the model built on the all-descriptor dataset, the D/Dtr05 descriptor was selected as the root node, which is a Ring descriptor that describes shortest distance between single rings at a certain position in the molecule. Moreover, the minssO, CATS2D_06_DL, WiA_D/Dt, MaxsOH, minssN and N% descriptors were selected as internal nodes. This shows that from 1179 descriptors in the data, these seven are integral for the classification of CYP3A4 inhibitors and noninhibitors.

While the performance measures in Table 11 show that the model displays good accuracy of 79.5%, it shows a low MCC score of 35.8%. This may be due to the class imbalance present in the data, where noninhibitors outnumber inhibitors by almost 3 times. Since Decision Trees are a cost sensitive learning and rely on information gain, it is sensitive to class imbalance and will make more accurate predictions for the majority class, and less accurate ones for the minority class. The confusion matrix in Figure 20 shows that the model resulted in relatively high correctly predicted values, but also a high number of False Negative predictions as a result of the class imbalance

issue. Figure 21 shows the ROC curve which has an AUC value of 0.613. The good accuracy evaluation indicates that the Decision Tree model created using all descriptors and with the chosen parameters results in a good classification model.

Table 11. Decision Tree Model Evaluation: 1179 Descriptor Set

Data	Accuracy	Specificity	Sensitivity	Precision	FP Rate	AUC	Classification Error	MCC
Training	0.806	0.837	0.803	0.981	0.163	0.632	0.194	0.412
Testing	0.795	0.792	0.776	0.975	0.208	0.613	0.222	0.358

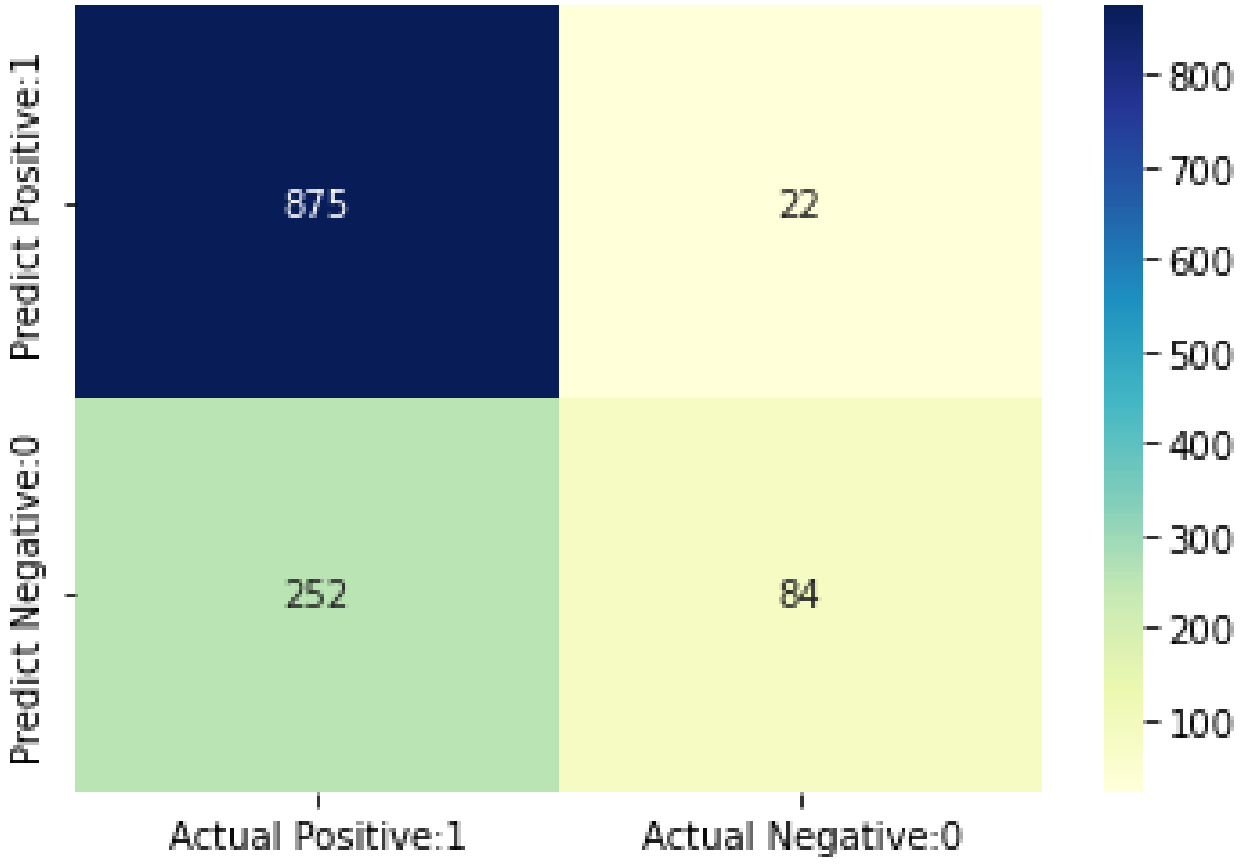

Figure 19. Decision Tree Confusion Matrix: 1179 Descriptor Set

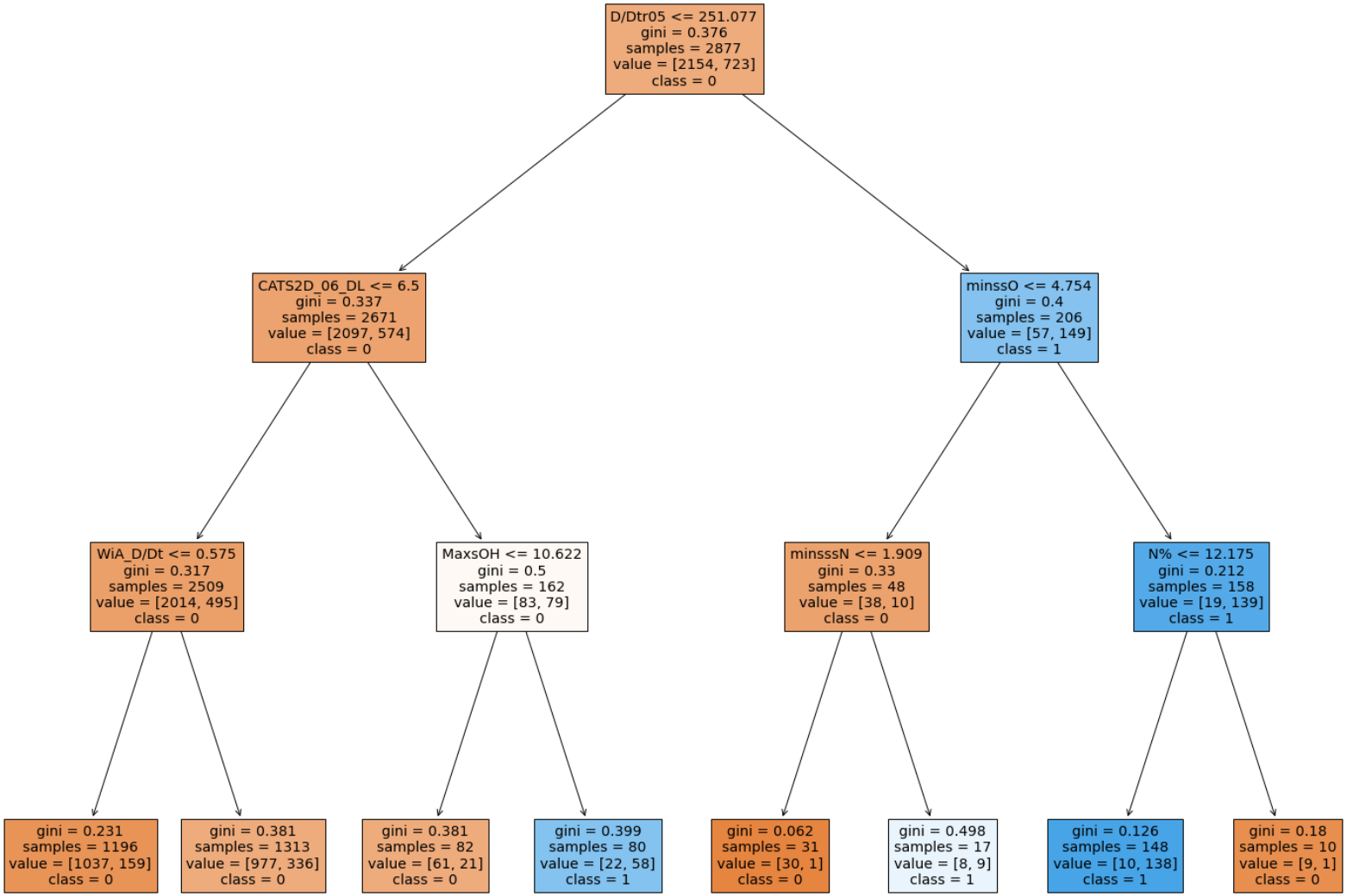

Figure 20. Decision Tree: 1179 Descriptor Se

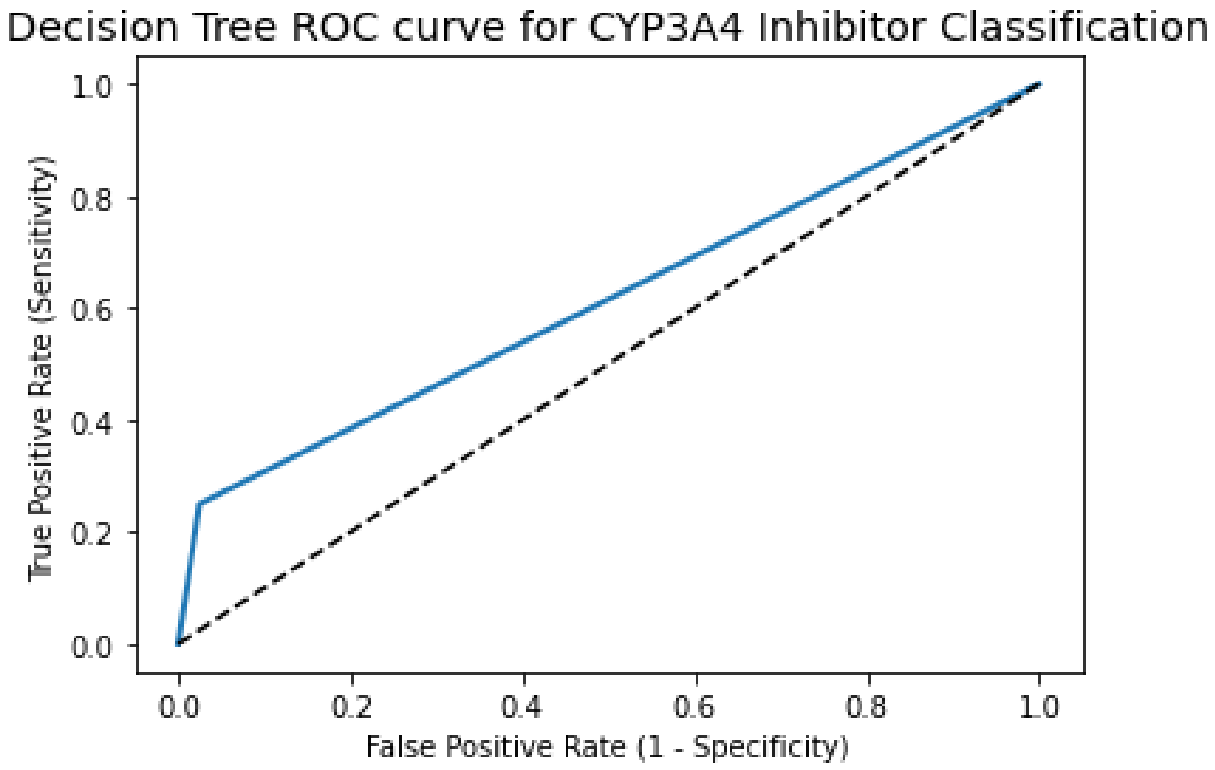

Figure 21. Decision Tree ROC: 1179 Descriptor Set

4.2.3.2 Model Performance for 20 Descriptors

The Decision Tree built on the 20-descriptor data set resulted in the tree in Figure 22. The tree shows that the H-048 descriptor was selected as the root node, which is Atom-centred fragments descriptor that quantifies the number of hydrogen atoms attached to C2(sp3)/C1(sp2)/C0(sp). Moreover, the S3_numRotBonds, CATS2D_06_DL, qnmax, and P_VSA_ppp_con descriptors were selected as internal nodes. This shows that from 20 descriptors in the data, these five are integral for the classification of CYP3A4 inhibitors and noninhibitors. The model built with the second set of 20 descriptors again yielded higher accuracies than the model built on all descriptors, as seen in Table 12. The specificity, sensitivity, and precision rates are also seen to be high, while the FP rate is lower, indicating an even better model. The confusion matrix in Figure 23 shows the high number of correctly predicted compounds and comparatively low number of incorrectly predicted compounds. The ROC curve in Figure 24 has an AUC value

of 0.592 indicating that the predictions are likely not random. The model performances for the two descriptor sets can be compared in Table.

Table 12. Decision Tree Model Evaluation: 20 Descriptor Set

Data	Accuracy	Specificity	Sensitivity	Precision	FP Rate	AUC	Classification Error	MCC
Training	0.796	0.857	0.791	0.987	0.143	0.606	0.204	0.370
Testing	0.792	0.557	0.782	0.902	0.442	0.592	0.254	0.281

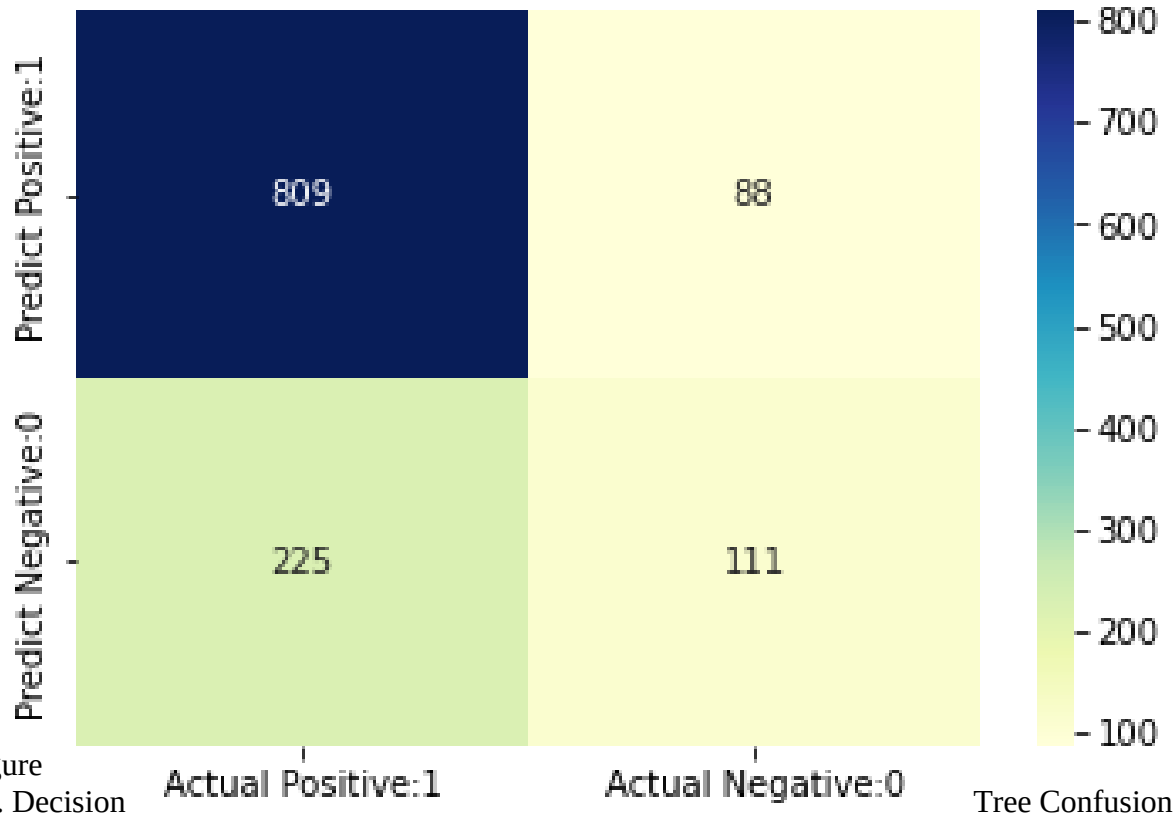

Figure 22. Decision Tree Confusion Matrix: 20 Descriptor Set

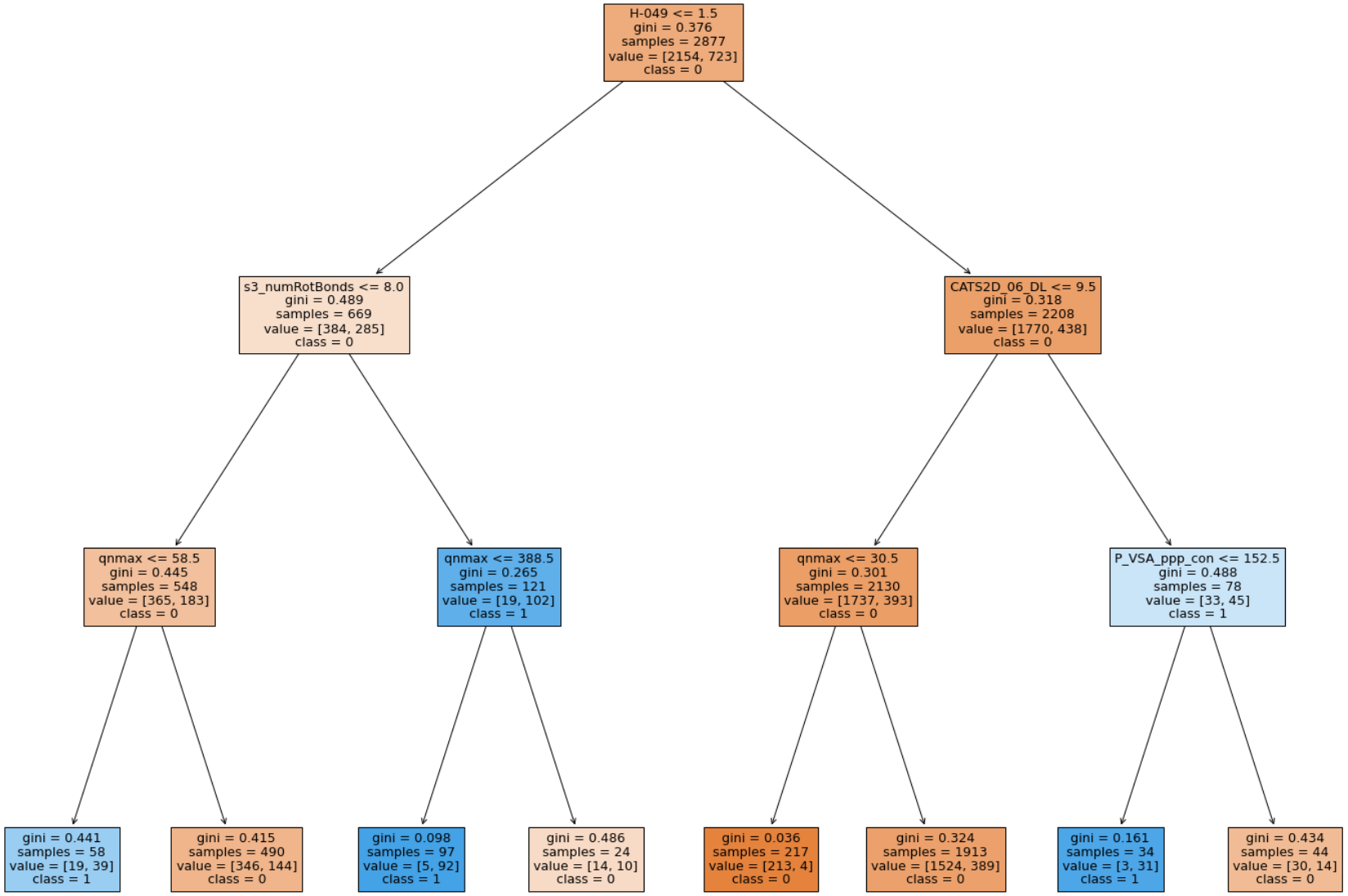

Figure 23. Decision Tree: 20 Descriptor Set

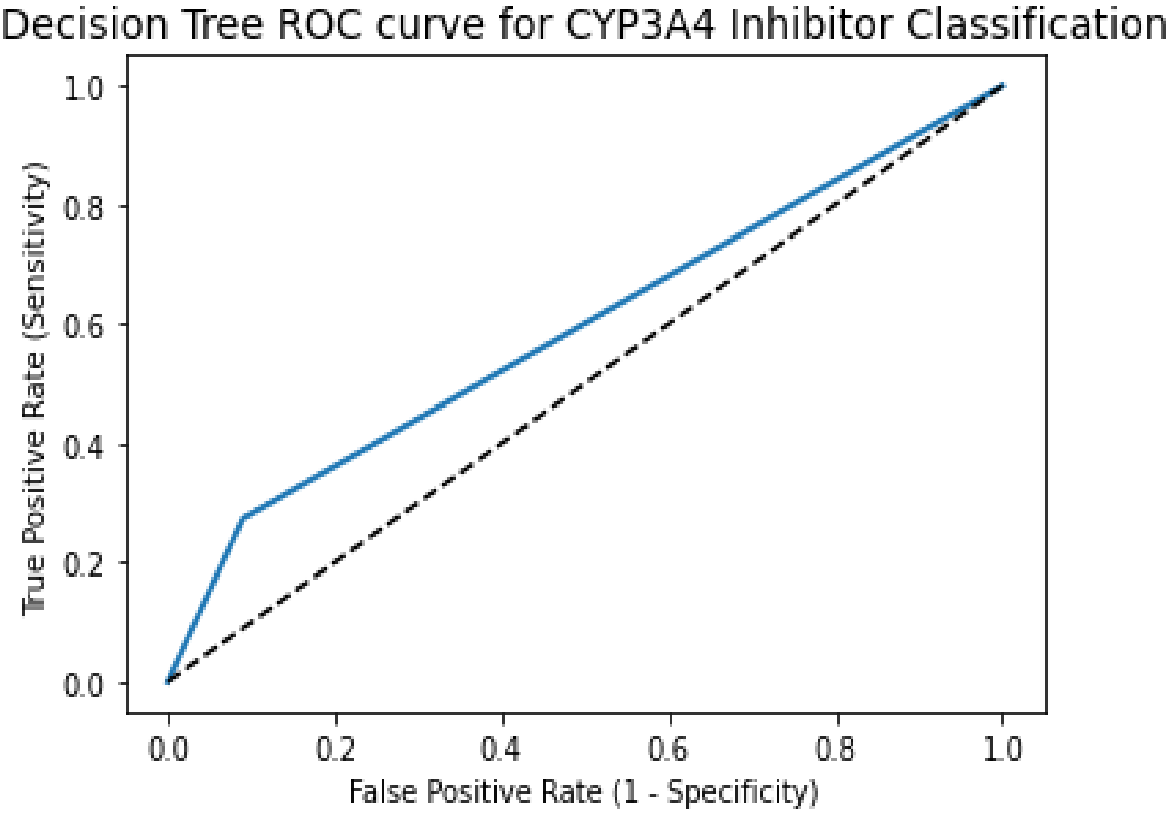

Figure 24. Decision Tree ROC: 20 Descriptor Set

4.2.4 Random Forest

4.2.4.1 Model Performance for All Descriptors

For the Random Forest model, the performance measures in Table 13 compares the accuracies for the training and test sets. The high training and testing set accuracies are indicative of a good model, as well as the low FP rate. The confusion matrix in Figure 25 shows that the model resulted in relatively high correctly predicted values and relatively low incorrectly predicted values. As Random Forests are built from multiple Decision Trees, it is susceptible to the same class imbalance problems as Decision Trees. This can be seen in the less than optimal ROC Curve in Figure 26, with the AUC of 0.6726, and the low MCC of 0.3999.

Table 13. Random Forest Model Evaluation: 1179 Descriptor Set

Data	**Accuracy**	**Specificity**	**Sensitivity**	**Precision**	**FP Rate**	**AUC**	**Classification Error**	**MCC**
Training	0.999	0.994	0.997	0.998	0.006	0.995	0.004	0.990
Testing	0.806	0.694	0.802	0.936	0.306	0.6726	0.214	0.399

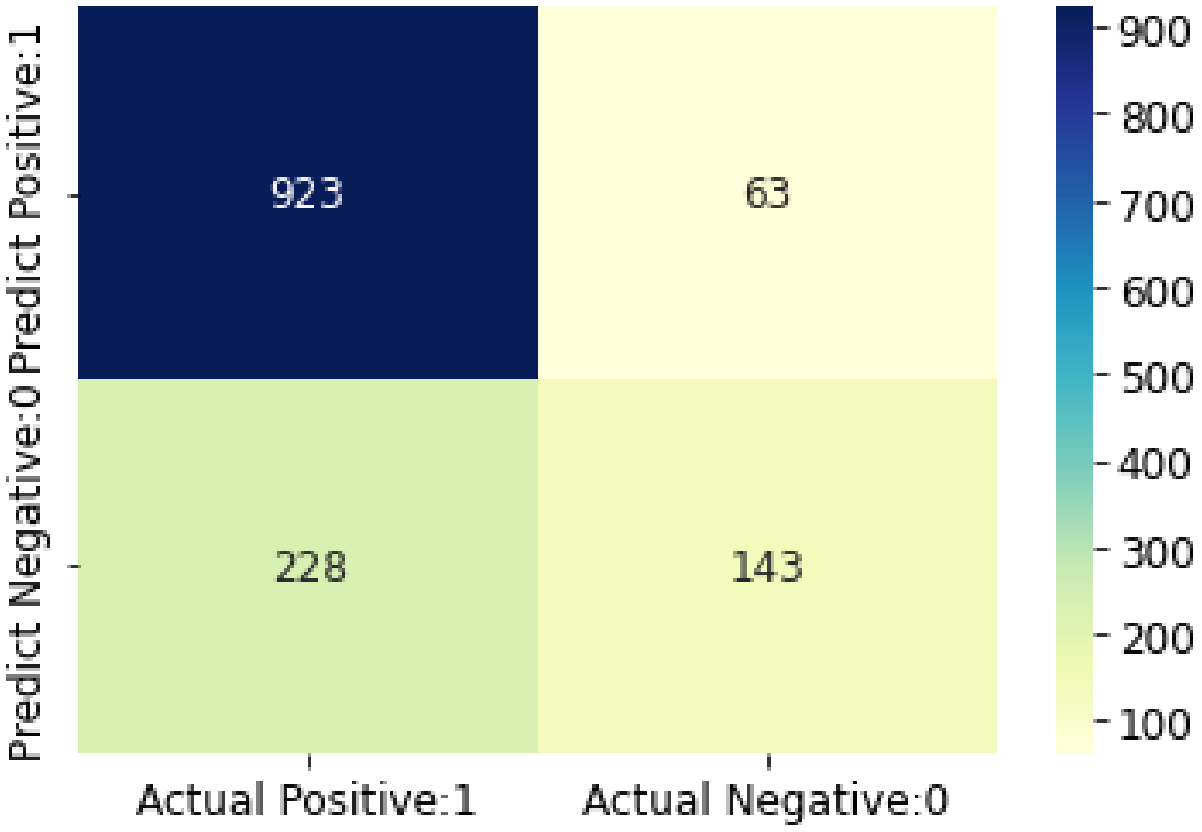

Figure 25. Random Forest Confusion Matrix: 1179 Descriptors

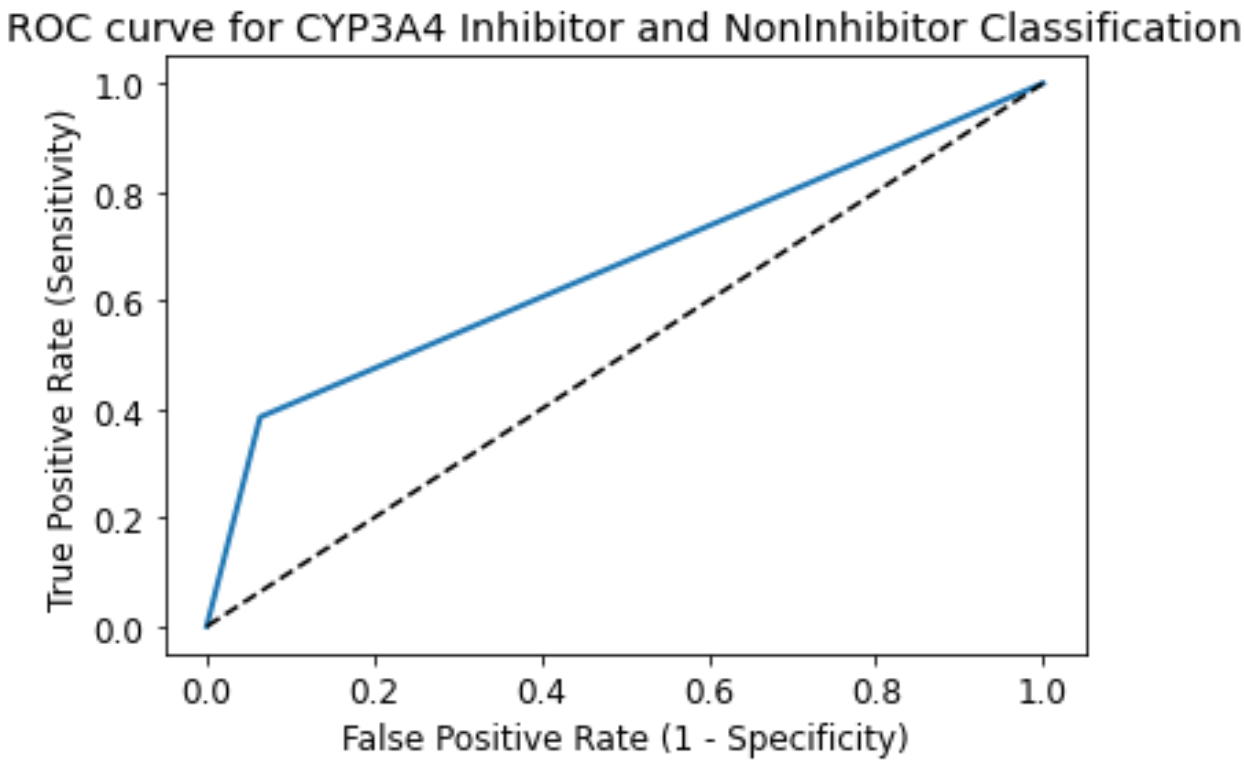

Figure 26. Random Forest ROC: 1179 Descriptors

4.2.4.2 Model Performance for 20 Descriptors

The 20-descriptor dataset yielded higher performance measures than the Random Forest model built on the 1179 descriptor dataset, as seen by the higher accuracies and overall performance measures in Table 14. Similarly, the lower FP rate and classification error are indicative of satisfactory performance. The confusion matrix in Figure 27 shows the high number of correctly predicted compounds and the small number of incorrectly predicted compounds. The ROC curve in Figure 28 has an AUC value of 0.7034, indicating that the predictions are not random. This is also demonstrated in the MCC value of 0.517. The model performances for the two descriptor sets can be compared in Table.

Table 14. Random Forest Model Evaluation: 20 Descriptor Set

Data	Accuracy	Specificity	Sensitivity	Precision	FP Rate	AUC	Classification Error	MCC
Training	0.996	0.994	0.997	0.998	0.006	0.995	0.004	0.990
Testing	0.833	0.858	0.826	0.972	0.142	0.703	0.166	0.540

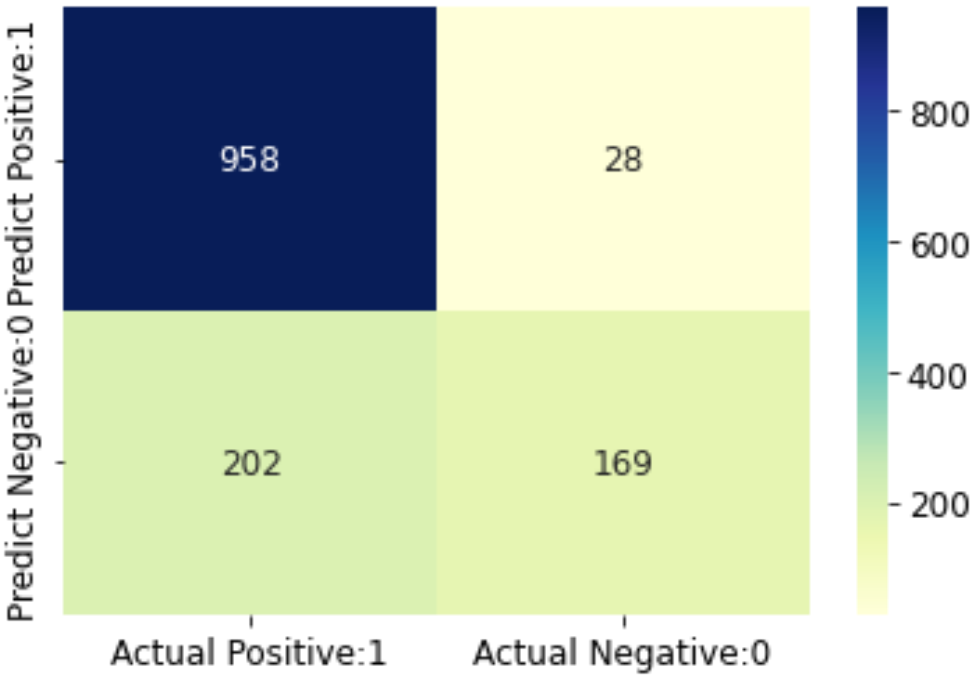

Figure 27. Random Forest Confusion Matrix: 20 Descriptor Set

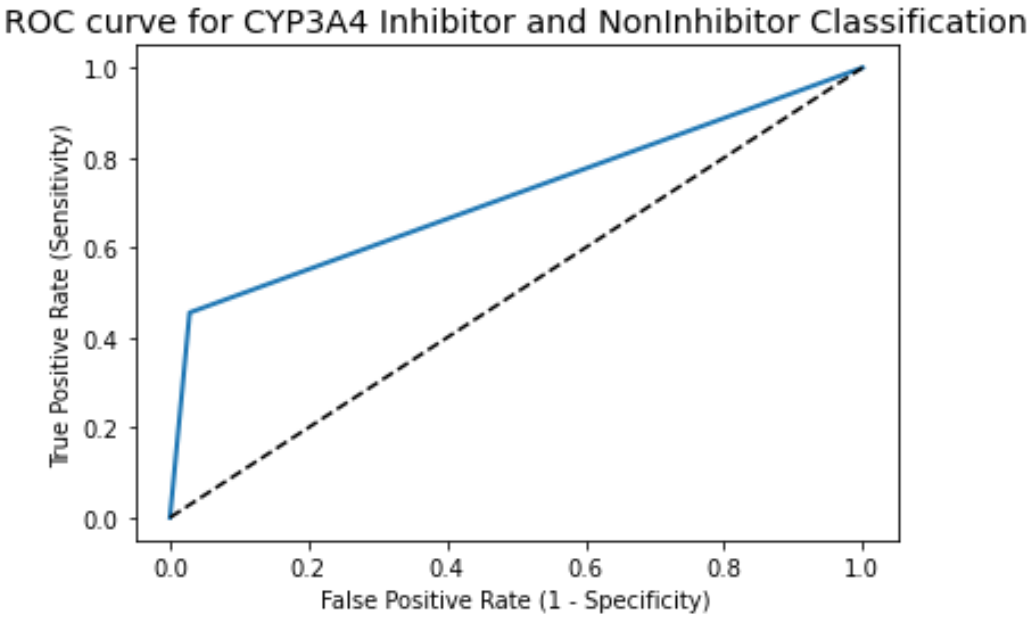

Figure 28. Random Forest ROC: 20 Descriptor Set

4.2.5 Multilayer Perceptron

4.2.5.1 Model Performance for All Descriptors

The Multilayer perceptron built on the 1179 descriptor dataset yielded excellent performance as indicated by the performance measure results in Table 15. The confusion matrix in Figure 29 also shows that the model resulted in relatively high correctly predicted values and relatively low incorrectly predicted values. The ROC curve in Figure 30 has an AUC of 0.779, indicating that the predictions made by the model are not random. The performance evaluation indicates that the Multilayer Perceptron created using all descriptors and with the chosen parameters results in a good classification model.

Table 15. MLP Model Evaluation: 1179 Descriptor Set

Data	**Accuracy**	**Specificity**	**Sensitivity**	**Precision**	**FP Rate**	**AUC**	**Classification Error**	**MCC**
Training	0.977	0.973	0.978	0.991	0.027	0.963	0.023	0.977
Testing	0.950	0.725	0.863	0.918	0.275	0.779	0.167	0.950

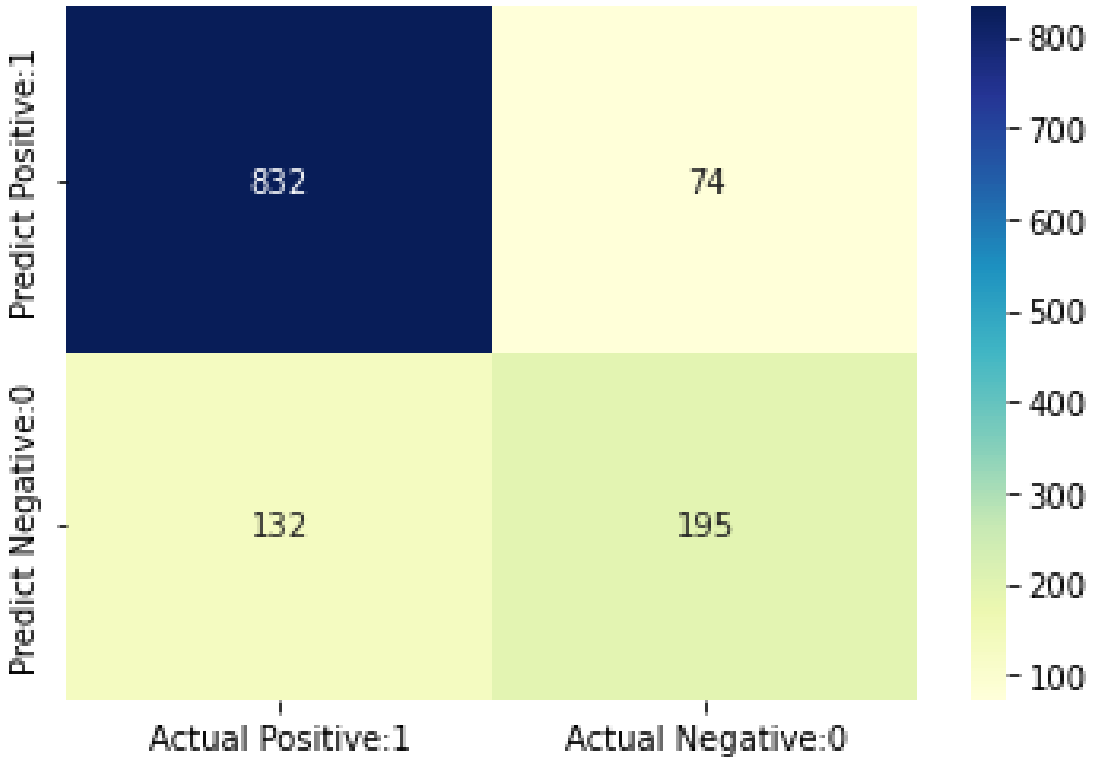

Figure 29. MLP Confusion Matrix: 1179 Descriptor Set

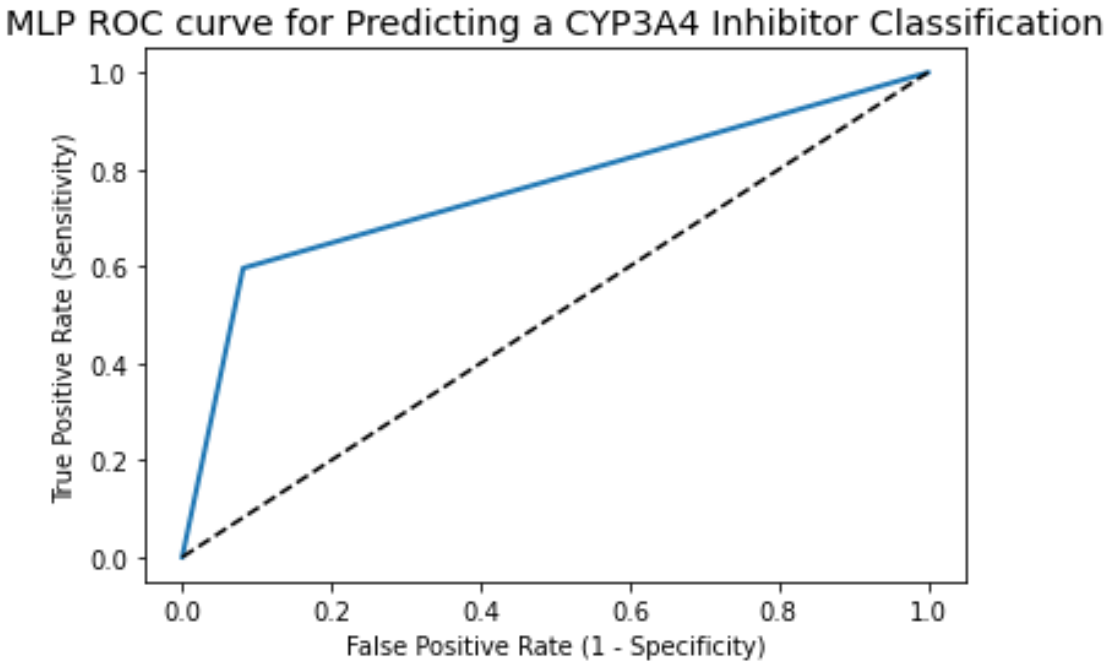

Figure 30. MLP ROC: 1179 Descriptor Set

4.2.5.2 Model Performance for 20 Descriptors

The Multilayer Perceptron built on the second set of 20 descriptors yielded higher accuracies than the model built on all descriptors, as seen in Table 16. The specificity, sensitivity, and precision rates are also seen to be high, while the FP rate is very low, indicating an even better model. The confusion matrix in Figure 31 shows the high number of correctly predicted compounds and the low number of incorrectly predicted compounds. The ROC curve in Figure 32 has an AUC value of 0.766, indicating that the predictions are not random. This is also demonstrated in the MCC value of 0.573. The model performances for the two descriptor sets can be compared in Table 17.

Table 16. MLP Model Evaluation: 20 Descriptor Set

Data	Accuracy	Specificity	Sensitivity	Precision	FP Rate	AUC	Classification Error	MCC
Training	0.9802	0.981	0.980	0.994	0.019	0.967	0.020	0.947
Testing	0.976	0.75	0.867	0.927	0.250	0.766	0.158	0.573

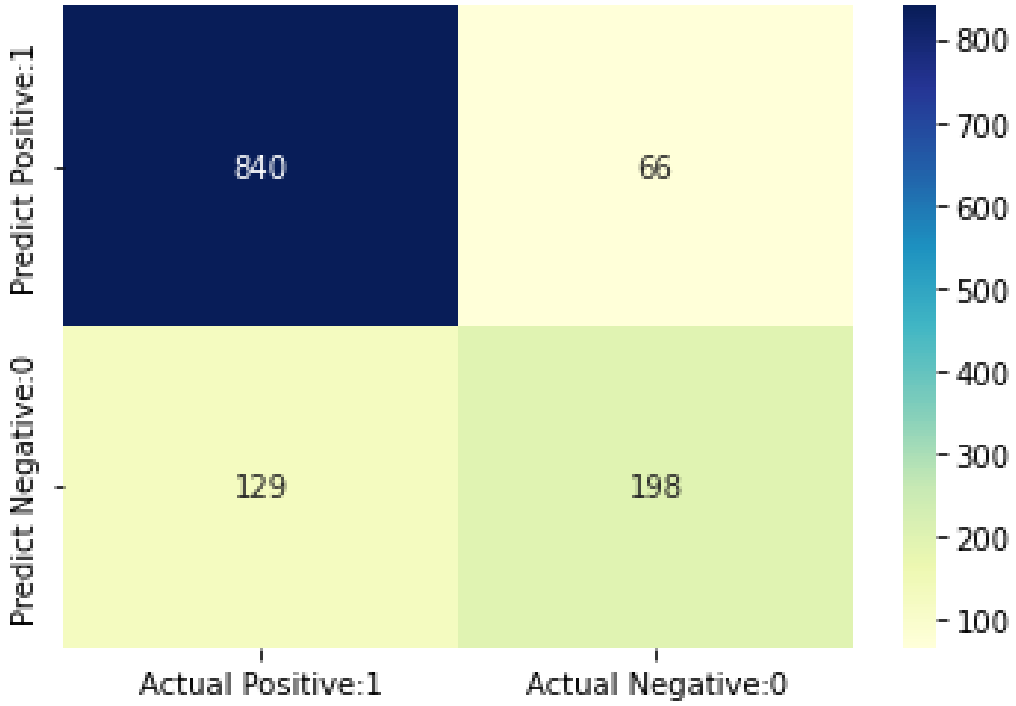

Figure 31. MLP Confusion Matrix: 20 Descriptor Set

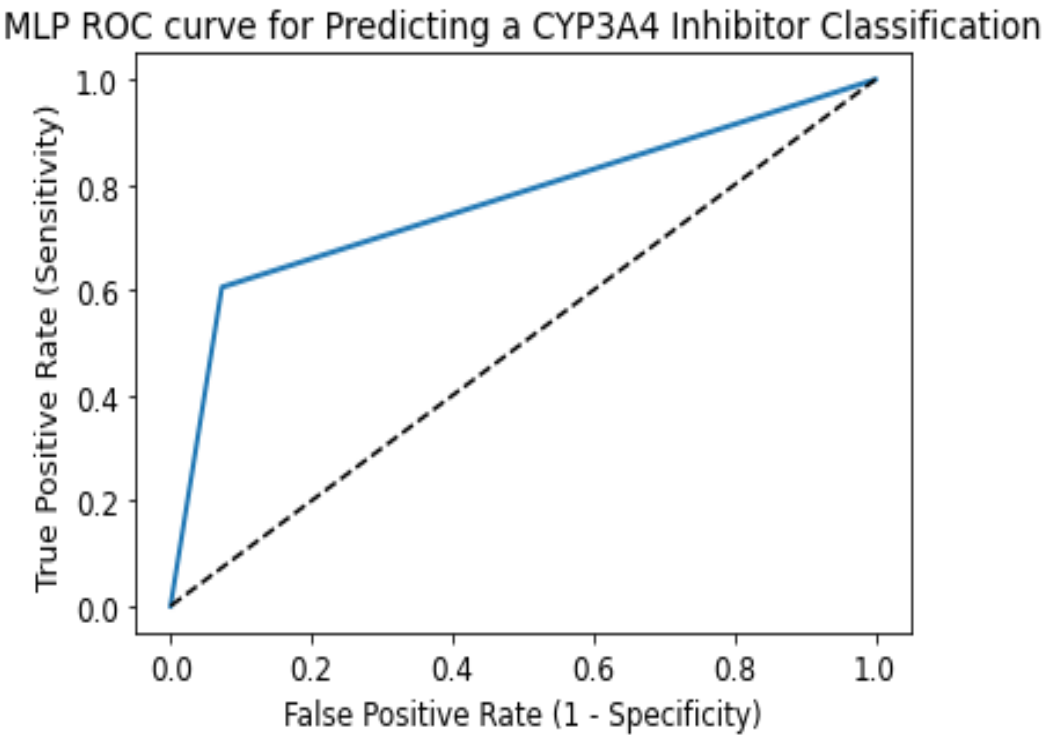

Figure 32. MLP ROC: 20 Descriptor Set

Table 17. Model Performance Evaluation for all Models

Cross Validated Machine Learning Results with 1079 Descriptors									
Sr	**Model**	**Accuracy**	**Specificity**	**Sensitivity**	**Precision**	**FP Rate**	**AUC**	**Classification Error**	**MCC**
1	SVM	0.848	0.718	0.895	0.897	0.282	0.805	0.153	0.612
2	MLP	0.950	0.725	0.863	0.918	0.275	0.757	0.167	0.550
3	Logistic Regression	0.818	0.663	0.876	0.876	0.337	0.769	0.182	0.538
4	Decision Tree	0.795	0.792	0.776	0.975	0.208	0.613	0.222	0.358
5	Random Forest	0.806	0.694	0.802	0.936	0.306	0.6726	0.214	0.399
Cross Validated Machine Learning Results 20 Important Descriptors									
Sr	**Model**	**Accuracy**	**Specificity**	**Sensitivity**	**Precision**	**FP Rate**	**AUC**	**Classification Error**	**MCC**
1	SVM	0.982	0.940	0.999	0.977	0.060	0.987	0.018	0.956
2	MLP	0.976	0.75	0.867	0.927	0.250	0.766	0.158	0.573
3	Logistic Regression	0.962	0.922	0.988	0.970	0.078	0.968	0.031	0.923
4	Decision Tree	0.792	0.557	0.782	0.902	0.442	0.592	0.254	0.281
5	Random Forest	0.833	0.858	0.826	0.972	0.142	0.703	0.166	0.540

CHAPTER 5

DISCUSSION

5: Discussion

CYP3A4 plays a vital role in the metabolism of majority of clinically available drugs. The large and flexible active site is accountable for the accommodation and metabolism of a wide range and number of substrates. Due to this, drug-drug interactions can occur that have a variety of adverse effects including decreased metabolism of a drug, ADRs, prolonged hospital stays and even mortality. Therefore, there is a need for screening protocols for new drug candidates against CYP3A4 to reduce the potential DDIs owing to inhibition, promising a higher chance of success of the drug candidate while also saving resources. Recently, Machine Learning techniques have been applied to solve drug design problems by predicating the biological activities of new chemical entities as well as predicting their tendencies to inhibit important biological compounds within the body[105]. In addition, machine learning models have been applied to extract features that are important in prediction[106]. The aim of this study was to build the best machine learning model on CYP3A4 datasets that would be able to classify and predict whether a new drug compound has the potential to inhibit CYP3A4, as well as to see which descriptors and structural properties of CYP3A4 inhibitors tend to lead to inhibition.

The Machine Learning models built for the purpose of CYP3A4 inhibition prediction in current literature are mostly built using experimental datasets[84,88,107]. This study made use of publicly available CYP3A4 inhibitor data from the ChEmbl and PubChem libraries for the purpose of CYP3A4 inhibitor prediction and classification. Studies that made use of similar datasets often make use of the PubChem AID 1851, which consists of inhibition data for multiple CYP isoforms instead of just CYP3A4[51,89]. These studies also build one type of machine learning model and aim to optimize the available model. This has its limitations as not every model works as well on every type of data[108], as seen with the lower MCC values in the Decision Tree, Random Forest, and Multilayer Perceptron models in this study, as a result of class imbalance. The advantage of using multiple machine learning models is that the model that can perform best according to the data is available and provides the best results, instead working around the limitations of the model. This study also has the advantage of using and comparing 2 refined datasets, built on a relatively large dataset. The first dataset consisted of 1179 refined descriptors while the second was even further refined on the basis of feature importance. Feature refining has proved to be essential in machine learning, especially during the training process to increase the quality of the data and make it easier for the model to determine features important for decision making, as well as identify patterns in

the data[109]. Thus, the high-performance measures seen in all the models in this study can be attributed towards the refined data. The application of hyperparameter optimization also ensures that the model performs as best as possible and gives it the groundwork to make the best possible decision.

Of the machine learning models built, the models built on the 20-descriptor dataset outperformed those built using the same classifiers but with all 1179 descriptors. This is because, while the 1179 descriptor dataset contains more information, not all this information is useful in classification, some may be useless, and some may in fact introduce discrimination into the model. The refined dataset contains only relevant information, making it easier for the model to identify key patterns for classification.

Within the models generated using the 20-descriptor dataset, the MLP model provided one of the best accuracies of 0.976. The MLP model also outperformed the other models when all descriptors were used. However, while the accuracies provided by these models were ideal, they tended to result in slightly less than ideal MCC and AUC values. This could be due to the relatively higher number of False Negative (FN) and False Positives (FP) results as seen above in the confusion matrixes in Figures 29 and 31. This increase in FNs and FP can be attributed to the class imbalance in the data, as there are nearly 4 times as many actives, or inhibitors, as compared to inactive, or noninhibitors. Studies have shown that class imbalance tends to deteriorate the MCC value in classification problems, and this has been seen in our study as well[110]. Similarly, other studies have also shown that deep neural networks tend to have high accuracies but also high FNs and FPs, leading to lower detection efficiency[111].

Resulting in only a slight increase in accuracy, SVMs provided the best accuracies (0.848 and 0.982 respectively), followed by the Logistic Regression Models (0.818 and 0.962). However, as opposed to the MLP models, the SVMs and Logistic Regression models provided high performance evaluation results all around, including for MCC and AUC. SVMs have been used in previous studies to predict CYP3A4 inhibition, often with reliable results as seen in Table 1. However, the overall performance is better in all regards for the SVM model generated using the 20-descriptor dataset in this study with the given hyperparameters. This means that this model is useful in identifying compounds with the potential of inhibiting the CYP3A4 enzyme and can be used early in the drug development process to screen new drug candidates for CYP3A4 inhibition.

On the other hand, Logistic Regression is not often used to predict CYP3A4 inhibition alone and is used in consensus with other machine learning models if at all. Even when used in consensus modeling, the logistic regression accuracies tend to be relatively low, with kappa values nearing 0.5 whereas kappa values of 0.7 and above are considered to show good agreement[112]. Our logistic regression model is also useful in telling us which descriptors are most relevant during classification (Table 8).

The decision tree models also provided us with vital information involving the relevance and importance of individual descriptors for classification. Similar to other studies, the model determined that the most important descriptor for classification was the H-048 or number of hydrogen bonds descriptor[113]. Our study also identified that the number of donor lipophilic atoms, rotatable bonds, maximum negative charge, and Vander Waals surface area for potentially pharmacophoric points are relevant features for classification. Overall, our decision tree also shows higher accuracy than decision trees used to predict CYP3A4 inhibition in literature and proved to be a reliable model[90].

According to already available literature, the presence of high overall negative charge in CYP3A4 inhibitors has been associated with more tightly bound ligands[114]. Increase in the amount of nitrogen, and hydrogen bonding has also been seen to be a determinant of CYP3A4 inhibition according to literature and is confirmed through the research performed in this study. Similarly, the study confirms what is known about lipophilicity and number of donor lipophilic atoms being a key characteristic in predicting inhibitor binding affinity and validates the presence of a high number of rotatable bonds being associated with high affinity CYP3A4 inhibitors, allowing the molecule to be more flexible and thus fit better within the binding pocket. The presence of a high number of aromatic moieties has been previously associated with higher affinity CYP3A4 inhibition as well and is also indicated as an important descriptor during this study[38]. This increase in affinity is due to the increase in pi-pi interactions within the binding pocket, which has previously been determined as a common interaction between CYP3A4 and its inhibitors[69]. On the other hand, while the study has decreed the importance of Vander Waals surface area in potentially pharmacophoric points, not much information is available on this in the literature, making this a potentially novel insight into CYP3A4 inhibitors.

In general, all the models were useful in identifying compounds with the potential to inhibit CYP3A4. The SVM model and the logistic regression model built on the 20-descriptor dataset had the best all round results indicating that these models can be used for screening projects early in drug development. This also indicates the usefulness and importance of the descriptors present in Table 3 in CYP3A4 inhibition prediction and classification.

CHAPTER 6

CONCLUSION

6: Conclusion

One of the major challenges of drug design and development is overcoming the 90% failure rate of drugs, of which pharmacokinetics and drug metabolism problems play a huge role. Of all the enzymes involved in drug metabolism, the CYP3A4 isoform of the Cytochrome P450 enzyme has a major contribution towards the metabolism of drugs, as it is responsible for the metabolism of nearly 50% of clinical drugs. The presence of the enzymes large and highly promiscuous binding site makes CYP3A4 prone to inhibition and induction by other molecules, leading to drug-drug interactions as the inhibition of CYP3A4 by one drug can lead to altered metabolism of another drug leading to toxicity, ADRs, and mortality. For these reasons CYP3A4 is a valid target for screening purposes early in the drug development process and this also demonstrates the vital importance for the pharmaceutical industry and professionals to identify the likely effect of a drug candidate in terms of interactions with any CYP isoform well before time to avoid the losses associated with drug developmental failures at the later stages.

This study aimed to build machine learning models to classify and predict CYP3A4 inhibitors and noninhibitors, to be used for the early predication of CYP3A4 inhibitors in new chemical entities and thus determine the success of a drug early on in the drug development process while also gaining insight into which features are commonly present in CYP3A4 inhibitors. This was done by generating multiple machine learning models including Logistic Regression Models, SVMs, Decision Trees, Random Forests, and Multilayer perceptron on two sets of descriptor data, and the best model was selected. The descriptor set was generated using Alvadesc and refined using the feature importance function in the Python Scikitlearn Library to select the 20 most important descriptors that made up the second descriptor set. Hyperparameter optimization was used to select the parameters that had the best performance for each model, and the accuracies were cross validated for more reliable results. Model performance was evaluated by calculating accuracy, sensitivity, specificity, precision, FP rate, AUC, and MCC. These metrices were used to select the best model. All the model accuracies fell within the range of 79.2%-98.2%.

While all the models performed well, the SVM and Logistic Regression models built on the 20-descriptor dataset performed the best with overall high results in all the evaluation metrices including accuracies of 98.2% and 96.2% respectively. This signifies that these models are the best

predictors of CYP3A4 inhibition and are valuable screening tools in the drug discovery process to accurately screen large datasets of CYP3A4 inhibitors and new chemical entities.

References

1. Drug discovery - Latest research and news | Nature. Accessed September 12, 2022. https://www.nature.com/subjects/drug-discovery

2. Song LG, Xie QX, Lao HL, Lv ZY. Human Coronaviruses and Therapeutic Drug Discovery. *Infect Dis Poverty*. 2021;10(1). doi:10.1186/s40249-021-00812-9

3. Antimicrobial Resistance. World Health Organization. Published 2021. Accessed September 13, 2022. https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance

4. Sun D, Gao W, Hu H, Zhou S. Why 90% of Clinical Drug Development Fails and How to Improve it? *Acta Pharm Sin B*. Published online July 1, 2022. doi:10.1016/j.apsb.2022.02.002

5. Tamimi NAM, Ellis P. Drug Development: From Concept to Marketing! *Nephron Clin Pract*. 2009;113(3). doi:10.1159/000232592

6. Hughes JP, Rees S, Kalindjian SB, Philpott KL, Philpott K, Building H. Principles of Early Drug Discovery. *Br J Pharmacol*. 2011;10:1476-5381. doi:10.1111/j.1476-5381.2010.01127.x

7. Smith C. Drug Target Validation: Hitting the Target. *Nature*. 2003;422(6929):342-345. doi:10.1038/422341a

8. Mohs RC, Greig NH. Drug Discovery and Development: Role of Basic Biological Research. *Alzheimer's and Dementia: Translational Research and Clinical Interventions*. 2017;3(4):651-657. doi:10.1016/j.trci.2017.10.005

9. Lipsky MS, Sharp LK. From Idea to Market: The Drug Approval Process. *Journal of the American Board of Family Medicine*. 2001;14:362-367. http://www.jabfm.org/

10. Norman GA van. Drugs, Devices, and the FDA: Part 1 An Overview of Approval Processes for Drugs. *J Am Coll Cardiol*. 2016;1(3).

11. Kraljevic S, Stambrook PJ, Pavelic K. Accelerating Drug Discovery. *EMBO Rep*. 2004;5(9):837-842.

12. Moein MM, el Beqqali A, Abdel-Rehim M. Bioanalytical Method Development and Validation: Critical Concepts and Strategies. *J Chromatogr B Analyt Technol Biomed Life Sci*. 2017;1043:3-11. doi:10.1016/j.jchromb.2016.09.028

13. Réda C, Kaufmann E, Delahaye-Duriez A. Machine Learning Applications in Drug Development. *Comput Struct Biotechnol J*. 2020;18:241-252. doi:10.1016/j.csbj.2019.12.006

14. Parasrampuria DA, Benet LZ, Sharma A. Why Drugs Fail in Late Stages of Development: Case Study Analyses from the Last Decade and Recommendations. *AAPS J*. 2018;20(46). doi:10.1208/s12248-018-0204-y

15. Kennedy T. Managing the Drug Discovery/Development Interface. *Drug Discov Today*. 1997;2(10):436-444. doi:10.1016/S1359-6446(97)01099-4

16. Alavijeh Pharmidex M, Palmer AM, Alavijeh MS. The Pivotal role of Drug Metabolism and Pharmacokinetics in the Discovery and Development of New Medicines. *Current Opinion in Investigational Drugs*. 2004;5(7). https://www.researchgate.net/publication/8378671

17. Palmer ME, Andrews LJ, Abbey TC, et al. The Importance of Pharmacokinetics and Pharmacodynamics in Antimicrobial Drug Development and their Influence on the Success of Agents Developed to Combat Resistant Gram Negative Pathogens: A Review. *Front Pharmacol*. 2022;13. doi:10.3389/fphar.2022.888079

18. Yengi LG, Leung L, Kao J. The Evolving Role of Drug Metabolism in Drug Discovery and Development. *Pharm Res*. 2007;24(5). doi:10.1007/s11095-006-9217-9

19. Mann B, Melton R, Thompson D. Drug Metabolism in Drug Discovery and Preclinical Development. *Acta Pharm Sin B*. 2018;8(5):721-732. www.intechopen.com

20. Abass KM, Turpeinen M, Rautio A, Abass K, Hakkola J, Pelkonen O. Metabolism of Pesticides by Human Cytochrome P450 Enzymes In Vitro: A Survey. doi:10.13140/2.1.3501.5689

21. Shanu-Wilson J, Evans L, Wrigley S, Steele J, Atherton J, Boer J. Biotransformation: Impact and Application of Metabolism in Drug Discovery. *ACS Med Chem Lett*. 2020;11(11):2087-2107. doi:10.1021/acsmedchemlett.0c00202

22. Kumar GN, Surapaneni S. Role of Drug Metabolism in Drug Discovery and Development. *Med Res Rev*. 2001;21(5).

23. Nelson SD. Perspective Metabolic Activation and Drug Toxicity. *J Med Chem*. 1982;25(7).

24. Wang YK, Li WQ, Xia S, Guo L, Miao Y, Zhang BK. Metabolic Activation of the Toxic Natural Products From Herbal and Dietary Supplements Leading to Toxicities. *Front Pharmacol*. 2021;12. doi:10.3389/fphar.2021.758468

25. Williams JA, Hyland R, Jones BC, et al. Drug-Drug Interactions for UDP-Glucuronosyltransferase Substrates: A Pharmacokinetic Explanation for Typically Observed Low Sxposure (AUC 1/AUC) Ratios. *Drug Metabolism and Disposition*. 2004;32(11):1201-1208. doi:10.1124/dmd.104.000794

26. Tanaka E. Clinically Important Pharmacokinetic Drug Drug Interactions Role of Cytochrome P450 Enzymes. *J Clin Pharm Ther*. 1998;23:403-416.

27. Zhang Z, Tang W. Drug Metabolism in Drug Discovery and Development. *Acta Pharm Sin B*. 2018;8(5):721-732. doi:10.1016/j.apsb.2018.04.003

28. Mclean KJ, Sabri M, Marshall KR, et al. *Biodiversity of Cytochrome P450 Redox Systems*. Vol 33.; 2005.

29. Sevrioukova IF, Poulos TL. Understanding the Mechanism of Cytochrome P450 3A4: Recent Advances and Remaining Problems. *Dalton Transactions*. 2013;42(9):3116-3126. doi:10.1039/c2dt31833d

30. de Groot MJ. Designing Better Drugs: Predicting Cytochrome P450 Metabolism. *Drug Discov Today*. 2006;11(13-14):601-606. doi:10.1016/j.drudis.2006.05.001

31. Nebert DW, Nelson DR, Coon MJ, et al. The P450 Superfamily: Update on New Sequences, Gene Mapping, and Recommended Nomenclature. *DNA Cell Biol*. 1991;10(1):1-14. www.liebertpub.com

32. Denisov IG, Makris TM, Sligar SG, Schlichting I. Structure and Chemistry of Cytochrome P450. *Chem Rev*. 2005;105(6):2253-2277. doi:10.1021/cr0307143

33. Rupasinghe S, Schuler MA, Kagawa N, et al. The Cytochrome P450 Gene Family CYP157 Does Not Contain EXXR in the K-helix Reducing the Absolute Conserved P450 Residues to a Single Cysteine. *FEBS*. 2006;580(27):6338-6342. doi:10.1016/j.febslet.2006.10.043

34. Gregg CR. Cytochrome P450. *Encyclopedia of Gastroenterology*. Published online 2004:542.

35. Inoue K, Inazawa J, Suzuki Y, et al. Fluorescence in Situ Hybridization Analysis of Chromosomal Localization of Three Human Cytochrome P450 2C Genes (CYP2C8, 2C9, and 2C10) at 10q24.1. *Jpn J Hum Genet*. 1994;39(3):337-343. doi:10.1007/BF01874052

36. Williams PA. Crystal Structures of Human Cytochrome P450 3A4 Bound to Metyrapone and Progesterone. *Science (1979)*. 2004;305(5684):683-686.

37. Kandel SE, Han LW, Mao Q, Lampe JN. Digging Deeper into CYP3A Testosterone Metabolism: Kinetic, Regioselectivity, and Stereoselectivity Differences between CYP3A4/5 and CYP3A7. *Drug Metabolism and Disposition*. 2017;45:1266-1275. doi:10.1124/dmd.117.078055

38. Beck TC, Beck KR, Morningstar J, Benjamin MM, Norris RA. Descriptors of Cytochrome Inhibitors and Useful Machine Learning Based Methods for the Design of Safer Drugs. *Pharmaceuticals*. 2021;14(5). doi:10.3390/ph14050472

39. Tornio A, Backman JT. Cytochrome P450 in Pharmacogenetics: An Update. In: *Advances in Pharmacology*. Vol 83. Academic Press Inc.; 2018:3-32. doi:10.1016/bs.apha.2018.04.007

40. Drug Metabolism - The Importance of Cytochrome P450 3A4. Accessed August 16, 2022. https://www.medsafe.govt.nz/profs/puarticles/march2014drugmetabolismcytochromep4503a4.htm

41. Zhou SF, Xue CC, Yu XQ, Li C, Wang G. Clinically Important Drug Interactions Potentially Involving Mechanism-based Inhibition of Cytochrome P450 3A4 and the Role of Therapeutic Drug Monitoring. *Therapeutic Drug Monitoring* . 2007;29(6):687-710.

42. Maher RL, Hanlon J, Hajjar ER. Clinical consequences of polypharmacy in elderly. *Expert Opin Drug Saf*. 2014;13(1):57-65. doi:10.1517/14740338.2013.827660

43. Hajar E, Hajjar ER, Hanlon JT, et al. Adverse Drug Reaction Risk Factors in Older Outpatients. *Am J Geriatr Pharmacother*. 2003;1(2):82-89.

44. Kennedy C, Brewer L, Williams D. Drug interactions. *Medicine*. 2020;48(7):450-455. http://www.factsandcomparisons.com

45. Marcath LA, Coe TD, Hoylman EK, Redman BG, Hertz DL. Prevalence of drug-drug interactions in oncology patients enrolled on National Clinical Trials Network oncology clinical trials. *BMC Cancer*. 2018;18(1). doi:10.1186/s12885-018-5076-0

46. Hakkola J, Hukkanen J, Turpeinen M, Pelkonen O. Inhibition and Induction of CYP Enzymes in Humans: an Update. *Arch Toxicol*. 2020;94:3671-3722.

47. Zakrzewski-Jakubiak H, Doan J, Lamoureux P, Singh D, Turgeon J, Tannenbaum C. Detection and prevention of drug-drug interactions in the hospitalized elderly: Utility of new cytochrome P450-based software. *American Journal Geriatric Pharmacotherapy*. 2011;9(6):461-470. doi:10.1016/j.amjopharm.2011.09.006

48. He K, Iyer KR, Hayes RN, Sinz MW, Woolf TF, Hollenberg PF. Articles Inactivation of Cytochrome P450 3A4 by Bergamottin, a Component of Grapefruit Juice. *Chem Res Toxicol*. 1998;11:252-259. https://pubs.acs.org/sharingguidelines

49. Krayenbuhl JC, Vozeh S, Kondo-Oestrieicher M, Dayer P. Drug Drug Interactions of New Active Substances Mibefradil Example. *Eur J Clin Pharmacol*. 1999;55:559-565.

50. Dresser GK, Spence JD, Bailey DG. Pharmacokinetic-Pharmacodynamic Consequences and Clinical Relevance of Cytochrome P450 3A4 Inhibition. *Clin Pharmacokinet*. 2000;38(1):41-57.

51. Plonka W, Stork C, Šícho M, Kirchmair J. CYPlebrity: Machine Learning Models for the Prediction of Inhibitors of Cytochrome P450 Enzymes. *Bioorg Med Chem*. 2021;46. doi:10.1016/j.bmc.2021.116388

52. Meng XY, Zhang HX, Mezei M, Cui M. Molecular Docking: A Powerful Approach for Structure-Based Drug Discovery. *Current Computer-Aided Drug Design* . 2011;7(2):146-157.

53. Maréchal JD, Yu J, Brown S, et al. In Silico and in Vitro Screening for Inhibition of Cytochrome P450 CYP3A4 by Comedications Commonly used by Patients with Cancer. *Drug Metabolism and Disposition*. 2006;34(4):534-538. doi:10.1124/dmd.105.007625

54. Panneerselvam S, Yesudhas D, Durai P, Ayaz Anwar M, Gosu V, Choi S. A Combined Molecular Docking/Dynamics Approach to Probe the Binding Mode of Cancer Drugs with Cytochrome P450 3A4. *Molecules*. 2015;20:14915-14935. doi:10.3390/molecules200814915

55. Bren U, Fuchs JE, Oostenbrink C. Cooperative Binding of Aflatoxin B1 by Cytochrome P450 3A4: A Computational Study. *Chem Res Toxicol*. 2014;27(12):2136-2147. doi:10.1021/tx5004062

56. Subhani S, Jamil K. Molecular Docking of Chemotherapeutic Agents to CYP3A4 in Non-Small Cell Lung Cancer. *Biomedicine and Pharmacotherapy*. 2015;73:65-74. doi:10.1016/j.biopha.2015.05.018

57. Zhou X, Wang Y, Hu T, et al. Enzyme Kinetic and Molecular Docking Studies for the Inhibitions of Miltirone on Major Human Cytochrome P450 Isozymes. *Phytomedicine*. 2013;20(3-4):367-374. doi:10.1016/j.phymed.2012.09.021

58. Ashour M. Inhibition of Cytochrome P450 (CYP3A4) Activity by Extracts from 57 Plants Used in Traditional Chinese Medicine (TCM). *Pharmacognosy Magazine*. 2017;13(50):300-308. doi:10.4103/0973-1296.204561

59. Hollingsworth SA, Dror RO. Molecular Dynamics Simulation for All. *Neuron*. 2018;99(6):1129-1143. doi:10.1016/j.neuron.2018.08.011

60. Park H, Lee S, Suh J. Structural and Dynamical Basis of Broad Substrate Specificity, Catalytic Mechanism, and Inhibition of Cytochrome P450 3A4. *J Am Chem Soc*. 2005;127(39):13634-13642. doi:10.1021/ja053809q

61. Han M, Qian J, Ye Z, et al. Functional Assessment of the Effects of CYP3A4 Variants on Acalabrutinib Metabolism in Vitro. *Chem Biol Interact*. 2021;345. doi:10.1016/j.cbi.2021.109559

62. Bren U, Oostenbrink C. Cytochrome P450 3A4 Inhibition by Ketoconazole: Tackling the Problem of Ligand Cooperativity Using Molecular Dynamics Simulations and Free-Energy Calculations. *J Chem Inf Model*. 2012;52(6):1573-1582. doi:10.1021/ci300118x

63. Teixeira VH, Ribeiro Vera V, Martel PJ. Analysis of Binding Modes of Ligands to Multiple Conformations of CYP3A4. *Biochim Biophys Acta Proteins Proteom*. 2010;1804(10):2036-2045. doi:10.1016/j.bbapap.2010.06.008

64. Kiani YS, Ranaghan KE, Jabeen I, Mulholland AJ. Molecular Dynamics Simulation Framework to Probe the Binding Hypothesis of CYP3A4 Inhibitors. *Int J Mol Sci*. 2019;20(18). doi:10.3390/ijms20184468

65. Verma J, Khedkar VM, Coutinho EC. 3D-QSAR in Drug Design-A Review. *Curr Top Med Chem*. 2010;10:95-115.

66. Didziapetris R, Dapkunas J, Sazonovas A, Japertas P. Trainable Structure-Activity Relationship Model for Virtual Screening of CYP3A4 Inhibition. *J Comput Aided Mol Des*. 2010;24(11):891-906. doi:10.1007/s10822-010-9381-1

67. Roy K, Pratim Roy P. Comparative Chemometric Modeling of Cytochrome 3A4 Inhibitory Activity of Structurally Diverse Compounds using Stepwise MLR, FA-MLR, PLS, GFA, G/PLS and ANN techniques. *Eur J Med Chem*. 2009;44(7):2913-2922. doi:10.1016/j.ejmech.2008.12.004

68. Riley RJ, Parker AJ, Trigg S, Manners CN. Development of a Generalized, Quantitative Physicochemical Model of CYP3A4 Inhibition for Use in Early Drug Discovery. *Pharm Res*. 2001;18(5).

69. Lewis DFV, Lake BG, Dickins M. Quantitative Structure-Activity Relationships (QSARs) in CYP3A4 Inhibitors: The Importance of Lipophilic Character and Hydrogen Bonding. *J Enzyme Inhib Med Chem*. 2006;21(2):127-132. doi:10.1080/14756360500532747

70. Guttman Y, Kerem Z. Computer-Aided (In Silico) Modeling of Cytochrome P450-Mediated Food-Drug Interactions (FDI). *Int J Mol Sci*. 2022;23(15). doi:10.3390/ijms23158498

71. Kaur P, Chamberlin AR, Poulos TL, Sevrioukova IF. Structure-Based Inhibitor Design for Evaluation of a CYP3A4 Pharmacophore Model. *J Med Chem*. 2016;59(9):4210-4220. doi:10.1021/acs.jmedchem.5b01146

72. Schuster D, Laggner C, Steindl TM, Langer T. Development and Validation of an In Silico P450 Profiler Based on Pharmacophore Models. *Curr Drug Discov Technol*. 2006;3:1-48.

73. Ekins S, Stresser DM, Williams JA. In Vitro and Pharmacophore Insights into CYP3A Enzymes. *Trends Pharmacol Sci*. 2003;24(4):161-166. doi:10.1016/S0165-6147(03)00049-X

74. Tyzack JD, Kirchmair J. Computational Methods and Tools to Predict Cytochrome P450 Metabolism for Drug Discovery. *Chem Biol Drug Des*. 2019;93(4):377-386.

75. Hammann F, Gutmann H, Baumann U, Helma C, Drewe J. Classification of cytochrome P450 activities using machine learning methods. *Mol Pharm*. 2009;6(6):1920-1926. doi:10.1021/mp900217x

76. Kiani YS, Jabeen I. Exploring the chemical space of cytochrome P450 inhibitors using integrated physicochemical parameters, drug efficiency metrics and decision tree models. *Computation*. 2019;7(2). doi:10.3390/computation7020026

77. Hu B, Zhou X, Mohutsky MA, Desai VP. Property Relationships and Machine Learning Models for Addressing CYP3A4 Mediated Victim Drug Drug Interaction Risk in Drug Discovery. *Mol Pharm*. 2020;17:3600-3608.

78. Cortes C, Vapnik V, Saitta L. Support-Vector Networks. *Mach Learn*. 1995;20:273-297.

79. Quinlan JR. Learning Decision Tree Classifiers. *ACM Comput Surv*. 1996;28(1):71-72.

80. Olier I, Sadawi N, Bickerton GR, et al. Meta-QSAR: a Large-Scale Application of Meta-Learning to Drug Design and Discovery. *Mach Learn*. 2018;107(1):285-311. doi:10.1007/s10994-017-5685-x

81. Zhang S, Li X, Zong M, Zhu X, Cheng D. Learning k for kNN Classification. *ACM Trans Intell Syst Technol*. 2017;8(3). doi:10.1145/2990508

82. Sperandei S. Understanding Logistic Regression Analysis. *Biochem Med (Zagreb)*. 2014;24(1):12-18. doi:10.11613/BM.2014.003

83. Yap CW, Chen YZ. Prediction of Cytochrome P450 3A4, 2D6, and 2C9 Inhibitors and Substrates by Using Support Vector Machines. *J Chem Inf Model*. 2005;45(4):982-992. doi:10.1021/ci0500536

84. Kriegl JM, Arnhold T, Beck B, Fox T. Prediction of human cytochrome P450 inhibition using support vector machines. In: *QSAR and Combinatorial Science*. Vol 24. ; 2005:491-502. doi:10.1002/qsar.200430925

85. Sun H, Veith H, Xia M, Austin CP, Huang R. Predictive Models for Cytochrome P450 Isozymes Based on Quantitative High Throughput Screening Data. *J Chem Inf Model*. 2011;51(10):2474-2481. doi:10.1021/ci200311w

86. Pang X, Zhang B, Mu G, et al. Screening of Cytochrome P450 3A4 Inhibitors via In Silico and In Vitro Approaches. *Royal Society of Chemistry Advances*. 2018;8. doi:10.1039/c8ra06311g

87. Hammann F, Gutmann H, Baumann U, Helma C, Drewe J. Classification of Cytochrome P 450 Activities Using Machine Learning Methods. *Mol Pharm*. 2009;6(6):1920-1926. doi:10.1021/mp900217x

88. Ekins S, Berbaum J, Harrison RK. Generation and Validation of Rapid Computational Filters for CYP2D6 and CYP3A4. *The American Society for Pharmacology and Experimental Therapeutics*. 2003;31(9):1077-1080.

89. Su BH, Tu YS, Lin C, Shao CY, Lin OA, Tseng YJ. Rule-Based Prediction Models of Cytochrome P450 Inhibition. *J Chem Inf Model*. 2015;55(7):1426-1434. doi:10.1021/acs.jcim.5b00130

90. Choi I, Kim SY, Kim H, et al. Classification Models for CYP450 3A4 Inhibitors and Non-Inhibitors. *Eur J Med Chem*. 2009;44(6):2354-2360. doi:10.1016/j.ejmech.2008.08.013

91. Lee JH, Basith S, Cui M, Kim B, Choi S. In Silico Prediction of Multiple-Category Classification Model for Cytochrome P450 Inhibitors and Non-Inhibitors Using Machine-Learning Method. *SAR QSAR Environ Res*. 2017;28(10):863-874. doi:10.1080/1062936X.2017.1399925

92. Wu Z, Lei T, Shen C, Wang Z, Cao D, Hou T. Reliable Prediction of Human Cytochrome P450 Inhibition Using Artificial Intelligence Approaches. *J Chem Inf Model*. 2019;59(11):4587-4601. doi:10.1021/acs.jcim.9b00801

93. Grebner C, Matter H, Kofink D, Wenzel J, Schmidt F, Hessler G. Application of Deep Neural Network Models in Drug Discovery Programs. *ChemMedChem*. 2021;16:3771-3786. doi:10.1002/cmdc.202100418

94. Li X, Xu Y, Lai L, Pei J. Prediction of Human Cytochrome P450 Inhibition Using a Multitask Deep Autoencoder Neural Network. *Mol Pharm*. 2018;15:4336-4345. doi:10.1021/acs.molpharmaceut.8b00110

95. Mauri A. AlvaDesc: A Tool to Calculate and Analyze Molecular Descriptors and Fingerprints. *Methods in Pharmacology and Toxicology*. Published online 2020:801-820. doi:10.1007/978-1-0716-0150-1_32

96. Sitthiyot T, Holasut K. A Simple Method for Measuring Inequality. *Palgrave Commun*. 2020;6(1). doi:10.1057/s41599-020-0484-6

97. 1.10. Decision Trees — scikit-learn 1.1.3 documentation. Accessed October 31, 2022. https://scikit-learn.org/stable/modules/tree.html

98. Smola AJ, Schölkopf B. A Tutorial on Support Vector Regression. *Stat Comput*. 2004;14(3):199-222. doi:10.1023/B:STCO.0000035301.49549.88

99. sklearn.neural_network.MLPClassifier — scikit-learn 1.1.3 documentation. Accessed October 31, 2022. https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html

100. Khan HA, Jabeen I. Combined Machine Learning and GRID-Independent Molecular Descriptor (GRIND) Models to Probe the Activity Profiles of 5-Lipoxygenase Activating Protein Inhibitors. *Front Pharmacol*. 2022;13. doi:10.3389/fphar.2022.825741

101. de Diego IM, Redondo AR, Fernández RR, Navarro J, Moguerza JM. General Performance Score for Classification Problems. *Applied Intelligence*. 2022;52(10):12049-12063. doi:10.1007/s10489-021-03041-7

102. Jung Y, Hu J. A K-fold Averaging Cross-validation Procedure. *Journal of Nonparametric Statistics*. 2015;27(2):167-179. doi:10.1080/10485252.2015.1010532

103. Nti IK, Nyarko-Boateng O, Aning J. Performance of Machine Learning Algorithms with Different K Values in K-fold CrossValidation. *International Journal of Information Technology and Computer Science*. 2021;13(6):61-71. doi:10.5815/ijitcs.2021.06.05

104. Oh S. Predictive Case-Based Feature Importance and Interaction. *Inf Sci (N Y)*. 2022;593:155-176. doi:10.1016/j.ins.2022.02.003

105. Stephenson N, Shane E, Chase J, et al. Survey of Machine Learning Techniques in Drug Discovery. *Curr Drug Metab*. 2018;20(3):185-193. doi:10.2174/1389200219666180820112457

106. Hakak S, Alazab M, Khan S, Gadekallu TR, Maddikunta PKR, Khan WZ. An Ensemble Machine Learning Approach Through Effective Feature Extraction to Classify Fake News. *Future Generation Computer Systems*. 2021;117:47-58. doi:10.1016/j.future.2020.11.022

107. Choi I, Kim SY, Kim H, et al. Classification Models for CYP450 3A4 Inhibitors and Non-inhibitors. *Eur J Med Chem*. 2009;44(6):2354-2360. doi:10.1016/j.ejmech.2008.08.013

108. Ali A, Ralescu A, Shamsuddin SM, Ralescu AL. Classification with Class Imbalance Problem: A Review. *Classification Int J Advance Soft Compu Appl*. 2013;5(3). https://www.researchgate.net/publication/288228469

109. Why Data Refinement Matters More Than Model Complexity | SpringML, Inc. Accessed October 31, 2022. https://www.springml.com/blog/why-data-refinement-matters-more-than-model-complexity/

110. Zhu Q. On the Performance of Matthews Correlation Coefficient (MCC) for Imbalanced Dataset. *Pattern Recognit Lett*. 2020;136:71-80. doi:10.1016/J.PATREC.2020.03.030

111. Mijalkovic J, Spognardi A. Reducing the False Negative Rate in Deep Learning Based Network Intrusion Detection Systems. *Algorithms*. 2022;15(8):258. doi:10.3390/a15080258

112. Arimoto R, Prasad MA, Gifford EM. Development of CYP3A4 Inhibition Models: Comparisons of Machine-Learning Techniques and Molecular Descriptors. *J Biomol Screen*. 2005;10(3):197-205. doi:10.1177/1087057104274091

113. Jones DR, Ekins S, Li L, Hall SD. Computational Approaches that Predict Metabolic Intermediate Complex Formation with CYP3A4 (+b5). *Drug Metabolism and Disposition*. 2007;35(9):1466-1475. doi:10.1124/dmd.106.014613

114. Samuels ER, Sevrioukova I. Inhibition of Human CYP3A4 by Rationally Designed Ritonavir-Like Compounds: Impact and Interplay of the Side Group Functionalities. *Mol Pharm*. 2018;15(1):279-288. doi:10.1021/acs.molpharmaceut.7b00957