CHALMERS
UNIVERSITY OF TECHNOLOGY

UNIVERSITY OF GOTHENBURG

Predictive AI for Hepatic Safety

A dual analysis of CYP450 time-dependent inhibition and trapping assays using supervised learning models

Master's thesis in Computer science and engineering

MARIA VIRGINIA RAMOS MARCA

Department of Computer Science and Engineering
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg, Sweden 2025

Master's thesis 2025

Predictive AI for Hepatic Safety

A dual analysis of CYP450 time-dependent inhibition and trapping assays using supervised learning models

MARIA VIRGINIA RAMOS MARCA

UNIVERSITY OF
GOTHENBURG

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Computer Science and Engineering
Chalmers University of Technology
University of Gothenburg
Gothenburg, Sweden 2025

Predictive AI for Hepatic Safety
A dual analysis of CYP450 time-dependent inhibition and trapping assays using supervised learning models
MARIA VIRGINIA RAMOS MARCA

Supervisor: Vigneshwari Subramanian, AstraZeneca
Examiner: Ola Engkvist, Department of Computer Science and Engineering

Master's Thesis 2025
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Typeset in LaTeX
Gothenburg, Sweden 2025

Predictive AI for Hepatic Safety
A dual analysis of CYP450 time-dependent inhibition and trapping assays using supervised learning models
MARIA VIRGINIA RAMOS MARCA
Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg

Abstract

This work explores the development and evaluation of machine learning models for predicting toxicity-related endpoints, focusing on time-dependent inhibition of cytochrome P450 enzymes and reactivity in trapping assays (glutathione, potassium cyanide, and methoxylamine). A variety of modeling strategies were assessed, including decision trees and Chemprop neural networks in both single-task and multitask configurations. Model performances were estimated using temporally split datasets to better reflect real-world prediction scenarios. While tree-based models consistently delivered more stable and balanced results, Chemprop models showed greater sensitivity to class imbalance, data partitioning, and representation. Attempts to mitigate these issues using data resampling techniques, additional molecular descriptors, and scaffold-based data reduction led to limited improvements. Further analysis of feature distributions and chemical space connectivity highlighted key challenges, such as weak class separation in descriptor values and structural isolation of test compounds, especially under temporal splits. In the case of trapping assays, multitask learning failed to improve generalization, likely due to the biological heterogeneity of the endpoints. Overall, results emphasize that data limitations are the primary bottleneck. Enhancing chemical diversity, improving feature representations, and tailoring models to specific endpoint properties appear critical for achieving more robust predictions in toxicity modeling.

Keywords: CYP450 inhibition, Trapping Assays, Machine Learning, Decision Trees, Chemprop, Toxicology Prediction, Imbalanced Data.

Acknowledgements

I would like to thank my supervisor, Vigneshwari Subramanian, for her continuous support throughout this journey. She has not only guided me academically, but has also been a personal support, offering advice, encouragement, and perspective whenever I needed it. I am sincerely grateful for her help in guiding me to this point in my path. She has been more than just a supervisor, she is someone who now has a place in my life.

I would also like to thank Sara Amberntsson, Bhavik Chouhan, Anna-Pia Palmgren, and Ulrik Jurva for always being approachable and willing to help. Their readiness to answer questions and provide support whenever needed made a big difference during this journey.

Finally, I want to thank all the people around me who, even without fully understanding what I was doing, were always there to listen, to help, and to offer their unconditional support. To my family, for always caring, hoping for the best, and giving me more than I could ever ask for. To my friends who are far away but still manage to feel close, thank you for your messages, calls and for always being present in your own way. And to those who are here with me, sharing my everyday life, making each day easier and helping me feel at home, with special thanks to Sophie You. I am truly grateful to have you in my life.

Virginia Ramos, Gothenburg, June 2025

List of Acronyms

Below is the list of acronyms that have been used throughout this thesis, listed in alphabetical order:

AI	Artificial Intelligence
CYP450	Cytochrome P450 enzymes
DDI	Drug-Drug Interaction
DILI	Drug-Induced Liver Injury
ECFP	Extended Connectivity Fingerprints
FN	False Negative
FP	False Positive
GSH	Glutathione trapping assay
KCN	Cyanide trapping assay
MA	Methoxylamine trapping assay
MCC	Matthews Correlation Coefficient
ML	Machine Learning
MPNN	Message Passing Neural Network
RF	Random Forest
RUS	Random Under-Sampling
SMILES	Simplified Molecular Input Line Entry System
SMOTE	Synthetic Minority Over-sampling Technique
TDI	Time-Dependent Inhibition
TN	True Negative
TP	True Positive
XGBOOST	Extreme Gradient Boosting

Contents

List of Figures

List of Tables

1

Introduction

Drug-Induced Liver Injury (DILI) refers to the pathological alterations in liver function caused by exposure to xenobiotic compounds, including pharmaceutical drugs. It arises when the body's normal metabolic processing of these compounds is disrupted, leading to adverse outcomes such as oxidative stress, inflammation, mitochondrial damage, and cellular apoptosis or necrosis [1]. DILI represents a major obstacle in drug development and is one of the leading causes of late-stage failure in clinical pipelines.

One of the primary drivers of DILI is the formation of reactive metabolites, which result from biotransformation processes intended to detoxify xenobiotics. While these metabolic reactions typically serve to facilitate excretion, certain compounds have the potential to undergo bioactivation and generate electrophilic intermediates that can covalently bind to essential cellular macromolecules such as proteins, lipids, or DNA [2]. These interactions can trigger toxic responses and organ dysfunction.

Reactive metabolites can arise from both Phase I and Phase II reactions, and their classification often follows their electrophilic nature. For example, soft electrophiles like quinones or Michael acceptors tend to react with nucleophilic centers such as cysteine residues, while harder electrophiles, such as imines or aldehydes, may interact with DNA or other hard nucleophiles [3].

To better understand the metabolic context in which reactive metabolites are generated, it is useful to examine the canonical pathways of hepatic drug processing. Figure 1.1 illustrates these pathways, beginning with enzymatic oxidation of the administered drug via cytochrome P450 enzymes (P450). From there, the compound may follow different metabolic routes: it can undergo conjugation reactions that render it more water-soluble and suitable for excretion, or it may form reactive intermediates. These reactive species can be further detoxified through trapping mechanisms or, if not neutralized, may contribute to cellular damage and hepatotoxicity.

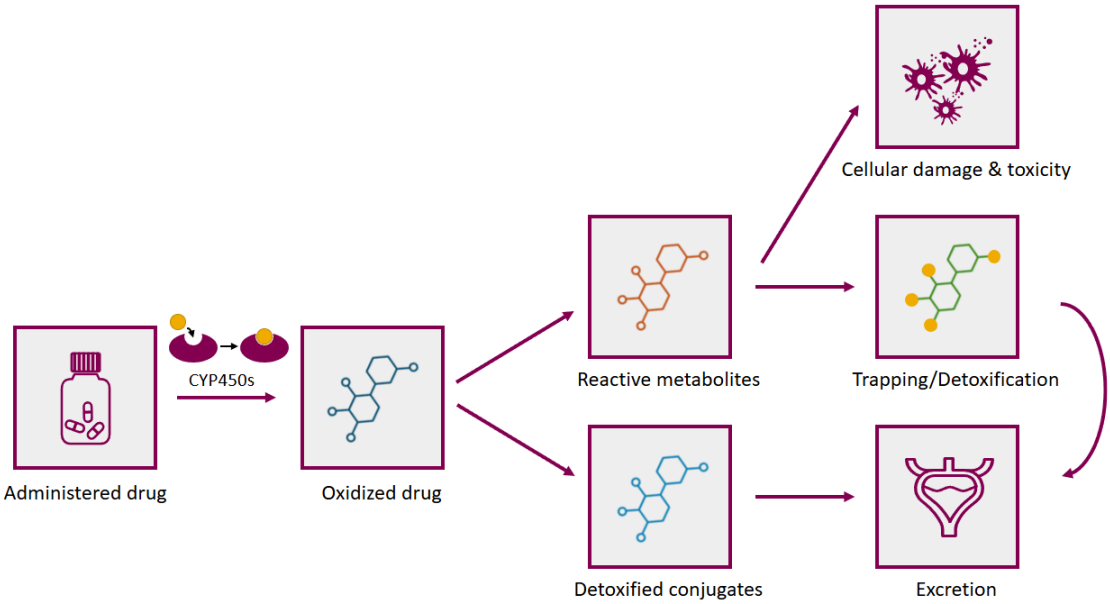


Figure 1.1: General pathways of drug metabolism by liver, including oxidation by CYP450 enzymes, the formation of reactive intermediates, detoxification routes, and possible excretion pathways.

Given the critical impact of liver toxicity on pharmaceutical attrition rates, early identification of hepatotoxic potential is essential to guide compound optimization and reduce downstream risks. Several experimental assays and predictive markers have been developed to support this goal, including studies on CYP450 enzyme inhibition, trapping assays for reactive metabolite detection, and complementary indicators such as BSEP inhibition, mitochondrial toxicity, THP-1 cytotoxicity, and spheroid-based assays.

1.1 Aim

This project aims to develop and benchmark multitask Artificial Intelligence (AI) models to predict parameters that influence liver toxicity of pharmaceutical compounds. Having the potential to predict these parameters at the point of compound design is expected to enhance the success rates of compound progression and thereby save significant time and resources, which would otherwise be invested on the wrong compounds. The project will explore classical and deep learning machine learning classification approaches to generate predictive models. These models will focus on two critical endpoints related to drug-induced liver toxicity: CYP450 enzyme interactions and trapping assays, including glutathione (GSH), potassium cyanide (KCN), and methoxylamine (MA) assays. In particular, the models will consider the formation of reactive metabolites that can interact with essential cellular macromolecules, leading to liver damage. The goal is to identify potential risk earlier in the process by using these predictive models, enhancing drug development efficiency and reducing the likelihood of late-stage failures.

1.2 Challenges

The development of predictive models for drug metabolism and toxicity involves a number of challenges that influence the design, implementation, and interpretation of computational experiments. These challenges are both data-related and methodological in nature, and they shaped the modeling strategies adopted in this thesis. Addressing them is essential to ensure the reliability and applicability of the results.

- **Limited data availability.** Although this study relies on high-quality internal experimental datasets, the total number of available data points (especially for certain isoforms or assay types) is limited. This data scarcity can impact the robustness of model training and the generalizability of the results.
- **Imbalanced class distribution.** Both CYP450 inhibition and trapping assay datasets are highly imbalanced, with a much smaller number of positive cases compared to negatives. This imbalance affects model training by biasing predictions toward the majority class and complicates the evaluation of real-world applicability.
- **Biological complexity of endpoints.** The mechanisms behind enzyme inhibition and bioactivation of metabolites are biologically complex, often involving subtle structural variations and multiple interacting factors. Capturing these processes from the structure alone presents a significant modeling challenge.
- **Inconsistent performance across tasks.** Preliminary results revealed that some endpoints are considerably more difficult to predict than others, despite using advanced algorithms and representations. This variability highlights the limitations of conventional modeling approaches when applied to heterogeneous or noisy data.
- **Integration of multiple tasks.** Predicting multiple endpoints simultaneously introduces architectural and optimization challenges. Sharing representations across tasks can lead to conflicts when tasks are not strongly correlated, requiring careful design of multitask strategies to avoid performance degradation.

These challenges underscore the need for robust and flexible modeling approaches. In particular, they motivate the use of multitask learning to exploit potential shared patterns among endpoints, as well as the inclusion of various molecular representations to capture complementary aspects of chemical information.

1.3 Thesis Outline

In order to guide the reader through the different components of this research, the thesis is organized into six chapters, each addressing a specific aspect of the work. The structure is designed to move progressively from context and theoretical foundations to methodology, results, and conclusions.

- **Chapter 1 Introduction**
  Provides an overview of the motivation behind the study, its scientific and practical relevance, and the main objectives pursued. It also outlines the scope and limitations of the project, and introduces the methodological framework.
- **Chapter 2 Related Works**
  Reviews previous research efforts related to the prediction of drug metabolism and toxicity using computational approaches. Special attention is given to machine learning applications in CYP450 inhibition modeling, as well as the evolution from single-task to multitask learning paradigms.
- **Chapter 3 Theoretical Background**
  Presents the theoretical foundations supporting this work. It is divided into two main sections:
  Biological Background, which introduces the role of cytochrome P450 enzymes in drug metabolism, discusses the formation of reactive metabolites, and describes experimental methods such as time-dependent inhibition (TDI) and trapping assays.
  Computational Background, which explains molecular representations (e.g., SMILES, descriptors, fingerprints), describes machine learning algorithms used (including classical models and deep learning architectures), and defines the evaluation metrics applied throughout the study.
- **Chapter 4 Methods**
  Details the datasets used, preprocessing steps, feature extraction, model architectures, training strategies, and evaluation procedures. It also describes how single-task and multitask settings were implemented and compared.
- **Chapter 5 Results and Discussion**
  Presents the experimental results obtained and analyzes them in light of the research objectives. This chapter includes performance comparisons between different models, the effect of multitask learning, and the interpretation of relevant molecular features.
- **Chapter 6 Conclusions**
  Summarizes the key findings of the study and discusses their implications in the context of toxicity prediction.
- **Chapter 7 Future Work**
  Proposes directions for further research, including improvements in data integration, expansion to broader CYP450 inhibition endpoints, and validation using external datasets.

2

Related Works

The use of machine learning in drug development has grown rapidly in the last decade, especially in the prediction of pharmacokinetic and toxicological parameters. Many studies have used molecular descriptors, fingerprints, and graphical representations to model the interaction between chemical structure and biological activity [4]–[6].

In the study by Mayr et al. [7], a large-scale evaluation of classical machine learning methods for molecular bioactivity prediction was conducted. The authors assessed models such as Random Forest (RF) and Support Vector Machines, trained from crafted features derived from chemical structures. These approaches showed competitive performance in several binary classification tasks. However, their predictive power was highly dependent on the relevance of the input descriptors, and they required training separate models for each endpoint, which limited their ability to capture information shared between related tasks.

Recent studies have evaluated the advantages of deep learning architectures based on learned molecular representations over traditional approaches using fixed descriptors. In a comprehensive comparative study, Yang et al. compared graph-based models, including directed message passing neural network, against classical models such as random forests and feed-forward neural networks on 19 public and 16 proprietary datasets. The results showed that the directed message passing neural network model consistently equaled or outperformed the benchmark methods on most public datasets and outperformed all benchmark methods on all private datasets. By directly learning molecular structure representations without the need for predefined features, these results demonstrate that graph neural networks offer a robust and scalable alternative for property prediction tasks, making them suitable for real-world applications in drug discovery.

Another key methodological evolution is the shift from single-task to multitask learning frameworks. In single-task learning, a model is trained independently for each endpoint, which can be effective when sufficient data is available. However, this approach does not exploit potential correlations between related prediction tasks and may lead to suboptimal generalization. In molecular property prediction, single-task models have traditionally been used, but often show large gaps between training and test performance, suggesting a tendency to overfit [8].

In contrast, multitask learning enables a shared representation to be learned across

multiple outputs, improving performance, especially in data-scarce or imbalanced scenarios. Ramsundar et al. [9] demonstrated the effectiveness of massively multitask neural networks for virtual screening, showing that performance improves as both the number of tasks and the amount of training data increase. Although some limitations were observed in generalizing to unseen tasks, the study highlighted multitask learning as a promising strategy to leverage shared patterns across datasets. The authors also emphasized the need for standardized benchmarks and broader data sharing to advance the field. It has also been shown to reduce overfitting and improve generalization by capturing shared chemical and biological patterns among tasks [7].

However, some studies have pointed out that multitask learning may introduce trade-offs between individual task performance and overall model performance. Moon and Kim [10] proposed a method that mitigates this limitation by incorporating group selection and knowledge distillation within a multitasking framework. Their results showed that this approach outperformed both single-task and traditional multitask models. Interestingly, they found that tasks with initially lower performance under single-task learning benefited the most from multitask training, while knowledge distillation helped recover performance in tasks negatively affected by multitask integration.

Despite the progress achieved in recent years, many studies still limit their scope to single-task models or focus exclusively on a single class of algorithms. In this work, a hybrid strategy is adopted, combining classical machine learning approaches such as RF and XGBoost with deep learning architectures like Chemprop, within a multitask learning framework. This design enables the evaluation of how algorithmic choices influence predictive performance across distinct tasks, as well as the extent to which shared learning improves generalization, particularly in endpoints related to CYP450 inhibition and reactive metabolite formation. By integrating multiple modeling paradigms and targeting biologically relevant endpoints through multitask learning, this study contributes a novel and comprehensive perspective to the field, extending beyond the boundaries of previous single-focus approaches.

3

Theorical Background

An understanding of the theoretical foundations underlying both the biological and computational aspects of this work is key to interpreting the reasoning and methodology applied. This chapter is divided into two main sections: one focused on the biological mechanisms involved in drug metabolism and toxicity, and the other dedicated to the computational techniques employed for modeling and analysis.

3.1 Biological Background

The biotransformation of xenobiotics, especially therapeutic drugs, is a critical process that influences both pharmacological efficacy and potential toxicity. Drug metabolism occurs primarily in the liver and involves a highly complex network of enzymes, transporters, and cofactors that work together to modify the chemical structure of exogenous compounds. These modifications are usually classified into two phases: Phase I reactions, which introduce or unmask functional groups by oxidation, reduction, or hydrolysis; and Phase II reactions, involving conjugation with endogenous molecules to increase hydrophilicity and facilitate excretion. Some authors also refer to a subsequent Phase III, which encompasses active transport and elimination of the resulting metabolites [11].

Hepatic metabolism plays a central role, particularly through the action of enzymes such as CYP450, which catalyze the majority of Phase I reactions [12]. The activity of these enzymes is subject to considerable variability across individuals, influenced by genetic polymorphisms, age, diet, environmental exposures, and disease states [13], [14]. Such variability can alter the rate of drug clearance, leading to subtherapeutic effects or increased risk of adverse drug reactions.

In recent years, the formation of reactive metabolites during hepatic metabolism has received more and more attention, given their potential to cause cellular damage, covalent binding to proteins, and long-term toxicity. Consequently, the evaluation of drug-metabolizing enzymes, their regulation, and their potential to generate harmful intermediates has become a critical component of preclinical safety assessment and rational drug design.

3.1.1 Cytochrome P450 Enzymes

The CYP450 system constitutes a superfamily of heme-containing enzymes widely distributed in all domains of life, of particular relevance in higher eukaryotic organisms due to their vital role in the metabolism of both endogenous and exogenous compounds. In humans, these enzymes are actively implicated in the biotransformation of xenobiotics, including most drugs in clinical use, as well as in the synthesis and degradation of steroid hormones, bile acids, fatty acids, and even cell signaling molecules [15].

CYP450 enzymes are located primarily in the smooth endoplasmic reticulum of liver cells, although they have also been identified in extrahepatic tissues such as the intestine [16], lungs [17], kidneys [18], and brain [19], among others. Structurally, all isoforms share a conserved domain with a heme group at their active site, containing an iron atom capable of alternating between different oxidation states, which facilitates the activation of molecular oxygen to carry out monooxygenase oxidation reactions [15].

Functionally, reactions catalyzed by CYP450 mainly involve the insertion of an oxygen atom into an organic substrate (usually lipophilic), converting it into a more polar metabolite and therefore one that is more easily excreted by the body. These reactions primarily include hydroxylations, dealkylations, epoxidations, oxidations, and N- and S-oxidations [12]. These modifications are part of the Phase I of drug metabolism, in which the chemical structure of a compound is altered to facilitate its conjugation in Phase II and its subsequent elimination.

Hepatic CYP450s are responsible for the metabolism of approximately 75% of all drugs currently in clinical use, making them a key determinant in drug pharmacokinetics [20]. A thorough understanding of both the qualitative and quantitative aspects of CYP450-mediated metabolism is therefore essential, particularly in patients with advanced liver diseases, where enzymatic activity may be significantly altered.

In terms of nomenclature, CYP450 isoforms are classified according to a standard convention based on amino acid sequence similarity. They are grouped into families (e.g. CYP1, CYP2, CYP3) that share at least 40% sequence identity and subfamilies (e.g. CYP1A, CYP2C, CYP3A) with more than 55% identity [21]. Within each subfamily, individual members are numbered (e.g., CYP1A2, CYP2D6, CYP3A4).

The CYP450 system shows considerable interindividual variability, both in terms of expression and catalytic activity. This variability is due to genetic (polymorphisms), epigenetic, environmental (diet, exposure to inducers or inhibitors), physiological (age, sex), and pathological (liver diseases, infections) factors [22]. This diversity partially explains the differences observed in drug responses between individuals and populations, as well as susceptibility to adverse effects or toxicity.

A key aspect of the CYP450 system is its involvement in phenomena such as enzymatic inhibition, induction, and time-dependent metabolism, all of which can alter the pharmacokinetics of compounds. In particular, time-dependent inhibition (TDI) represents a major challenge in predictive toxicology, as it can lead to irreversible

or prolonged decreases in enzyme activity, with clinically relevant consequences. As it plays a central role in hepatic metabolism, functional assessment of the CYP450 system has become a keystone in clinical pharmacology, rational drug design, and preclinical toxicology.

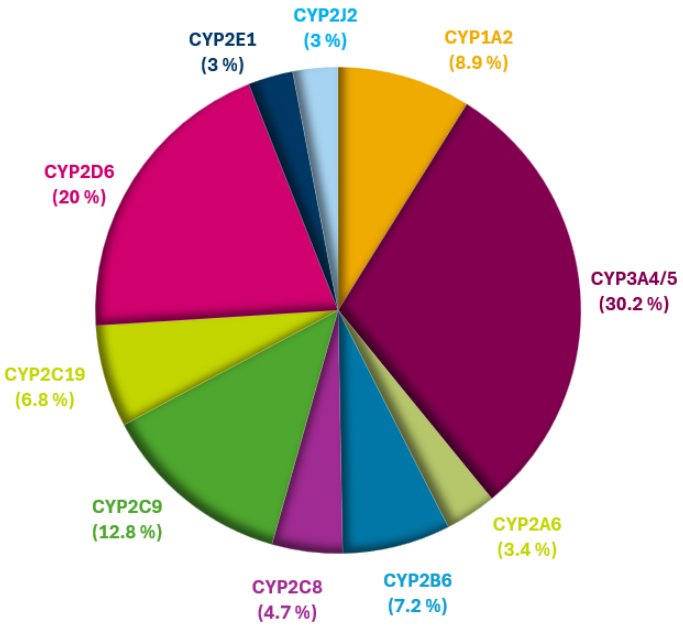


Figure 3.1: Proportion of CYP450 isoforms in drug metabolism pathways. Adapted and redrawn from the original source [15].

3.1.1.1 CYP1 Family: CYP1A2

The CYP1 family plays a key role in the metabolism of structurally complex compounds such as polycyclic aromatic hydrocarbons and aromatic amines [15]. These enzymes are highly inducible via the aryl hydrocarbon receptor, a ligand-activated transcription factor that responds to environmental contaminants. Among this family, CYP1A2 is constitutively expressed in the human liver and can be significantly induced by aryl hydrocarbon receptor ligands [23].

CYP1A2 is responsible for the metabolism of various therapeutic drugs, including analgesics, antipyretics, and antipsychotics. In addition, it contributes to the bioactivation of certain procarcinogens. For instance, compounds such as arylarenes present in grilled food or formed during industrial combustion can be metabolized by CYP1A2 into reactive intermediates capable of damaging DNA. This enzymatic function has been associated with increased cancer risk and has been extensively studied in toxicological evaluations of dietary and environmental exposures [15].

3.1.1.2 CYP2 Family: CYP2C9, CYP2C19, and CYP2D6

The CYP2 family encompasses a large and diverse group of enzymes with both hepatic and extrahepatic expression. Many of its members are critical for drug

metabolism, although the family also includes orphan enzymes whose physiological roles are not yet well understood. A distinctive feature of pharmacologically relevant CYP2 isoforms is their high degree of genetic polymorphism, which results in marked interindividual variability in enzymatic activity [24].

CYP2C9 plays a key role in the metabolism of weakly acidic compounds such as the anticoagulant warfarin, the anticonvulsant phenytoin, and several NSAIDs. It also contributes to the metabolism of endogenous steroids. Polymorphisms in CYP2C9 can lead to altered drug clearance rates and are particularly important in the clinical management of anticoagulant therapies [15], [25].

CYP2C19 is involved in the biotransformation of several classes of drugs, including proton pump inhibitors (e.g., omeprazole, pantoprazole), antidepressants, and antiepileptics. Like CYP2C9, this isoform is highly polymorphic, and its allelic variants influence both the efficacy and safety profiles of drugs, especially those requiring metabolic activation, such as clopidogrel [26].

CYP2D6, despite its relatively low hepatic abundance, is one of the most pharmacologically significant CYP450s. It metabolizes around 15–25% of all clinically used drugs, including beta-blockers, opioids, and tricyclic antidepressants. Over a hundred known allelic variants contribute to phenotypic diversity, classifying individuals as poor, intermediate, extensive, or ultrarapid metabolizers. This variation has critical implications for personalized medicine and therapeutic monitoring [15].

3.1.1.3 CYP3 Family: CYP3A4

The CYP3 family is of particular importance in clinical pharmacology due to its broad substrate specificity and high expression levels in both the liver and intestine. CYP3A4, the most abundant isoform of this family, is involved in the metabolism of nearly 30% of all marketed drugs. Its large and flexible active site allows it to accommodate a wide range of structurally diverse and lipophilic compounds, including immunosuppressants, anticancer agents, statins, and benzodiazepines [27].

Beyond its role in xenobiotic metabolism, CYP3A4 also participates in the hydroxylation of endogenous steroids such as testosterone and progesterone. The enzyme shares more than 85% sequence identity with CYP3A5, and although their substrate preferences are largely overlapping, their expression levels vary significantly between individuals and populations.

CYP3A4 activity can be modulated by a range of external compounds. It is strongly induced by drugs such as rifampicin and carbamazepine and inhibited by substances like ketoconazole, ritonavir, and components of grapefruit juice. These interactions make CYP3A4 a central mediator of drug-drug interactions and a critical target in early-stage pharmacokinetic and safety assessments [15].

3.1.2 Inhibition of CYP450 Enzymes

During the metabolism of xenobiotics, CYP450 enzymes can generate reactive metabolites that, in some cases, inhibit the enzymes themselves. This inhibition can be re-

versible or irreversible, depending on the nature of the inhibitor and its interaction with the enzymes active site. Reversible inhibition involves a temporary binding of the inhibitor to the enzyme, often through non-covalent interactions [28]. In such cases, enzyme activity can be restored upon the dissociation or elimination of the inhibitor from the system, and its pharmacological consequences are generally transient.

Irreversible inhibition, by contrast, leads to the permanent inactivation of the enzyme. This occurs when a compound is metabolized into a reactive intermediate that forms a covalent bond with the enzyme, typically with the heme group or a nucleophilic amino acid residue in the active site, resulting in structural modifications that render the enzyme catalytically inactive. Because this process cannot be reversed through simple dissociation, the restoration of enzymatic function relies on the synthesis of new enzyme molecules. During this regeneration period, liver capacity to metabolize substrates is diminished, which can lead to drug accumulation, impaired detoxification, and hepatotoxicity [29].

Both types of CYP450 enzyme inhibition are depicted in Figure 3.2. These schematics demonstrate the mechanisms by which drugs can inhibit CYP450 enzymes in the liver. Reversible inhibition is temporary and allows for enzyme recovery, while irreversible inhibition involves metabolic activation leading to permanent enzyme inactivation.

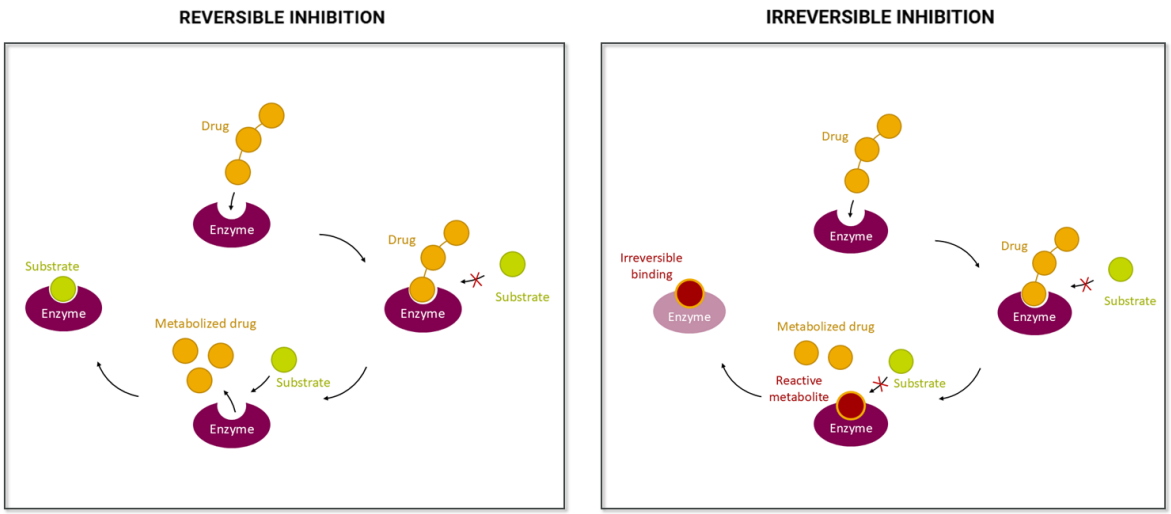


Figure 3.2: Mechanisms of CYP450 enzyme inhibition. The left diagram illustrates reversible inhibition, and the right diagram illustrates irreversible inhibition.

TDI is a particular case of iireversible inhibition, where the inhibitory potency increases progressively with time, usually as a result of metabolic activation. In the context of CYP450, TDI is of high concern during drug development, as it can lead to unexpected pharmacokinetic behavior and adverse clinical outcomes [30]. The enzymes most commonly affected include CYP3A4, CYP1A2, CYP2D6, and members of the CYP2C subfamily, though the extent and consequences of TDI vary according

to the chemical structure, reactivity of metabolites, and enzyme involved.

In cases where reversible inhibition is also assessed, the inhibitory potency of a compound is often expressed using the half-maximal inhibitory concentration (IC_{50}), which reflects the concentration required to reduce enzyme activity by 50% under specific experimental conditions [31]. For comparative purposes, the negative logarithm of this value, known as pIC_{50}, is frequently used:

$$pIC_{50} = -\log_{10}(IC_{50}) \tag{3.1}$$

Although IC_{50} values are commonly applied in reversible inhibition studies, they are less informative in the context of TDI, where kinetic parameters such as k_{inact} and K_I provide a more mechanistic and accurate description of the inhibition process [31].

The kinetic parameters that define the process include [32]:

$$K_I = \text{inhibitor concentration at half-maximal } k_{\text{inact}}$$

$$k_{\text{inact}} = \text{maximum rate constant for enzyme inactivation}$$

These values allow for the quantitative characterization of the inhibitors potency and efficiency, and are essential for predicting in vivo risk through physiologically based pharmacokinetic modeling.

TDI is particularly relevant in the context of drug-drug interactions (DDIs). When two drugs are co-administered, one may act as a time-dependent inhibitor of the CYP450 enzyme responsible for metabolizing the other. If this inhibition is irreversible or quasi-irreversible, the second drug may remain unmetabolized in the system for a prolonged period, until sufficient de novo synthesis of the enzyme occurs. This can lead to the accumulation of the affected drug, enhancing its pharmacodynamic effects or triggering toxic responses. In drugs with narrow therapeutic windows, such interactions can have serious clinical implications [33].

From a regulatory perspective, both the U.S. Food and Drug Administration and the European Medicines Agency recommend the early identification and evaluation of TDI potential during drug development [34], [35]. Compounds that exhibit TDI are flagged for additional in vivo assessment and may require dose adjustments or contraindications when used in combination with other medications. The accurate prediction of CYP-mediated DDIs, including those involving TDI, remains one of the most critical aspects of ensuring drug safety and efficacy in clinical settings.

3.1.3 Trapping Assays for Reactive Metabolites

In addition to CYP-mediated inhibition, a second key mechanism contributing to drug-induced toxicity involves the formation of reactive metabolites. These unstable and electrophilic intermediates can interact covalently with cellular macromolecules,

potentially triggering hepatotoxic, mutagenic, or immunogenic effects. Because of their transient nature, direct detection is difficult, and alternative strategies have been developed to assess their presence. For this, trapping assays are essential tools for evaluating the bioactivation potential of drug candidates, particularly in the early stages of drug development [3].

One of the most widely used approaches involves the use of reduced GSH as a trapping agent. GSH is a tripeptide naturally synthesized in human cells, particularly in the liver, and plays a central role in maintaining redox homeostasis and detoxifying reactive intermediates. Its biological relevance, combined with its high nucleophilicity, makes it especially suitable for capturing soft electrophiles formed during cytochrome P450-mediated metabolism. In typical assays, the test compound is incubated with liver microsomes in the presence of NADPH or an NADPH-generating system to support CYP activity, along with an excess of reduced GSH. The resulting GSH conjugates are then detected by liquid chromatography-mass spectrometry, providing a qualitative indication of bioactivation potential [3], [36], [37].

Figure 3.3 illustrates the metabolic pathway of clopidogrel, where the compound is bioactivated by CYP450 enzymes to form an active thiol metabolite. This nucleophilic intermediate contains a free SH group, which can undergo conjugation with GSH to form a disulfide adduct, facilitating detoxification and excretion. In the figure, the active thiol metabolite of clopidogrel is highlighted, with nucleophilic SH and adjacent CH groups shown in red. Upon conjugation, GSH binds via a disulfide bond (SS), forming a stable adduct. The conjugated sulfur and the surrounding GSH structure are highlighted in green and blue, respectively. This process represents a key detoxification pathway following CYP450-mediated activation.

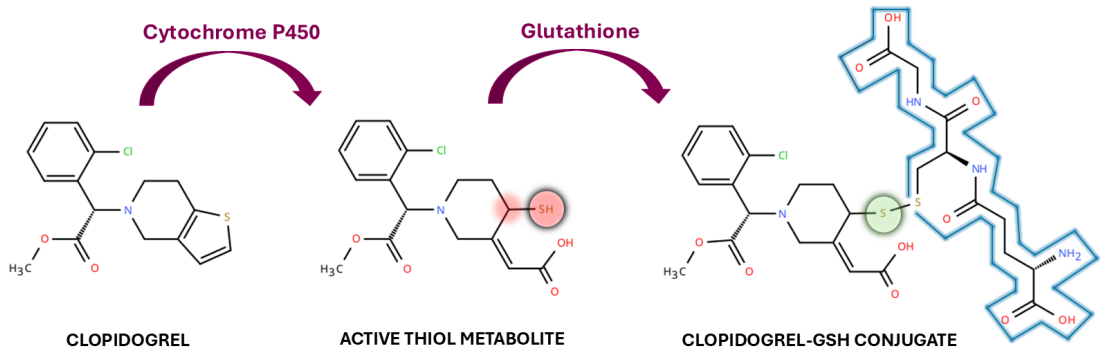


Figure 3.3: Bioactivation and GSH conjugation of clopidogrel. Reworked from [38].

To overcome the limitations of qualitative analysis, more sensitive and quantitative variants have been developed. One such example is the Dansyl-GSH trapping assay, which introduces a fluorescent dansyl group onto the glycine amino group of GSH. This modification allows for enhanced detection and quantification of GSH adducts using high-performance liquid chromatography coupled with fluorescence and mass spectrometric detection [39]. This assay enables both the identification of reactive metabolite structures and a more precise estimation of their formation.

In addition to GSH-based assays, other trapping agents are employed to broaden the detection of metabolite types. KCN is used as an alternative nucleophile when GSH is insufficient, particularly for detecting hard electrophiles that are less reactive with thiols. KCN trapping has proven effective in identifying reactive intermediates in metabolic pathways that are undetectable through conventional GSH analysis. For example, the bioactivation of compounds such as nicotine and nefazodone has been successfully studied using this strategy [3].

MA is another useful trapping agent, primarily employed to detect aldehyde intermediates. These are often formed during oxidative metabolism and are not effectively trapped by GSH or KCN. In this assay, the test compound is incubated with liver microsomes and MA, and the resulting oxime adducts are analyzed by high-performance liquid chromatography-mass spectrometry. A related technique involves the use of radiolabeled semicarbazide as a selective aldehyde-trapping reagent, enabling quantitative analysis through radiometric detection or liquid scintillation counting [3].

Trapping assays contribute to a more comprehensive understanding of metabolic pathways and the identification of structural features associated with reactive metabolite formation. By incorporating a variety of trapping agents with different chemical reactivities, it is possible to detect a broader spectrum of electrophilic species. This information is valuable not only for assessing safety risks but also for guiding medicinal chemistry efforts aimed at minimizing bioactivation and improving the toxicological profile of new compounds.

3.2 Computational Background

The integration of *in silico* methods in drug discovery has become increasingly important for predicting the pharmacokinetic and toxicological properties of chemical compounds. In particular, the development of machine learning (ML) models trained on molecular representations has enabled researchers to anticipate biological behaviors from chemical structure alone, reducing the need for extensive *in vitro* experimentation and accelerating the early stages of compound evaluation. This section outlines the theoretical foundations underlying the computational strategy used in this work, including molecular representations, descriptor generation, learning algorithms, and model evaluation metrics.

3.2.1 Molecular Representations and Descriptors

In cheminformatics, chemical structures must be converted to formats that are interpretable by computational algorithms. One of the most widely used textual representations is the SMILES (Simplified Molecular Input Line Entry System) notation. SMILES encodes molecular graphs into strings by representing atoms and bonds through a linear syntax. Due to its simplicity and compatibility with many cheminformatics libraries, SMILES has become a standard input format for molecular modeling tasks. However, in its raw form, SMILES may present inconsistencies due to stereoisomers, tautomers, or differences in protonation states [40]. Therefore,

molecules are commonly preprocessed and standardized before further analysis to ensure that equivalent structures are treated uniformly.

To extract meaningful features from molecules, a variety of molecular descriptors can be computed. These descriptors are numerical values that reflect different aspects of the molecular structure, such as atom composition, bond types, electronic distribution, geometrical configuration, and topological characteristics [41]. The use of descriptors allows chemical structures to be mapped into a numerical feature space, enabling the application of classical ML techniques. Toolkits like RDKit [42] offer a broad selection of general purpose descriptors, while extended packages such as Mordred compute hundreds to thousands of descriptors grouped into families (e.g., constitutional, topological, geometrical, or electronic) [43]. Including physicochemical properties such as pKa and log D enhances the interpretability of the input space and often provides critical information on solubility, ionization, and membrane permeability, all of which are relevant for drug metabolism and transport [44].

Beyond traditional descriptors, molecular fingerprints offer an alternative representation based on structural patterns. One widely used method is the Extended-Connectivity Fingerprint (ECFP), commonly referred to as Morgan fingerprints. ECFP generates circular fingerprints by iteratively hashing local atomic environments up to a defined radius, encoding the presence or absence of substructural motifs. These fingerprints produce high-dimensional binary vectors that are particularly suitable for similarity-based searches and classification tasks [45]. Unlike global descriptors, which often summarize whole-molecule properties, fingerprints capture fine-grained structural detail and have demonstrated high performance in modeling biological endpoints [46].

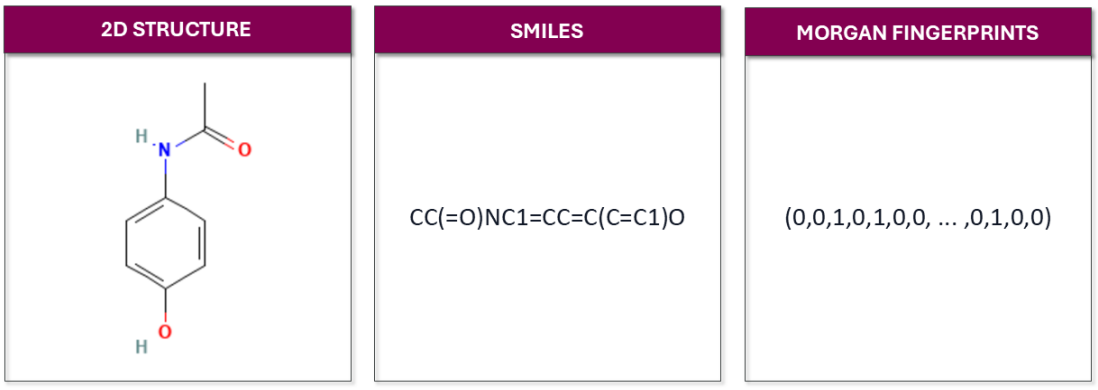


Figure 3.4: 2D, SMILES and Morgan fingerprints representations of Acetaminophen [47].

3.2.2 Machine Learning in Molecular Property Prediction

Machine learning provides a data-driven approach to predicting molecular properties, allowing for the modeling of complex relationships between chemical features and biological outcomes. Classical supervised learning algorithms, such as Ran-

dom Forests (RF) [48] and Extreme Gradient Boosting (XGBoost) [49], are particularly well suited for tasks involving high-dimensional and heterogeneous data. Random Forests operate by constructing an ensemble of decision trees trained on bootstrapped subsets of the data and aggregating their predictions. This approach reduces variance and improves generalizability, especially when input features exhibit collinearity or noise. XGBoost, on the contrary, builds additive models in a forward stage-wise manner, optimizing a differentiable loss function through gradient descent. Its ability to capture non-linear interactions, along with built-in regularization, makes it one of the most powerful algorithms in current practice.

In addition to tree-based models, deep learning methods have gained prominence in cheminformatics due to their ability to learn directly from molecular structure. Chemprop [50] is an example of such a framework, utilizing a Message Passing Neural Network (MPNN) to model chemical graphs. In this approach, atoms are treated as nodes and bonds as edges, allowing the network to iteratively propagate information between neighboring atoms. Unlike traditional ML methods, which rely on predefined features, MPNNs learn internal molecular representations during training, potentially capturing relevant patterns that are not explicitly encoded in conventional descriptors. These models are particularly advantageous when working with large datasets and when feature engineering is infeasible or limiting [51].

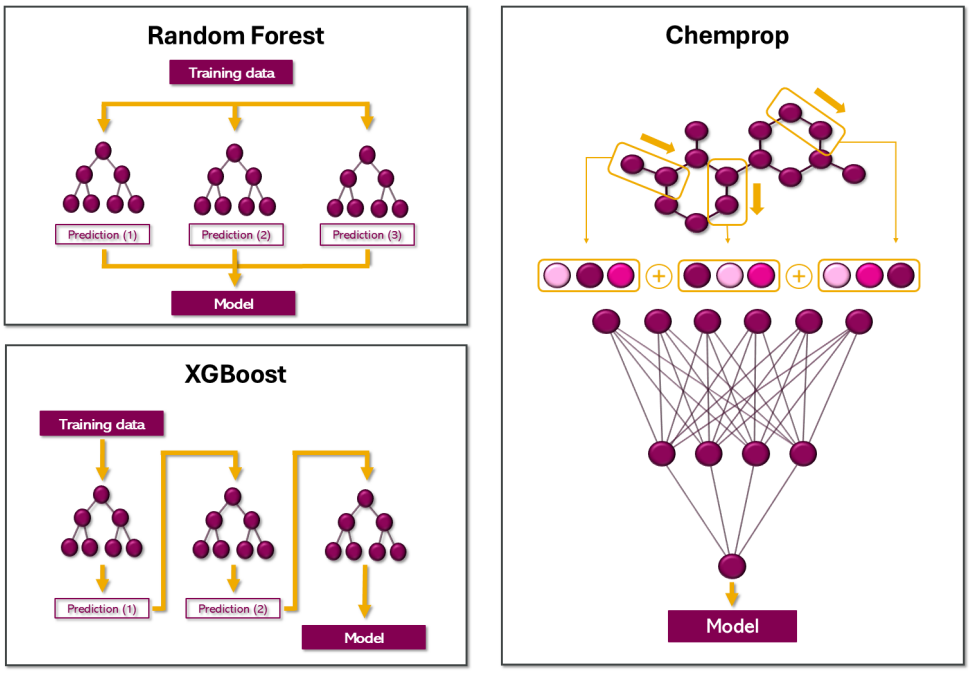


Figure 3.5: Schematic overview of RF, XGBoost, and Chemprop architectures. Rework from [52], [53].

Each modeling approach offers different strengths. Tree-based models are generally more interpretable and less sensitive to hyperparameter tuning, while deep learning models often outperform tasks involving subtle or complex patterns in molecular

topology. The choice between them depends on the dataset size, the diversity of the chemical space, and the desired balance between performance and interpretability.

3.2.3 Model Evaluation Metrics

To assess the predictive performance of classification models, a variety of evaluation metrics are used. These metrics allow for a detailed understanding of how the model behaves in terms of identifying both positive and negative classes.

Classification performance is commonly evaluated using the four components of the confusion matrix [54]:

- **True Positives (TP):** instances correctly predicted as positive
- **True Negatives (TN):** instances correctly predicted as negative
- **False Positives (FP):** negative instances incorrectly predicted as positive
- **False Negatives (FN):** positive instances incorrectly predicted as negative

Based on these definitions, the following evaluation metrics [54] were used:

Precision Measures the proportion of predicted positive cases that are actually positive.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{3.2}$$

Recall (Sensitivity) Measures the proportion of actual positive cases that are correctly predicted.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3.3}$$

Specificity (True Negative Rate) Measures the proportion of actual negative cases that are correctly predicted.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{3.4}$$

F1-score Harmonic mean of precision and recall. Provides a balance between both metrics.

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3.5}$$

Balanced Accuracy Averages the true positive rate and true negative rate, accounting for class imbalance.

$$\text{BACC} = \frac{\text{sensitivity} + \text{specificity}}{2} \tag{3.6}$$

Matthews Correlation Coefficient (MCC) A robust single-value metric that accounts for all elements of the confusion matrix, giving a general overview of how the model works. This metric ranges from -1 to 1, being -1 the inverse predictions and, 1 a perfect prediction.

$$\text{MCC} = \frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.7}$$

Accuracy Measures the overall proportion of correct predictions. In imbalanced datasets, this metric can be misleading, as high accuracy may simply reflect correct classification of the majority class while ignoring the minority class.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.8}$$

Each of these metrics captures different aspects of model performance, and their combined interpretation provides a more comprehensive evaluation of predictive reliability, especially in the presence of class imbalance.

4

Methods

This chapter details the methodology followed throughout the project, including data preparation, processing, model design, and evaluation. All datasets were curated and processed to ensure data consistency, reduce noise, and prepare for machine learning workflows.

4.1 Dataset Characteristics

The datasets used in this study were extracted from the in-house measurements made in AstraZeneca. For all endpoints, isomeric SMILES were available and used as the primary molecular representation. The TDI dataset included percentage inhibition values, the compound concentration in µM, and an update date, which was later used for temporal data splitting.

Compound inhibition values was preprocessed based on a set of predefined criteria provided by experts. Initially, negative inhibition percentages were removed, as these likely reflect assay noise or error. 20% inhibition was used as the threshold to group the compounds into active/ inactive class. If the value was exactly 20%, the binary label was assigned based on the activity flags captured by the assays. Values with concentration above 50 µM were excluded from the active class, since very high concentrations may produce non-specific inhibition and are generally not biologically meaningful. Furthermore, if the inhibition values included data qualifiers ("<", ">", or "="), a set of standardized rules were applied for preprocessing (Table 4.1).

Table 4.1: Labeling rules for TDI classification; Active: 1 and Inactive: 0

Inhibition value	Flag	Label
>20	Active	1
<20 and ≥0	Not Active	0
<0	Not Active	0
<0	Active / None	Removed
=20	Not Active	0
=20	Active	1
>20 at >50 µM	Any	Removed

The class distribution after applying these rules is shown in Figure 4.1. As illustrated, the dataset is notably imbalanced, with fewer active cases compared to inactive ones across most CYP450 isoforms. Among them, CYP3A4 displays the most balanced distribution, likely due to its more frequent experimental testing and relevance in drug metabolism.

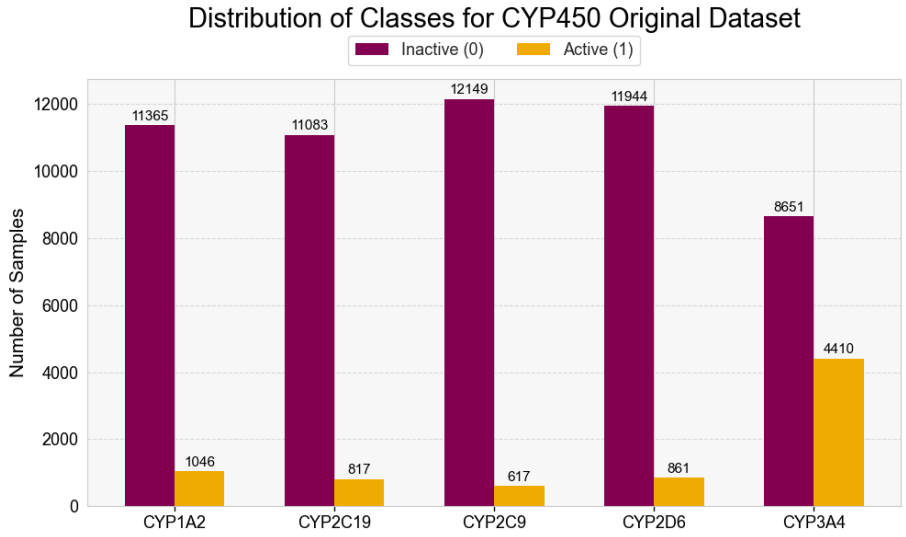


Figure 4.1: Distribution of TDI data after cleaning and labeling.

In the case of trapping assays, the datasets already contained binary labels ('Yes' for active, 'No' for inactive), and no additional thresholding was required. However, the three datasets underwent the same cleaning pipeline described in section 4.2.

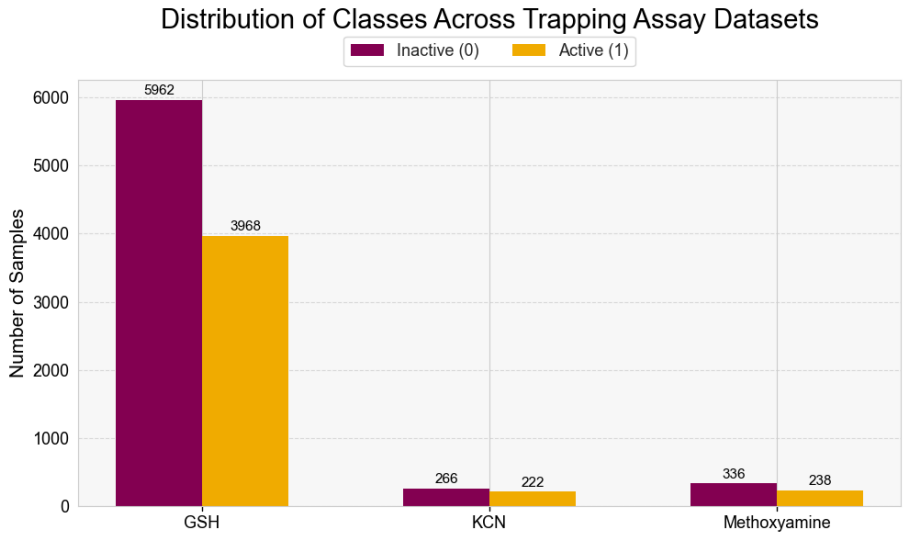


Figure 4.2: Number of data points per trapping assay type: GSH, KCN, and Methoxyamine.

As shown in Figure 4.2, the number of data points differs substantially between the

three trapping assay types. In particular, the GSH assay is more widely represented in the dataset. This likely reflects its broader application in early compound screening, as GSH is a well-established nucleophilic trapping agent used to detect reactive electrophilic species in drug metabolism studies. In contrast, the KCN and MA assays have significantly fewer data points, which limits the amount of information available for model training. However, these two assays are less affected by class imbalance, presenting a more even distribution of actives and inactives.

4.2 Dataset Preprocessing

4.2.1 Standardization and Feature Generation

The initial step in data preprocessing involved standardizing all SMILES strings using ChEMBL structure pipeline, an open source package for chemical structure curation. This ensured uniform and valid molecular representations across all compounds. Molecules containing salt fragments were desalted to retain only the main active species, as salts may introduce redundancy or noise without contributing to meaningful structural information. Invalid SMILES, including unparsable strings or entries returning NaN values, were removed. Duplicate compounds were also discarded to prevent bias during model training and evaluation.

After the cleaning process, molecular features were generated using two distinct representations. ECFPs were computed as binary vectors with a radius of 2 and a size of 2048 bits. Additionally, 2D physicochemical descriptors were calculated using RDKits descriptor module to capture broader chemical characteristics. These representations were used independently to evaluate model performance with different feature sets.

4.2.2 Feature Selection

To reduce redundancy and improve model generalization, a feature selection step was applied prior to model training. This process was performed separately for both the fingerprint and descriptor feature sets.

Initially, we removed features exhibiting near-zero variance by employing a variance thresholding technique from Scikit-learn. This process involved fitting the filter on the training set and then consistently applying it to the test set, ensuring uniform preprocessing across data sets. This step was crucial in eliminating variables that displayed minimal variation across compounds, as such features contribute little to model discrimination.

Next, highly correlated features were removed to reduce redundancy and mitigate multicollinearity. The Pearson correlation coefficient was computed between all pairs of features using the training set which returns the full pairwise correlation matrix. To avoid redundant comparisons and self-correlation, only the upper triangle of the matrix (excluding the diagonal) was considered. Any pair of features with a Pearson correlation coefficient greater than 0.95 was flagged, and one feature from

each pair was removed. This filtering process was applied independently to both the fingerprint and descriptor feature sets to ensure consistency across representations.

For the descriptor-based models, feature values were additionally standardized using Scikit-learn's `StandardScaler`, which transformed the data to have zero mean and unit variance. This normalization step was essential for models sensitive to feature scales, such as gradient boosting algorithms.

4.2.3 Data Splitting Strategy

Two types of dataset splitting strategies were implemented: temporal splitting and stratified random splitting.

To emulate realistic deployment conditions, the primary strategy adopted was temporal splitting. Each compound in the dataset included an experimental update date, which was used to chronologically sort the data. Based on the distribution of compound update dates, two cutoff points were selected for CYP450 compounds: April 2024 (one year of the latest data available) and October 2024 (six months of the latest data), as shown in Figure 4.3. Throughout this thesis, these temporal splits will be referred to as April 2024 and October 2024, respectively. Initially, October 2023 was also considered as a potential split; however, the final decision was to proceed with one-year and six-month horizons to better represent near-future deployment scenarios. For completeness, the results for October 2023 are included in the Appendix A.

In the case of trapping assays, only one temporal split (October 2024) was selected due to the reduced datasets. For each split, all compounds dated from the corresponding cutoff onward were assigned to the test set, while the remaining data were used for training. From the training set, 10% was further held out for validation. This approach ensures that the model is evaluated on future, unseen data, mimicking how it would perform in prospective use.

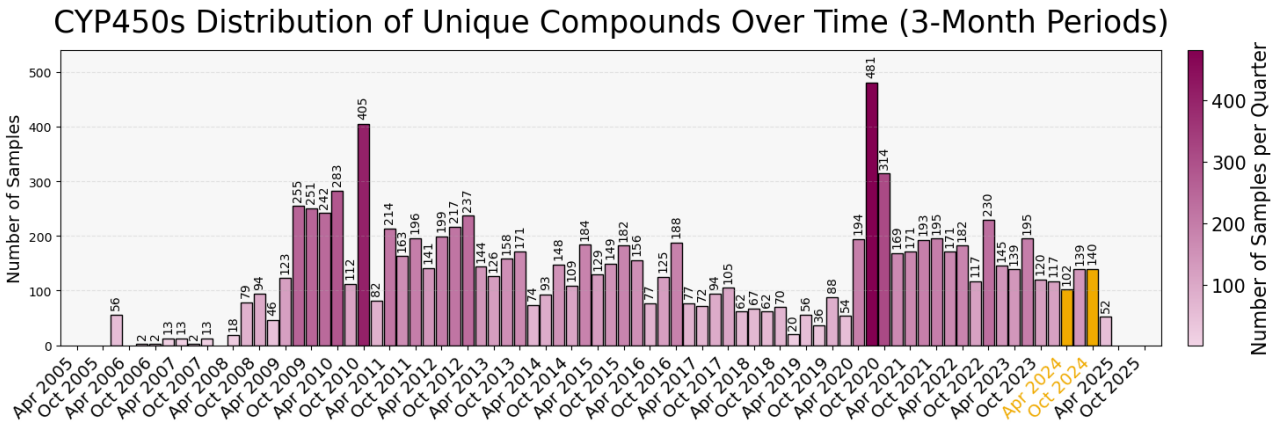


Figure 4.3: Distribution of CYP450 samples over time grouped by 3-month periods. The color intensity represents the number of samples per quarter. Gold bars indicate the selected temporal split points: April and October 2024.

In addition to the main experiments with temporal splits, a set of exploratory analyses was conducted using stratified random splitting. While random splitting maintains the proportion of active and inactive compounds across training and test sets and is commonly used in machine learning, it may overestimate model performance by mixing data that would not realistically be available during inference. Nonetheless, comparing performance between the two strategies helps contextualize the robustness of the models. For each CYP450 isoform, the random split used a test set of equivalent size to that of the corresponding temporal split, ensuring fair comparison. As with the temporal strategy, 10% of the training data was reserved for validation.

4.2.4 Handling Data Imbalance

Most of the datasets used in this study were affected by varying degrees of class imbalance, with a smaller proportion of active compounds compared to inactive ones. To mitigate this, several resampling techniques were explored. Synthetic Minority Over-sampling Technique (SMOTE) [55] was applied to increase the number of active samples in the training data, while Random Under-Sampling (RUS) [56] was used to reduce the number of majority class examples. These approaches have been widely compared in previous studies, and although results can vary by context, they often lead to improved model performance when dealing with strongly imbalanced datasets [57], [58].

For both methods, multiple resampling ratios were evaluated to investigate their impact on model performance. The following sampling strategies values were tested: 0.3, 0.5, 0.7, and 1.0, representing different levels of balancing between the minority and majority classes.

In addition, scaffold-based filtering was used to remove structurally redundant inactive compounds and improve the chemical diversity of the training set. Specifically, BemisMurcko scaffolds were computed using RDKit to identify the molecular frameworks of each compound. This approach defines a scaffold as the union of a molecules ring systems and linkers, excluding side chains and substituents [59]. By retaining only one representative per scaffold, this filtering step ensures that highly similar compounds do not dominate the training data. As a result, the model is encouraged to learn more generalizable patterns rather than overfitting to recurring chemotypes.

A visual comparison between the original dataset and the scaffold-reduced version is shown in Figure 4.4, highlighting the effect of this filtering step on class distribution and compound diversity.

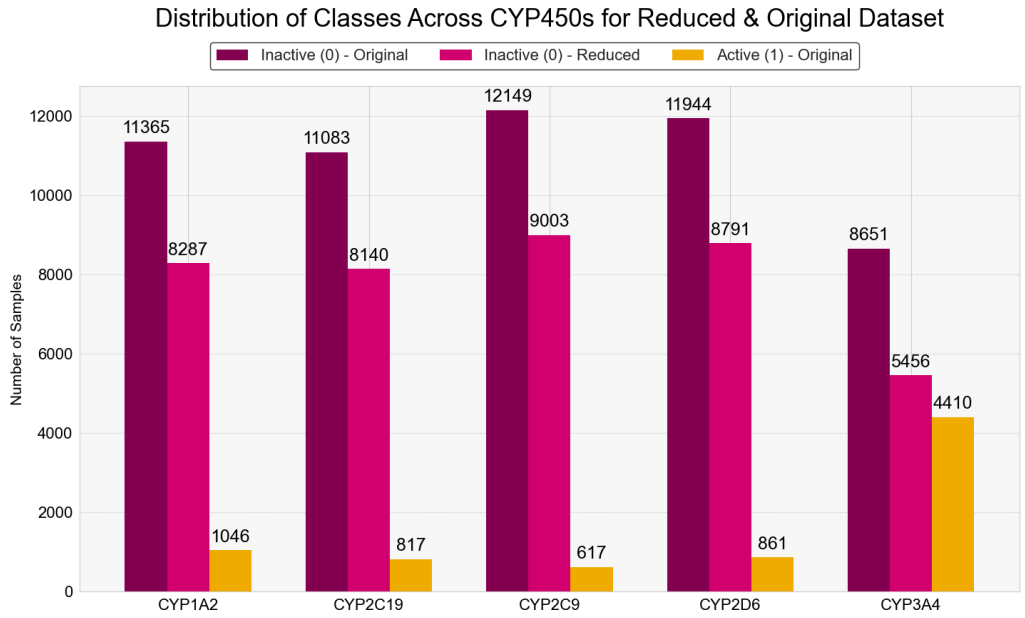


Figure 4.4: Comparison of original vs scaffold reduced dataset.

4.3 Software and Infrastructure

All data processing, molecular standardization, and model development were implemented in Python 3.13.2. Molecular structures were handled using ChEMBL [60] structure pipeline, which was employed for parsing and validating SMILES, removing salts and standardizing molecules. For generating both 2D descriptors and ECFP fingerprint, RDKIT was employed. RDKit provided a consistent and efficient framework for molecular feature extraction across all endpoints.

Tree-based machine learning models, specifically RF and XGBoost classifiers, were implemented using Scikit-learn and XGBoost libraries, respectively. Scikit-learn also supported preprocessing steps such as variance filtering, correlation analysis, and feature scaling.

Chemprop was used to model compounds directly from their SMILES representations. Chemprop handles feature extraction internally by learning molecular representations from graph-structured input data, and includes built-in functionality for training, validation, and hyperparameter optimization.

All experiments were executed on a high-performance computing cluster managed by the SLURM workload manager. Batch submission scripts were used to allocate computational resources and parallelize runs. Tree-based models were trained on CPU nodes, which were sufficient for their computational requirements. In contrast, single and multitask Chemprop models were trained on GPU-accelerated nodes to support deep learning frameworks and reduce training time.

4.4 Modeling Approaches

Two types of models were used in this thesis: single-task models, which were trained independently for each endpoint, and multitask models, which attempted to leverage shared information across multiple endpoints simultaneously.

4.4.1 Single-Task Models

For each endpoint, a series of single-task classification models were trained using three tree-based algorithms: RF, a class-balanced variant of RF, and XGBoost. All models were evaluated using both ECFP and 2D molecular descriptors to assess performance across molecular representations.

Hyperparameter optimization was performed using randomized search guided by F1-score, as this metric offered a better balance in the presence of class imbalance. For Chemprop models, 10% of the training data was held out as an internal validation set to monitor overfitting. In contrast, for tree-based models, a 5-fold cross-validation was conducted on the subset of compounds labeled as "train" and "val" in the temporal split, in order to evaluate model performance and perform hyperparameter tuning efficiently.

Chemprop models were trained on the same datasets to allow a consistent comparison. Single-task and multitask architectures were evaluated, both with and without additional descriptors (RDKit or ECFP). As with the tree-based models, F1-score was used as the optimization criterion. The number of training epochs was fixed at 50 after preliminary experiments showed no meaningful improvement from training for 100 or 200 epochs.

The best-performing models from the training phase were selected and evaluated on the test sets using the full set of evaluation metrics introduced in subsection 3.2.3. These included F1-score, precision, recall, specificity, and MCC, providing a comprehensive view of classification quality, especially under class imbalance. The same evaluation protocol was applied to both tree-based and Chemprop models to ensure fair and consistent comparisons across architectures and molecular representations.

To gain further insight into the behavior of the best-performing tree-based models, additional analyses were conducted. For the RDKit-based model trained on CYP3A4, the six most important descriptors were identified based on feature importance scores, and their distributions were visualized separately for active and inactive compounds. This helped assess whether the features used by the model offered meaningful class separation.

In parallel, the ECFP representation was used to construct a chemical similarity network based on Tanimoto coefficients (using 0.7 as threshold). The resulting graph was visualized using the Fruchterman-Reingold algorithm, a force-directed layout technique that positions nodes in two-dimensional space by simulating attractive forces between connected nodes and repulsive forces between all others [61], [62]. This layout helps reveal structural clusters and neighborhood relationships. The network enabled a qualitative and quantitative exploration of structural connectivity

between training and test compounds, as well as how that connectivity related to prediction outcomes, particularly in terms of false positives and false negatives across different regions of the chemical space.

4.4.2 Multi-Task Models

Additionally, multitask models were trained to simultaneously predict CYP time-dependent inhibition across 5 isoforms and 3 trapping assay outcomes. Given that these endpoints reflect different underlying biological mechanisms of enzyme inhibition versus reactive metabolite formation this configuration aimed to test whether shared molecular patterns could still be leveraged despite limited task similarity. This setup allowed for a broader evaluation of the generalization capacity of multi-task learning across heterogeneous but potentially related toxicity endpoints.

All multitask models were implemented as classification problems for the TDI and trapping assay endpoints, which were provided with binary activity labels. Multitask models were initially trained using Chemprops default hyperparameters. To improve performance, hyperparameter optimization was later performed using Chemprops built-in hyperopt module. The tuning process focused on maximizing task-specific metrics such as F1-score and MCC, aiming to improve performance particularly on minority classes.

Model performance was evaluated independently for each task, using the same evaluation metrics as in the single-task setting. Results were compared to those of the single-task models to assess whether multitask learning led to consistent improvements or introduced trade-offs in predictive accuracy.

5

Results and Discussion

This chapter presents the main findings from the models developed throughout the project. Results are divided into two sections based on the endpoint type: the first section focuses on TDI prediction for the five selected CYP450 isoforms, and the second section summarizes the modeling outcomes in predicting the trapping assays. Details regarding evaluation of model performance across different molecule representations and machine learning approaches together with a discussion on key trends and limitations are presented below.

5.1 TDI CYP450s Results

5.1.1 Influence of Scaffold-Based Dataset Reduction on Model Performance

This section compares the performance of models trained on the processed dataset before and after applying scaffold-based reduction. The comparison focuses on the models built on temporal datasets considering 2 cut-off dates (April 2024 and October 2024) in both single-task and multitask setups.

Among the different metrics evaluated throughout the project, only MCC is shown in Figure 5.1. This metric was selected because it offers a balanced view of model performance in imbalanced classification settings and is particularly suitable for comparing overall quality across datasets and model types.

Training with the scaffold-reduced data generally led to equal or better performance compared to using the original datasets. In April 2024, MCC improved for CYP1A2 in both single-task (from 0.166 to 0.194) and multitask models (from 0.197 to 0.321). Similar trends were observed for CYP3A4, where MCC increased from 0.154 to 0.164 in the single-task setup, and from 0.164 to 0.276 in the multitask setup. CYP2D6 also showed consistent results, with comparable or slightly higher MCC when using the reduced dataset.

In October 2024, the same pattern was observed. For CYP1A2, the MCC increased from 0.084 to 0.145 in single-task and remained stable in multitask models. CYP3A4 showed the most consistent gain, with multitask MCC increasing from 0.285 to 0.345. CYP2D6 also improved slightly in both model types. Although predicting CYP2C19 and CYP2C9 remained challenging due to the low number of actives, the

reduced dataset did not negatively affect model performance and even produced small improvements in some cases.

Figure 5.1 shows a visual comparison of MCC values for each CYP across both time splits and model types. Each pair of bars compares the performance using the original dataset (purple) and the scaffold-reduced version (gold). While some differences are modest, the reduced dataset generally led to more stable or slightly improved MCC values, especially in the multitask setting.

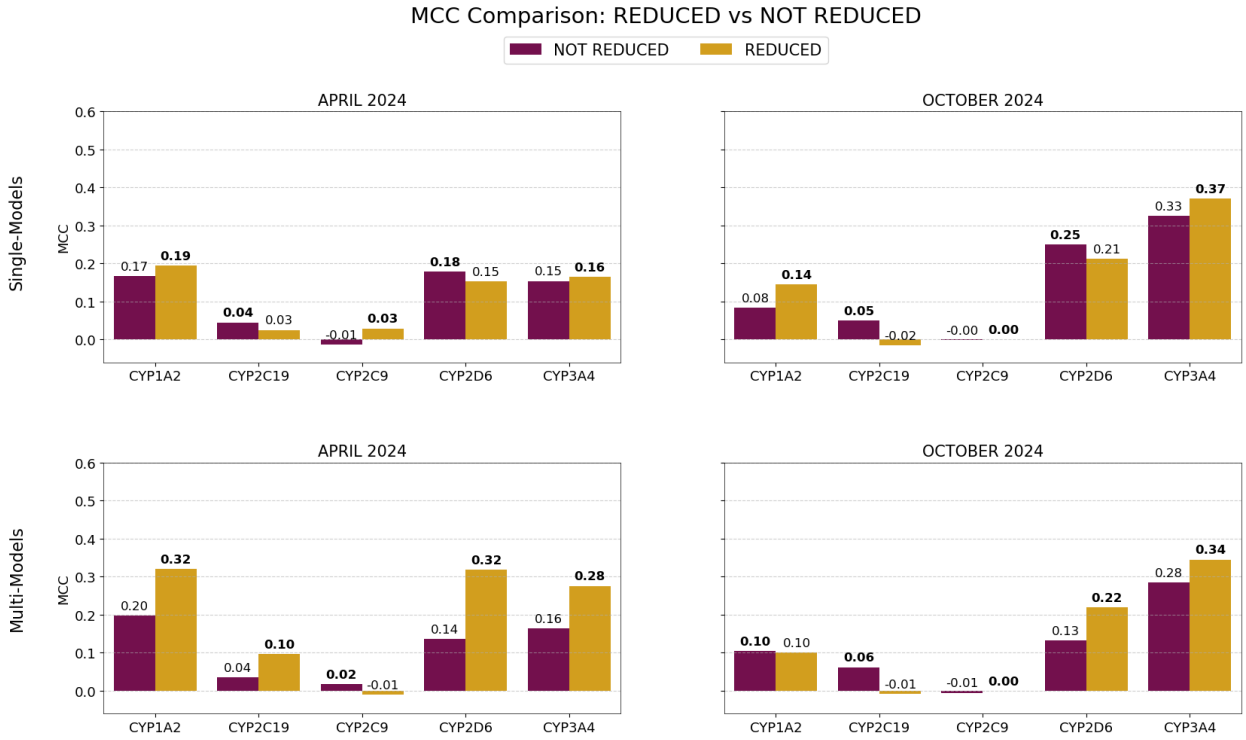


Figure 5.1: MCC comparison between models trained with the original dataset and the scaffold-reduced dataset for April 2024 and October 2024 splits. Results are shown for both single-task (RF and Chemprop) and multitask Chemprop models. The best MCC per CYP isoform within each setting is highlighted in bold.

Considering these results, the scaffold-reduced dataset was selected for the rest of the experiments. Although not all isoforms showed clear gains, this version allowed for a more balanced setup without compromising diversity or model performance.

5.1.2 Models Performance on Temporal Splits

Before presenting the detailed results, it is useful to clarify how the models shown in this section were selected. For the single-task Chemprop models, the default version without additional descriptors was always included as a baseline. For the tree-based models, only the best-performing configuration was selected for each CYP450 and time split. To allow a fair comparison, a second Chemprop variant was then chosen to match the descriptor type (RDKit or ECFP) used in that tree model. The

multitask Chemprop model shown is always the default version without descriptors. This setup provides a consistent and balanced view of model behavior across cyp isoforms and modeling strategies. All the other tested models and their evaluation metrics are available in section A.1.

5.1.2.1 Model Performance Summary on April 2024 Split

Figure 5.2 summarizes the April 2024 performance across all five CYP450 isoforms, comparing four modeling strategies using recall, precision, specificity, F1-score, and MCC, offering a detailed perspective on their strengths and weaknesses.

Across isoforms, the single-tree model showed the most consistent and balanced behavior. In CYP1A2, for example, it achieved the highest recall and precision, resulting in stronger F1 and MCC scores compared to the other approaches. The Chemprop models, especially those trained with added descriptors, often showed a tendency to favor specificity over recall. This implies that they were good at identifying negatives but frequently missed true positives, which negatively affected their F1 and MCC values.

CYP2C19 proved particularly challenging. The tree model detected most positives (high recall) but with low precision, leading to poor F1. All Chemprop variants struggled to identify any positives, resulting in near-zero metrics, though the multitask version showed a slight improvement.

For CYP2C9, performance was generally poor. Only the Chemprop model without added descriptors achieved a non-zero recall, although its precision remained very low. All other models failed to detect any positives, and even though their specificity was high, this did not translate into meaningful predictive performance. These results suggest that CYP2C9 is particularly affected by class imbalance or lacks distinctive features that the models can learn from.

CYP2D6 showed modest improvements. The tree model again achieved the best recall and F1, while the multitask Chemprop model stood out for its higher precision. The descriptor-enhanced Chemprop model failed again, highlighting its sensitivity to imbalance. Here, multitask learning appeared to help Chemprop generalize better, though the tree model remained superior.

Finally, CYP3A4 was the isoform with the most promising results across all models. The Chemprop model with added descriptors achieved the highest recall, while the tree model delivered the strongest MCC and overall balance between precision and recall. All models performed relatively well in this case, suggesting that CYP3A4 may have a more informative or better-distributed feature space that makes the task easier to learn.

Taken together, these results show that the decision tree remains the most reliable option across isoforms, especially when working with imbalanced datasets. Chemprop models tend to overfit to the majority class when trained individually and often fail to identify positives unless enough signal or data is present. The multitask setting offers some benefit in a few cases but is not consistently superior.

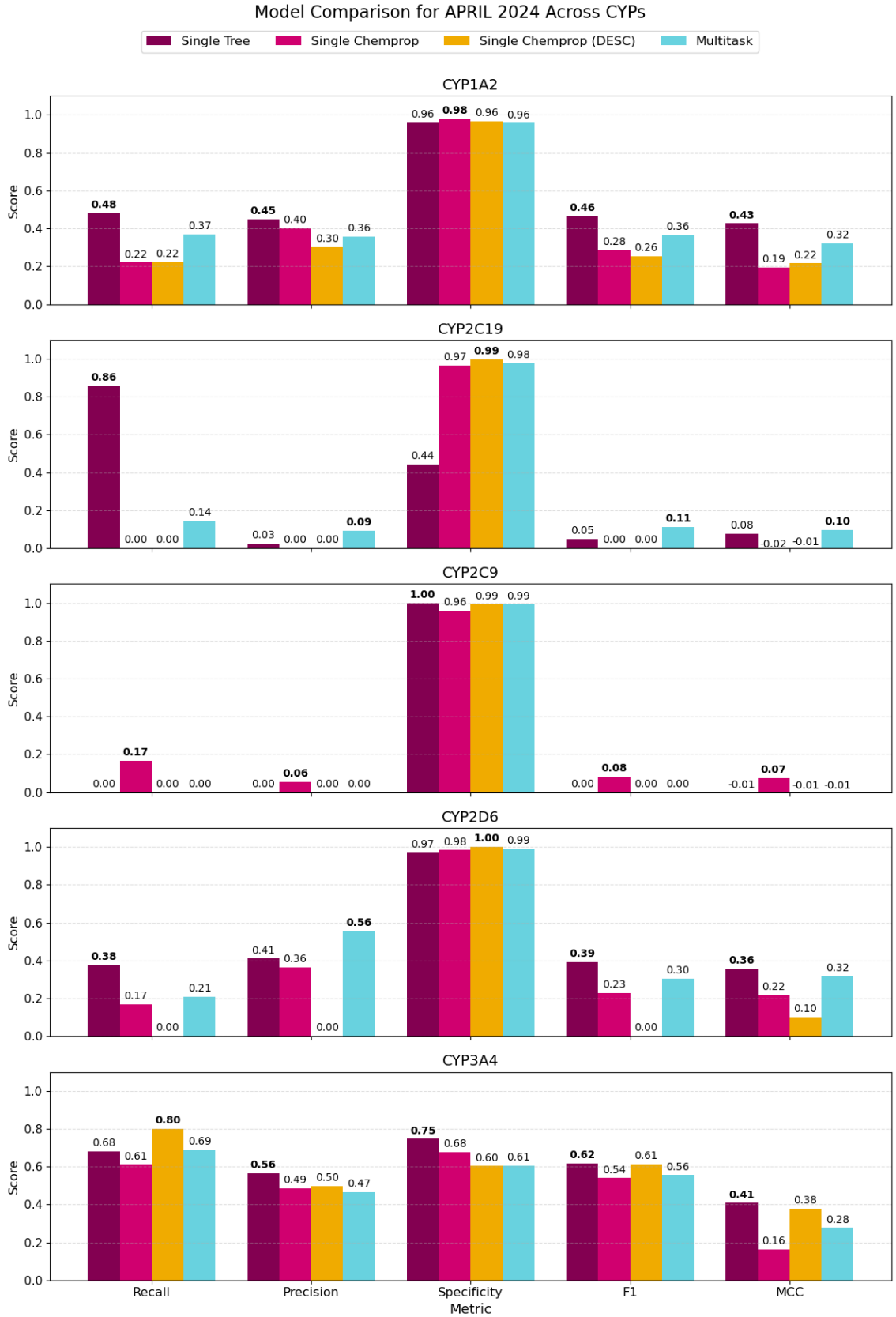


Figure 5.2: Comparison of model performance for April 2024 split across the five CYP450 isoforms. The best value per metric within each isoform is highlighted in bold.

5.1.2.2 Model Performance Summary on October 2024 Split

Figure 5.3 reports the October 2024 performance across all CYP450 isoforms, evaluated using the same metrics and modeling configurations as in the April assessment.

For CYP1A2, the single-task decision tree remained the strongest model overall. It achieved the highest recall and outperformed the others in both F1-score and MCC. The Chemprop models, especially the descriptor-enhanced version, lagged behind in recall and precision, although all maintained high specificity. Multitask Chemprop showed limited improvement and failed to match the stability of the tree-based approach.

As already observed in April 2024, CYP2C19 and CYP2C9 remained the most challenging isoforms in October. None of the models were able to detect active compounds, resulting in zero recall, precision, and F1-scores, along with negative or near-zero MCC values. Although specificity remained high, this was due to models consistently predicting the majority (inactive) class, offering no meaningful predictive power. These results confirm the extremely limited learnability of this task in its current form, possibly due to a severe lack of active compounds.

CYP2D6 showed modest improvements compared to the other isoforms. The tree and descriptor-enhanced Chemprop models achieved the highest recall values (both at 0.222), while the descriptor-based Chemprop model also had the highest F1-score and MCC among the four approaches. Interestingly, this was one of the few cases where adding descriptors to Chemprop improved performance meaningfully. However, absolute values remained relatively low, highlighting persistent difficulties in achieving strong generalization.

CYP3A4 continued to stand out with comparatively high and balanced performance. All models had recall above 0.55, and the tree model reached the highest F1-score and MCC. Although the descriptor-based Chemprop model had slightly higher recall, it showed a drop in precision. Multitask Chemprop offered competitive results, though it did not outperform the simpler single-task alternatives.

Overall, October 2024 results reinforced earlier trends. The decision tree model remained the most reliable and stable across isoforms, especially in case of extreme class imbalance. Chemprop models had some challenges, especially when trained in a single-task setting, and the inclusion of descriptors led to inconsistent gains. Multitask learning showed limited advantages, performing comparably to the single Chemprop model but rarely surpassing it.

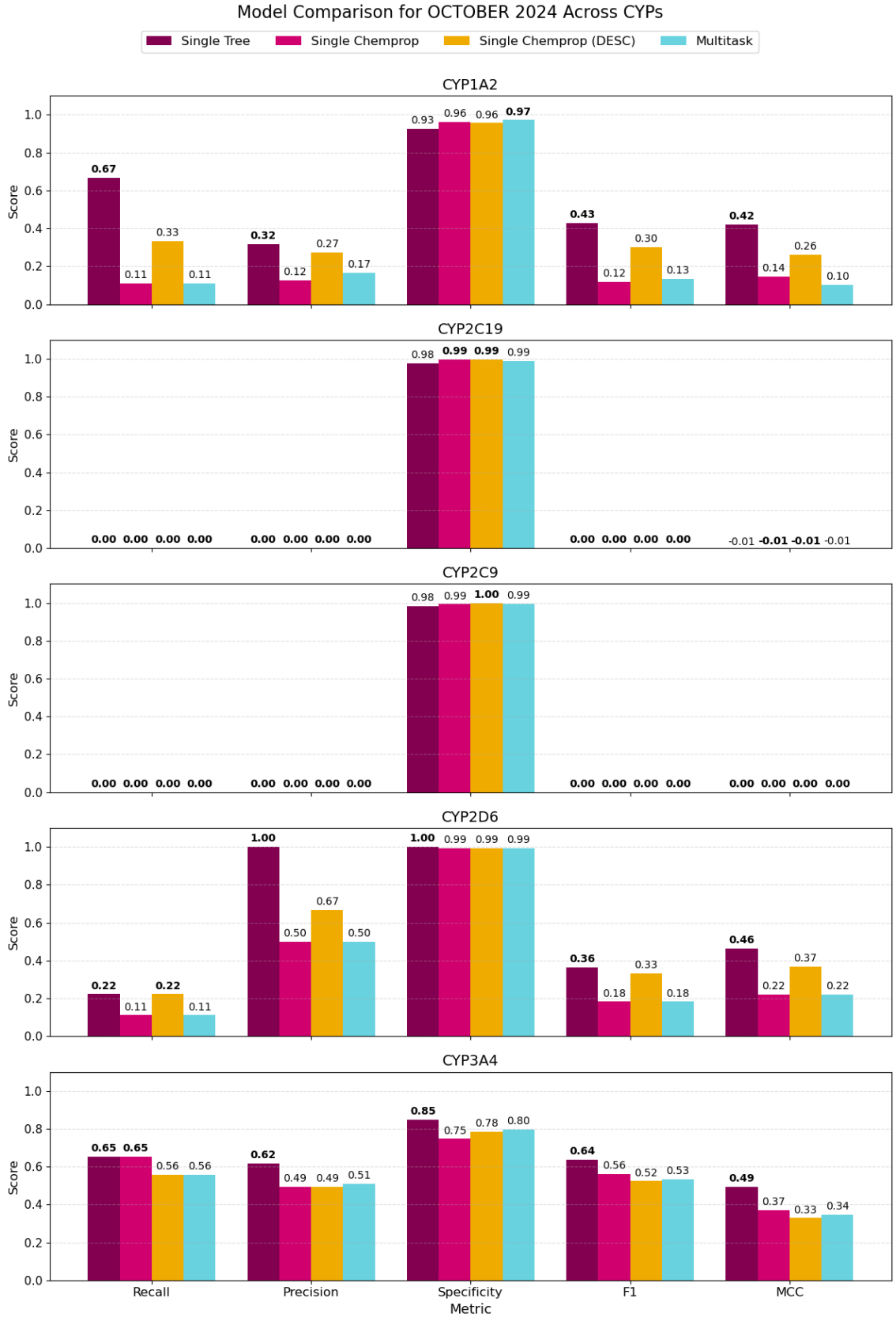


Figure 5.3: Comparison of model performance for October 2024 split across the five CYP450 isoforms. The best value per metric within each isoform is highlighted in bold.

5.1.2.3 Impact of Class Imbalance on Model Performance

The class distribution in the test sets, shown in Table 5.1, provide insights on some of the observed trends. In particular, CYP2C19 and CYP2C9 had very few positive samples (October 2024: 1 CYP2C19 active and no CYP2C9 actives; April 2024: 7 CYP2C19 actives and 6 CYP2C9 actives). This severe imbalance makes it very difficult for the models to generalize or detect active compounds, often resulting in zero recall and near-zero or negative MCC values. Even the tree-based models, which tend to handle imbalance better, were unable to capture these few positives consistently. These results underline a limitation in the dataset itself. When targets have so few active examples, the metrics become unstable and hard to interpret, regardless of the modeling approach or the descriptors used.

Table 5.1: Counts of active (1) and inactive (0) compounds per CYP450 in April and October temporal test sets.

Test set				
CYP	**April (1)**	**April (0)**	**October (1)**	**October (0)**
CYP1A2	27	400	9	180
CYP2C19	7	417	1	186
CYP2C9	6	417	0	187
CYP2D6	24	403	9	181
CYP3A4	140	287	52	138

These results also help explain why Chemprop models, despite their more complex architecture, often underperformed compared to the simpler tree-based models. Although neural networks are expected to capture subtle molecular patterns, this ability was likely limited by the datasets size and imbalance. The low number of actives in some CYP450 isoforms made generalization difficult, and recall and precision rarely balanced well. Even in the best case, CYP3A4, the F1-score reached only 0.62, while for most isoforms it stayed below 0.4.

This imbalance between recall and precision indicates that the models struggled to simultaneously detect true positives and avoid false positives. Chemprop models often leaned towards higher specificity and precision, showing a tendency to over-predict the inactive class. In extreme cases, some variants predicted only inactives, leading to high specificity but no potential to identify actives at all. In contrast, tree-based models managed to maintain a more stable trade-off, even if absolute values were not always high.

These observations raise an important question about model priorities in this context. Since no model achieved strong performance across all metrics, it becomes necessary to consider what matters most for the intended use. If the goal is to recover as many actives as possible, high recall may be preferred. If the aim is to reduce false positives, high precision and specificity are more relevant. In early screening stages, recall may

be critical to avoid missing potential candidates, while in later phases, precision may take priority. Understanding this trade-off is key when evaluating and selecting the most appropriate modeling strategy.

5.1.3 Exploratory Comparison with Random Splits

This section examines how model performance changes when using a stratified random split instead of a temporal one. Although random splits are frequently used, they tend to give overestimated results in modeling tasks. Because random splits mix compounds from different time periods, they increase the chance of including structurally similar molecules in both the training and test sets, making the task easier for the model. The goal here is to analyze how metrics such as MCC, recall, and precision behave under this easier setup and to compare the impact across models and isoforms.

Figure 5.4 compares results for October 2024 using temporal and random splits across five CYP450 isoforms, focusing on single-task models (decision tree and Chemprop). In all cases, performance improved with the random split, as shown by higher MCC scores. The gains were especially marked in Chemprop models, particularly for CYP2C19 and CYP2C9, where performance increased substantially. These isoforms had very few actives in the temporal test sets, so even small distribution shifts led to large metric differences. The stratified random split ensured a more even distribution of actives, improving their recovery and boosting performance.

The difference between temporal and random splits was also evident in the decision tree models, although the improvements were less dramatic. In CYP1A2, CYP2D6, and CYP3A4, the tree models already performed relatively well with the temporal split, but still reached MCC values of 0.5 or higher under the random setting. This indicates that the models benefited from the more favorable data distribution, even if the relative change was smaller than in Chemprop. Chemprop, in contrast, showed greater sensitivity to the partitioning strategy, likely because it relies more on the presence of structurally similar compounds during training.

These results are consistent with previous observations. The low performance of Chemprop in temporal splits is not strictly due to its architecture, but rather to its sensitivity to class imbalance and the limited chemical overlap between training and test sets. Tree-based models appear more robust in this regard, possibly due to their ability to handle small and imbalanced datasets more effectively. The extreme case of CYP2C9, with zero actives in the October 2024 temporal test set, further illustrates how random splitting can inflate model performance by ensuring that a few actives are always available during training.

Although random splits help reveal model potential under ideal conditions, they do not reflect how these models would behave in real-world scenarios. For this reason, the rest of the analysis focuses only on temporal splits, which offer a more realistic and challenging evaluation framework.

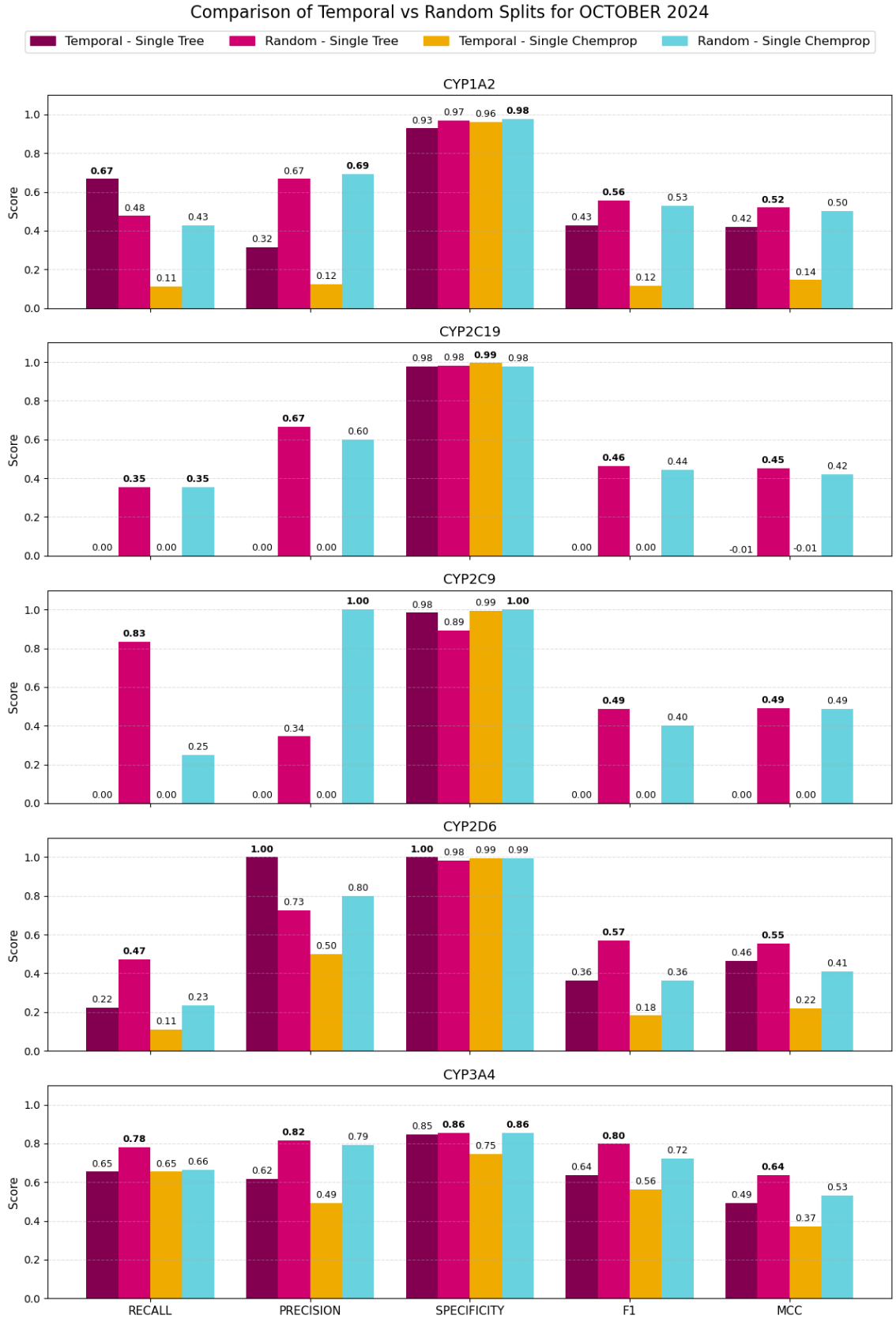


Figure 5.4: Comparison of model performance for October 2024 across the five CYP450 isoforms with temporal and random splits. The best value per metric within each isoform is highlighted in bold.

5.1.4 Effect of Resampling Techniques

After selecting tree-based models as the most reliable approach, this section evaluates whether applying resampling techniques improves their performance considering class imbalance. Specifically, SMOTE and RUS were tested at different sampling rates, and their impact was assessed using MCC as the main metric. Figure 5.5 shows the results for both April and October 2024 splits across all CYP450 isoforms.

For CYP1A2, some improvement was observed, particularly in April 2024. The best result was achieved with SMOTE at a 0.5 ratio, increasing the active class to reach half the size of the inactive class. This configuration led to a higher MCC compared to the baseline, suggesting that moderate oversampling can help in this case. In contrast, October results were more mixed, with some configurations underperforming the original model.

In CYP2C19 and CYP2C9, resampling had no positive effect. Across all configurations and both splits, MCC remained low or negative, indicating that the models were still unable to recover true positives. These targets continue to be highly challenging, and balancing the data did not meaningfully improve predictive power.

For CYP2D6, several resampling configurations led to clear performance gains. The highest MCC was observed in October 2024 using RUS at a 0.5 ratio, reaching 0.58 compared to 0.46 in the original model. April results also showed improvement in some settings, supporting the idea that resampling can be helpful when the imbalance is moderate but not extreme.

In the case of CYP3A4, most resampling configurations could not be applied due to the relatively small class imbalance. Since the number of active compounds was already close to the number of inactives, applying SMOTE or RUS at lower ratios like 0.5 was not feasible. As a result, only the 1:1 configurations were tested. These maintained MCC values similar to or slightly better than the original model, suggesting that further balancing was neither necessary nor helpful for this target.

Overall, resampling techniques did not produce the expected impact on model performance. While there were some improvements in specific cases, especially in CYP1A2 and CYP2D6, the results were inconsistent across isoforms. One possible explanation is that synthetic samples generated by SMOTE may not accurately reflect the original data distribution, leading to noise rather than signal. Similarly, reducing the dataset through undersampling may discard useful information, limiting the models ability to generalize. These limitations suggest that resampling alone is not sufficient to address the challenges posed by imbalanced and sparse datasets in this context.

Figure 5.5: Model performance across different SMOTE and RUS sampling ratios compared to the original data. The best MCC value per split (April 2024 and October 2024) and CYP isoform is highlighted in bold. Missing bars indicate configurations where model training was not feasible due to data distribution incompatibilities.

5.1.5 Overview of Model Limitations

Despite testing different model architectures, training strategies, and resampling techniques, performance remained limited for several CYP targets. However, in many cases these limitations could be attributed to extreme class imbalance or lack of data. In contrast, CYP3A4 had a more balanced dataset and consistently higher overall performance, yet the models still failed to reach strong or consistent metrics.

For this reason, the following analysis focuses specifically on CYP3A4. The goal is to investigate potential reasons behind these limitations by exploring two main aspects: the informativeness of the molecular descriptors and the structure of the chemical space defined by ECFP fingerprints. Together, these perspectives aim to shed light on why the models lacked the potential to generalize, even in a more favorable data scenario.

5.1.5.1 Descriptor-Based Study

To better understand why model performance remains limited even in CYP3A4, the best-performing decision tree model based on the April 2024 temporal split was analyzed. This model used RDKit descriptors, making it suitable for exploring the relevance of input features in classification. Based on the models feature importance, the six most informative descriptors were selected and visualized according to their distribution in actives and inactives (Figure 5.6).

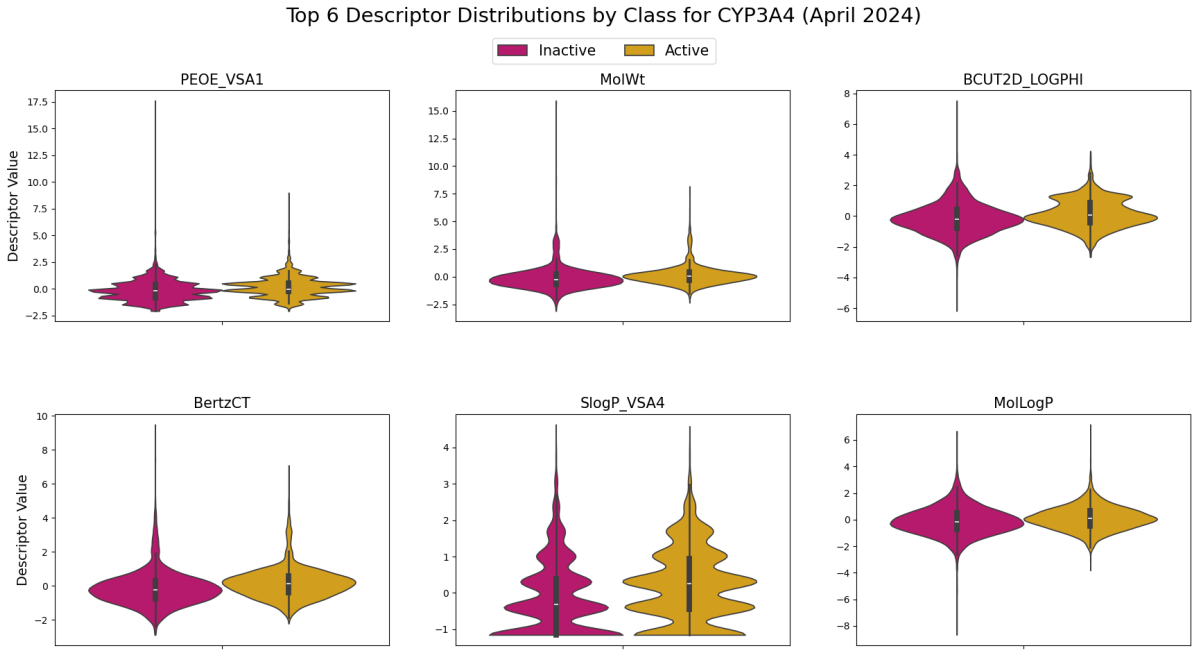


Figure 5.6: Top six descriptor distributions for active and inactive compounds in CYP3A4 tree-based models on April 2024 temporal dataset .

Overall, the violin plots reveal a weak separation between classes. While some descriptors, such as `PEOE_VSA1` or `MolWt`, show slight shifts in their distributions,

actives and inactives still largely overlap. For a descriptor to be informative, we would expect to see more distinct, non-overlapping distributions between classes, or at least clear differences in central tendency or spread. In this case, the overlap suggests that these features alone are not sufficient to define consistent decision boundaries.

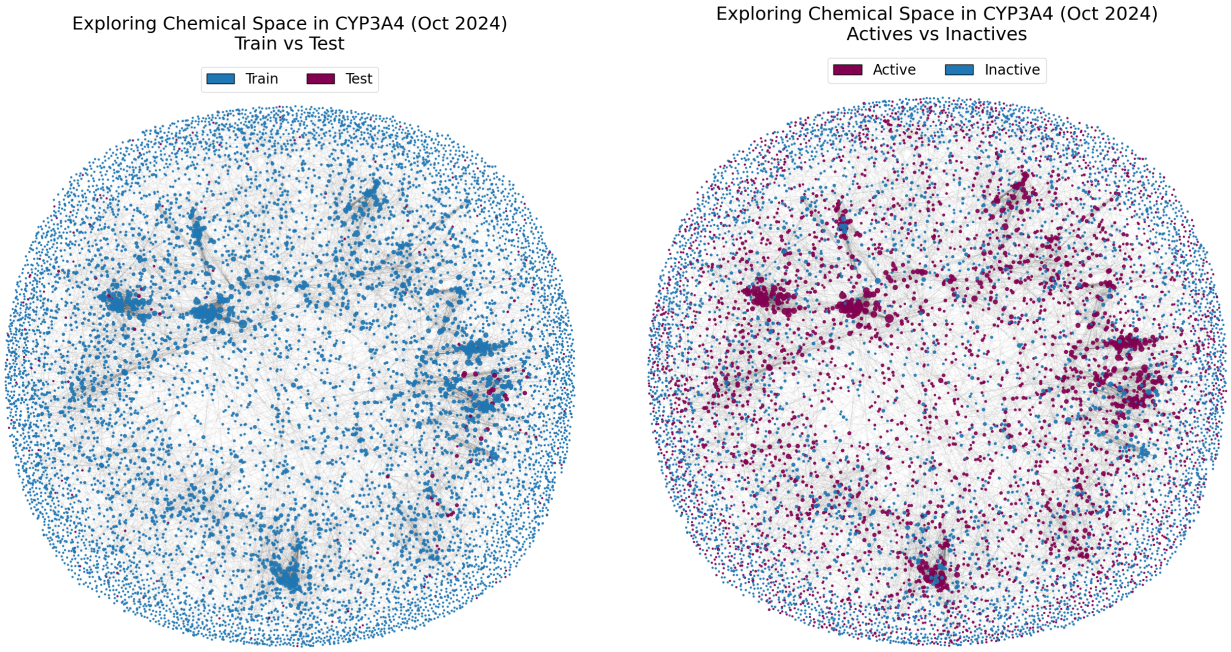
In addition to RDKit descriptors, other molecular representations were also tested, including Mordred descriptors and physicochemical properties such as pKa and logD. However, these alternatives did not lead to better models. One possible reason is that Mordred generates a very high number of descriptors, which can introduce noise and redundancy, especially with limited data. On the other hand, while pKa and logD are chemically meaningful, they may not provide enough structural variation across the dataset to help the model differentiate actives from inactives effectively.

These observations support the idea that, even in a relatively balanced dataset like CYP3A4, the descriptors used may not capture the relevant patterns for classification. This limits the learning potential of the models and highlights the need to explore alternative representations or additional sources of information in future work.

5.1.5.2 ECFP Representations and Chemical Space Exploration

To further investigate the limitations in terms of model performance, especially for CYP3A4, we analyzed the underlying structure of the chemical space using ECFP fingerprints. This analysis focused on the CYP3A4 model from October 2024, which was trained using ECFP representations. A similarity network was constructed based on Tanimoto coefficients and visualized using the Fruchterman-Reingold algorithm, as shown in Figure 5.7. This network offers a way to examine molecular similarity and neighborhood structure, although it represents only a partial view of chemical space. Other factors not captured by ECFP, such as 3D conformation or electronic properties, may also influence model predictions. Two views of the network were considered: one showing the distribution of training versus test compounds, and another comparing actives and inactives.

The network colored by data split shows that many test compounds lie in peripheral or disconnected regions of the graph, with limited overlap with the training set. This suggests that the model faces a domain shift in the test set: molecules to be predicted often differ structurally from those seen during training, which may explain the drop in performance under temporal splits. In contrast, the actives vs inactives view reveals that although some clusters of actives are visible, they are not clearly separated from inactives. This indicates that ECFP similarity does not yield chemically homogeneous regions aligned with biological activity, reducing the discriminative potential of the representation.



(a) Training vs Test.

(b) Actives vs Inactives.

Figure 5.7: Chemical space networks for CYP3A4 in the October 2024 dataset, constructed using Tanimoto similarity and visualized with Fruchterman-Reingold algorithm.

To quantify these patterns, we analyzed the prediction outcomes by grouping test compounds according to their connectivity within the chemical similarity network, shown in Figure 5.8. Four categories were defined: compounds connected only to the training set (test compounds with 70% inter-similarity), connected to both training and test sets (intra- and inter-similarity), connected only to other test compounds (compounds only with intra-similarity), and completely isolated nodes with no neighbors. One of the most striking observations was that nearly 44% of the test compounds fell into the isolated category. These molecules showed no measurable similarity to any compound in the training or test set based on the Tanimoto threshold. As expected, this group had the poorest performance: recall and F1-score were both zero, and MCC was negative. The model predicted all of them as inactive, which artificially increased specificity but failed to identify any true positives. In the absence of structural context, the model defaulted to the majority class, revealing the lack of generalizability to unfamiliar regions of chemical space.

The best results came from compounds connected to both the training and test sets. This group had the highest recall (96.2%) and the most true positives (25), which suggests that seeing similar structures during training helped the model identify actives. However, it also generated the most false positives (12), indicating that the model may have learned some relevant patterns but applies them too broadly, predicting activity even in similar but inactive compounds.

Compounds connected only to the training set behaved differently. In this group,

the model achieved perfect recall (1.0), identifying all actives, but at the cost of low precision (0.5). This suggests that the model learns to predict actives based on similar structural patterns captured by ECFP fingerprints. However, predicting a compound as active/inactive needs not be necessarily driven by structural features, but also other physicochemical properties can have an influence.

Perhaps the most surprising result came from compounds connected only to other test compounds. Despite having no link to the training data, this group still showed moderate MCC and F1 scores. However, a closer look reveals that this was largely driven by correctly predicting inactives. The model may simply be defaulting to the negative class in these regions, which is easier when majority of the compounds are inactive. So, while the metrics look reasonable, they do not reflect meaningful learning about activity.

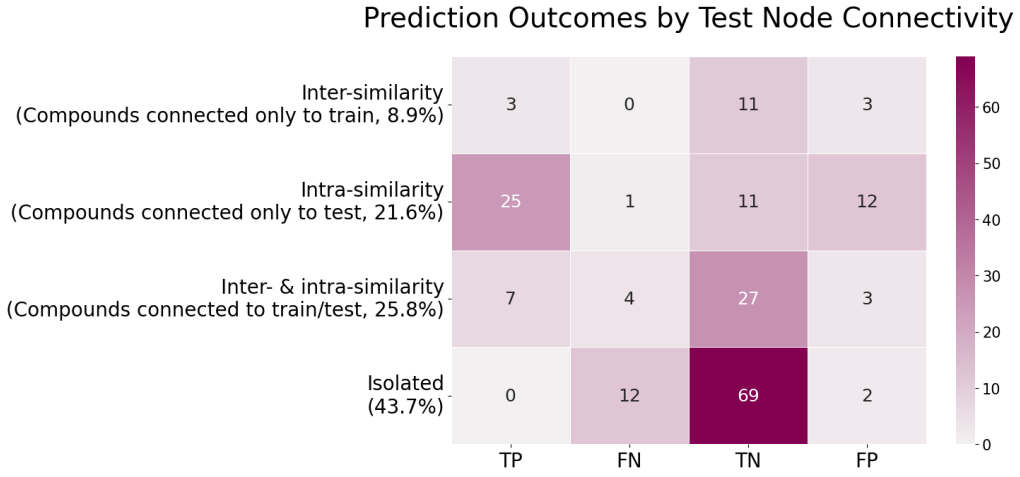


Figure 5.8: Distribution of true/false positives and negatives across CYP3A4 test compound groups based on their chemical connectivity.

In summary, this analysis highlights that model performance is strongly influenced by the presence of structurally similar examples in the training set, at least as defined by ECFP and Tanimoto similarity. When the test compounds are disconnected from the training space, especially in structurally novel areas, the model has difficulties in generalization. And even when some structural similarity is present, distinguishing actives from inactives remains challenging. These results suggest that expanding chemical diversity in the training data or adopting representations that generalize better across different regions of chemical space could help improve model robustness.

5.2 Trapping Assays Results

One of the major drivers of hepatic safety is the potential of a molecule to generate reactive metabolites. So, we investigated the possibilities to predict the outcomes of the three trapping assays: GSH, KCN, and MA. These assays capture distinct mechanisms of chemical reactivity, providing a way to assess whether models can

generalize across endpoints that exhibit different mechanism of action. Figure 5.9 compares the performance of five models: a single-task decision tree (Single Tree), a single-task Chemprop model without descriptors (Single Chemprop), its descriptor-enhanced variant (Single Chemprop (DESC)), and a multitask Chemprop model (Multitask). For clarity, only the best-performing tree configuration is shown; full results are available in section A.2.

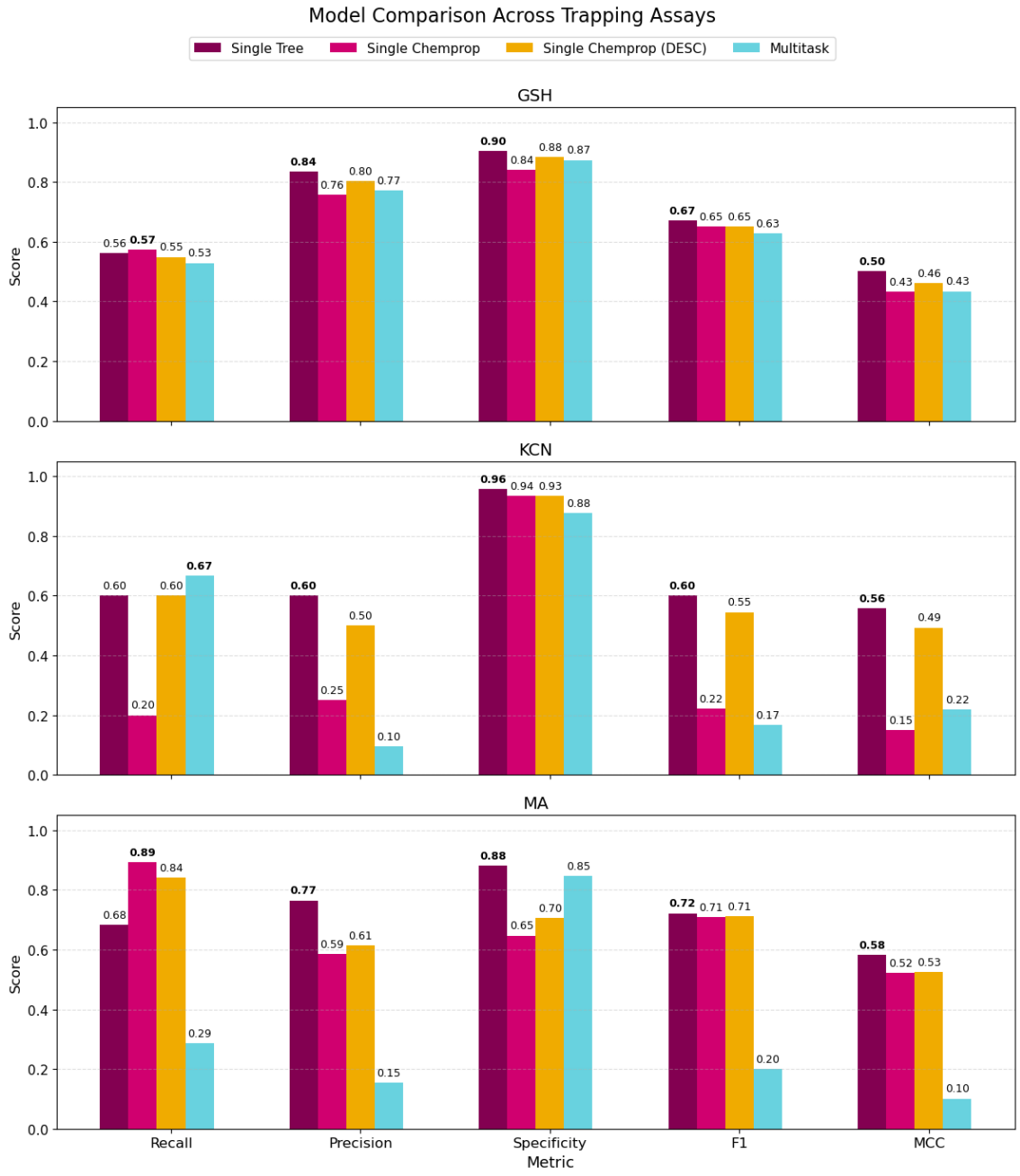


Figure 5.9: Performance comparison of single-task and multitask models across the three trapping assays (GSH, KCN, MA) using five evaluation metrics. The best value per metric within each endpoint is highlighted in bold.

Across all assays, the tree-based single-task model consistently delivered the most balanced performance. It achieved the highest F1 and MCC scores for the three endpoints, reflecting a strong ability to capture both sensitivity and precision while minimizing false predictions. In GSH and MA, it outperformed all Chemprop variants, particularly in specificity and an overall stability was observed across metrics.

The single-task Chemprop models showed reasonable performance, particularly when using RDKit descriptors. However, their precision and MCC scores were generally lower than those of the tree model. This suggests that, while Chemprop is able to predict more active compounds in some cases, it does so at the cost of introducing more false positives.

Interestingly, the multitask models exhibited consistently high recall but very low precision. This trade-off resulted in poor MCC values, especially for the MA and KCN assays. These results were expected to some extent, as the three trapping assays are biologically distinct and likely involve different mechanisms of reactivity. Training a single multitask model to simultaneously predict endpoints with limited shared signal can lead to overgeneralization. In this case, the model appears to favor sensitivity over specificity, producing a high number of false positives.

These observations highlight a key limitation of multitask learning in this context. While multitask models can be advantageous when tasks are closely related or share common patterns, their performance can deteriorate when the endpoints differ substantially, as seen here. Thus, for the prediction of trapping assays with heterogeneous biological mechanisms, single-task models may provide more robust and reliable predictions.

Overall, the tree-based single-task approach emerged as the most consistent and interpretable option across all three assays. The performance of Chemprop-based models remains promising, especially when tailored to single endpoints, but their generalization across unrelated assays remains a challenge.

6

Conclusion

Throughout this work, several modeling strategies were explored to predict toxicity-related endpoints involving CYP450 inhibition and trapping assays. Although tree-based models consistently emerged as the most stable and effective approach, their overall performance remained modest and, in many cases, insufficient for fully reliable application. Chemprop models, in both single-task and multitask configurations, showed more variable behavior and were particularly sensitive to dataset sizes. Their performance was hindered by severe class imbalance, the scarcity of active compounds, and limited structural diversity in the training data. While neural networks should, in principle, be capable of capturing more complex patterns, this potential was not reflected in the results, likely due to the low signal contained in the molecular descriptors and the poor coverage of chemical space.

Resampling techniques such as SMOTE and RUS led to only marginal improvements in isolated cases. Overall, they were not sufficient to overcome the limitations imposed by the original data distribution. In CYP450s such as CYP2C19 and CYP2C9, where active compounds were extremely scarce, models consistently failed to identify positives even after resampling. This suggests that when available data lacks representativeness, neither class balancing nor model architecture can compensate for it. In the case of CYP3A4, where class imbalance was less severe, performance still plateaued. Neither resampling, nor usage of alternative descriptors, provided a substantial boost. Feature importance analysis showed that even the most relevant descriptors failed to clearly separate classes, and that many test molecules were structurally disconnected from the training set, limiting the model's ability to generalize.

Unlike the CYP models, the models on trapping assays showed a better predictive potential in a single-task set up. Although multitask learning was expected to help capture shared signals across endpoints, this benefit did not materialize. Instead, multitask models displayed high recall but very low precision, indicating a tendency to overpredict the active class. This behavior may stem from the biological heterogeneity between the assays (GSH, KCN, and MA), whose underlying mechanisms are likely not aligned for joint learning. Once again, tree-based models showed the most balanced behavior, achieving a better trade-off between sensitivity and specificity.

Taken together, these findings suggest that performance limitations are driven more by the dataset than by the models themselves. Class imbalance, limited chemical coverage, and underinformative representations present substantial challenges that

are not easily addressed through architecture or sampling alone. Nevertheless, adopting scaffold reduction and alternative splitting strategies such as random partitions, offer a reasonable boost to the overall performance of the CYP isoform models.

Ultimately, these results highlight the importance of not only developing robust modeling strategies but also improving how molecular datasets are curated and represented. Expanding chemical diversity, incorporating richer structural or biological information, and tailoring models to the specific characteristics of each endpoint will be essential steps toward achieving more accurate and generalizable toxicity predictions in practical settings.

In summary, it is well-known that the data availability is quite limited for toxicity-related end points and this in fact is reflected in the modeling outcomes. Irrespective of the challenges in making accurate predictions for all compounds, the advantage of the models in predicting chemical spaces similar to the training set should be highlighted. We believe that these models have the potential to flag early safety risks and can guide compound prioritization in projects.

7

Future work

Several directions emerge as promising extensions of this study. First, future efforts could explore additional modeling strategies, including more recent or specialized machine learning architectures, to better capture the complexity of molecular toxicity. Given the limited performance observed for TDI, a logical next step would be to expand the modeling framework to include general CYP450 inhibition. This endpoint benefits from larger and more diverse datasets, which could help identify more consistent predictive patterns and support more robust model development.

For TDI in particular, a deeper investigation into the structural or physicochemical features that distinguish active from inactive compounds could provide valuable insights. This may involve the identification of structural alerts, reactivity patterns, or substructure-based rules that are associated with time-dependent inhibition. Combining this knowledge with existing features could improve both model interpretability and performance.

In case of trapping assays, further improvements may be achieved by applying advanced data scaling techniques and conducting a more in-depth analysis of molecular properties. Understanding which physicochemical features correlate with chemical reactivity could help guide model design and feature engineering.

Additionally, it would be valuable to validate the current models using external public datasets. This step is essential for assessing the generalization capacity of the models beyond the internal temporal splits used throughout the study. Evaluating performance on unseen, independently collected data would provide a more realistic estimate of their potential in real-world screening applications. It would also help uncover potential sources of overfitting or dataset-specific biases that may have gone unnoticed during internal validation.

Altogether, future work should focus on expanding both the data landscape and the methodological tools to enhance the reliability of toxicity prediction models. This includes increasing the chemical and biological diversity of the training data, improving molecular representations, incorporating domain-specific knowledge, and tailoring modeling strategies to the characteristics of each endpoint. Together, these efforts could lead to more robust, interpretable, and generalizable models suitable for deployment in practical drug development workflows.

Bibliography

[1] R. Allison, A. Guraka, I. T. Shawa, G. Tripathi, W. Moritz, and A. Kermanizadeh, "Drug induced liver injury a 2023 update," *Journal of Toxicology and Environmental Health, Part B*, vol. 26, no. 8, pp. 442–467, Oct. 2023, ISSN: 1521-6950. DOI: 10.1080/10937404.2023.2261848. [Online]. Available: http://dx.doi.org/10.1080/10937404.2023.2261848.

[2] A. Srivastava, J. L. Maggs, D. J. Antoine, D. P. Williams, D. A. Smith, and B. K. Park, "Role of reactive metabolites in drug-induced hepatotoxicity," in *Adverse Drug Reactions*. Springer Berlin Heidelberg, Sep. 2009, pp. 165–194, ISBN: 9783642006630. DOI: 10.1007/978-3-642-00663-0_7. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-00663-0_7.

[3] A. V. Stachulski, T. A. Baillie, B. Kevin Park, *et al.*, "The generation, detection, and effects of reactive drug metabolites: Reactive drug metabolites," *Medicinal Research Reviews*, vol. 33, no. 5, pp. 985–1080, Oct. 2012, ISSN: 0198-6325. DOI: 10.1002/med.21273. [Online]. Available: http://dx.doi.org/10.1002/med.21273.

[4] D. E. V. Pires, T. L. Blundell, and D. B. Ascher, "Pkcsm: Predicting small-molecule pharmacokinetic and toxicity properties using graph-based signatures," *Journal of Medicinal Chemistry*, vol. 58, no. 9, pp. 4066–4072, Apr. 2015, ISSN: 1520-4804. DOI: 10.1021/acs.jmedchem.5b00104. [Online]. Available: http://dx.doi.org/10.1021/acs.jmedchem.5b00104.

[5] B. Oso, O. Agboola, A. Awotula, I. Olaoye, G. Akhigbe, and E. Nwokeocha, "Predictive analysis of the pharmacokinetic and toxicological endpoints of naphthalene and its derivatives," Mar. 2021. DOI: 10.21203/rs.3.rs-295918/v1. [Online]. Available: http://dx.doi.org/10.21203/rs.3.rs-295918/v1.

[6] J. Cremer, L. Medrano Sandonas, A. Tkatchenko, D.-A. Clevert, and G. De Fabritiis, "Equivariant graph neural networks for toxicity prediction," *Chemical Research in Toxicology*, Sep. 2023, ISSN: 1520-5010. DOI: 10.1021/acs.chemrestox.3c00032. [Online]. Available: http://dx.doi.org/10.1021/acs.chemrestox.3c00032.

[7] A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, "Large-scale comparison of machine learning methods for drug target prediction on chembl," *Chemical Science*, vol. 9, no. 24, pp. 5441–5451, 2018. DOI: 10.1039/C8SC00148K.

[8] Z. Wu, B. Ramsundar, E. N. Feinberg, *et al.*, "Moleculenet: A benchmark for molecular machine learning," *Chemical Science*, vol. 9, no. 2, pp. 513–530, 2018. DOI: 10.1039/C7SC02664A.

[9] B. Ramsundar, S. Kearnes, P. Riley, D. Webster, D. Konerding, and V. Pande, "Massively multitask networks for drug discovery," *arXiv preprint arXiv:1502.02072*, 2015.

[10] C. Moon and D. Kim, "Prediction of drugtarget interactions through multi-task learning," *Scientific Reports*, vol. 12, no. 1, Oct. 2022, ISSN: 2045-2322. DOI: 10.1038/s41598-022-23203-y. [Online]. Available: http://dx.doi.org/10.1038/s41598-022-23203-y.

[11] J. V. Castell, R. Jover, C. P. Martnez-Jimnez, and M. J. Gmez-Lechn, "Hepatocyte cell lines: Their use, scope and limitations in drug metabolism studies," *Expert Opinion on Drug Metabolism amp; Toxicology*, vol. 2, no. 2, pp. 183–212, Mar. 2006, ISSN: 1744-7607. DOI: 10.1517/17425255.2.2.183. [Online]. Available: http://dx.doi.org/10.1517/17425255.2.2.183.

[12] Z. Zhang and W. Tang, "Drug metabolism in drug discovery and development," *Acta Pharmaceutica Sinica B*, vol. 8, no. 5, pp. 721–732, Sep. 2018, ISSN: 2211-3835. DOI: 10.1016/j.apsb.2018.04.003. [Online]. Available: http://dx.doi.org/10.1016/j.apsb.2018.04.003.

[13] T. Lynch and A. Price, "The effect of cytochrome p450 metabolism on drug response, interactions, and adverse effects," en, *American family physician*, vol. 76, no. 3, pp. 391–396, 2007, ISSN: 0002-838X.

[14] A.-C. LUCA, "Mini-review on the implications of gene polymorphism in the metabolism of xenobiotics," *FARMACIA*, vol. 70, no. 4, pp. 573–582, Jun. 2022, ISSN: 0014-8237. DOI: 10.31925/farmacia.2022.4.1. [Online]. Available: http://dx.doi.org/10.31925/farmacia.2022.4.1.

[15] U. M. Zanger and M. Schwab, "Cytochrome p450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation," *Pharmacology amp; Therapeutics*, vol. 138, no. 1, pp. 103–141, Apr. 2013, ISSN: 0163-7258. DOI: 10.1016/j.pharmthera.2012.12.007. [Online]. Available: http://dx.doi.org/10.1016/j.pharmthera.2012.12.007.

[16] K. Thelen and J. B. Dressman, "Cytochrome p450-mediated metabolism in the human gut wall," *Journal of Pharmacy and Pharmacology*, vol. 61, no. 5, pp. 541–558, May 2009, ISSN: 2042-7158. DOI: 10.1211/jpp.61.05.0002. [Online]. Available: http://dx.doi.org/10.1211/jpp.61.05.0002.

[17] J. Hukkanen, O. Pelkonen, J. Hakkola, and H. Raunio, "Expression and regulation of xenobiotic-metabolizing cytochrome p450 (cyp) enzymes in human lung," *Critical Reviews in Toxicology*, vol. 32, no. 5, pp. 391–411, Jan. 2002, ISSN: 1547-6898. DOI: 10.1080/20024091064273. [Online]. Available: http://dx.doi.org/10.1080/20024091064273.

[18] X. Zhao and J. Imig, "Kidney cyp450 enzymes: Biological actions beyond drug metabolism," *Current Drug Metabolism*, vol. 4, no. 1, pp. 73–84, Feb. 2003, ISSN: 1389-2002. DOI: 10.2174/1389200033336892. [Online]. Available: http://dx.doi.org/10.2174/1389200033336892.

[19] W. Kuban and W. A. Daniel, "Cytochrome p450 expression and regulation in the brain," *Drug Metabolism Reviews*, vol. 53, no. 1, pp. 1–29, Dec. 2020, ISSN: 1097-9883. DOI: 10.1080/03602532.2020.1858856. [Online]. Available: http://dx.doi.org/10.1080/03602532.2020.1858856.

[20] S. Ali, "Phenoconversion and in vivo phenotyping of hepatic cytochrome p450: Implications in predictive precision medicine and personalized therapy," *Hepatology Forum*, 2024, ISSN: 1307-5888. DOI: 10.14744/hf.2023.2023.0047. [Online]. Available: http://dx.doi.org/10.14744/hf.2023.2023.0047.

[21] D. R. NELSON, T. KAMATAKI, D. J. WAXMAN, *et al.*, "The p450 superfamily: Update on new sequences, gene mapping, accession numbers, early trivial names of enzymes, and nomenclature," *DNA and Cell Biology*, vol. 12, no. 1, pp. 1–51, Jan. 1993, ISSN: 1557-7430. DOI: 10.1089/dna.1993.12.1. [Online]. Available: http://dx.doi.org/10.1089/dna.1993.12.1.

[22] T. S. Tracy, A. S. Chaudhry, B. Prasad, *et al.*, "Interindividual variability in cytochrome p450mediated drug metabolism," *Drug Metabolism and Disposition*, vol. 44, no. 3, pp. 343–351, Mar. 2016, ISSN: 0090-9556. DOI: 10.1124/dmd.115.067900. [Online]. Available: http://dx.doi.org/10.1124/dmd.115.067900.

[23] P. Sulem, D. F. Gudbjartsson, F. Geller, *et al.*, "Sequence variants at cyp1a1cyp1a2 and ahr associate with coffee consumption," *Human Molecular Genetics*, vol. 20, no. 10, pp. 2071–2077, Feb. 2011, ISSN: 0964-6906. DOI: 10.1093/hmg/ddr086. [Online]. Available: http://dx.doi.org/10.1093/hmg/ddr086.

[24] M. Ingelman-Sundberg and S. C. Sim, "Intronic polymorphisms of cytochromes p450," *Human Genomics*, vol. 4, no. 6, p. 402, 2010, ISSN: 1479-7364. DOI: 10.1186/1479-7364-4-6-402. [Online]. Available: http://dx.doi.org/10.1186/1479-7364-4-6-402.

[25] V. M. Pratt, L. H. Cavallari, A. L. Del Tredici, *et al.*, "Recommendations for clinical warfarin genotyping allele selection," *The Journal of Molecular Diagnostics*, vol. 22, no. 7, pp. 847–859, Jul. 2020, ISSN: 1525-1578. DOI: 10.1016/j.jmoldx.2020.04.204. [Online]. Available: http://dx.doi.org/10.1016/j.jmoldx.2020.04.204.

[26] U. Klotz, M. Schwab, and G. Treiber, "Cyp2c19 polymorphism and proton pump inhibitors," *Basic amp; Clinical Pharmacology amp; Toxicology*, vol. 95, no. 1, pp. 2–8, Jul. 2004, ISSN: 1742-7843. DOI: 10.1111/j.1600-0773.2004.pto950102.x. [Online]. Available: http://dx.doi.org/10.1111/j.1600-0773.2004.pto950102.x.

[27] A. N. Werk and I. Cascorbi, "Functional gene variants of cyp3a4," *Clinical Pharmacology amp; Therapeutics*, vol. 96, no. 3, pp. 340–348, Jun. 2014, ISSN: 1532-6535. DOI: 10.1038/clpt.2014.129. [Online]. Available: http://dx.doi.org/10.1038/clpt.2014.129.

[28] J. D. Lutz and N. Isoherranen, "In vitro-to-in vivopredictions of drugdrug interactions involving multiple reversible inhibitors," *Expert Opinion on Drug Metabolism amp; Toxicology*, vol. 8, no. 4, pp. 449–466, Mar. 2012, ISSN: 1744-7607. DOI: 10.1517/17425255.2012.667801. [Online]. Available: http://dx.doi.org/10.1517/17425255.2012.667801.

[29] M. Mohutsky and S. D. Hall, "Irreversible enzyme inhibition kinetics and drugdrug interactions," in *Enzyme Kinetics in Drug Metabolism*. Springer US, 2021, pp. 51–88, ISBN: 9781071615546. DOI: 10.1007/978-1-0716-1554-6_3. [Online]. Available: http://dx.doi.org/10.1007/978-1-0716-1554-6_3.

[30] R. J. Riley, K. Grime, and R. Weaver, "Time-dependent cyp inhibition," *Expert Opinion on Drug Metabolism amp; Toxicology*, vol. 3, no. 1, pp. 51–66, Feb. 2007, ISSN: 1744-7607. DOI: 10.1517/17425255.3.1.51. [Online]. Available: http://dx.doi.org/10.1517/17425255.3.1.51.

[31] B.-F. Krippendorff, R. Neuhaus, P. Lienau, A. Reichel, and W. Huisinga, "Mechanism-based inhibition: Deriving ki and kinact directly from time-dependent ic50 values," *SLAS Discovery*, vol. 14, no. 8, pp. 913–923, Sep. 2009, ISSN: 2472-5552. DOI: 10.1177/1087057109336751. [Online]. Available: http://dx.doi.org/10.1177/1087057109336751.

[32] S. Zhou, S. Yung Chan, B. Cher Goh, *et al.*, "Mechanism-based inhibition of cytochrome p450 3a4 by therapeutic drugs," *Clinical Pharmacokinetics*, vol. 44, no. 3, pp. 279–304, 2005, ISSN: 0312-5963. DOI: 10.2165/00003088-200544030-00005. [Online]. Available: http://dx.doi.org/10.2165/00003088-200544030-00005.

[33] C. Lu and L. Di, "In vitro and in vivo methods to assess pharmacokinetic drug drug interactions in drug discovery and development," *Biopharmaceutics amp; Drug Disposition*, vol. 41, no. 12, pp. 3–31, Jan. 2020, ISSN: 1099-081X. DOI: 10.1002/bdd.2212. [Online]. Available: http://dx.doi.org/10.1002/bdd.2212.

[34] S.-M. Huang, J. M. Strong, L. Zhang, *et al.*, "New era in drug interaction evaluation: Us food and drug administration update on cyp enzymes, transporters, and the guidance process," *The Journal of Clinical Pharmacology*, vol. 48, no. 6, pp. 662–670, Jun. 2008, ISSN: 1552-4604. DOI: 10.1177/0091270007312153. [Online]. Available: http://dx.doi.org/10.1177/0091270007312153.

[35] T. Prueksaritanont, X. Chu, C. Gibson, *et al.*, "Drugdrug interaction studies: Regulatory guidance and an industry perspective," *The AAPS Journal*, vol. 15, no. 3, pp. 629–645, Mar. 2013, ISSN: 1550-7416. DOI: 10.1208/s12248-013-9470-x. [Online]. Available: http://dx.doi.org/10.1208/s12248-013-9470-x.

[36] K. Samuel, W. Yin, R. A. Stearns, *et al.*, "Addressing the metabolic activation potential of new leads in drug discovery: A case study using ion trap mass spectrometry and tritium labeling techniques," *Journal of Mass Spectrometry*, vol. 38, no. 2, pp. 211–221, Jan. 2003, ISSN: 1096-9888. DOI: 10.1002/jms.434. [Online]. Available: http://dx.doi.org/10.1002/jms.434.

[37] H. Sies, "Glutathione and its role in cellular functions," *Free Radical Biology and Medicine*, vol. 27, no. 910, pp. 916–921, Nov. 1999, ISSN: 0891-5849. DOI: 10.1016/s0891-5849(99)00177-x. [Online]. Available: http://dx.doi.org/10.1016/S0891-5849(99)00177-X.

[38] K. Hagihara, M. Kazui, A. Kurihara, T. Ikeda, and T. Izumi, "Glutaredoxin is involved in the formation of the pharmacologically active metabolite of clopidogrel from its gsh conjugate," *Drug metabolism and disposition: the biological fate of chemicals*, vol. 40, pp. 1854–9, Jun. 2012. DOI: 10.1124/dmd.112.045914.

[39] J. Gan, T. W. Harper, M.-M. Hsueh, Q. Qu, and W. G. Humphreys, "Dansyl glutathione as a trapping agent for the quantitative estimation and identification of reactive metabolites," *Chemical Research in Toxicology*, vol. 18, no. 5,

pp. 896–903, Apr. 2005, ISSN: 1520-5010. DOI: 10.1021/tx0496791. [Online]. Available: http://dx.doi.org/10.1021/tx0496791.

[40] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *Journal of Chemical Information and Computer Sciences*, vol. 28, no. 1, pp. 31–36, Feb. 1988, ISSN: 1520-5142. DOI: 10.1021/ci00057a005. [Online]. Available: http://dx.doi.org/10.1021/ci00057a005.

[41] V. Consonni and R. Todeschini, "Molecular descriptors," in *Recent Advances in QSAR Studies*. Springer Netherlands, Oct. 2009, pp. 29–102, ISBN: 9781402097836. DOI: 10.1007/978-1-4020-9783-6_3. [Online]. Available: http://dx.doi.org/10.1007/978-1-4020-9783-6_3.

[42] G. Landrum, *Rdkit*, en. [Online]. Available: https://www.rdkit.org/.

[43] H. Moriwaki, Y.-S. Tian, N. Kawashita, and T. Takagi, "Mordred: A molecular descriptor calculator," *Journal of Cheminformatics*, vol. 10, no. 1, Feb. 2018, ISSN: 1758-2946. DOI: 10.1186/s13321-018-0258-y. [Online]. Available: http://dx.doi.org/10.1186/s13321-018-0258-y.

[44] S. S. Bharate, V. Kumar, and R. A. Vishwakarma, "Determining partition coefficient (log p), distribution coefficient (log d) and ionization constant (pka) in early drug discovery," *Combinatorial Chemistry amp; High Throughput Screening*, vol. 19, no. 6, pp. 461–469, Jun. 2016, ISSN: 1386-2073. DOI: 10.2174/1386207319666160502123917. [Online]. Available: http://dx.doi.org/10.2174/1386207319666160502123917.

[45] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, Apr. 2010, ISSN: 1549-960X. DOI: 10.1021/ci100050t. [Online]. Available: http://dx.doi.org/10.1021/ci100050t.

[46] M. Xu, Z. Lu, Z. Wu, *et al.*, "Development of in silico models for predicting potential time-dependent inhibitors of cytochrome p450 3a4," *Molecular Pharmaceutics*, vol. 20, no. 1, pp. 194–205, Dec. 2022, ISSN: 1543-8392. DOI: 10.1021/acs.molpharmaceut.2c00571. [Online]. Available: http://dx.doi.org/10.1021/acs.molpharmaceut.2c00571.

[47] PubChem, *Acetaminophen*, en. [Online]. Available: https://pubchem.ncbi.nlm.nih.gov/compound/Acetaminophen.

[48] en. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

[49] en. [Online]. Available: https://xgboost.readthedocs.io/en/release_3.0.0/.

[50] C. Developers, *Welcome to chemprops documentation! chemprop 2.1.2 documentation*, en. [Online]. Available: https://chemprop.readthedocs.io/en/latest/.

[51] E. Heid, K. P. Greenman, Y. Chung, *et al.*, "Chemprop: A machine learning package for chemical property prediction," *Journal of Chemical Information and Modeling*, vol. 64, no. 1, pp. 9–17, Dec. 2023, ISSN: 1549-960X. DOI: 10.1021/acs.jcim.3c01250. [Online]. Available: http://dx.doi.org/10.1021/acs.jcim.3c01250.

[52] D. Suenaga, Y. Takase, T. Abe, G. Orita, and S. Ando, "Prediction accuracy of random forest, xgboost, lightgbm, and artificial neural network for shear resistance of post-installed anchors," *Structures*, vol. 50, pp. 1252–1263, Apr. 2023, ISSN: 2352-0124. DOI: 10.1016/j.istruc.2023.02.066. [Online]. Available: http://dx.doi.org/10.1016/j.istruc.2023.02.066.

[53] D. B. Catacutan, J. Alexander, A. Arnold, and J. M. Stokes, "Machine learning in preclinical drug discovery," *Nature Chemical Biology*, vol. 20, no. 8, pp. 960–973, Jul. 2024, ISSN: 1552-4469. DOI: 10.1038/s41589-024-01679-1. [Online]. Available: http://dx.doi.org/10.1038/s41589-024-01679-1.

[54] D. Berrar, "Performance measures for binary classification," in *Encyclopedia of Bioinformatics and Computational Biology*. Elsevier, 2019, pp. 546–560, ISBN: 9780128114322. DOI: 10.1016/b978-0-12-809633-8.20351-8. [Online]. Available: http://dx.doi.org/10.1016/B978-0-12-809633-8.20351-8.

[55] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, ISSN: 1076-9757. DOI: 10.1613/jair.953. [Online]. Available: http://dx.doi.org/10.1613/jair.953.

[56] D. Devi, S. K. Biswas, and B. Purkayastha, "A review on solution to class imbalance problem: Undersampling approaches," in *2020 International Conference on Computational Performance Evaluation (ComPE)*, IEEE, Jul. 2020, pp. 626–631. DOI: 10.1109/compe49325.2020.9200087. [Online]. Available: http://dx.doi.org/10.1109/ComPE49325.2020.9200087.

[57] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: Overview study and experimental results," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, IEEE, Apr. 2020, pp. 243–248. DOI: 10.1109/icics49469.2020.239556. [Online]. Available: http://dx.doi.org/10.1109/ICICS49469.2020.239556.

[58] T. Wongvorachan, S. He, and O. Bulut, "A comparison of undersampling, oversampling, and smote methods for dealing with imbalanced classification in educational data mining," *Information*, vol. 14, no. 1, p. 54, Jan. 2023, ISSN: 2078-2489. DOI: 10.3390/info14010054. [Online]. Available: http://dx.doi.org/10.3390/info14010054.

[59] G. W. Bemis and M. A. Murcko, "The properties of known drugs. 1. molecular frameworks," *Journal of Medicinal Chemistry*, vol. 39, no. 15, pp. 2887–2893, Jan. 1996, ISSN: 1520-4804. DOI: 10.1021/jm9602928. [Online]. Available: http://dx.doi.org/10.1021/jm9602928.

[60] A. P. Bento, A. Hersey, E. Félix, *et al.*, "An open source chemical structure curation pipeline using rdkit," *Journal of Cheminformatics*, vol. 12, no. 1, Sep. 2020, ISSN: 1758-2946. DOI: 10.1186/s13321-020-00456-1. [Online]. Available: http://dx.doi.org/10.1186/s13321-020-00456-1.

[61] V. F. Scalfani, V. D. Patel, and A. M. Fernandez, "Visualizing chemical space networks with rdkit and networkx," *Journal of Cheminformatics*, vol. 14, no. 1, Dec. 2022, ISSN: 1758-2946. DOI: 10.1186/s13321-022-00664-x. [Online]. Available: http://dx.doi.org/10.1186/s13321-022-00664-x.

[62] J. G. M. Conn, J. W. Carter, J. J. A. Conn, *et al.*, "Blinded predictions and post hoc analysis of the second solubility challenge data: Exploring training data and feature set selection for machine and deep learning models," *Journal of Chemical Information and Modeling*, vol. 63, no. 4, pp. 1099–1113, Feb. 2023, ISSN: 1549-960X. DOI: 10.1021/acs.jcim.2c01189. [Online]. Available: http://dx.doi.org/10.1021/acs.jcim.2c01189.

A

Appendix 1

In all tables presented in this appendix, the models that were selected as the best-performing ones in the main results section are highlighted in bold.

A.1 CYP450 Models

A.1.1 Single-Task Tree Models

Table A.1: Performance and hyperparameters of CYP3A4 models across temporal splits

Model	Input	TP	TN	FP	FN	Recall	Precis.	Specific.	F1	MCC
CYP3A4 OCT 2023										
RF	DESC	160	301	99	100	0.615	0.618	0.753	0.617	0.368
{'n_estimators': 100, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': True}										
XGB	DESC	155	257	143	105	0.596	0.520	0.643	0.556	0.234
{'n_estimators': 400, 'max_depth': 3, 'learning_rate': 0.01}										
RF	ECFP	123	354	46	137	0.473	0.728	0.885	0.573	0.401
{'n_estimators': 400, 'max_features': 'sqrt', 'max_depth': None, 'bootstrap': True}										
XGB	ECFP	130	312	88	130	0.500	0.596	0.780	0.544	0.291
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
CYP3A4 APRIL 2024										
RF	**DESC**	**95**	**214**	**73**	**45**	**0.679**	**0.565**	**0.746**	**0.617**	**0.408**
{'n_estimators': 100, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': True}										
XGB	DESC	106	182	105	34	0.757	0.502	0.634	0.604	0.367
{'n_estimators': 100, 'max_depth': 3, 'learning_rate': 0.1}										
RF	ECFP	79	227	60	61	0.564	0.568	0.791	0.566	0.356
{'n_estimators': 100, 'max_features': 'sqrt', 'max_depth': None, 'bootstrap': True}										
XGB	ECFP	82	216	71	58	0.586	0.536	0.753	0.560	0.331
{'n_estimators': 400, 'max_depth': 3, 'learning_rate': 0.1}										
CYP3A4 OCT 2024										
RF	DESC	39	102	36	13	0.750	0.520	0.739	0.614	0.446
{'n_estimators': 200, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': True}										
XGB	DESC	40	95	43	12	0.769	0.482	0.688	0.593	0.411
{'n_estimators': 400, 'max_depth': 3, 'learning_rate': 0.01}										
RF	**ECFP**	**34**	**117**	**21**	**18**	**0.654**	**0.618**	**0.848**	**0.636**	**0.493**
{'n_estimators': 200, 'max_features': 'sqrt', 'max_depth': None, 'bootstrap': True}										
XGB	ECFP	33	110	28	19	0.635	0.541	0.797	0.584	0.412
{'n_estimators': 400, 'max_depth': 3, 'learning_rate': 0.1}										

Table A.2: Performance and hyperparameters of CYP1A2 models across temporal splits

Model	Input	TP	TN	FP	FN	Recall	Precision	Specificity	F1	MCC
CYP1A2 OCT 2023										
RF	DESC	40	547	39	34	0.541	0.506	0.933	0.523	0.461
{'n_estimators': 100, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': False}										
RFB	DESC	42	536	50	32	0.568	0.457	0.915	0.506	0.439
{'sampling_strategy': 0.8, 'n_estimators': 200, 'max_features': 'log2', 'max_depth': None, 'bootstrap': False}										
XGB	DESC	18	574	12	56	0.243	0.600	0.980	0.346	0.337
{'n_estimators': 400, 'max_depth': 3, 'learning_rate': 0.1}										
RF	ECFP	50	524	62	24	0.676	0.446	0.894	0.538	0.479
{'n_estimators': 100, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': True}										
RFB	ECFP	54	509	77	20	0.730	0.412	0.869	0.527	0.473
{'sampling_strategy': 0.6, 'n_estimators': 200, 'max_features': 'log2', 'max_depth': None, 'bootstrap': True}										
XGB	ECFP	36	570	16	38	0.486	0.692	0.973	0.571	0.538
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
CYP1A2 APRIL 2024										
RF	**RDKIT**	**13**	**384**	**16**	**14**	**0.481**	**0.448**	**0.958**	**0.464**	**0.427**
{'n_estimators': 100, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': False}										
RFB	RDKIT	15	371	29	12	0.556	0.341	0.928	0.423	0.387
{'sampling_strategy': 0.8, 'n_estimators': 200, 'max_features': 'log2', 'max_depth': None, 'bootstrap': False}										
XGB	RDKIT	2	390	10	25	0.074	0.167	0.975	0.103	0.072
{'n_estimators': 400, 'max_depth': 3, 'learning_rate': 0.1}										
RF	ECFP	20	348	52	7	0.741	0.278	0.870	0.404	0.397
{'n_estimators': 400, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': True}										
RFB	ECFP	20	349	51	7	0.741	0.282	0.873	0.408	0.401
{'sampling_strategy': 0.5, 'n_estimators': 200, 'max_features': 'log2', 'max_depth': None, 'bootstrap': True}										
XGB	ECFP	12	385	15	15	0.444	0.444	0.963	0.444	0.407
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
CYP1A2 OCT 2024										
RF	DESC	1	171	9	8	0.111	0.100	0.950	0.105	0.058
{'n_estimators': 400, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': False}										
RFB	DESC	2	166	14	7	0.222	0.125	0.922	0.160	0.111
{'sampling_strategy': 0.8, 'n_estimators': 400, 'max_features': 'sqrt', 'max_depth': None, 'bootstrap': True}										
XGB	DESC	0	177	3	9	0.000	0.000	0.983	0.000	-0.028
{'n_estimators': 100, 'max_depth': 6, 'learning_rate': 0.1}										
RF	ECFP	6	161	19	3	0.667	0.240	0.894	0.353	0.353
{'n_estimators': 100, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': False}										
RFB	**ECFP**	**6**	**167**	**13**	**3**	**0.667**	**0.316**	**0.928**	**0.429**	**0.421**
{'sampling_strategy': 0.6, 'n_estimators': 400, 'max_features': 'sqrt', 'max_depth': None, 'bootstrap': True}										
XGB	ECFP	1	178	2	8	0.111	0.333	0.989	0.167	0.170
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										

Table A.3: Performance and hyperparameters of CYP2C19 models across temporal splits

Model	Input	TP	TN	FP	FN	Recall	Precision	Specificity	F1	MCC
CYP2C19 OCT 2023										
RF	RDKIT	14	574	52	16	0.467	0.212	0.917	0.292	0.266
{'n_estimators': 100, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': False}										
RFB	RDKIT	26	256	370	4	0.867	0.066	0.409	0.122	0.118
{'sampling_strategy': 0.8, 'n_estimators': 200, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': True}										
XGB	RDKIT	6	623	3	24	0.200	0.667	0.995	0.308	0.351
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
RF	ECFP	17	544	82	13	0.567	0.172	0.869	0.264	0.254
{'n_estimators': 400, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': False}										
RFB	ECFP	30	21	605	0	1.000	0.047	0.034	0.090	0.040
{'sampling_strategy': 0.5, 'n_estimators': 100, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': False}										
XGB	ECFP	7	614	12	23	0.233	0.368	0.981	0.286	0.267
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
CYP2C19 APRIL 2024										
RF	RDKIT	0	398	19	7	0.000	0.000	0.954	0.000	-0.028
{'n_estimators': 100, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': False}										
RFB	**RDKIT**	**6**	**185**	**232**	**1**	**0.857**	**0.025**	**0.444**	**0.049**	**0.077**
{'sampling_strategy': 0.8, 'n_estimators': 100, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': True}										
XGB	RDKIT	0	413	4	7	0.000	0.000	0.990	0.000	-0.013
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
RF	ECFP	0	362	55	7	0.000	0.000	0.868	0.000	-0.050
{'n_estimators': 100, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': False}										
RFB	ECFP	0	0	0	0	1.000	0.017	0.000	0.032	0.000
{'sampling_strategy': 0.8, 'n_estimators': 400, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': True}										
XGB	ECFP	0	414	3	7	0.000	0.000	0.993	0.000	-0.011
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
CYP2C19 OCT 2024										
RF	RDKIT	0	182	4	1	0.000	0.000	0.978	0.000	-0.011
{'n_estimators': 100, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': False}										
RFB	RDKIT	0	91	95	1	0.000	0.000	0.489	0.000	-0.075
{'sampling_strategy': 0.6, 'n_estimators': 200, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': True}										
XGB	RDKIT	0	0	0	0	0.000	0.000	0.000	0.000	0.000
{'n_estimators': 400, 'max_depth': 3, 'learning_rate': 0.1}										
RF	ECFP	0	175	11	1	0.000	0.000	0.941	0.000	-0.018
{'n_estimators': 400, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': False}										
RFB	**ECFP**	**0**	**0**	**0**	**0**	**1.000**	**0.005**	**0.000**	**0.011**	**0.000**
{'sampling_strategy': 0.8, 'n_estimators': 400, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': True}										
XGB	ECFP	0	0	0	0	0.000	0.000	0.000	0.000	0.000
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										

Table A.4: Performance and hyperparameters of CYP2C9 models across temporal splits

Model	Input	TP	TN	FP	FN	Recall	Precision	Specificity	F1	MCC
CYP2C9 OCT 2023										
RF	RDKIT	7	621	18	10	0.412	0.280	0.972	0.333	0.318
{'n_estimators': 200, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': False}										
RFB	RDKIT	5	619	20	12	0.294	0.200	0.969	0.238	0.218
{'sampling_strategy': 0.6, 'n_estimators': 200, 'max_features': 'log2', 'max_depth': None, 'bootstrap': True}										
XGB	RDKIT	4	634	5	13	0.235	0.444	0.992	0.308	0.311
{'n_estimators': 200, 'max_depth': 6, 'learning_rate': 0.1}										
RF	ECFP	9	608	31	8	0.529	0.225	0.951	0.316	0.319
{'n_estimators': 200, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': False}										
RFB	ECFP	9	586	53	8	0.529	0.145	0.917	0.228	0.242
{'sampling_strategy': 0.8, 'n_estimators': 200, 'max_features': 'log2', 'max_depth': None, 'bootstrap': True}										
XGB	ECFP	1	638	1	16	0.059	0.500	0.998	0.105	0.165
{'n_estimators': 200, 'max_depth': 6, 'learning_rate': 0.1}										
CYP2C9 APRIL 2024										
RF	RDKIT	0	409	8	6	0.000	0.000	0.981	0.000	-0.017
{'n_estimators': 100, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': False}										
RFB	RDKIT	0	369	48	6	0.000	0.000	0.885	0.000	-0.043
{'sampling_strategy': 0.8, 'n_estimators': 400, 'max_features': 'log2', 'max_depth': None, 'bootstrap': False}										
XGB	**RDKIT**	**0**	**416**	**1**	**6**	**0.000**	**0.000**	**0.998**	**0.000**	**-0.006**
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
RF	ECFP	0	408	9	6	0.000	0.000	0.978	0.000	-0.018
{'n_estimators': 100, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': True}										
RFB	ECFP	0	389	28	6	0.000	0.000	0.933	0.000	-0.032
{'sampling_strategy': 0.6, 'n_estimators': 200, 'max_features': 'sqrt', 'max_depth': None, 'bootstrap': True}										
XGB	ECFP	0	0	0	0	0.000	0.000	0.000	0.000	0.000
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
CYP2C9 OCT 2024										
RF	**RDKIT**	**0**	**184**	**3**	**0**	**0.000**	**0.000**	**0.984**	**0.000**	**0.000**
{'n_estimators': 400, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': False}										
RFB	RDKIT	0	182	5	0	0.000	0.000	0.973	0.000	0.000
{'sampling_strategy': 0.5, 'n_estimators': 100, 'max_features': 'sqrt', 'max_depth': None, 'bootstrap': False}										
XGB	RDKIT	0	0	0	0	0.000	0.000	0.000	0.000	0.000
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
RF	ECFP	0	171	16	0	0.000	0.000	0.914	0.000	0.000
{'n_estimators': 100, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': False}										
RFB	ECFP	0	181	6	0	0.000	0.000	0.968	0.000	0.000
{'sampling_strategy': 0.6, 'n_estimators': 400, 'max_features': 'log2', 'max_depth': None, 'bootstrap': True}										
XGB	ECFP	0	0	0	0	0.000	0.000	0.000	0.000	0.000
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										

Table A.5: Performance and hyperparameters of CYP2D6 models across temporal splits

Model	Input	TP	TN	FP	FN	Recall	Precision	Specificity	F1	MCC
CYP2D6 OCT 2023										
RF	RDKIT	20	599	16	25	0.444	0.556	0.974	0.494	0.464
{'n_estimators': 100, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': False}										
RFB	RDKIT	39	371	244	6	0.867	0.138	0.603	0.238	0.239
{'sampling_strategy': 0.5, 'n_estimators': 100, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': True}										
XGB	RDKIT	4	612	3	41	0.089	0.571	0.995	0.154	0.207
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
RF	ECFP	26	576	39	19	0.578	0.400	0.937	0.473	0.435
{'n_estimators': 100, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': False}										
RFB	ECFP	45	35	580	0	1.000	0.072	0.057	0.134	0.064
{'sampling_strategy': 0.5, 'n_estimators': 200, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': False}										
XGB	ECFP	8	608	7	37	0.178	0.533	0.989	0.267	0.281
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
CYP2D6 APRIL 2024										
RF	**RDKIT**	**9**	**390**	**13**	**15**	**0.375**	**0.409**	**0.968**	**0.391**	**0.357**
{'n_estimators': 100, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': False}										
RFB	RDKIT	23	216	187	1	0.958	0.110	0.536	0.197	0.228
{'sampling_strategy': 0.6, 'n_estimators': 400, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': True}										
XGB	RDKIT	2	402	1	22	0.083	0.667	0.998	0.148	0.223
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
RF	ECFP	12	379	24	12	0.500	0.333	0.940	0.400	0.365
{'n_estimators': 200, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': True}										
RFB	ECFP	14	354	49	10	0.583	0.222	0.878	0.322	0.300
{'sampling_strategy': 0.8, 'n_estimators': 100, 'max_features': 'log2', 'max_depth': None, 'bootstrap': True}										
XGB	ECFP	2	402	1	22	0.083	0.667	0.998	0.148	0.223
{'n_estimators': 200, 'max_depth': 6, 'learning_rate': 0.1}										
CYP2D6 OCT 2024										
RF	RDKIT	2	175	6	7	0.222	0.250	0.967	0.235	0.200
{'n_estimators': 400, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': False}										
RFB	RDKIT	9	125	56	0	1.000	0.138	0.691	0.243	0.309
{'sampling_strategy': 0.5, 'n_estimators': 400, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': True}										
XGB	**RDKIT**	**2**	**181**	**0**	**7**	**0.222**	**1.000**	**1.000**	**0.364**	**0.463**
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
RF	ECFP	4	172	9	5	0.444	0.308	0.950	0.364	0.332
{'n_estimators': 400, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': True}										
RFB	ECFP	9	11	170	0	1.000	0.050	0.061	0.096	0.055
{'sampling_strategy': 0.6, 'n_estimators': 200, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': False}										
XGB	ECFP	1	181	0	8	0.111	1.000	1.000	0.200	0.326
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										

A.1.2 Single & Multitask Chemprop Models

Table A.6: Performance of CYP1A2 with Chemprop models across temporal splits

CYP1A2 OCT 2023						
Model	**Input**	**Recall**	**Precision**	**Specificity**	**F1**	**MCC**
SINGLE	NONE	0.350	0.660	0.978	0.460	0.441
SINGLE	ECFP	0.189	0.608	0.984	0.288	0.299
SINGLE	DESC	0.256	0.703	0.986	0.376	0.293
MULTI	NONE	0.459	0.641	0.968	0.535	0.497
MULTI	ECFP	0.000	0.000	1.000	0.000	0.000
MULTI	DESC	0.351	0.634	0.974	0.452	0.426
CYP1A2 APRIL 2024						
SINGLE	**NONE**	**0.220**	**0.400**	**0.977**	**0.285**	**0.194**
SINGLE	ECFP	0.296	0.381	0.968	0.333	0.297
SINGLE	**DESC**	**0.222**	**0.300**	**0.965**	**0.255**	**0.216**
MULTI	**NONE**	**0.370**	**0.357**	**0.956**	**0.364**	**0.321**
MULTI	ECFP	0.333	0.360	0.961	0.346	0.305
MULTI	DESC	0.259	0.350	0.968	0.298	0.262
CYP1A2 OCT 2024						
SINGLE	**NONE**	**0.111**	**0.125**	**0.961**	**0.117**	**0.145**
SINGLE	**ECFP**	**0.333**	**0.273**	**0.956**	**0.300**	**0.263**
SINGLE	DESC	0.111	0.333	0.989	0.167	0.170
MULTI	**NONE**	**0.111**	**0.167**	**0.972**	**0.133**	0.101
MULTI	ECFP	0.000	0.000	1.000	0.000	0.000
MULTI	DESC	0.000	0.000	0.978	0.000	-0.033

Table A.7: Performance of CYP2C19 with Chemprop models across temporal splits

CYP2C19 OCT 2023						
Model	**Input**	**Recall**	**Precision**	**Specificity**	**F1**	**MCC**
SINGLE	NONE	0.167	0.217	0.971	0.189	0.189
SINGLE	ECFP	0.167	0.625	0.995	0.263	0.308
SINGLE	DESC	0.300	0.429	0.981	0.352	0.333
MULTI	NONE	0.300	0.250	0.957	0.273	0.235
MULTI	ECFP	0.000	0.000	1.000	0.000	0.000
MULTI	DESC	0.200	0.222	0.967	0.211	0.175
CYP2C19 APRIL 2024						
SINGLE	**NONE**	**0.000**	**0.000**	**0.966**	**0.000**	**-0.024**
SINGLE	ECFP	0.000	0.000	0.978	0.000	-0.019
SINGLE	**DESC**	**0.000**	**0.000**	**0.995**	**0.000**	**-0.009**
MULTI	**NONE**	**0.143**	**0.091**	**0.977**	**0.111**	**0.096**
MULTI	ECFP	0.000	0.000	0.986	0.000	-0.015
MULTI	DESC	0.000	0.000	0.988	0.000	-0.014
CYP2C19 OCT 2024						
SINGLE	**NONE**	**0.000**	**0.000**	**0.995**	**0.000**	**-0.005**
SINGLE	ECFP	0.000	0.000	1.000	0.000	0.000
SINGLE	**DESC**	**0.000**	**0.000**	**0.995**	**0.000**	**-0.005**
MULTI	**NONE**	**0.000**	**0.000**	**0.989**	**0.000**	**-0.008**
MULTI	ECFP	0.000	0.000	1.000	0.000	0.000
MULTI	DESC	0.000	0.000	1.000	0.000	0.000

Table A.8: Performance of CYP2C9 with Chemprop models across temporal splits

CYP2C9 OCT 2023						
Model	**Input**	**Recall**	**Precision**	**Specificity**	**F1**	**MCC**
SINGLE	NONE	0.059	0.143	0.991	0.083	0.076
SINGLE	ECFP	0.000	0.000	0.997	0.000	-0.009
SINGLE	DESC	0.294	0.278	0.980	0.285	0.366
MULTI	NONE	0.000	0.000	0.988	0.000	-0.018
MULTI	ECFP	0.000	0.000	1.000	0.000	0.000
MULTI	DESC	0.000	0.000	0.995	0.000	-0.011
CYP2C9 APRIL 2024						
SINGLE	**NONE**	**0.167**	**0.056**	**0.959**	**0.083**	**0.074**
SINGLE	ECFP	0.000	0.000	1.000	0.000	0.000
SINGLE	**DESC**	**0.000**	**0.000**	**0.993**	**0.000**	**-0.010**
MULTI	**NONE**	**0.000**	**0.000**	**0.993**	**0.000**	**-0.010**
MULTI	ECFP	0.000	0.000	0.970	0.000	-0.021
MULTI	DESC	0.000	0.000	0.991	0.000	-0.011
CYP2C9 OCT 2024						
SINGLE	**NONE**	**0.000**	**0.000**	**0.993**	**0.000**	**0.000**
SINGLE	ECFP	0.000	0.000	1.000	0.000	0.000
SINGLE	**DESC**	**0.000**	**0.000**	**1.000**	**0.000**	**0.000**
MULTI	**NONE**	**0.000**	**0.000**	**0.995**	**0.000**	**0.000**
MULTI	ECFP	0.000	0.000	1.000	0.000	0.000
MULTI	DESC	0.000	0.000	0.995	0.000	0.000

Table A.9: Performance of CYP2D6 with Chemprop models across temporal splits

CYP2D6 OCT 2023						
Model	**Input**	**Recall**	**Precision**	**Specificity**	**F1**	**MCC**
SINGLE	NONE	0.333	0.500	0.976	0.400	0.374
SINGLE	ECFP	0.111	1.000	1.000	0.200	0.323
SINGLE	DESC	0.044	0.667	0.998	0.083	0.160
MULTI	NONE	0.156	0.538	0.990	0.241	0.265
MULTI	ECFP	0.000	0.000	1.000	0.000	0.000
MULTI	DESC	0.044	0.333	0.994	0.078	0.101
CYP2D6 APRIL 2024						
SINGLE	**NONE**	**0.167**	**0.364**	**0.983**	**0.229**	**0.217**
SINGLE	**ECFP**	**0.000**	**0.000**	**1.000**	**0.000**	**0.101**
SINGLE	DESC	0.292	1.000	1.000	0.154	0.281
MULTI	**NONE**	**0.208**	**0.556**	**0.990**	**0.303**	**0.318**
MULTI	ECFP	0.250	0.286	0.963	0.267	0.227
MULTI	DESC	0.083	0.400	0.993	0.138	0.163
CYP2D6 OCT 2024						
SINGLE	**NONE**	**0.111**	**0.500**	**0.994**	**0.182**	**0.220**
SINGLE	ECFP	0.111	0.333	0.989	0.167	0.176
SINGLE	**DESC**	**0.222**	**0.667**	**0.994**	**0.333**	**0.369**
MULTI	**NONE**	**0.111**	**0.500**	**0.994**	**0.182**	**0.220**
MULTI	ECFP	0.000	0.000	1.000	0.000	0.000
MULTI	DESC	0.111	1.000	1.000	0.200	0.326

Table A.10: Performance of CYP3A4 with Chemprop models across temporal splits

CYP3A4 OCT 2023						
Model	**Input**	**Recall**	**Precision**	**Specificity**	**F1**	**MCC**
SINGLE	NONE	0.611	0.486	0.675	0.539	0.164
SINGLE	ECFP	0.657	0.458	0.620	0.540	0.261
SINGLE	DESC	0.800	0.496	0.603	0.612	0.379
MULTI	NONE	0.688	0.465	0.606	0.555	0.276
MULTI	ECFP	0.667	0.468	0.623	0.550	0.273
MULTI	DESC	0.799	0.494	0.592	0.610	0.369
CYP3A4 APRIL 2024						
SINGLE	**NONE**	**0.611**	**0.486**	**0.675**	**0.539**	**0.164**
SINGLE	ECFP	0.657	0.458	0.620	0.540	0.261
SINGLE	**DESC**	**0.800**	**0.496**	**0.603**	**0.612**	**0.379**
MULTI	**NONE**	**0.688**	**0.465**	**0.606**	**0.555**	**0.276**
MULTI	ECFP	0.667	0.468	0.623	0.550	0.273
MULTI	DESC	0.799	0.494	0.592	0.610	0.369
CYP3A4 OCT 2024						
SINGLE	**NONE**	**0.654**	**0.493**	**0.746**	**0.562**	**0.371**
SINGLE	**ECFP**	**0.558**	**0.492**	**0.783**	**0.523**	**0.328**
SINGLE	DESC	0.808	0.512	0.710	0.627	0.466
MULTI	**NONE**	**0.558**	**0.509**	**0.797**	**0.532**	**0.345**
MULTI	ECFP	0.558	0.274	0.442	0.367	0.235
MULTI	DESC	0.692	0.324	0.457	0.442	0.135

A.1.3 Comparison Temporal vs Random Split

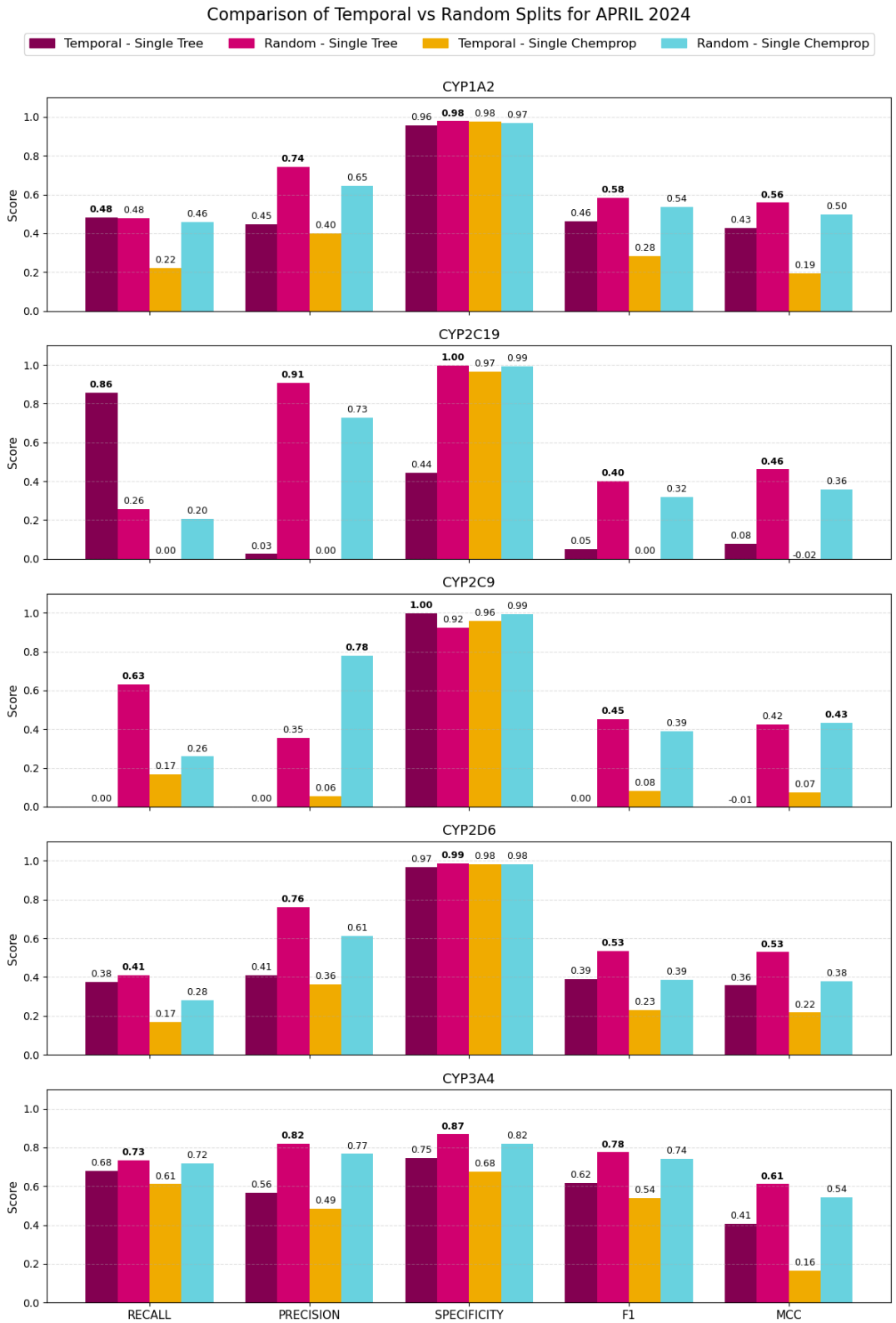


Figure A.1: Comparison of model performance for April 2024 across the five CYP enzymes with temporal and random splits. The best value per metric within each isoform is highlighted in bold.

A.2 Trapping Assays

Table A.11: Performance and hyperparameters of single-task tree-based models for each trapping assay (October 2024).

Model	Input	TP	TN	FP	FN	Recall	Precision	Specificity	F1	MCC
GSH										
RF	DESC	42	84	10	40	0.512	0.808	0.894	0.627	0.444
{'n_estimators': 200, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': False}										
RFB	DESC	63	59	35	19	0.768	0.643	0.628	0.700	0.398
{'sampling_strategy': 'not minority', 'n_estimators': 400, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': False}										
XGB	**DESC**	**46**	**85**	**9**	**36**	**0.561**	**0.836**	**0.904**	**0.672**	**0.501**
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
RF	ECFP	54	75	19	28	0.659	0.740	0.798	0.697	0.462
{'n_estimators': 400, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': False}										
RFB	ECFP	61	59	35	21	0.744	0.635	0.628	0.685	0.372
{'sampling_strategy': 'not minority', 'n_estimators': 400, 'max_features': 'sqrt', 'max_depth': None, 'bootstrap': False}										
XGB	ECFP	49	81	13	33	0.598	0.790	0.862	0.681	0.480
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										
KCN										
RF	DESC	2	43	3	3	0.400	0.400	0.935	0.400	0.335
{'n_estimators': 400, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': True}										
RFB	DESC	2	43	3	3	0.400	0.400	0.935	0.400	0.335
{'sampling_strategy': 'all', 'n_estimators': 400, 'max_features': 'sqrt', 'max_depth': None, 'bootstrap': True}										
XGB	**DESC**	**3**	**44**	**2**	**2**	**0.600**	**0.600**	**0.957**	**0.600**	**0.557**
{'n_estimators': 100, 'max_depth': 3, 'learning_rate': 0.01}										
RF	ECFP	2	42	4	3	0.400	0.333	0.913	0.364	0.289
{'n_estimators': 400, 'max_features': 'log2', 'max_depth': None, 'bootstrap': True}										
RFB	ECFP	1	43	3	4	0.200	0.250	0.935	0.222	0.149
{'sampling_strategy': 'auto', 'n_estimators': 200, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': False}										
XGB	ECFP	3	34	12	2	0.600	0.200	0.739	0.300	0.221
{'n_estimators': 400, 'max_depth': 3, 'learning_rate': 0.1}										
MA										
RF	DESC	13	22	12	6	0.684	0.520	0.647	0.591	0.318
{'n_estimators': 200, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': False}										
RFB	DESC	17	13	21	2	0.895	0.447	0.382	0.596	0.295
{'sampling_strategy': 'not minority', 'n_estimators': 100, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': False}										
XGB	DESC	11	27	7	8	0.579	0.611	0.794	0.595	0.378
{'n_estimators': 200, 'max_depth': 3, 'learning_rate': 0.1}										
RF	**ECFP**	**13**	**30**	**4**	**6**	**0.684**	**0.765**	**0.882**	**0.722**	**0.582**
{'n_estimators': 400, 'max_features': 'sqrt', 'max_depth': 10, 'bootstrap': False}										
RFB	ECFP	19	2	32	0	1.000	0.373	0.059	0.543	0.148
{'sampling_strategy': 'not minority', 'n_estimators': 100, 'max_features': 'log2', 'max_depth': 10, 'bootstrap': True}										
XGB	ECFP	14	25	9	5	0.737	0.609	0.735	0.667	0.457
{'n_estimators': 400, 'max_depth': 6, 'learning_rate': 0.1}										