Learning *in silico* Reactant and Bond-of-Metabolism Predictors for Human Cytochrome P450 Enzymes

by

Siyang Tian

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

University of Alberta

Abstract

Human beings are exposed to many chemicals through their routine interactions with the environment, such as food/drug consumption, household or workplace activities, industrial or transportation activities, and even common environmental processes. Once absorbed, these chemicals are usually further biologically transformed into metabolites. Hence it is important to understand and predict the metabolism of those endogenous chemicals in our body. We decompose this *in silico* metabolism prediction task into three subtasks: given a compound m and a specific metabolizing enzyme α, (1) predicting whether m is a substrate of α, (2) if so, predicting what part of m is changed (here, the "bond of metabolism") and (3) predicting the resulting terminal metabolite. This dissertation addresses the first two of these subtasks, for the nine most important human cytochrome P450 (CYP450) enzymes – CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP2E1, CYP3A4. (1) Given an arbitrary molecule m and one of these nine CYP450 enzymes α, CypReact accurately predicts whether m will react with α. On a dataset of 1632 molecules, CypReact's (cross-validation) AUROCs (area under the receiver operating characteristic curves) vary from 0.83 to 0.92. (2) Given one of the nine enzymes α and its substrate m, CypBoM$_{\eta-\eta}$ accurately predicts *where* m is metabolized by α – which of its η-η bonds (each a bond between two non-Hydrogen atoms) is a "bond of metabolism". Over a dataset of 679 compounds, CypBoM$_{\eta-\eta}$'s (cross-validation) Jaccard scores ranged from 0.401 to 0.594. Our empirical studies, on datasets disjoint from our training sets, demonstrated that CypReact and CypBoM$_{\eta-\eta}$ performed significantly better than related tools (eg, ADMET Predictor and Meteor Nexus), over several evaluation metrics, such as Jaccard score and MCC (Matthews correlation coefficient). As both tools are freely available, we anticipate many

future researchers and developers will use them to better understand human metabolism.

Acknowledgements

Firstly, I would like to express my sincere gratitude to my advisor Prof. Russel Greiner, for the continuous support of my Master studies. He is very patient and always willing to help whenever needed. Working with him was a valuable experience in my life and I learned a lot from it.

Secondly, I would like to thank my co-supervisor Prof. David Wishart. His great knowledge of bioinformatic and suggestions helped a lot in my Master's studies.

I would also like to thank my colleagues, Yannick Djoumbou, Maheswor Gautam and Xuan Cao, for their help, including discussions and suggestions in my research.

Finally, I would like to thank my family for their consistent support and encouragement.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

On a daily basis, humans are exposed to many chemicals through our routine interactions with the environment. These exposures can occur as a result of food/drug consumption, household or workplace activities, industrial or transportation activities, and even common environmental processes. Once absorbed, these chemicals usually undergo further biologically mediated transformations. These biotransformations can be beneficial or detrimental, depending on the type of chemicals (*e.g.*, food supplements vs pesticides), the length of the exposure (short-term vs long-term), and the amount absorbed. If our bodies have absorbed or produced a toxic metabolite, [1] it is very important that it be deactivated (through various metabolic processes) and/or excreted from our body quickly.

Therefore, understanding how a molecule can be transformed (aka metabolized) is crucial for the assessment of its bioavailability, bioactivity, and toxicology. As a result, identifying the metabolites of a compound through chemical experiments along with *in silico* metabolite prediction have become increasingly important research activities for a number of life science disciplines, including drug development, drug testing, pharmaceutics, pharmacology, toxicology, environmental monitoring, metabolomics, food science and personalized medicine [1].

In humans, many chemicals are extensively metabolized by cytochrome P450 (CYP450) enzymes. CYP450-mediated metabolism, which is a major component of Phase I metabolism, occurs primarily in the liver and kidneys. In humans, among the >50 known CYP450 variants (also known as CYP450 *isozymes* [2], [3]), nine – CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, 2C19, CYP2D6, CYP2E1 and CYP3A4 – are most expressed

[1] See terms defined in the Glossary, Appendix A.

and responsible for most of the known Phase I metabolism of drugs [4], as well as the Phase I metabolism of a number of food compounds, environmental pollutants, and other xenobiotic molecules. Hence, it is important to understand the Phase I CYP450 metabolism of a compound and to develop prediction tools to help with the relevant study.

In silico metabolism prediction is a field of metabolite analysis that involves predicting the likely metabolites from a given starting molecule. It was initially developed in the early 1960's to help identify drug metabolites generated through Phase I metabolism based on observed mass spectrometry and/or NMR spectroscopy data [5]. Since then, *in silico* metabolism prediction has expanded to include not only the prediction of drug metabolism, but also the prediction of environmental/microbial metabolism [6], promiscuous enzyme metabolism [7] and many other kinds of xenobiotic and exogenous metabolic processes [8]. Typically, *in silico* metabolism prediction can be decomposed into three general steps (see Figure 1.1):

1. predicting whether a molecule will react with an enzyme ("reactant" prediction);[2]
2. predicting where this interaction will occur (typically viewed as "site of metabolism" prediction – but "bond of metabolism" prediction in this work); and
3. predicting the result of this interaction (structure prediction).

Section 1.4 below summarizes some relevant related projects, that address some of the steps, or variants outlined above.

1.2 My Contributions

This dissertation explores two hypotheses: (1) is it possible to learn a model that can accurately predict whether a given small molecule will react with a specific CYP450 isozyme? and (2) is it possible to predict where within the molecule, the reaction will take place? The second task requires defining what a reaction is and providing a clear, unambiguous way to identify the appropriate location within a molecule. In particular, we divide chemical bonds into three different types, define a new term BoM (bond of metabolism) that clearly describes the location of a metabolic reaction in terms of bonds, and introduce two *in silico* metabolism prediction tools, CypReact and CypBoM$_{\eta-\eta}$, that use

[2]Here, we classify an inhibitor as a non-reactor.

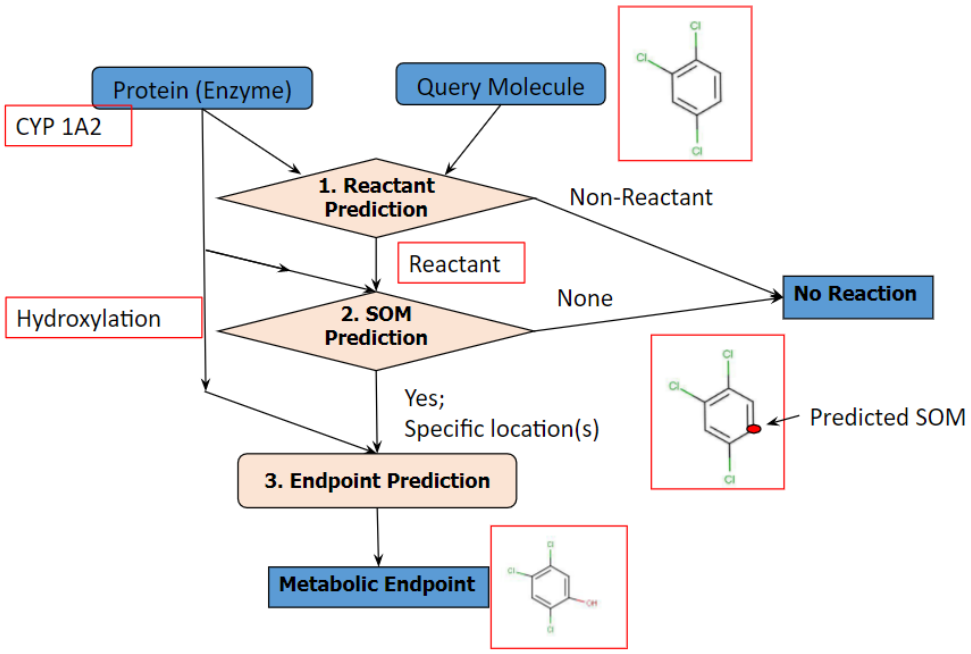


Figure 1.1: Overview of the overall Reaction-Prediction process.

machine learning approaches to produce models that can predict the CYP450-mediated metabolism of chemical compounds. Given a small molecule m and and CYP450 isoform α, CypReact predicts whether m will react with α. Our empirical results demonstrated that this system is effective – with cross-validation AUROC ranging from 0.83 to 0.92 (for different isoforms α) on the training set of 1632 relevant molecules. CypBoM$_{\eta-\eta}$ is a crucial component of CypBoM that predicts a very common type of reaction in Phase I CYP450-mediated metabolism: modification of bonds between two non-Hydrogen atoms; here called η-η, for each of the nine CYP450 enzymes. Over a dataset of 679 relevant molecules (that included 829 reactive η-η sites), CypBoM$_{\eta-\eta}$'s (cross-validation) average Jaccard score was 0.47. Another contribution of this work are the datasets we mentioned above: we created new datasets for substrate and BoM predictions, that we used for training, and then validating, our models. The datasets are publicly available on `https://drive.google.com/open?id=1NQPFKVnJC8f0XXV9lpeAzW4YXDmrWMdU`.

1.3 Outline

Chapter 2 gives the foundations about chemical compounds and machine learning.

Chapter 3 describes CypReact, including how it is learned, its performance and the

dataset used.

Chapter 4 defines the term BoM (bond of metabolism), describes how the BoM dataset is created and explains CypBoM$_{\eta-\eta}$.

Finally, Chapter 5 discusses the knowledge we want to share and the future work.

The rest of this chapter summarizes 4 related metabolism prediction tools.

1.4 Related Work

This need for *in silico* metabolism prediction tools has led to a number of specific programs implementing specific individual steps in the process shown at the end of Section 1.1 (or something similar to one or more of those steps) [1]. For example, WhichCyp [9] predicts whether a given molecule *inhibits* a specified CYP450 enzyme, which is similar to predicting reactants (step 1). SmartCyp [10], FAME2 [11] and MetaPrint2D [12] each take a molecule and an enzyme as input, then predict the site(s) where the interaction occurs – *i.e.*, the site(s) of metabolism (SOM), which is similar to our step 2.

There are also several commercial programs, such as ADMET Predictor [1] (developed by Simulations Plus, Inc., Lancaster, California, USA), Meteor Nexus [13](Lhasa Limited, UK) and StarDrop [14](Optibrium Ltd., Cambridge, UK), that combine all three steps to predict whether a given compound is a substrate of several general CYP450 enzymes, if so, then the sites of metabolism and the corresponding chemical structures are predicted.

In this work, we will compare SmartCyp and ADMET Predictor with CypReact in predicting reactants, and compare Meteor Nexus, ADMET Predictor and FAME2 with CypBoM$_{\eta-\eta}$ in predicting reactive η-η bonds.

1.4.1 SmartCyp

SmartCyp [15] is a traditional *in silico* metabolism tool for predicting the SoMs (sites of metabolism) of drug-like compounds for CYP2C9, CYP2D6 and CYP3A4, which are the three most important enzymes involved in drug metabolism. It uses the 2D structure of the compound and makes predictions based on scores mainly calculated according to the energy required for oxidation at every atom and the distance between atoms. We will later compare our CypReact to a modification of this SmartCyp system for predicting reactants for those three CYP450 enzymes.

1.4.2 Meteor Nexus

Meteor Nexus (Lhasa Limited, UK) [13] is a commercial *in silico* metabolism prediction software package that predicts the metabolic fate of compounds. It uses a knowledge base, a dictionary of biotransformations and a reasoning method to predict the metabolites of a given compound. Because Meteor Nexus (v.3.0.1) predicts SoMs and metabolites for the entire CYP450 enzyme family, rather than individual isozymes, we will later show how to convert SoMs to BoMs and how to compare it with a variant of our tool, CypBoM$_{\eta-\eta}$-All, which claims a η-η bond is reactive if it is modified in a reaction catalyzed by any of the the nine major CYP450 enzymes.

1.4.3 ADMET Predictor

ADMET Predictor (Simulation Plus, Lancaster, CA, USA) [16] is a commercial *in silico* software that predicts the ADMET (Absorption, Distribution, Metabolism, Excretion and Toxicity) properties of a compound. Its Metabolism Module allows the users to predict the SoMs and metabolites of a given molecule for each of the nine major CYP450 enzymes, using the corresponding SoM models built with atomic descriptors. We will compare CypReact and CypBoM with ADMET Predictor (v.8.5.1.1) for each of the nine CYP450 enzymes.

1.4.4 FAME2

FAME2 [11] is a free *in silico* metabolism tool for predicting the SoMs for CYP450 enzymes, using chemical descriptors to represent the properties of atoms and their environments. In FAME2, a site location in a specified molecule is predicted as a SoM for CYP450 enzymes if and only if it is predicted so for one of CYP2C9, CYP2D6 and CYP3A4. We will later compare FAME2 with the variant of our CypBoM$_{\eta-\eta}$, CypBoM$_{\eta-\eta}$-Tri, that claims a η-η bond is reactive if it is modified in a reaction catalyzed by any of CYP2C9, CYP2D6 and CYP3A4 enzymes.

Chapter 2

Chemical and Machine Learning Foundations

2.1 Chemical foundation

2.1.1 Representing a molecule using SDF format

A molecule is a group of atoms connected by chemical bonds (see Figure 2.1[left]). From a computational point of view, the structure of a molecule stores the natural information about atoms and bonds, and is used to generate informative features, such as structure-based features, for metabolism prediction tools. A "chemical reaction", in general, transforms one molecule to another, changing the properties of its atoms and the chemical bonds between them.

In this dissertation, we use the SDF (structure-data file) format, which is developed by Molecular Design Limited (MDL) [17], to store the information, including structure and reaction information, of molecules. The SDF format is a widely used standard format that allows a user to represent the structures of multiple molecules with optional fields in one file. Figure 2.1[right] shows how the dichlorotrifluoroethane molecule is stored in a SDF file used in CypBoM. The first two blocks store the name of the molecule, the total number of atoms and η-η bonds of the molecule. The AtomInformation block stores the 3D coordinates and element type of each atom. The BondInformation block stores the actual atoms connected and the bond type for each bond between two non-Hydrogen atoms. The Identification block stores the identification information of the compound using InChiKey – the hashed version of full InChI (International ChemicalIdentifier), and the PubChemID – the identification number which is used to retrieve the compound from the PubChem database [18]. We also

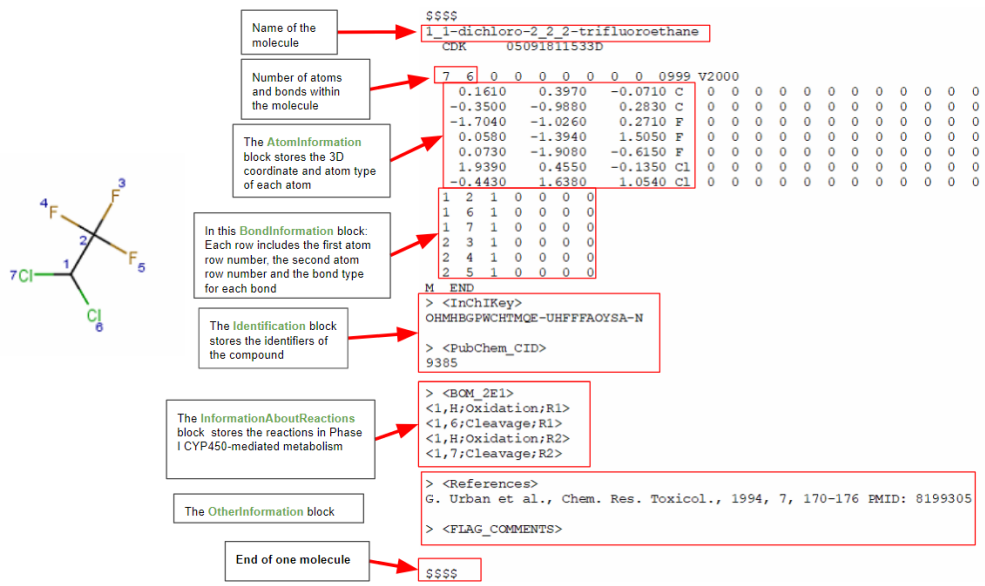


Figure 2.1: The structure of a dichlorotrifluoroethane molecule (left); and how it is stored in a SDF file (right).

generate the InformationAboutReactions block, to store the BoMs of the molecule for each of the nine major CYP450 enzymes. Here, we show the BoMs for CYP2E1, but not the eight other entries for the other isoforms. Note that the SDF file used in CypReact has a similar block that includes nine entries, each storing whether the molecule is a reactant for one of the nine CYP450 enzymes. The OtherInformation block stores the reference where the BoMs of the molecule are found and additional information about the molecule, such as low concentration of some metabolite.

2.1.2 Representing molecules with numeric values

In this section, we introduce the attributes we used to describe the molecular and atomic properties of a molecule for creating features for CypReact and CypBoM, with the example of the anti-inflammatory drug, *lornoxicam*, given in Figure 2.2.

Chemical Descriptors: Chemical descriptors represent the physico-chemical properties of a molecule as a set of numbers. We use molecular and atomic descriptors to describe the characteristics of the whole molecule and the atoms within it, respectively. For example, the molecular weight descriptor in Figure 2.2 shows the weight of the *lornoxicam* molecule is 371.81 daltons; as this is < 900 daltons, we see that it is a small molecule [19]. The Atom Degree of atom C.18 is 1 means that this C.18 atom is connected to one non-Hydrogen atom.

Fingerprint and structural patterns: A functional group is an atom or a group of connectedatoms within molecules that usually behave similarly to one another in chemical reaction(s) [20]. A structural pattern is an extension of a functional group that the structural pattern in different molecules may behave similarly in chemical reactions. A fingerprint is a binary vector that encodes the information about different structural patterns within a molecule; note we use these fingerprints extensively in our work. A molecule fingerprint expresses the presence ("1") or absence ("0") of each chosen strutural pattern within the given molecule. Each bond within a molecule is associated with one or more elements of a bond fingerprint, each of which represents whether that chemical bond is part of a specific structural pattern, using "1" for "Yes" and "0" for "No".

For example, Figure 2.2 shows two structural patterns, one carbonyl and one hydroxyl group, highlighted using orange and green circles, respectively. The molecule fingerprint table shows that there are hydroxyl and carbonyl groups, but no benzene rings within the *lornoxicam* molecule. The bond fingerprint for bond $\langle C.7, O.15 \rangle$ shows that this double bond is part of a carbonyl group and not within hydroxyl nor benzene groups.

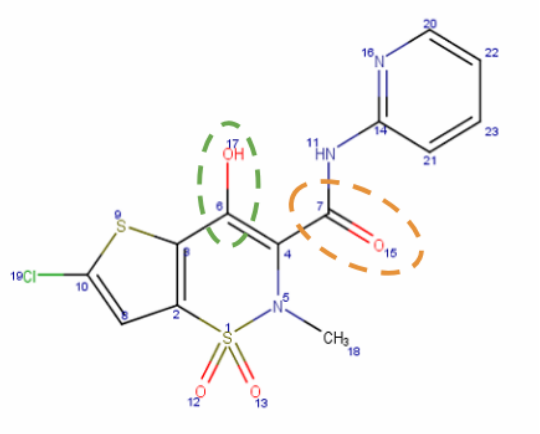
Atom Type: An atom type attribute describes the type of a atom, which is used to compute the properties of that atom. In *in silico* metabolism tools, the atom types are usually encoded into a binary vector using "1" to indicate which atom type the given atom matches. The atom type vector in Figure 2.2 shows that the carbon atom with index 18 is a sp3 hybridized carbon.

Atom environment: The atom environment shows the information of the neighbors of a atom within a molecule, such as the Atom Type and Electronegativity (the tendency of attracting electrons), etc., and can affect the behavior of that atom. For example, the nitrogen atom with index 16 is connected with a carbon atom by a *pi* bond, which means it is likely that this N.16 atom will form a N-Oxide by sharing its lone pair electrons.

We use the above attributes to represent a molecule with informative numeric values; we will see that this allows our learning algorithms to produce effective classifiers.

2.2 Machine Learning Foundations

Machine learning is a modern, scientific approach that allows a computer to learn to perform a specific task, often to make predictions about specific instances, from a dataset of many



Descriptor features for C.18		
Molecular Weight	Atom Degree	...
371.81	1	...

Molecule fingerprint			
Hydroxyl	Carbonyl	Benzene	...
1	1	0	...

Bond fingerprint for <C.7,O.15>			
Hydroxyl	Carbonyl	Benzene	...
0	1	0	...

Atom Type for C.18			
C.sp3	C.sp2	S	...
1	0	0	...

Figure 2.2: The structure of lornoxicam, showing 4 categories of descriptions.

labeled instances. The machine learning approach usually involves a learning process and a performance process. Figure 2.3 shows an example of how machine learning is used in solving the cancer prediction task. It first uses a learning algorithm to learn from the training data to produce a learned classifier and then use the learned classifier to predict whether a person (an novel instance, not in the training set) has cancer. In the learning process, a learning algorithm attempts to find the parameters that lead to a model that performs well on the validation data; this is the vertical line in Figure 2.3. Afterwards, a user can use that learned model to make prediction about novel instances; see the horizontal process in that figure. There are many learning algorithms [21], such as SVM (support vector machine), Naive Bayes, Random Forests, etc. Below, we first briefly introduce some concepts and methods used in machine learning and later present how we use machine learning approach to create our two *in silico* metabolism prediction tools, CypReact and CypBoM.

2.2.1 Feature generation and selection

Standard machine learning algorithms assume that each instance is described as a vector, whose components are values of certain "features". For our task, a good feature is one that

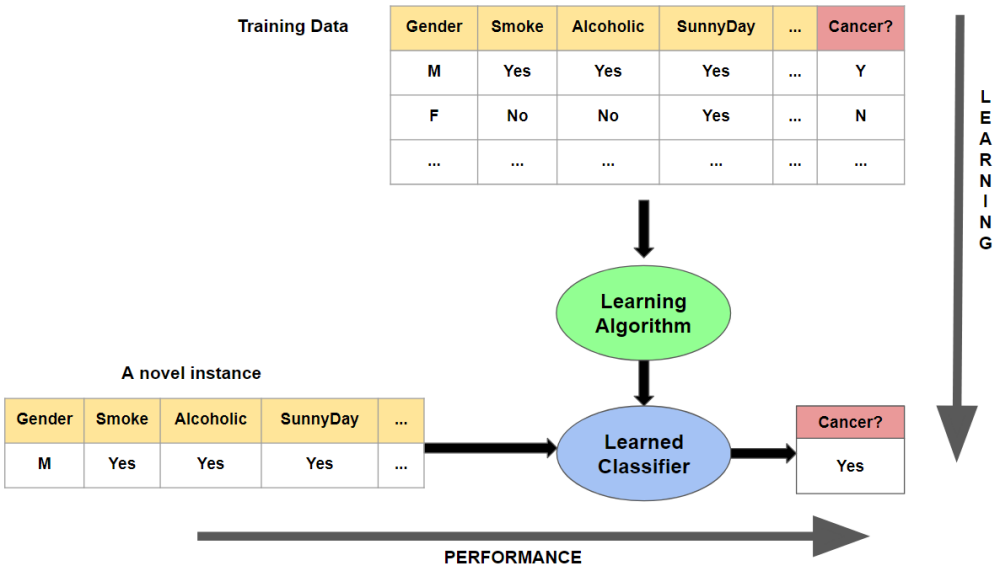


Figure 2.3: Overview of machine learning processes: performance (left to right) and learning (top to bottom).

can help discriminate between the classes and the quality of the features used in the dataset has a major impact of the quality of the classifier built on it.

Feature selection is a technique, often used in machine learning, to select a subset of features that are most relevant to the task we want to solve, for improving the efficiency of the learning algorithm and the quality of the learned model. For example, in Figure 2.3, whether the weather is sunny obviously does not contribute in determining whether the patient has cancer or not, and thus the SunnyDay feature should not be selected during the feature selection process. There are many feature selection methods, such as mRMR (Minimum Redundancy and Maximum Relevance) [22], Dragonfly Algorithm [23], etc.

Sections 3.1.3 and 4.3.1 will describe how we generate features for CypReact and CypBoM, and Sections 3.1.4 and 4.3.2 will present how we select features by the information gain value of each features.

2.2.2 Cross-validation

Cross-validation is the most popular method used in machine learning to estimate the performance of the learned model on novel instances, given limited training data [24]. Here, we apply the learning algorithm $L(\cdot)$ to a labeled dataset D, to produce a model $\theta = L(D)$. We now want to estimate the quality of this learned model θ. (This quality is typically "accu-

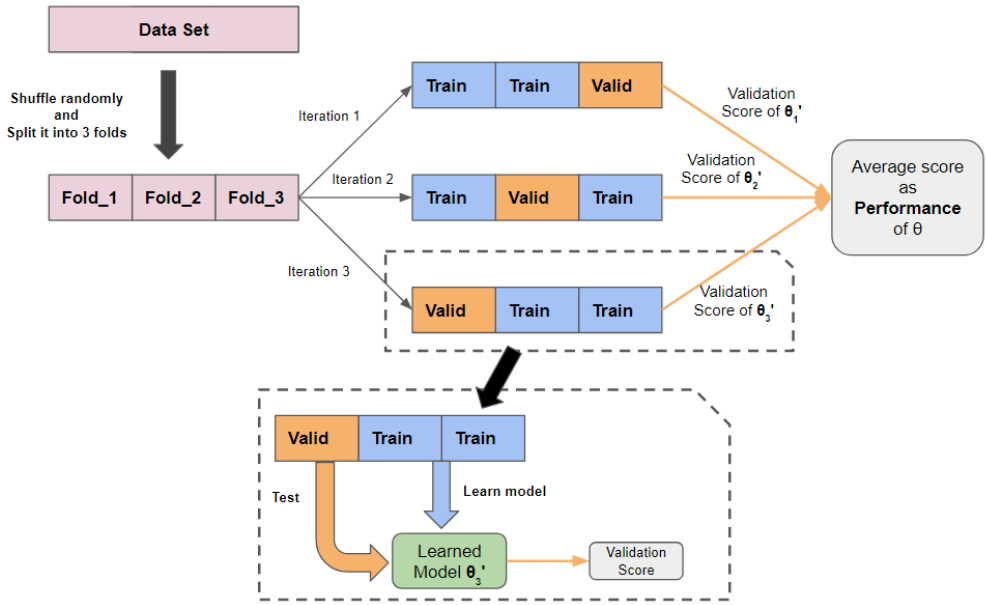


Figure 2.4: An example of 3-fold cross validation.

racy", but we also used other evaluation metrics that are described in Sections 3.4 and 4.4.) Unfortunately, running this θ on the training data D will not produce accurate estimates; this is overfitting. Instead, we produce $k = 5$ similar models, each produced by running the same $L(\cdot)$ on a dataset D' that is similar to D, producing a model θ', then evaluating that resulting θ' on a dataset D'' that is also similar to D, but is disjoint from D'. This typically produces a reasonable estimate of the quality of the leanred model θ. In particular, this D' is a random 80% of D, and D'' is the remaining subset – called the validation set. 5-fold cross-validation actually does this 5 times, where each 1/5 appears as the validation set, once. Figure 2.4 shows 3-fold cross-validation.

That explains "external cross-validation", for estimating the quality of a learned classifier. We can also use a similar "internal cross-validation" to estimate the best values of some parameters, including the parameters used in feature selection procedure.

In this dissertation, we used an improved version of k-fold cross validation – nested-k-fold cross validation [25] – which is explained in Section 3.4.1 for both of these steps.

Chapter 3

CypReact: A Software Tool for Predicting Reactants for Human Cytochrome P450 Enzymes

CypReact is a *in silico* metabolism prediction tool that predicts whether a given compound is a reactant(substrate) for each of the nine major CYP450 enzymes based on our published paper "CypReact: A Software Tool for *in silico* Reactant Prediction for Human Cytochrome P450 Enzymes" [26]. In this section, we will describe the learning process of CypReact and present its performance, including a comparison with other tools.

3.1 Materials and Methods

3.1.1 Approach

Because of the difficulty of the problem we are attempting to solve, we decided to pursue a machine learning approach, which is based on learning the relevant predictors from a large, high quality set of training data; see Figure 3.1. As each of the nine most important CYP450 enzymes has its own set of reactants, we built nine separate predictors – one for each CYP450 isoform. Below, we will let CypReact(α, $\cdot$) refer to the predictor for the isoform $\alpha \in \{$ CYP1A2, CYP2A6, ..., CYP3A4 $\}$, where CypReact(α, m) is 1 ("True") if the molecule m is a reactant to the isoform α, and otherwise is 0 ("False").

3.1.2 Dataset Creation

CYP450 isozymes have a very broad substrate specificity and are responsible for most of the oxidative reactions seen in the Phase I metabolism of small molecule xenobiotics [2].

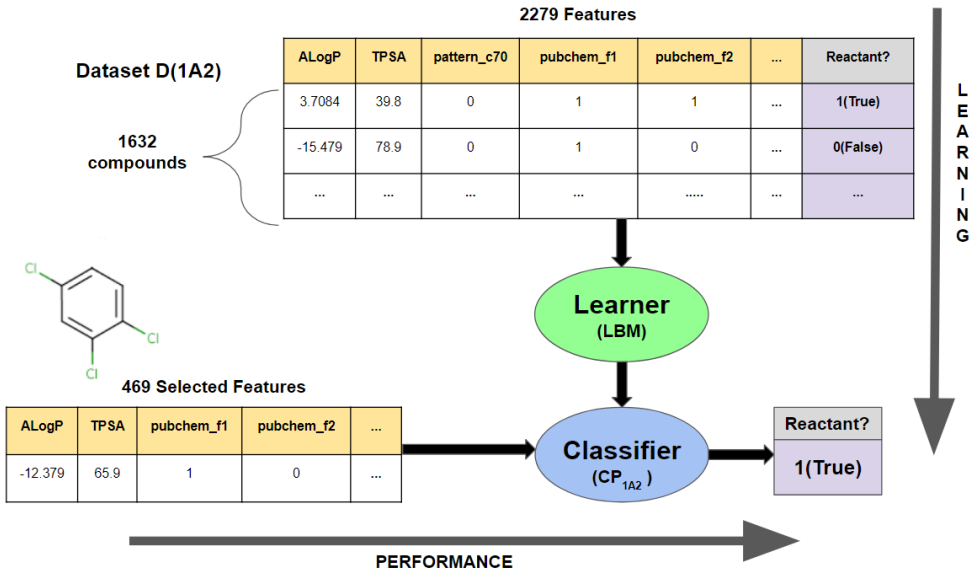


Figure 3.1: Basic Machine Learning Paradigm, with learning algorithm LBM (Learning Base Model) using the D(1A2) dataset to produce a classifier CP_{1A2} (top-to-bottom), where this resulting CP_{1A2} can then make a prediction about an input molecule (left to right). Note the classifier uses a reduced set of features. Also, the datasets for the 8 other isoforms are slightly different (with different "Reactant?" labels), leading to 8 different classifiers.

However, small changes in the chemical structure of a molecule can significantly alter its bioactivity or its metabolic profile [2]. Therefore, in order to train and test our models, it is very important to use a large and diverse dataset that captures the molecular patterns and chemical features responsible for the specific interaction between a given CYP450 and its substrates. To be useful, this dataset should include just the molecules that a biochemist would consider as possible reactants – *i.e.*, just the molecules that a researcher would consider plausible, and therefore worth sending to the resulting CypReact prediction system.

We built a dataset with 1632 compounds, including 679 known CYP450 reactants from the set provided by Zaretzki *et al.* [27], each of which is metabolized by at least one of the nine CYP450 isozymes.[1] To provide a sufficiently large and relevant training set, we manually collected an additional set of 1,053 non-reactant compounds that were "plausible" metabolites – *i.e.*, small molecules that are structurally similar to known substrates, in terms of structural classification, functional classification, and size. We included these 1,053 non-reactant "decoys" to enrich the existing set of "Zaretzki *et al.* non-reactants"[2], and to span a greater

[1] That paper claimed 680 CYP450 substrates; however one of them (phenanthrene) appeared twice.

[2] Recall that only some of those 679 molecules will react with any specific CYP450 isoform; see Table 3.1.

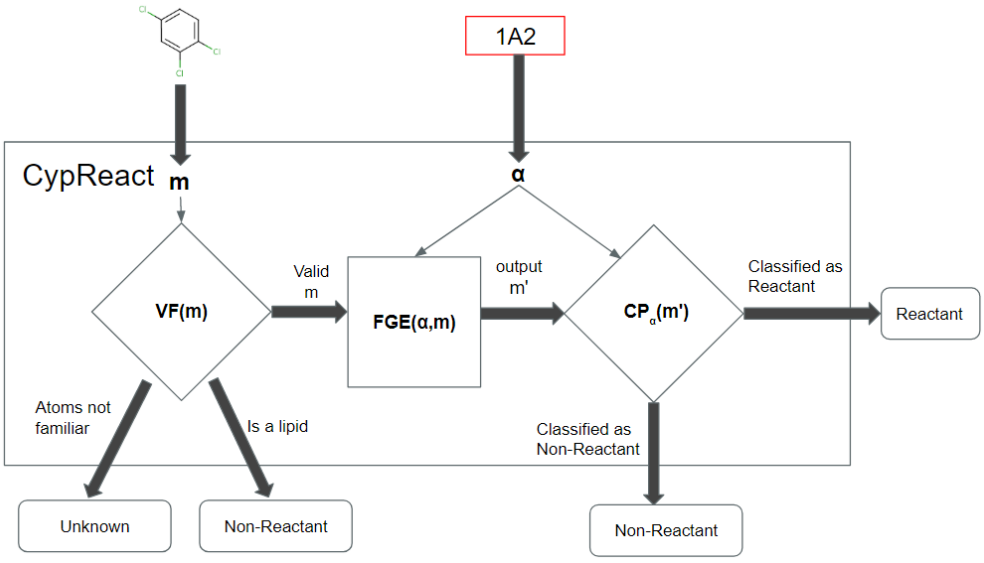


Figure 3.2: Components of the CypReact performance process.

portion of the relevant chemical space of small molecules. These compounds include known drugs, pesticides, food compounds, pollutants, endogenous metabolites and a variety of other compounds that, while plausible CYP450 reactants, are all known *not* to be metabolized by any of the nine selected CYP450 isozymes. We extracted these non-reactants from various databases, including the Human Metabolome Database [28], the KEGG database [29], DrugBank [30], and the PubChem database [31]. In selecting the set of non-reactants, we explicitly avoided molecules that are obviously not metabolized by CYP450 isozymes – *e.g.*, glycerolipids, glycerophospholipids, sphingolipids, inorganic compounds [3], [32]. To be robust, the CypReact performance system handles these molecules separately, using a simple rule-based filter; see Figure 3.2.

We formed a training set for each of the nine selected CYP450 isozymes, consisting of the same 1632 compounds, but with different reactant/non-reactant labels, as a given compound might be a reactant for one CYP soform, but not for another. For instance, the anti-inflammatory drug, amodiaquine (DrugBank ID DB00613) is labeled as a reactant for CYP2C8, CYP2C19, CYP2D6, and CYP3A4, but labelled as a non-reactant for CYP1A2, CYP2A6, CYP2B6, CYP2C9, and CYP2E1. As different CYP450 isoforms react with different molecules, the class distribution (reactant vs. non-reactant) varied from one CYP450 isozyme to another. Table 3.1 shows the number of reactants, and non-reactants, for each of the 9 datasets, as well as the union over all 9, labeled "All". We will let D(α) denote the

Table 3.1: Data distribution of the nine CYP450 isoforms. The light-cyan colored rows correspond to the training datasets; note these datasets contain the same set of 1632 instances for each CYP450 isoform, but different labels. The Hold-Out Testing Datasets (in yellow) have different reactant sets, but the same non-reactant set.

	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4	All
Training Dataset Data Distribution										
#Reactants	271	105	151	142	226	218	270	145	475	679
#Non-Reactants	1361	1527	1481	1490	1406	1414	1362	1487	1157	953
#R / #Total	0.17	0.06	0.09	0.09	0.14	0.13	0.17	0.09	0.29	0.42
Hold-out Testing Dataset Data Distribution										
#Reactants	24	6	4	12	28	20	21	6	32	69
#Non-Reactants	100	100	100	100	100	100	100	100	100	100

dataset associated with the "isoform" $\alpha \in \{$ 1A2, 2A6, ..., 3A4, All $\}$.

3.1.3 Feature Generation

Standard machine learning algorithms assume that each instance is described as a vector, whose components are values of certain "features". Here, we want to identify which properties or features associated with a molecule m are useful for determining whether m is a reactant versus a non-reactant.

We first performed several standardization operations to each of the 1632 compounds, to produce a precise description of each molecule. This involved removing salts, explicitly adding hydrogen atoms, and generating a geometrically correct 3D structure for each molecule. Here, we used the Molconvert command-line tool from ChemAxon's Marvin Suite [33].

Our LBM learning algorithm then considered a set of 2,279 features for each molecule – selected based on their reported effect on the metabolism and the bio-availability of small molecules [27], [34], [35]. This included 36 physico-chemical properties (such as molecular weight, and XLogP – each computed using the Chemistry Development Kit (CDK) [36]) and 2,243 structure-based features, which includes the MACCS 166 fingerprint [37], and 881 PubChem fingerprints [31]. Additionally, LBM used a ClassyFire [38] fingerprint, which consists of 1196 structural features encoded in the SMARTS language [39]. These include (1) functional group/chemical class definitions provided by ClassyFire, (2) structural patterns reported by the literature to correlate with reactivity to, or inhibition of CYP450

isozymes, (3) structural patterns of length 3 to 18 atoms obtained by mining the chemical structures of known CYP450 reactants and non-reactants and (4) the MACCS 322 fingerprint (provided by Sud *et al.* [40]). The MACCS 166 fingerprint and the PubChem fingerprint are calculated using MACCSFingerprinter and the PubchemFingerprinter modules of the CDK library, respectively. The ClassyFire fingerprint was computed using the SMARTSQueryTool module of the CDK library. While the physicochemical properties were represented as numerical features, the structural features were represented as binary features to express the presence "1" or absence "0" of a specific structural feature within the molecule of interest.

3.1.4 Feature Selection

Feature selection is a technique, often used in machine learning, to select a subset of the features that the learner will use, to produce a classifier that uses only these features. Once identified, this makes the training phase faster and more efficient (as it involves fewer features) while also reducing the chance that the learner will overfit, as this means the learned model will involve relatively few parameters.

Recall that we initially selected 2,279 features that are potentially useful for our task – *e.g.*, the number of hydrogen bond acceptors, the sum of atomic polarizabilities, etc. However, some features contribute very little information. For example, while fingerprint features in general are potentially useful for our task, certain ones had values that were the same for all the molecules in the dataset. As such features do not distinguish any molecules from one another, they of course cannot help in classification. Moreover, different features may have different degrees of importance for predicting the substrate specificity for each of the nine CYP450s – *e.g.*, features that are critical to CYP1A2's substrate specificity, might be irrelevant to CYP2B6's substrate specificity.

Hence, in order to reduce the chance of overfitting, and also to improve the computational efficiency, for each D(α) dataset, our learning algorithm computed the information gain [41] of each feature with respect to the "reactant/non-reactant" label. This measures how important that feature is, for the given isoform, α. It then removed the features that appeared to be relatively uninformative – specifically removing all of the features with an information gain less than a threshold γ, which was learned by internal cross-validation; see below. Hence each CYP450 has its own unique feature set; Table 3.2 provides the numbers of features for each CYP450 reactant predictor.

Table 3.2: Number of features selected by CypReact with respect to each CYP450 enzyme. (Note the "All" value corresponds to the union of the features over all 9 isoforms.)

	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4	All
# of features	469	421	274	563	536	509	495	263	934	1082
γ [‡]	0.0075	0.001	0.0075	0.001	0.005	0.005	0.0075	0.0075	0.001	

[‡] $\gamma \in \{0.001, 0.005, 0.0075, 0.01, 0.03\}$ is the information gain threshold, found in the cross-validation process, used to find the number of features to use.

LBM also normalized each feature f_i in each D(α) dataset: Assume the values of f_i in D(α) are $\{ x_i^j \}_j$. First let $b_i = \max_j \{x_i^j\}$ (resp., $s_i = \min_j \{x_i^j\}$) be the maximum (resp., minimum) of these values. It then replaced each x_i^j with its normalized value

$$\hat{x}_i^j \quad \leftarrow \quad \frac{x_i^j - s_i}{b_i - s_i}$$

which is by construction in the range [0,1].

Each D(α) dataset uses these features to describe each molecule. (We will soon see that the FGE(α, m) process translates a molecule m, in SMILES or Structure (sdf) format, into a vector of values for these features.)

3.1.5 Cost-Sensitive Learner

Most machine learning algorithms are designed to work best when the data set is relatively balanced – *i.e.*, when the number of positive and negative cases (here, reactants vs non-reactants) is nearly the same. Our dataset is, however, very imbalanced, as the number of reactants (∼11%) is much less than the number of non-reactants (∼89%). This is intentional, as it reflects the performance task that we anticipate for most of the scientists using our reactant predictor. In particular, we expect that very few of the molecules they will consider will actually be reactants. For instance, of the more than 400,000 known natural products, metabolite and drugs, less than 10,000 molecules have been tested, of which fewer than 1,000 are actually CYP450 reactants. In addition to this imbalance, we anticipate most users will consider false negatives (predicting a reactant to be a non-reactant) to be worse than false positives (predicting a non-reactant to be a reactant). Such users will prefer tools that rarely predict a reactant to be a non-reactant, even if this means (as an unavoidable side-effect) that those tools incorrectly predict several non-reactants to be reactants. After all, each false positive means the researcher may need to do a bit of extra work (*e.g.*, run an extra

experiment), before finding this mistake. However, each false negative means the researcher will (probably) just ignore this molecule, which might mean s/he may not bother to look for a metabolite. In the world of drug research, not knowing about a reaction means the researcher may miss a potential toxic metabolite, or a potentially beneficial drug byproduct.

To emphasize the importance of false negatives over false positives, LBM uses a *cost sensitive learner* [42], which involves a base learner (for instance, a support vector machine [43] or a neural network) and a cost matrix (such as Table 3.3[right]). It trains the base learner, seeking a classifier that minimizes the *total weighted cost*, which is the dot product of the given cost matrix and the confusion matrix, where a confusion matrix presents the number of each type of classification results produced by the classifier $C(\cdot)$ on the test data D – in particular, the number of true positives, false positives, false negatives and true negatives; see Table 3.3[left]. (Note that "Reactant" is considered "True" and "Non-reactant" is considered "False".) A cost matrix presents the cost of each of these types of classification results as seen in Table 3.3[right]. Note that true positives and true negatives each cost 0 while the cost of each false positive is set to 1, and the cost of each false negative is set to β.

Given this cost matrix, the "(Weighted) Cost" of a classifier $C(\cdot)$, based on its confusion matrix on a set of test data D, simplifies to the sum of the number of false positives, plus β times the number of false negatives.

$$\text{Cost}_\beta(\, C(\cdot),\, D\,) \quad = \quad \frac{(1 \times \#\text{False\_Positives} \;+\; \beta \times \#\text{False\_Negatives})}{|D|} \tag{3.1}$$

(We divide by the number of instances, $|D|$, to "normalize" the cost.)

Hence, this parameter β quantifies the trade-off between false-positives to false-negatives. For example, standard machine learning algorithms try to minimize the total (unweighted) number of mistakes, which is the sum of the number of false positives and false negatives. Hence, they implicitly assume that $\beta = 1$. As noted above, this is not appropriate here. Setting $\beta = 3.1$ means the learning algorithm would rather mistakenly claim that 3 non-reactants are reactants, rather than claim 1 reactant is a non-reactant.

To determine the appropriate value for β, we consulted with experts in the field, who collectively suggested we use a β between 3 and 7. Our subsequent sensitivity studies (*e.g.*, using Cost Curves; see below) showed that the resulting classifiers were not particularly sensitive to the precise value in that range. We therefore selected the midpoint $\beta = 5$ – that

is, our system treats each false negative as five times as bad as a false positive. (While this paper focuses on this setting, our code-base allows the user to set this β parameters as s/he wishes.)

Our learning algorithm LBM($\cdot$) takes as input a labeled dataset, here D_α (see top portion of Table 3.1), and implicitly the cost matrix shown in Figure 3.3[right], and returns a classifier. This learned classifier, called CP_α[3], takes a representation of a molecule, and returns {1, 0} (and occasionally "Unknown"; see below). We will see that this CP_α is the main part of the CypReact($\alpha, \cdot$) system but there are also several other important components; see Figure 3.2.

For each isoform α, using the dataset D(α), LBM considers five candidate base learners for the cost sensitive classifier: support vector machine SVM, logistic regression LR [21], decision tree DT [44], random forest RF [45] and an ensemble method ES [46] that returns the majority class of the learned weak classifiers. Given the various parameter settings for some learners, there are 31 different learners+parameters. LBM first identifies the best base learner, and also the best setting for its parameters, as well as the best threshold $\gamma \in \{0.001, 0.005, 0.0075, 0.01, 0.03\}$ for the feature selection process, by running an internal cross-validation process on its given entire dataset D(α). This process involved dividing the given dataset into five disjoint subsets. It then trains each of these learners on four of these five subsets, to produce $155 = 31 \times 5$ models (one for each of pair of [base_learner+parameter, value of γ]). It then evaluates each of these models on the remaining subset, which produced a single score (Equation 3.1) for each of the models. It does this five times, each time holding-out a different subset, then computes the average score (over these five iterations) for each of the 155 base_learner+parameter+γ settings. For each D_α, LBM found that the most accurate method was RF (random forest) for $\alpha \in$ {CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C19, CYP2E1, CYP3A4} and ES (ensemble methods) for $\alpha \in$ {CYP2C9, CYP2D6}. Table 3.2 shows the number of features selected, for each isoform. Note that both of these base learners, RF and ES, involve consensus voting [47]. LBM then ran the selected base learner on the entire D(α) dataset, which generated the model we will use – called CP_α.

3.1.6 Implementation (see Figure 3.2)

Recall our CypReact tool was trained on only compounds that were "plausible" CYP450

[3] This CP_α represents the function $CP(\alpha, \cdot)$.

Table 3.3: Confusion Matrix of classifier $C(\cdot)$ on dataset D (left); and Cost Matrix (right)

Truth⇒ Prediction⇓	R	N
R	#True_Positives	#False_Positives
N	#False_Negatives	#True_Negatives

Truth⇒ Prediction⇓	R	N
R	0.0	1.0
N	β	0.0

substrates – the set of 1632 summarized above. As noted, our training data intentionally did not include any molecule from classes of compounds that are obviously not CYP450 reactants – which means we ignored very large and hydrophobic molecules such as lipids (glycerolipids, glycerophospholipids, and sphingolipids) as well as inorganic compounds. We also noted that the training set included only molecules that contain only the following atoms: {H, C, O, N, S, F, Cl, P, Br, I}, which means we know the pre-processing can correctly handle those atoms.

To make our system more robust, we want to allow users to enter any molecule. For most molecules, CypReact will be able to make an accurate assessment. But for some – *e.g.*, the ones that include atoms that did not appear in any molecule in the training set – we cannot be as confident. We therefore wrote a molecular filter program, called *VF*(m), that makes a 3-way decision, for any molecule m:

1. If m is in an excluded class (currently, any lipid), *VF* returns "No" (not a reactant) and exits.

2. If m includes any atom that is not "familiar" (*i.e.*, not in the list above), *VF* returns "Unknown", and exits.

3. Otherwise, m is considered valid, and *VF* passes it to the main part of the CypReact process, to be labeled.

If the molecule m is valid (#3 above), it will be passed to the $FGE(\alpha, m)$ function,[4] which will re-express m as a set of values associated with molecular features relevant to the CYP α (such as "PubChem fingerprints" [31]). The resulting description, m', will be input into the trained CP_α model and classified. Our implementation is written in Java using the WEKA [48] APIs.

[4]FGE stands for FeatureGeneration&Extraction.

3.2 Related Systems

In general, a good way to understand how well a system works is to compare its performance to that of other similar systems. Below we describe two systems: one that performs the same task as our CypReact, and another that performs a similar function.

3.2.1 ADMET Predictor

ADMET Predictor (Simulations Plus, Inc., Lancaster, California, USA) is a commercial software tool for predicting properties of chemical compounds, including whether a molecule is a reactant for a specific CYP450 enzyme – *i.e.*, the same function as CypReact. We can therefore compare our tool directly to ADMET Predictor. (Of course, as we do not know the dataset on which ADMET Predictor was trained, we do not know whether that training set included our test set; this means we do not know whether our estimate of ADMET Predictor's accuracy is optimistic as we may be testing its performance on its training set.)

3.2.2 A Reactant-predictor variant of SmartCyp

We also compare our tool with a reactant predictor variant of SmartCyp [10], which is a site-of-metabolism (SOM) predictor. In general, SmartCyp(α, m, s) generates a score for a site s of a given molecule m, for any of three isoforms $\alpha \in \{$CYP3A4, CYP2D6, CYP2C9$\}$, where lower scores means SmartCyp thinks it more likely that that site will be a SOM. We can use SmartCyp to produce a tool that predicts whether a given molecule is a reactant: Given that a molecule is a reactant if and only if at least one of its sites is a SOM, we created a tool SmartCyp-React(α, m) that predicts whether m is a reactant of the isoform α, which is TRUE whenever SmartCyp$_\tau(\alpha, m, s)$ is below some learned threshold τ, for any site s.

We use a learning algorithm to learn τ by internal cross-validation – *i.e.*, the learning algorithm considers various different thresholds to determine the threshold that has the best score. It then uses external cross-validation to estimate the weighted cost of SmartCyp-React$_{\tau^*}(\alpha, \cdot)$, with this best τ^*.

3.3 "All" Variants of the Predictors

Some users may just want to know whether a molecule will react with any CYP450 isoform, but not care which one. We therefore consider the CypReact-All variant that predicts a given molecule m as an "All-reactant" if and only if CP_α predicts it is a reactant, for at least one of the nine CYP450 isoforms α. (Note this uses the nine already-trained $\{CP_\alpha\}$ models – *n.b.*, it does not train a new CP_{All} model to optimize the weighted cost.) We used the same approach to create a combined model for ADMET Predictor-All, over all 9 isoforms, and also for SmartCyp-React-All, over its 3 isoforms: CY2C9, CYPD6, and CYP3A4.

3.4 Results and discussion

3.4.1 Evaluation criterion

As mentioned above, for each CYP isoform α, we first ran the LBM($\text{D}(\alpha)$) learning process to find the best model $\text{CP}_\alpha(\cdot)$, based on all of the training data. To evaluate the quality of this learned model, for each isoform α we then used a evaluation algorithm that ran this LBM$(\cdot)$ process five more times, as a form of external cross-validation [44]. That evaluation algorithm divided $\text{D} = \text{D}(\alpha)$ into five subsets, then it ran the entire LBM$(\cdot)$ process on four of these five subsets – recall this LBM process will run internal cross-validation to identify the best base learner. Note this might lead to different base learners, and different values of γ, in different iterations. It then ran the resulting learned classifier on the hold-out subset. It repeated this process five times, and reported the average score. Note this means our evaluation algorithm will run each base learner+parameter (*e.g.*, SVM) at least five times for the external cross-validation, and another $5 \times 5 = 25$ times for the internal cross-validation runs, each time on a slightly different subset of the $\text{D}(\alpha)$ dataset.

3.4.2 Average Weighted Cost

Based on the discussion above, our goal is to optimize the weighted cost (Equation 3.1); this section reports those scores, for each of our various classifiers: CypReact, Majority-Classifier (which just returns "No, not a reactant" for each molecule, and so serves as a baseline), SmartCyp-React (for the 3 CYP isoforms {CYP2C9, CYP2D6, CYP3A4} that it considers), and ADMET Predictor for all 9 isoforms. Notice we also consider the "All"

Table 3.4: The 5-fold cross-validation (top, in cyan; average±standard-deviation) and hold-out testing (bottom, yellow) Weighted Cost of the CypReact, SmartCyp, ADMET Predictor, and MajorityClassifier models, for each CYP450 enzyme. Recall that smaller values of Weighted Cost are better.

Classifier	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4	All
5-fold CV results										
CypReact	0.313 ± 0.05	0.207 ± 0.03	0.278 ± 0.03	0.290 ± 0.05	0.359 ± 0.07	0.343 ± 0.06	0.296 ± 0.02	0.247 ± 0.05	0.289 ± 0.05	0.218
ADMET†	0.347	0.331	0.369	0.430	0.400	0.393	0.309	0.339	0.478	0.408
SmartCyp -React					0.682 ± 0.03		0.740 ± 0.05		0.702 ± 0.01	0.629‡
Majority Classifier	0.830	0.322	0.463	0.435	0.692	0.668	0.827	0.444	1.455	2.496
Hold-out testing results										
CypReact	0.177	0.038	0.077	0.143	0.141	0.217	0.099	0.104	0.152	0.183
ADMET†	0.298	0.320	0.288	0.375	0.500	0.475	0.190	0.311	0.333	0.497
SmartCyp -React					1.032		0.831		0.752	0.669‡
Majority Classifier	0.935	0.098	0.098	0.536	1.094	0.833	0.868	0.146	1.154	2.450

†ADMET is the abbreviation for ADMET Predictor. ‡These results are based on only the 3 isoforms that SmartCyp covers: CYP2C9, CYP2D6 and CYP3A4.

situation (see below). These results appear in the top (cyan-color) portion of Table 3.4, and Figure 3.3. Note that lower score means better performance: a perfect result is 0, and the weighted cost of the baseline (MajorityClassifier) varies from 0.322 to 1.455. Paired two-sided t-tests showed that each CYP450 predictor in CypReact is statistically significantly better than the baseline, at $p < [1.91e^{-5}, 1.56e^{-3}, 1.02e^{-4}, 1.89e^{-3}, 1.68e^{-4}, 9.1e^{-6}, 1.95e^{-5}, 2.83e^{-5}, 8.29e^{-7}]$ over the 9 CYPs (in order shown in Table 3.4). After applying Bonferroni correction, we can claim that all are significantly ($p < 0.0056$) better than the baseline. We also see that our CypReact is statistically better than SmartCyp-React, for $\alpha \in$ {CYP2C9, CYP2D6, CYP3A4}, at p<$[3.38e^{-6}, 8.63e^{-7}, 1.06e^{-7}]$.

The final column of Tables 3.4 shows that CypReact-All performs better than ADMET Predictor-All and SmartCyp-React-All.

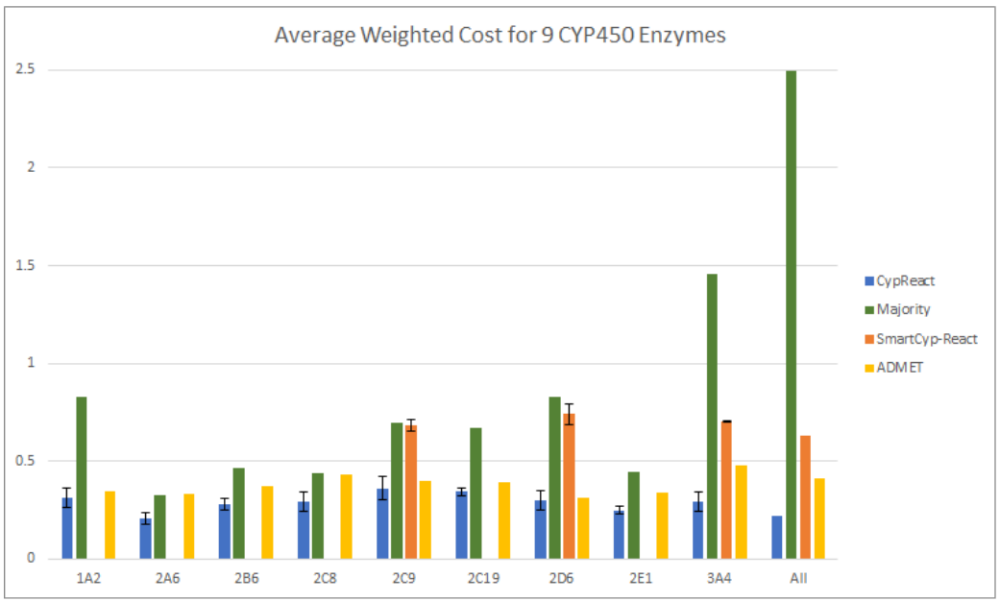


Figure 3.3: Average Weighted Cost for CypReact, SmartCyp-React and ADMET Predictor (lower is better).

3.4.3 Jaccard Scores

Another obvious measure to deal with imbalanced data is the Jaccard score, which is intersection over union, with respect to the minority class:

$$\text{Jaccard} \quad = \quad \frac{\#\texttt{True\_Positives}}{\#\texttt{True\_Positives} \; + \; \#\texttt{False\_Positives} \; + \; \#\texttt{False\_Negatives}} \; .$$

The closer to 1.0, the better the Jaccard score is. The top (cyan-color) portion of Table 3.5 reports the Jaccard score for each of these classifiers; note these are the same classifiers discussed above – *i.e.*, each is still trained to optimize the weighted loss function.

A simple paired t-test shows that CypReact is statistically significantly better than the baseline, at $p <$[$4.17e^{-6}$, $2.60e^{-4}$, $2.36e^{-5}$, $1.46e^{-4}$, $4.41e^{-5}$, $5.01e^{-6}$, $3.23e^{-5}$, $6.44e^{-6}$, $3.25e^{-6}$] over the 9 CYPs. CypReact is also statistically better than SmartCyp-React, for all three isoforms considered, at $p <$ [$4.90e^{-6}$, $5.25e^{-6}$, $1.54e^{-7}$].

The final column of Table 3.5 shows that CypReact-All performs better than ADMET Predictor-All and SmartCyp-React-All, in terms of this criterion as well.

3.4.4 Cost Curves

Above, we motivated the use of a cost-sensitive learner, and suggested we learn classifiers that optimize Equation 3.1, with $\beta = 5$. Below we show the confusion matrix for the

Table 3.5: The 5-fold cross-validation (top, cyan; average±standard-deviation) and hold-out testing (bottom, yellow) Jaccard score of the CypReact, SmartCypand ADMET Predictor models, for each CYP450 enzyme. We did not show the Majority Classifier as it was 0.0 for all isoforms. Recall that larger values of Jaccard score are better.

Classifier	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4	All
5-fold CV results										
CypReact	0.389 ± 0.03	0.275 ± 0.05	0.282 ± 0.03	0.251 ± 0.04	0.302 ± 0.04	0.304 ± 0.02	0.406 ± 0.04	0.306 ± 0.02	0.545 ± 0.03	0.687
ADMET†	0.379	0.157	0.201	0.157	0.286	0.278	0.448	0.211	0.463	0.506
SmartCyp -React					0.092 ± 0.03		0.164 ± 0.03		0.296 ± 0.01	0.369‡
Hold-out testing results										
CypReact	0.605	0.455	0.364	0.556	0.651	0.567	0.714	0.375	0.593	0.690
ADMET†	0.488	0.150	0.118	0.231	0.385	0.298	0.621	0.147	0.437	0.459
SmartCyp -React					0.094		0.143		0.248	0.331‡

†ADMET is the abbreviation for ADMET Predictor. ‡These results are based only on the 3 isoforms that SmartCyp covers: CYP2C9, CYP2D6 and CYP3A4.

CypReact classifier for the CYP2D6 isoform (see Table 3.3):

Truth⇒ Prediction⇓	R	N
R	#True_Positives$_{\beta=5}$ = 235	#False_Positives$_{\beta=5}$ = 308
N	#False_Negatives$_{\beta=5}$ = 35	#True_Negatives$_{\beta=5}$ = 1054

(3.2)

The previous sections evaluated this classifier, using the evaluation function Equation 3.1, with $\beta = 5$ – which we will write as Equation 3.1$[\beta = 5]$. We can also consider evaluating simple variants of this classifier, and others, with respect to other values of β.

To be more precise: the core component of each learned CypReact system actually returns a score for each input molecule m; the β value is used to set a threshold $\tau(\beta) = \frac{1}{\beta+1} \in [0, 1]$, for determining whether that molecule should be labeled Reactant – here m is labeled "Reactant" if that score is larger than $\tau(\beta)$, and otherwise, "NonReactant". Equation 3.2 corresponds to the performance-time value of $\beta = 5$; we clearly produce different confusion matrices for other values of β.

This idea motivates "Cost Curves" [49]: a curve of (x, y) pairs, where each x-value corresponds (indirectly) to a value of β, and the y-value measures how well this fixed classifier does, with respect to this β. The orange curve in Figure 3.4 corresponds to the CypReact(

2D6, $\cdot$) classifier, based on the points $(x_\beta,\, y_\beta)$, computed as

$$x_\beta = \frac{p(R) \;\times\; M(\,N\,|\,R\,)}{p(R) \times M(\,N\,|\,R\,) \;+\; [1-p(R)] \times M(\,R\,|\,N\,)} \;=\; \frac{0.17 \times \beta}{0.17 \times \beta \;+\; 0.83 \times 1} \tag{3.3}$$

$$y_\beta = y_\beta(C) = FN(C) \;\times\; x_\beta \;\;+\;\; FP(C) \;\times\; (1 - x_\beta) \tag{3.4}$$

where in general

- $p(R)$ is the ratio of reactants over all instances (which corresponds to the bottom cyan-color row of Table 3.1, "#R / #Total" – and so is 0.17 for our dataset)

- $M(\,N\,|\,R\,)$ is the misclassification cost of predicting an instance with real label "Reactant" as "Non-Reactant" – which recall we defined as β – and the other misclassification cost $M(\,R\,|\,N\,)$, here is set to 1

- $FN(C) \;=\; \frac{\texttt{\#False\_Negatives}}{\texttt{\#False\_Negatives + \#True\_Positives}}$ is the false negative rate for this classifier – which using Equation 3.2, is $\frac{35}{35+235} \approx 0.13$ for C =CypReact(2D6, $\cdot$) and
  $FP(C) \;=\; \frac{\texttt{\#False\_Positives}}{\texttt{\#False\_Positives + \#True\_Negatives}}$ is the false positive rate – here $\frac{308}{308+1054} \approx 0.23$

(Here, we include C as an argument of y_β, FN and FP, to show its dependence.)

With a little algebra, using Equation 3.1, we find that

$$y_\beta(C) \;=\; \frac{Cost_\beta(\,C,\, D\,)}{p(R) \times M(\,N\,|\,R\,) \;\;+\;\; [1-p(R)] \times M(\,R\,|\,N\,)} \;=\; \frac{Cost_\beta(\,C,\, D\,)}{0.17 \times \beta \;\;+\;\; 0.83 \times 1} \tag{3.5}$$

which is why it is often called "Normalized Expected Cost". Now notice that the denominator does not depend on the classifier, which means a classifier that optimizes Equation 3.1, will be optimizing this $y_\beta(C)$ value.

Note the x values are independent of the classifier itself, and so can vary independently. This allows us to compare different classifiers, over a range of different β-values, to see when each classifier is best.[5] This is why we consider the full range of values $x_\beta \in [0, 1]$ for the x-axis, then use Equation 3.4 to compute the associated normalized expected cost y_β (which is related to Cost($\cdot$); see Equation 3.5). In operation, the user would first identify the

[5] In addition, we could consider other "label distributions": While our training dataset had 17-to-83 mix of Reactants to NonReactants (see the bottom cyan-color row of Table 3.1), we could alternatively consider a dataset that had a 20-to-80 mix, or 50-to-50, or whatever, by varying the $p(R)$ value. However, we did not do this here.

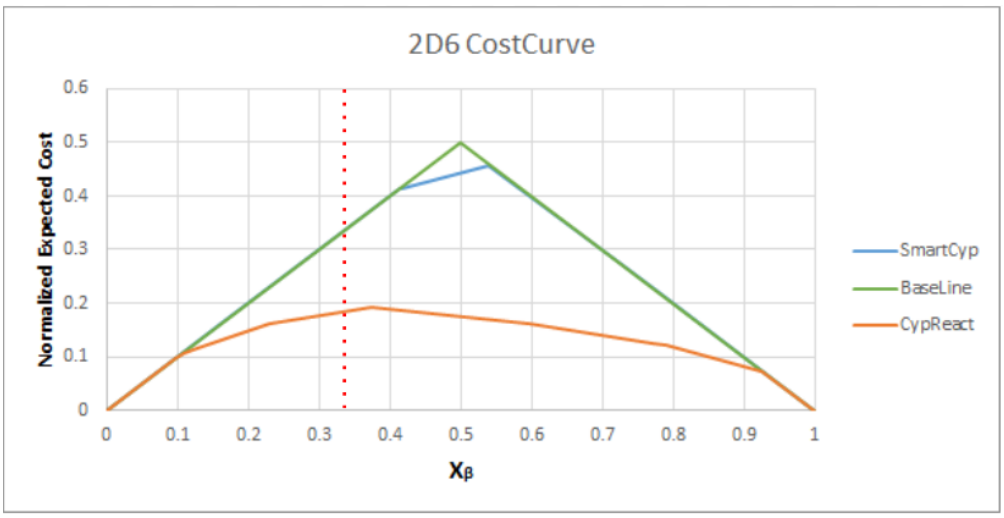


Figure 3.4: The CostCurves for CypReact(2D6, ·) in orange, SmartCyp-React(2D6, ·) in blue, and the baseline in green (covering much of SmartCyp-React(2D6, ·)). The red vertical dashed line corresponds to $\beta = 5$ here. We see that CypReact dominates SmartCyp-React over all x_β values – which means for all misclassification costs, β.

Cost Matrix (Table 3.3), which here means stating the β value. That user would then use Equation 3.3 to compute the x_β value, then adjust the classifier to this value of β – call it C^β – which updates the classifier's confusion matrix, which is then used to determine the associated $y_\beta(C^\beta)$ cost.

We can also see how well other classifiers would perform over the entire range of β values, which induces values for both x_β-values $\in [0, 1]$ and then y_β, based on x_β and the confusion matrix (based on β). We can consider some trivial classifiers: The "JustSayN" classifier just returns "NonReactant" for each instance; it is easy to see that, for any x, its Normalized-Expected-Cost (*i.e.*, its y-value) will be the $y = x$ line. There is no reason for any classifier to ever be above this line – *i.e.*, if for any x_β value, a classifier $C(\cdot)$ had a cost that was above this $y_\beta = x_\beta$ line, it would be silly to use $C(\cdot)$, as we would get a better score by just ignoring that $C(\cdot)$ classifier, and instead using the JustSayR classifier.

Similarly, the cost curve for the "JustSayR" classifier, which just returns "Reactant", would trivially be the $y = 1 - x$ line. Again, there is no reason to consider a classifier that is above that line. We consider the minimum of these two lines to be the "Baseline" – show as the GreenLine in Figure 3.4 – and for any classifier, will only show the cost-curve portion that appears below this curve.

The blue line in Figure 3.4 shows the curve for SmartCyp(Cyp2D6, ·). We see that

it matches the Baseline for much of the domain $x_\beta \in [0, 1]$, dipping below only around $x_\beta \in (0.41, 0.54)$. Moreover, we see that our CypReact(2D6, $\cdot$) system is strictly better (that is, smaller) than SmartCyp(Cyp2D6, $\cdot$) for many x_β values, and it is never worse.

This suggests that one should prefer the CYP2D6 model of CypReact over the one of SmartCyp as CypReact is always at least as good, and often better. (While it did not happen here, the curves for different classifiers could cross – meaning there would be a region of x_β-values where classifier#1 is best, and another where classifier#2 is best. Here, once we knew the β value for the target domain, we could compute the x_β value, then find which classifier is best here – that is, use $C_\beta = \arg\min_C\{y_\beta(C)\}$.)

We also found that CypReact is similarly superior to SmartCyp-React for CYP3A4 and CYP2C9; see the Cost Curves for CYP3A4 and CYP2C9 in the Supporting Information.

3.4.5 ROC and AUC

CostCurves allow the user to decide, for each β, which specific classifier to use – meaning one might use one 2D6 classifier for $\beta = 5$, here corresponding to the value $x_\beta = 0.51$, but another classifier for $\beta = 8$ (leading to $x_\beta = 0.62$). If one just wanted to use a single classifier, we could evaluate a classifier based on its AUROC (area under the ROC [receiver operating characteristic] curve), which essentially measures how well its performance "on average", over the entire range of β values. In general, a curve's ROC curve is a set of (x, y) points, where here x is the FalsePositiveRate and y is the TruePositiveRate, as you vary some natural parameter. Note that the shape of the ROC curve for a perfect classifier is essentially a Gamma "Γ", while the baseline is a diagonal line ("/") with a slope of one. This means the AUROC of a perfect classifier is 1.0, and of the baseline is 0.5.

Figure 3.5 shows the ROC curves for CypReact and SmartCyp-React for 2D6, as well as the baseline "random guess" classifier. We see that CypReact performs much better than SmartCyp-React here – with AUROC of 0.872 versus 0.490. Table 3.6 shows the AUROC values for all nine isoforms, showing they range from 83% to 92% for CypReact, and from 49% to 60% for SmartCyp. (The Supporting Information presents the CypReact ROC curves for the other eight CYP450 isoforms, and for SmartCyp where relevant.)

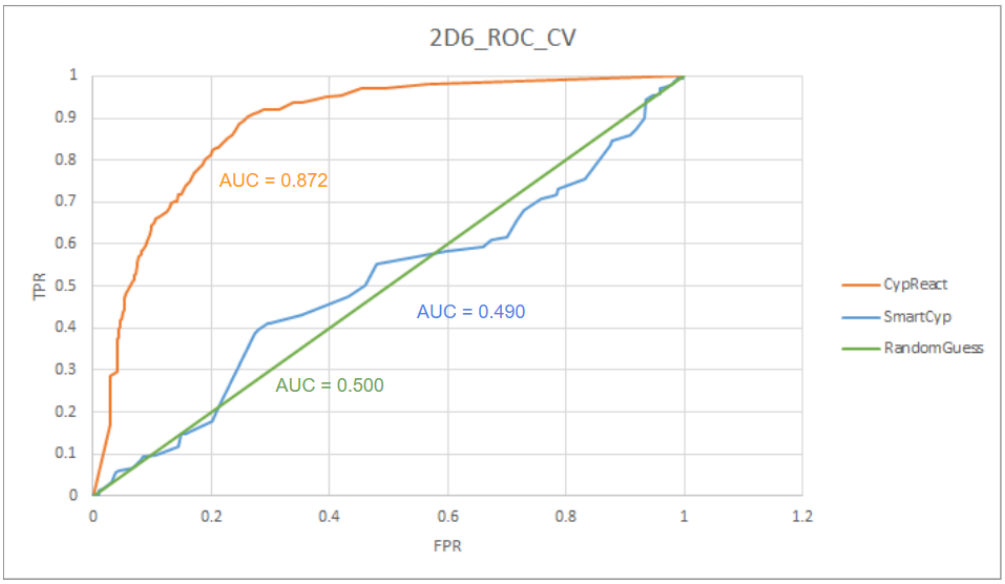


Figure 3.5: ROC curve of CypReact and SmartCyp-React for CYP2D6. (Note we did not take the convex hull, to better illustrate the shapes.)

Table 3.6: Area under ROC of CypReact on the nine CYP450 isoforms.

	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4
CypReact	86%	84%	86%	84%	83%	83%	87%	87%	92%
SmartCyp-React					51%		49%		60%
ADMET Predictor	79%	77%	74%	68%	74%	75%	81%	75%	75%

3.4.6 Results on a New Dataset

After computing the cross-validation scores on the training/testing set, LBM then learned nine CypReact models, each based on *all* 1632 molecules, then tested these learned models on new, disjoint datasets – one for each isoform α. We produced these datasets by first identifying 69 new molecules that were reactants to at least one isoform, and combining them with 100 molecules that are known to be non-reactants to all 9 isoforms; see bottom 3 rows (colored yellow) in Table 3.1.

The lower (yellow-colored) portions of Tables 3.4 and 3.5 shows the results of these learned CypReact algorithms on these validation sets – showing (respectively) average weighted cost and Jaccard scores. It also presents the results of SmartCyp-React and ADMET Predictor on these datasets.

These results confirm that CypReact works extremely well, and in particular, better than the other CYP450 reaction prediction systems considered.

3.4.7 Summary

CypReact is a family of CYP450 reaction predictors that contains nine subtools, each built for one CYP450 enzyme individually. Each CypReact classifier is trained to minimize the average weighted cost score for its associated CYP450 isoform, based on a weighted cost that penalizes each false negative five times more than each false positive. Our empirical results show that our classifiers exhibit very good weighted cost scores, and AUROC scores – here ranging from 83% to 92% – and that they significantly outperform SmartCyp-React and ADMET Predictor.

Chapter 4

CypBoM: A software tool for Predicting "Bond of Metabolism" for CYP450 Enzymes

This chapter has 3 contributions, which appear in the following 3 sections: Section 4.1 motivates and defines "Bond of Metabolism" (BoM); it also relates this to the more standard term "Site of Metabolism", and shows that there are 3 types of BoMs. Section 4.2 describes how we created two datasets, listing the BoMs of many hundreds of molecules, which are publicly available on `https://drive.google.com/open?id=1NQPFKVnJC8f0XXV9lpeAzW4YXDmrWMdU`.

Section 4.3 describes how the CypBoM$_{\eta-\eta}$ tool is learned and its performance.

4.1 Bond of Metabolism

As chemical reactions always involve breaking existing bonds between a pair of atoms, or forming new bonds, we define a new term, BoM (bond of metabolism), that explicitly describes the location where a chemical reaction occurs in terms of bonds and information about the reaction. Each BoM is specified by a 4-tuple:

$$\langle X, Y; \text{ReactionType; ReactionID} \rangle \tag{4.1}$$

The initial two components $\langle X, Y \rangle$ represent a pair of atoms, where the associated bond either already appears in the molecule, or is formed in a reaction. We consider 3 types of BoMs, which we think are sufficient to represent all changes to chemical bonds occurring in Phase I metabolism (illustrated in Figure 4.1):

1. η-η: written "$\langle i, j\rangle$": the existing or potential bond connecting two non-hydrogen atoms whose indices are i and j. For example, the $\langle 20, 19\rangle$ (resp., $\langle 4, 5\rangle$) pair represents the single bond between atom C.20 and atom N.19 – see the arc whose label ends with "R1" in Figure 4.1 (resp., the π bond between atom C.4 and C.5 – see R5). The $\langle 9, 21\rangle$ pair indicates a possible bond (not in the initial molecule) between atoms N.9 and C.21 (see R4).

2. η-H: written "$\langle i, H\rangle$": the bond or bonds between a non-hydrogen atom with index i and any number of its attached hydrogens. For example, $\langle 5, H\rangle$ represents the bond between the C.5 atom and its connected hydrogen atom – see Reaction R2.

3. η-SPN: written "$\langle i, S\rangle, \langle i, P\rangle$ or $\langle i, N\rangle$": a bond that is not present in the initial compound, but is formed with a Sulphur, Phosphorus or Nitrogen atom by sharing its lone pair electrons. For example, in Reaction R3, the N.19 atom is oxidized to form a N-O bond without modifying the existent bonds in the Olanzapien substrate; the new bond is recorded as $\langle 19, N\rangle$.

In Equation 4.1, the "ReactionType" records the type of the reaction occurs on the bond $\langle X, Y\rangle$. The reaction types can be either high level, low level, or a mix of them, based on the user's interest. For example, when a N-Dealkylation reaction occurs, the user can either record it as N-Deakylation (low level) or cleavage (high level). While there are an arbitrary number of possible reaction types, we will focus on the following mix of both level types: Oxidation, Cleavage, EpOxidation, Reduction, Hydroxylation, S(sulfur)-Oxidation, N(nitrogen)-Oxidation, P(phosphorus)-Oxidation and Cyclization.

To explain "ReactionID", note that we view a reaction as a mapping from a substrate to one of its stable, detectable metabolites; this can involve changes to more than one bond. We therefore use "ReactionID" to connect the individual bonds affected in a single reaction. To illustrate, note the $\langle 20, 19;$ Cleavage; R1$\rangle$ reaction (presented above) is actually one step of an N-demethylation reaction, which also includes an oxidation $\langle 20, H;$ Oxidation; R1$\rangle$. Here, both steps use the same ReactionID R1 to show they are part of the same reaction, which produces both N-desmethyl olanzapine and formaldehyde.

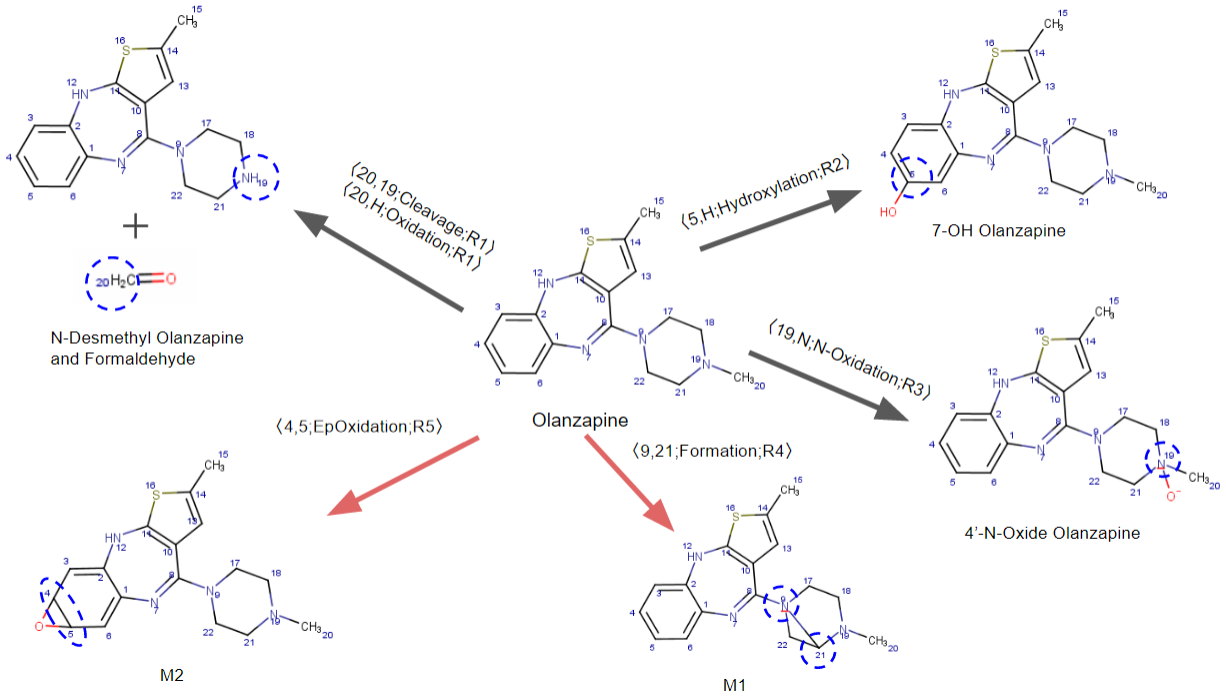


Figure 4.1: Three substrate-metabolite(s) pairs, showing the BoMs (beside each arrow) representing the associated reactions for olanzapine [50]. The blue circles indicate the locations where the reaction occurs. The red arrows and the corresponding metabolites M1, M2 are not real and used for illustration purposes only.

4.2 EBoMD Dataset

Zaretzki's dataset [27] is a public bioinformatic dataset that lists the SoMs for 679 substrates[1] for the nine highest expressed CYP450 isozymes – CYP1A2, CYP2A6, CYP2B6, CYP2C8, CYP2C9, CYP2C19, CYP2D6, CYP2E1 and CYP3A4 [51]. It has been widely used in CYP450 metabolism studies and was used in developing many *in silico* metabolism prediction tools, such as RS-Predictor [52], FAME2, etc. For our research, we converted Zaretzki's SoM dataset to a corresponding BoM dataset, by applying the following process:

- For every compound, we checked its entry in PubChem [18] and Drugbank [53], and read through the papers that reported its metabolic activities for CYP450 Phase I metabolism. We compared the substrate to its detected stable metabolites reported in the papers and then recorded the bonds changed in the reaction as BoMs. Note we did not include purported metabolites if they were not reported to be observable and stable.

[1]They claimed 680 CYP450 substrates; however one of them (phenanthrene) appeared twice.

- A reaction is treated as a pair between a substrate and its stable detectable metabolites, such as the olanzapine and 7-OH olanzapine pair in Figure 4.1. Note that there are often other downstream reactions – and in some cases, that further result may be better known. For example, the nicotine to norcotine reaction is well known. However, this process begins with nicotine to cotinine, where cotinine is stable and detectable [54]. We therefore view cotinine as the "result" of nicotine. In such cases, the intermediate metabolites are used in the representation of the reaction and their further metabolites are ignored.

  If more than one metabolite is produced in one reaction, all changed bonds will be recorded as BoMs by sharing the same ReactionID. Returning to Figure 4.1, we actually could have used $\langle 20,\ 19;\ \text{N-Demethylation};\ \text{R\#}\rangle$ to represent the upper left reaction. However, our dataset instead uses $\langle 20,\ H;\ \text{Oxidation};\ \text{R1}\rangle$ and $\langle 20,\ 19;\ \text{Cleavage};\ \text{R1}\rangle$, because this reaction actually produces two stable, detectable products: N-desmethyl olanzapine and formaldehyde.

- We include the BoM for a substrate and metabolite pair, as long as the metabolite is reported as detected in the paper, regardless of its concentration, amount or percentage, because we do not want to miss any plausible/potential metabolites.

- While most η-η reactions involve modifying an existing bond, some will instead form a new bond. For example, the "oxidative cyclization" reaction will form an η-η bond, which we record as $\langle C.i,\ N.j;\ \text{Cyclization};\ \text{R\#}\rangle$[2]. This was the only such η-η-forming reaction we encountered – *i.e.*, this occurred in only 4 of the 829 η-η-reactions in our EBoMD.

Our analysis found some reactions (for these 679 compounds) that were not in the original Zaretzki's dataset, and also fixed several mistakes. We let EBoMD ("Edmonton Bond-of-Metabolism Dataset") refer to resulting dataset, including 829 η-η BoMs out of 16418 η-η bonds from the 679 compounds in Zaretzki's dataset; see Table 4.1. We then created a hold-out dataset of 74 relevant compounds (called EBoMD2), including drugs, pesticides, etc., extracted from the DrugBank databse and publications [55], [56], from which we extract 115 η-η BoMs out of 1728 η-η bonds, using the methods shown above; see Table 4.2.

[2]The four molecules having oxidative cyclization reaction are JPC-2056, proguanil, chlorproguanil and PS-15.

Table 4.1: Distribution of the three different types of chemical bonds for nine CYP450 isoforms, in the EBoMD Dataset.

	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4
#Reactants	271	105	151	142	226	218	270	145	475
η-η bonds									
#BoMs	340	127	155	183	224	235	297	171	596
#Non-BoMs	5486	1656	2731	3166	5101	4655	6029	2068	12181
η-H bonds									
#BoMs	495	160	208	228	368	358	441	230	811
#Non-BoMs	2552	807	1394	1627	2458	2265	3090	1025	6192
η-SPN bonds									
#BoMs	28	13	12	11	26	20	33	13	68
#Non-BoMs	493	140	214	245	404	396	549	161	964

Table 4.2: Distribution of the η-η bonds for nine CYP450 isoforms. in the EBoMD2 Dataset.

	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4	All
# Compounds	21	14	13	10	15	17	29	11	43	74
η-η bonds										
#BoMs	29	21	14	12	24	23	37	21	71	115
#Non-BoMs	384	225	180	260	324	310	682	61	1049	1613

Both EBoMD and EBoMD2 datasets are publicly available on `https://drive.google.com/open?id=1NQPFKVnJC8f0XXV9lpeAzW4YXDmrWMdU`.

We believe that essentially all metabolic reactions involve a combination of 1 or more of these three BoMs. The eventual overall CypBoM(α, m) process will take a CYP450 enzyme α and a given molecule m (given as either a SMILES string or SDF file) as input, then sort each of its current bonds into those 3 types. It will then pass each such bond to one of 3 classifiers – one for each type – which then generates features appropriate for that type of bond, then uses a learned model (for this type of bond) to decide whether that specific bond is a BoM; see Figure 4.2. That figure shows how CypBoM would predict BoMs of the phenacetin molecule for CYP1A2.

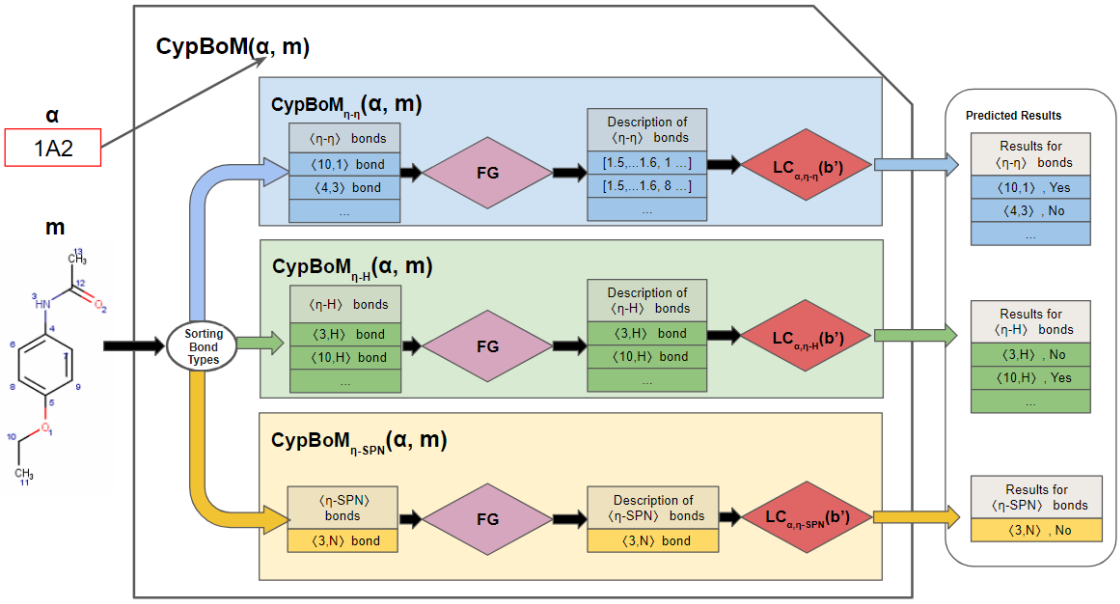


Figure 4.2: An overview of how CypBoM predicts the BoMs of phenacetin for CYP1A2.

4.3 The CypBoM$_{\eta-\eta}$ Classifier

Our current implementation only predicts whether a bond within the molecule is reactive, without giving the ReactionType and ReactionID. Moreover, it deals only with the η-η bonds; see the blue flow in Figure 4.2. This *in silico* metabolism prediction tool, CypBoM$_{\eta-\eta}(\alpha, m)$, uses a machine learned model that, given a molecule m and an isoform α, predicts which of m's η-η bonds are BoM$_{\eta-\eta}$s (the η-η bonds of η-η BoMs), with respect to that α isoform. This involves making a binary decision at each of the η-η bonds: Yes if the η-η bond is modified during a reaction and otherwise No. Note that our CypBoM$_{\eta-\eta}$ tool was not trained to handle the situation where a η-η bond is formed, as it is so rare (only 4 of the 829 η-η-reactions in our EBoMD).

This section will follow the flow of the blue line in Figures 4.2 and 4.3. We will describe the features we used, the process for learning the BoM$_{\eta-\eta}$ classifier, CypBoM$_{\eta-\eta}$, and the performance of that classifier.

4.3.1 Feature Generation

In general, a classifier assigns a label to each instance, described as a vector of values; here, each instance corresponds to both a compound, and one of its bonds – eg, Diuron and its C.6-N.5 bond in Figure 4.4. Each element in that vector corresponds to a feature, whose

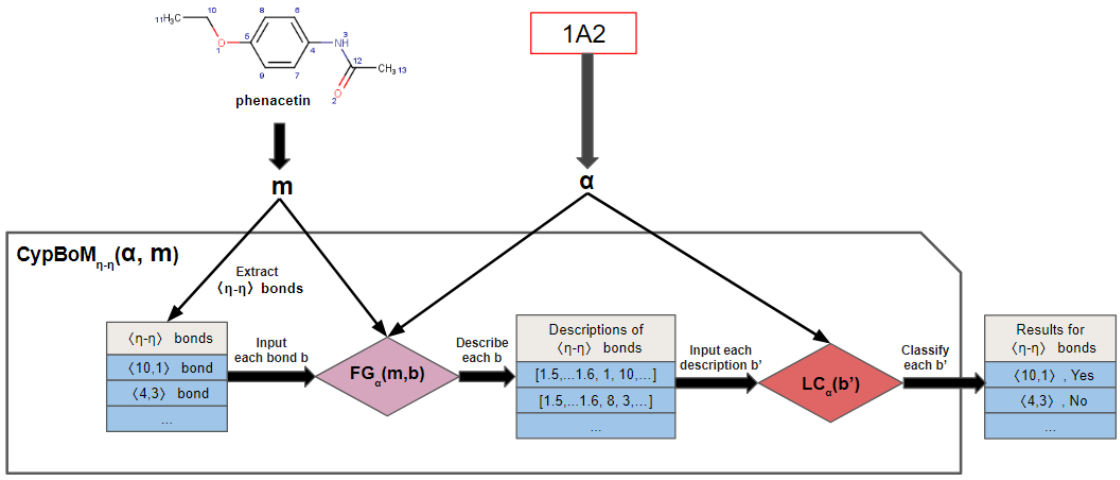


Figure 4.3: Implementation of the CypBoM$_{\eta-\eta}$.
For each CYP450 isoform $\alpha \in \{$ 1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1, 3A4 $\}$, FG$_\alpha(m, b)$ generates features for the bond b, and LC$_\alpha(b')$ classifies that (description of the instance) b' as either a BoM$_{\eta-\eta}$ or not.

value is calculated based on the properties, such as molecular weight, electronegativity, etc., of the corresponding bond and/or the compound.

Here, we generate the features for each η-η bond within the molecule based on the chemical descriptors, fingerprints, atom types and number of connected atoms. Note that $\langle 20, 19;$ Cleavage; R1$\rangle$ could also be written as $\langle 19, 20;$ Cleavage; R1$\rangle$ – *i.e.*, $\langle 20, 19\rangle$ and $\langle 19, 20\rangle$ are the same, which means the naive encoding would need to have two versions of each. To avoid duplication, we seek a canonical version by reordering the two atoms connected by a η-η bond. For any η-η bond $\langle i, j\rangle$, we first reorder the two connected atoms i and j, following: (1) if the atomic numbers of atom i and atom j are different, then the pair is reordered, if necessary, to start with the atom having the smaller atomic number; (2) otherwise, compute $\mathbf{CA}(i)$ (the connected atoms feature for atom i, which will be explained later) and $\mathbf{CA}(j)$, and the bond is reordered as $\langle j, i\rangle$ if and only if $\mathbf{CA}(j) < \mathbf{CA}(i)$. For example, in Figure 4.1, the bond between N.19 and C.20 is reordered as $\langle 20, 19\rangle$, as C.20's atomic number is 6 (as it is a Carbon) while N.19's is 7 (Nitrogen); and the bond between C.15 and C.14 as $\langle 15, 14\rangle$, as $\mathbf{CA}(15) < \mathbf{CA}(14)$.

Then the features of bond $\langle i, j\rangle$ are (see Table 4.3, from left to right):

- 21 molecular features computed from 7 molecular descriptors (see table 4.4) and 14 molecular fingerprints, such as carbonyl, amino, etc.

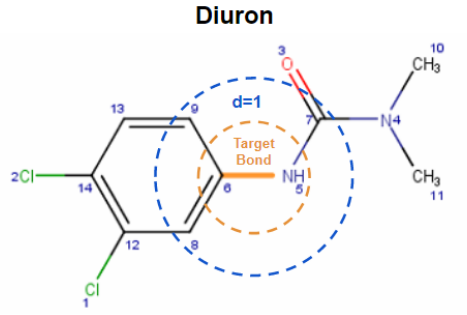
Table 4.3: The number of features, of each category, for each η-η instance.

Bond $\langle i, j \rangle$	Molecular DES‡	Molecular FP‡	Connected Atoms	Bond FP	Bond AtomTypes	Bond AtomDES	Neighbor AtomTypes	Neighbor AtomDES
1	7	14	2	30	23	10×2	23×4	10× 14 × 2

‡FP and DES are the abbreviations for FingerPrints and DEScriptors.

- 75 bond features including:
  - 2 **ConnectedAtoms** features, $\mathbf{CA}(i)$ and $\mathbf{CA}(j)$. Assume the bond $\langle i, j \rangle$ is removed from the structure of the molecule; then $\mathbf{CA}(i)$ is the total number of atoms in the remaining substructure that includes atom i minus 1. Note that $\mathbf{CA}(i)$ equals $\mathbf{CA}(j)$ if the $\langle i, j \rangle$ pair is part of a ring.
  - 30 **BondFingerprint** features.
  - 23 **BondAtomTypes** features that describe the atom types of atom i and j using 23 SYBYL atom types. For example, in Figure 4.4, the **BondAtomTypes** features C.ar and N.am for bond $\langle 6, 5 \rangle$ each is 1, as C.6 is a aromatic carbon and N.5 is the nitrogen in an amide.
  - 10×2 **BondAtomDescriptors** features that use 10 atomic descriptors to describe the physicochemical properties of atoms i and j.
- 372 environment features of the bond $\langle i, j \rangle$ including:
  - 23×4 **NeighborAtomTypes** features that use the same 23 atom types in bond atom types to describe the neighbor atoms' types at depth from 1 to 4 (see Figure 4.4).
  - $10 \times 14 \times 2$ **NeighborAtomDescriptor** features; each computed as the average value of one of the 10 atomic descriptors of the atoms that matches one of the 14 atom types at depth 1 or 2.

Note FAME2 [11] also included environment features, and all features are computed using the Chemistry Development Kit (CDK) [57].



ID	BondAtomType				NeighborAtomType					NeighborAtomDescriptors				
Bond	C.sp3	C.ar	N.am	...	$C.3_{d=1}$	$C.ar_{d=1}$	...	$Cl_{d=4}$	...	$C.noar_{d=1}$ AtomDegree	$C.noar_{d=1}$ Atom_ASA	...	$Cl_{d=2}$ AtomDegree	...
⟨6,5⟩	0	1	1	...	0	2	...	1	...	3	0.049917	...	0	...

Figure 4.4: Listing several bond atom types, neighbor atom types and descriptors and explaining how some are calculated.

The orange bond ⟨6, 5⟩ is the target bond. The value 1 of $Cl_{d=4}$ in the **NeighborAtomType** set indicates that there is only one chlorine atom among all atoms that are four bonds away from atoms C.6 (to the left, away from N.5), and from N.5 (to the right, away from C.6). Because there is only one non-aromatic carbon atom that is one bond away from the target bond and its atom degree is 3, the $C.noar_{d=1}$AtomDegree value in the **NeighborAtomDescriptors** set is calculated as $3/1 = 3$.

Descriptor name	Type	Description
ALOGPDescriptor	Real	the ALOGP value
APolDescriptor	Real	the APol value
HBondAcceptorCountDescriptor	Integer	the # of acceptors of hydrogen bonds
HBondDonorCountDescriptor	Integer	the # of donors of hydrogen bonds
MomentOfInertiaDescriptor	Real	MOMI value
RotatableBondsCountDescriptor	Integer	the # of rotatable bonds
TPSADescriptor	Real	the TPSA value
WeightDescriptor	Real	the weight of the molecule
XLogPDescriptor	Real	the xlogP value
ASA	Real	the accessible surface area

Table 4.4: The molecular descriptors calculated by the CDK toolkit

Descriptor name	Type	Description
AtomDegreeDescriptor	Integer	the atom degree
AtomHybridizationDescriptor	Integer	the hybridization of an atom
AtomValenceDescriptor	Integer	the valence of an atom
EffectiveAtomPolarizabilityDescriptor	Real	the effective atom polarizability value
PartialSigmaChargeDescriptor	Real	the sigma partial charge of an atom
PartialTChargeMMFF94Descriptor	Real	the total partial charges of an atom
PiElectronegativityDescriptor	Real	the π electronegativity of an atom
SigmaElectronegativityDescriptor	Real	the sigma electronegativity of an atom
StabilizationPlusChargeDescriptor	Real	the stabilization of the + charge

Table 4.5: The atomic descriptors calculated by the CDK toolkit

4.3.2 Feature Selection

Here, there are 473 features for each instance corresponding to a compound/bond pair. We then use a feature selection technique to reduce the number of features, by removing the "bad" features, such as the ones that are redundant or apparently irrelevant; this process is designed to improve the efficiency and performance of the learned model. This feature selection method is essentially the same as the one described in Section 3.1.4, that is: (1) remove those features whose values are the same for all the instances in the dataset, then (2) rank the remaining features according to their information gain values with respect to the label (which here is 1 for BoM and 0 otherwise), then retain the top-N attributes and remove the rest. Note that the number N is learned for each CYP450 enzyme by the learning algorithm; see Section 4.3.3.

4.3.3 Cost-Sensitive Learner

Because the number of reactive η-η bonds (8%) is much less than the number of non-reactive η-η bonds (92%) in the EBoMD dataset, again, we use a cost-sensitive learning algorithm **LBM**, similar to the one described in Chapter 3, to learn the classifiers that predict which of these η-η bonds are BoM$_{\eta-\eta}$s for the CYP450 enzymes, but with the following alterations:

- The classifier for each CYP450 enzyme is learned on the substrates of that enzyme (see Table 4.1).
- The target of the learning algorithm is now to achieve the optimal Jaccard score, which has been described in Section 3.4.3, rather than minimizing the average cost.
- We now use an internal cross-validation process to find the best values for three parameters:
  - Instead of using a fixed $\beta = 5$ in the cost matrix, we treat β as a parameter learned from its integer candidate set $\beta \in \{2, 3, ..., 10\}$.
  - **LBM** only considers a single base learner – random forest – and we use internal cross-validation to identify the appropriate batch size $t \in \{90, 95, 100\}$.
  - As mentioned in Section 3.1.4, we now learn the number of features to keep in the reduced dataset, $N \in \{100, 200, 300, 400, \text{All}\}$.
- We attempted to have roughly the same proportion of the various types of reactions, in the folds. The stratification of the cross-validation is based molecules following a subclass strategy: label each substrate with a reaction type, according to the η-η BoMs within that substrate, with the priorities: Reduction > EpOxidation > Cleavage > Oxidation. Note that all reactions of each molecule can only appear in either the training or the validation dataset.

4.3.4 Implementation

CypBoM$_{\eta-\eta}$ is a family of 9 CYP450 BoM$_{\eta-\eta}$ classifiers, one for each of isoform; see Figure 4.3. CypBoM$_{\eta-\eta}$ takes a molecule m and a CYP450 enzyme α as input. The η-η bonds within m are then extracted and each bond b is input to the FG$_\alpha(m, b)$ function that encodes b as a vector of values associated with features relevant to the CYP α. The resulting

vector b' is then passed to the learned classifier $\mathrm{LC}_\alpha(b')$, which returns either "Yes, BoM" or "No".

The implementation of CypBoM is written in Java using the WEKA [48] APIs.

4.4 Results and discussion

This section presents cross-validation results of $\mathrm{CypBoM}_{\eta-\eta}$ generated by applying the learner **LBM** to the EBoMD dataset and then comparing the performance of the learned $\mathrm{CypBoM}_{\eta-\eta}$ with ADMET Predictor(v.8.5.1.1) [16], FAME2 [11] and Meteor Nexus(v.3.0.1) [13] on three different hold-out test datasets: EBoMD2, HdFame and HdMeteor. Note that both HdFame and HdMeteor datasets are generated based on the EBoMD2 dataset.

The evaluation metrics used are AUROC (see Section 3.4.5), Jaccard score

$$\text{Jaccard} \quad = \quad \frac{TP}{TP + FP + FN} \tag{4.2}$$

and MCC (Matthews correlation coefficient)

$$\text{MCC} \quad = \quad \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{4.3}$$

which is a balanced measure of the quality of a binary classifier. Note that TP, TN, FP, FN are the numbers of true positives, true negatives, false positives and false negatives in the confusion matrix, respectively.

4.4.1 Cross-Validation Result

We use the internal-external cross-validation which is described in Section 3.4.1 to compute the cross-validation results of $\mathrm{CypBoM}_{\eta-\eta}$. Table 4.6 shows the Jaccard, MCC and AUROC scores.

4.4.2 Comparison with ADMET Predictor

ADMET Predictor (Simulations Plus, Inc., Lancaster, California, USA) is a commercial software that predicts over 140 properties, including Phase I site of metabolism and metabolites for molecules, for each of the nine major CYP450 enzymes. In order to compare our tool with ADMET Predictor (v.8.5.1.1) on the holdout test dataset EBoMD2, we focused on

Table 4.6: Cross-validation results compared with the random classifier

	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4
Jaccard Score									
CypBoM$_{\eta-\eta}$	0.523	0.401	0.574	0.467	0.443	0.594	0.543	0.350	0.516
Random †	0.055	0.067	0.051	0.052	0.040	0.046	0.045	0.071	0.045
MCC									
CypBoM$_{\eta-\eta}$	0.668	0.542	0.714	0.617	0.597	0.733	0.690	0.478	0.667
Random †	0	0	0	0	0	0	0	0	0
AUROC									
CypBoM$_{\eta-\eta}$	0.925	0.818	0.956	0.873	0.896	0.916	0.933	0.832	0.917
Random †	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500

†Random means the random classifier.

only BoM$_{\eta-\eta}$ and converted the ADMET Predictor's predicted results to η-η bonds by (1) checking every substrate-metabolite pair to determine whether the changed bond within the structure is a η-η bond or not, and (2) if so, checking whether the predicted η-η bond is true for each individual CYP450 enzyme. The Jaccard score, MCC value and AUROCs are presented in Figures 4.5, 4.6, 4.7, respectively; see Table B.1 in the Appendix for more details.

4.4.3 Comparison with Meteor Nexus

We also compared our CypBoM$_{\eta-\eta}$ tool with Meteor Nexus (v.3.0.1), which is a commercial tool for predicting the metabolic fate of a compound, on the HdMeteor dataset that includes the same 74 molecules in the EBoMD2 dataset, where each η-η bond within a molecule is labeled as BoM$_{\eta-\eta}$ if and only if it is a reactive bond for any of the nine CYP450 enzymes. Unlike ADMET Predictor, Meteor Nexus predicts SoMs and metabolites for the CYP450 enzymes[3] but not for every individual CYP450 enzyme, and thus we use the rules described in Section 4.4.2 to transform its results to the η-η ones. Similar to Section 3.3, we also used a variant version of CypBoM$_{\eta-\eta}$, CypBoM$_{\eta-\eta}$-All, that predicts a η-η bond is a reactive if it is predicted so for any of the nine major CYP450 isoforms.

The results are shown in Table 4.7 (left).

[3] Meteor Nexus predicts metabolites for CYP450 enzymes without specifying which CYP450 isoforms are used. This means it might be using CYP450 enzymes other than the nine major ones

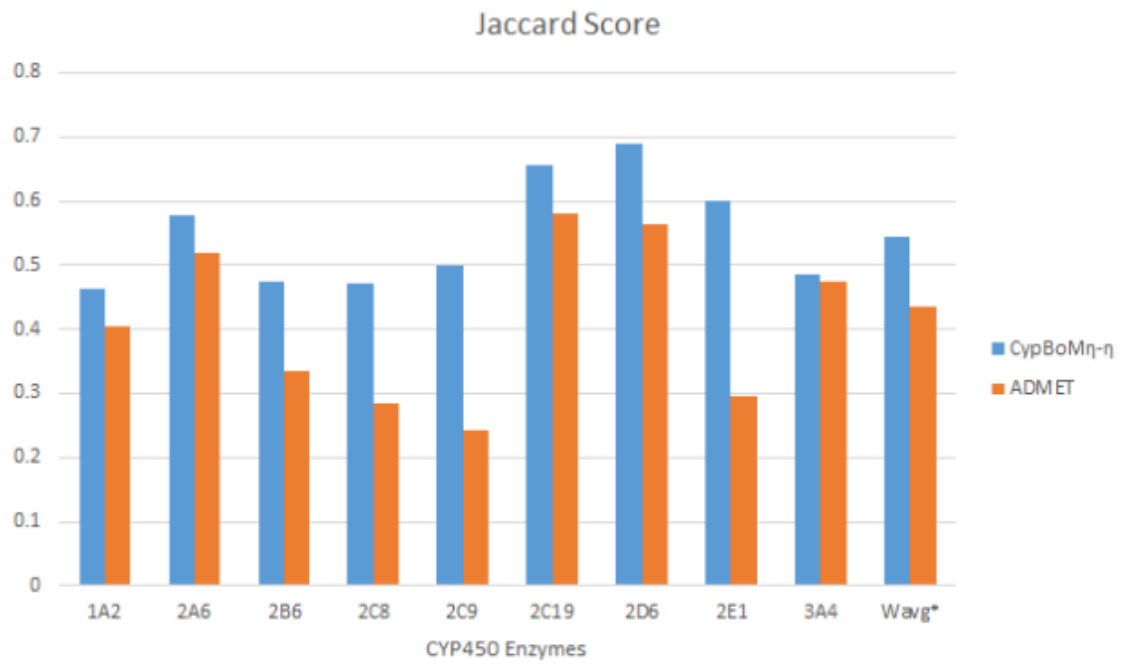


Figure 4.5: Jaccard scores for CypBoM and ADMET Predictor, on the EBoMD2 dataset. Note that Wavg* means "macro weighted average value".

4.4.4 Comparison with FAME2

FAME2 is a free software tool for predicting the sites of metabolism for a molecule, with respect to the three CYP450 isoforms: 2C9, 2D6 and 3A4. Here, we compare FAME2 with our CypBoM$_{\eta-\eta}$ on the HdFame dataset that contains 60 molecules in the EBoMD2 dataset, where each molecule is a reactant for at least one of the three CYP450 enzymes, and each η-η bond within a molecule is labeled as BoM$_{\eta-\eta}$ if and only if it is a reactive bond for any of the three isoforms. Because knowing which atoms are reactive without corresponding metabolites is not sufficient to identify the reactive bonds, we used the following set of rules to convert the predicted SoMs to η-η bonds. A predicted SoM is treated as a reactive η-η bond only if:

- the predicted SoM is a carbon atom that is connected to an oxygen or nitrogen.
- both carbon and sulfur atoms within a $\langle C, S\rangle$ bond are predicted as SoMs.
- the nitrogen and carbon on the diagonal of a ring are predicted as SoMs and the real reaction leads to a ring rearrangement reaction. (Note that all bonds within the ring

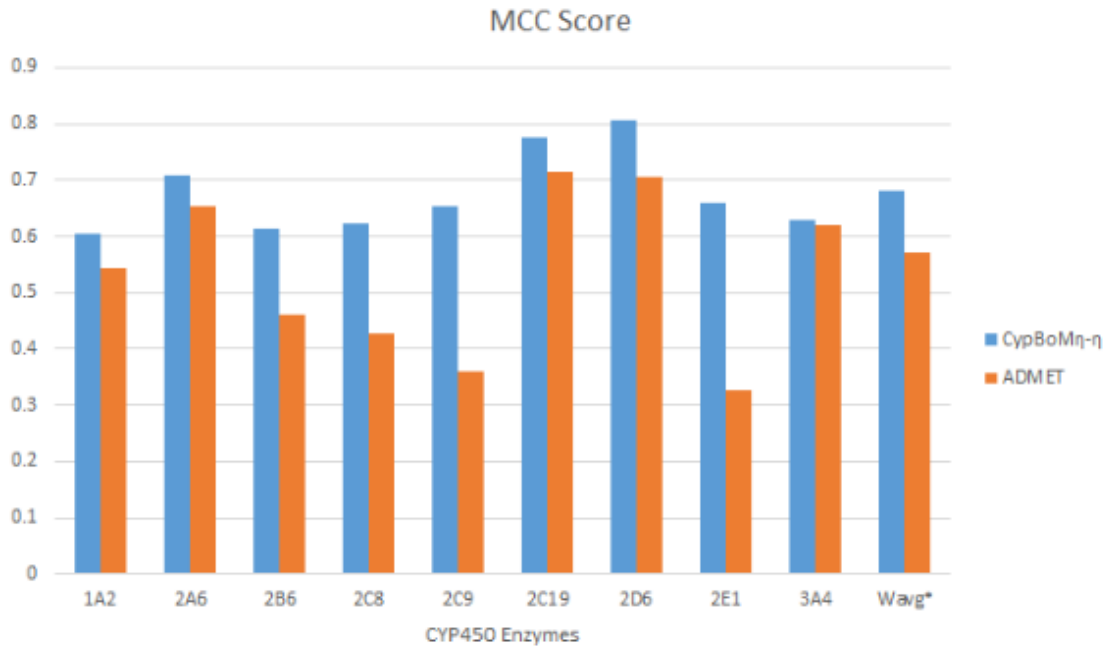


Figure 4.6: MCC score for CypBoM and ADMET Predictor, on the EBoMD2 dataset. Note that Wavg* means "macro weighted average value".

Table 4.7: Hold-out results for the CYP450 enzyme family compared with Meteor Nexus (left); and the hold-out results for CYP2C9, 2D6 and 3A4 compared with FAME2 (right);

	Jaccard	MCC
CypBoM$_{\eta-\eta}$	0.540	0.685
Meteor Nexus	0.417	0.565

	Jaccard	MCC
CypBoM$_{\eta-\eta}$	0.556	0.697
FAME2	0.543	0.684

are treated as BoM$_{\eta-\eta}$s in this case. An example of the ring rearrangement reaction can be found in the acetaminophen molecule in our EBoMD dataset.)

After converting all predicted SoMs of the 60 compounds in the HdFame dataset to η-η bonds, they are compared with the real BoM$_{\eta-\eta}$s to generate the confusion matrix. In order to compare with FAME2, we use a variant of CypBoM$_{\eta-\eta}$, CypBoM$_{\eta-\eta}$-Tri, where CypBoM$_{\eta-\eta}$-Tri predicts a η-η bond to be reactive if and only if it is predicted as reactive by any of the three LC$_\alpha$s for isoforms $\alpha \in$ {CYP2C9, CYP2D6, CYP3A4}. The results are shown in Table 4.7 (right).

4.4.5 Summary

CypBoM$_{\eta-\eta}$ is a family of 9 CYP450 BoM$_{\eta-\eta}$ classifiers, one for each of the CYP450

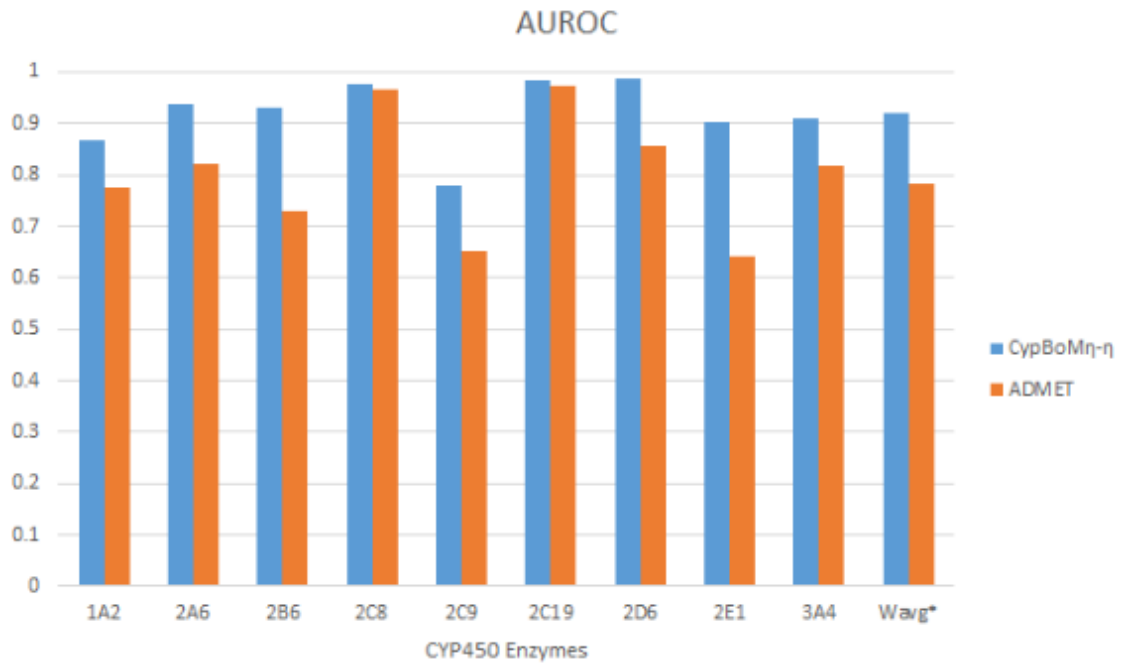


Figure 4.7: AUROC for CypBoM and ADMET Predictor, on the EBoMD2 dataset. Note that Wavg* means "macro weighted average value".

enzymes. Each CypBoM classifier is trained to maximize the Jaccard score for its associated CYP450 isoform. Our empirical results show that our classifiers exhibit very good Jaccard, MCC and AUROC scores, and they work better than ADMET Predictor, Meteor Nexus and FAME2 in predicting the η-η bonds of η-η BoMs for CYP450 enzymes.

Chapter 5

Conclusion

In this dissertation, we introduce two *in silico* metabolism prediction tools, CypReact and CypBoM, for predicting the substrates and reactive η-η bonds for the nine most highly expressed CYP450 enzymes. Our experimental results with these tools help confirm our two hypothesis: (1) it is possible to learn a model that can accurately predict which small molecules will react with various CYP450 enzymes, and (2) it is possible to predict where within the molecule, the reaction takes place. In order to predict the location where a metabolic reaction occurs, we need to declare what a reaction is and provide a clear, informative approach to describe the location within a molecule – this lead to our definition of BoM (bond of metabolism). Because we also needed an appropriate dataset that shows these BoMs, we developed our EBoMD dataset based on the Zaretzki's dataset; see `https://drive.google.com/open?id=1NQPFKVnJC8f0XXV9lpeAzW4YXDmrWMdU`.

Our empirical results show that both our tools outperform other relevant tools described earlier and thus, could be used as essential components of a suite of *in silico* metabolism prediction tools for accurately predicting the products of Phase I, Phase II and microbial metabolism in humans.

While our CypReact and CypBoM$_{\eta-\eta}$ work extremely well, there is still room for improvement.

Improve the quality of datasets: Both CypReact and CypBoM$_{\eta-\eta}$ are learned from variants of the Zaretzki's dataset, which is published years ago. Due to the low number of positive instances in the datasets, both our tools may be imperfect; we anticipate they would work even better if trained on a larger dataset, containing more relevant compounds.

Implement the complete CypBoM**:** As shown in Figure 4.2, the complete CypBoM tool

includes three components – each predicting the reactive bonds for one of the three bond types: η-η, η-H and η-SPN. This dissertation presented the CypBoM$_{\eta-\eta}$ component that predicts the reactive η-η bonds using corresponding features and found it worked well.

As η-H and η-SPN bonds have different physicochemical properties and the relevant reaction types are different, this may require finding other more relevant features to achieve high-quality classifiers.

In our future work, we will implement the CypBoM$_{\eta-H}$ and CypBoM$_{\eta-SPN}$, and finally implement the complete CypBoM.

Bibliography

[1] H. Van De Waterbeemd and E. Gifford, "Admet in silico modelling: Towards prediction paradise," *Nature reviews. Drug discovery*, vol. 2, no. 3, p. 192, 2003.

[2] K. A. Delaney and K. C. Kleinschmidt, "Chapter 12. biochemical and metabolic principles," in *Goldfrank's Toxicologic Emergencies, 9e*, L. S. Nelson, N. A. Lewin, M. A. Howland, R. S. Hoffman, L. R. Goldfrank, and N. E. Flomenbaum, Eds. New York, NY: The McGraw-Hill Companies, 2011. [Online]. Available: `accesspharmacy.mhmedical.com/content.aspx?aid=6504103`.

[3] L. L. Furge and F. P. Guengerich, "Cytochrome p450 enzymes in drug metabolism and chemical toxicology: An introduction," *Biochemistry and Molecular Biology Education*, vol. 34, no. 2, pp. 66–74, 2006.

[4] D. W. Nebert and D. W. Russell, "Clinical importance of the cytochromes p450," *The Lancet*, vol. 360, no. 9340, pp. 1155–1162, 2002.

[5] Z. Pan and D. Raftery, "Comparing and combining nmr spectroscopy and mass spectrometry in metabolomics," *Analytical and bioanalytical chemistry*, vol. 387, no. 2, pp. 525–527, 2007.

[6] *Eawag-bbd pathway prediction system.* Last visited 2017-09-20. [Online]. Available: `http://eawag-bbd.ethz.ch/predict.`.

[7] J. G. Jeffryes, R. L. Colastani, M. Elbadawi-Sidhu, T. Kind, T. D. Niehaus, L. J. Broadbelt, A. D. Hanson, O. Fiehn, K. E. Tyo, and C. S. Henry, "Mines: Open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics," *Journal of cheminformatics*, vol. 7, no. 1, p. 44, 2015.

[8] P. Anzenbacher and E. Anzenbacherová, "Cytochromes p450 and metabolism of xenobiotics," *Cellular and Molecular Life Sciences*, vol. 58, no. 5, pp. 737–747, 2001.

[9] M. Rostkowski, O. Spjuth, and P. Rydberg, "Whichcyp: Prediction of cytochromes p450 inhibition," *Bioinformatics*, vol. 29, no. 16, pp. 2051–2052, 2013.

[10] P. Rydberg, D. E. Gloriam, and L. Olsen, "The smartcyp cytochrome p450 metabolism prediction server," *Bioinformatics*, vol. 26, no. 23, pp. 2988–2989, 2010.

[11] B. Manavalan, R. G. Govindaraj, T. H. Shin, M. O. Kim, and G. Lee, "Ibce-el: A new ensemble learning framework for improved linear b-cell epitope prediction," *Frontiers in immunology*, vol. 9, 2018.

[12] S. E. Adams, "Molecular similarity and xenobiotic metabolism," PhD thesis, University of Cambridge, 2010.

[13] C. A. Marchant, K. A. Briggs, and A. Long, "In silico tools for sharing data and knowledge on toxicity and metabolism: Derek for windows, meteor, and vitic," *Toxicology mechanisms and methods*, vol. 18, no. 2-3, pp. 177–187, 2008.

[14] *Stardrop*, Last visited 2017-05-21. [Online]. Available: `https://www.optibrium.com/stardrop/`.

[15] P. Rydberg, D. E. Gloriam, and L. Olsen, "The smartcyp cytochrome p450 metabolism prediction server," *Bioinformatics*, vol. 26, no. 23, pp. 2988–2989, 2010.

[16] *Admet predictor (2018) simulations plus, inc., lancaster, california, usa.* Last visited 2019-03-26, 2018. [Online]. Available: `https://www.simulations-plus.com/software/admetpredictor/metabolism/`.

[17] A. Dalby, J. G. Nourse, W. D. Hounshell, A. K. Gushurst, D. L. Grier, B. A. Leland, and J. Laufer, "Description of several chemical structure file formats used by computer programs developed at molecular design limited," *Journal of chemical information and computer sciences*, vol. 32, no. 3, pp. 244–255, 1992.

[18] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, *et al.*, "Pubchem 2019 update: Improved access to chemical data," *Nucleic acids research*, vol. 47, no. D1, pp. D1102–D1109, 2018.

[19] M. J. Macielag, "Chemical properties of antimicrobials and their uniqueness," in *Antibiotic Discovery and Development*, Springer, 2012, pp. 793–820.

[20] A. D. McNaught and A. D. McNaught, *Compendium of chemical terminology*. Blackwell Science Oxford, 1997, vol. 1669.

[21] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[22] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 8, pp. 1226–1238, 2005.

[23] M. M. Mafarja, D. Eleyan, I. Jaber, A. Hammouri, and S. Mirjalili, "Binary dragonfly algorithm for feature selection," in *2017 International Conference on New Trends in Computing Sciences (ICTCS)*, IEEE, 2017, pp. 12–17.

[24] R. Kohavi *et al.*, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Ijcai*, Montreal, Canada, vol. 14, 1995, pp. 1137–1145.

[25] G. C. Cawley and N. L. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research*, vol. 11, no. Jul, pp. 2079–2107, 2010.

[26] S. Tian, Y. Djoumbou-Feunang, R. Greiner, and D. S. Wishart, "Cypreact: A software tool for in silico reactant prediction for human cytochrome p450 enzymes," *Journal of chemical information and modeling*, vol. 58, no. 6, pp. 1282–1291, 2018.

[27] J. Zaretzki, M. Matlock, and S. J. Swamidass, "Xenosite: Accurately predicting cyp-mediated sites of metabolism with neural networks," *Journal of chemical information and modeling*, vol. 53, no. 12, pp. 3373–3383, 2013.

[28] D. S. Wishart, T. Jewison, A. C. Guo, M. Wilson, C. Knox, Y. Liu, Y. Djoumbou, R. Mandal, F. Aziat, E. Dong, S. Bouatra, I. Sinelnikov, D. Arndt, J. Xia, P. Liu, F. Yallou, T. Bjorndahl, R. Perez-Pineiro, R. Eisner, F. Allen, V. Neveu, R. Greiner, and A. Scalbert, "Hmdb 3.0–the human metabolome database in 2013," *Nucleic acids research*, vol. 41, no. D1, pp. D801–D807, 2012.

[29] *Kegg database.* Last visited 2017-08-03. [Online]. Available: `http://www.genome.jp/kegg/kegg1.html.`.

[30] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A. C. Guo, Y. Liu, A. Maciejewski, D. Arndt, M. Wilson, V. Neveu, A. Tang, G. Gabriel, C. Ly, S. Adamjee, Z. T. Dame, B. Han, Y. Zhou, and D. S. Wishart, "Drugbank 4.0: Shedding new light on drug metabolism," *Nucleic acids research*, vol. 42, no. D1, pp. D1091–D1097, 2013.

[31] S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang, and S. H. Bryant, "Pubchem substance and compound databases," *Nucleic acids research*, vol. 44, no. D1, pp. D1202–D1213, 2015.

[32] F. D. Gunstone, J. L. Harwood, and A. J. Dijkstra, *The lipid handbook with CD-ROM*. CRC press, 2007.

[33] *Chemaxon's marvin suite.* Last visited 2017-11-25, 2017. [Online]. Available: `https://www.chemaxon.com/download/marvin-suite/`.

[34] C. Ioannides, *Cytochromes P450: role in the metabolism and toxicity of drugs and other xenobiotics*. royal society of chemistry, 2008.

[35] A. G. Wilson, *New Horizons in Predictive Drug Metabolism and Pharmacokinetics*. Royal Society of Chemistry, 2015.

[36] E. L. Willighagen, J. W. Mayfield, J. Alvarsson, A. Berg, L. Carlsson, N. Jeliazkova, S. Kuhn, T. Pluskal, M. Rojas-Chertó, O. Spjuth, G. Torrance, C. T. Evelo, R. Guha, and C. Steinbeck, "The chemistry development kit (cdk) v2. 0: Atom typing, depiction, molecular formulas, and substructure searching," *Journal of Cheminformatics*, vol. 9, no. 1, p. 33, 2017.

[37] (2011). Biovia: The keys to understanding mdl keyset technology. Last visited 2017-11-10, [Online]. Available: `http://accelrys.com/products/pdf/keys-to-keyset-technology.pdf`.

[38] Y. Djoumbou Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner, and D. S. Wishart, "Classyfire: Automated chemical classification with a comprehensive, computable taxonomy," *Journal of cheminformatics*, vol. 8, no. 1, p. 61, 2016.

[39] (2007). Smarts - a language for describing molecular patterns. Last visited 2017-01-25, [Online]. Available: `http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html`.

[40] M. Sud, "Mayachemtools: An open source package for computational drug discovery," *Journal of chemical information and modeling*, vol. 56, no. 12, pp. 2292–2297, 2016.

[41] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[42] C. Elkan, "The foundations of cost-sensitive learning," in *International joint conference on artificial intelligence*, Lawrence Erlbaum Associates Ltd, vol. 17, 2001, pp. 973–978.

[43] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their applications*, vol. 13, no. 4, pp. 18–28, 1998.

[44] E. Alpaydin, *Introduction to machine learning*. MIT press, 2014.

[45] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[46] T. G. Dietterich, "Ensemble methods in machine learning," *Multiple classifier systems*, vol. 1857, pp. 1–15, 2000.

[47] D. Ballabio, F. Biganzoli, R. Todeschini, and V. Consonni, "Qualitative consensus of qsar ready biodegradability predictions," vol. 99, pp. 1193–1216, Sep. 2017.

[48] E. Frank, M. A. Hall, and I. H. Witten, *The WEKA Workbench. Online Appendix for "Data Mining: Practical machine learning tools and techniques"*. Morgan Kaufmann, 2016.

[49] C. Drummond and R. C. Holte, "Cost curves: An improved method for visualizing classifier performance," *Machine learning*, vol. 65, no. 1, pp. 95–130, 2006.

[50] B. J. Ring, J. Catlow, T. J. Lindsay, T. Gillespie, L. K. Roskos, B. J. Cerimele, S. P. Swanson, M. A. Hamman, and S. A. Wrighton, "Identification of the human cytochromes p450 responsible for the in vitro formation of the major oxidative metabolites of the antipsychotic agent olanzapine.," *Journal of Pharmacology and Experimental Therapeutics*, vol. 276, no. 2, pp. 658–666, 1996.

[51] U. M. Zanger and M. Schwab, "Cytochrome p450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation," *Pharmacology & therapeutics*, vol. 138, no. 1, pp. 103–141, 2013.

[52] J. Zaretzki, C. Bergeron, P. Rydberg, T.-w. Huang, K. P. Bennett, and C. M. Breneman, "Rs-predictor: A new tool for predicting sites of cytochrome p450-mediated metabolism applied to cyp 3a4," *Journal of chemical information and modeling*, vol. 51, no. 7, pp. 1667–1689, 2011.

[53] D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, *et al.*, "Drugbank 5.0: A major update to the drugbank database for 2018," *Nucleic acids research*, vol. 46, no. D1, pp. D1074–D1082, 2017.

[54] M. Nakajima and T. Yokoi, "Interindividual variability in nicotine metabolism: C-oxidation and glucuronidation," *Drug metabolism and pharmacokinetics*, vol. 20, no. 4, pp. 227–235, 2005.

[55] S. Rendic, "Summary of information on human cyp enzymes: Human p450 metabolism data," *Drug metabolism reviews*, vol. 34, no. 1-2, pp. 83–448, 2002.

[56] S. Gad, *Preclinical Development Handbook: ADME and Biopharmaceutical Properties*, ser. Pharmaceutical Development Series. Wiley, 2008, ISBN: 9780470249024. [Online]. Available: https://books.google.ca/books?id=QtXXn%5C_pEI3MC.

[57] E. L. Willighagen, J. W. Mayfield, J. Alvarsson, A. Berg, L. Carlsson, N. Jeliazkova, S. Kuhn, T. Pluskal, M. Rojas-Chertó, O. Spjuth, G. Torrance, C. T. Evelo, R. Guha, and C. Steinbeck, "The chemistry development kit (cdk) v2. 0: Atom typing, depiction, molecular formulas, and substructure searching," *Journal of Cheminformatics*, vol. 9, no. 1, p. 33, 2017.

[58] Wikipedia contributors, *Receiver operating characteristic — Wikipedia, the free encyclopedia*, [Online; accessed 1-May-2019], 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Receiver_operating_characteristic&oldid=888671034.

[59] Wikipedia contributors, *Atomic mass unit — Wikipedia, the free encyclopedia*, [Online; accessed 10-April-2019], 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Atomic_mass_unit&oldid=886667601.

[60] Wikipedia contributors, *Metabolite — Wikipedia, the free encyclopedia*, [Online; accessed 10-April-2019], 2018. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Metabolite&oldid=859269996.

[61] Wikipedia contributors, *Functional group — Wikipedia, the free encyclopedia*, [Online; accessed 10-April-2019], 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Functional_group&oldid=889869762.

[62] Wikipedia contributors, *Drug metabolism — Wikipedia, the free encyclopedia*, [Online; accessed 10-April-2019], 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Drug_metabolism&oldid=878834763.

Appendix A

Glossary

AUC: area under the curve [58].

AUROC: area under the receiver operating characteristic (ROC) curve [58].

BoM**:** bond of metabolism that describes where a reaction occurs in terms of bonds.

$\text{BoM}_{\eta-\eta}$**:** the η-η bond of the η-η BoM, which is also called reactive η-η bond.

CYP450: Cytochrome P450.

CypReact**:** an *in silico* metabolism prediction tool that predicts the substrates for CYP450 enzymes.

CypBoM**:** an *in silico* metabolism prediction tool that predicts the locations of the BoMs; here we focus on the component that deals with η-η bonds.

Dalton: unified atomic mass unit: 1 dalton equals 1.66×10^{-27} kg [59].

Metabolite: the intermediate or terminal product of a compound in the metabolic reaction [60].

Experimental metabolite identification: identify the metabolites of a compound through chemical experiments.

Functional group: a group of atoms that undergo the same or similar chemical reaction [61].

Phase I metabolism and reaction: Phase I metabolism is a part of drug metabolism that converts a compound into its more polar metabolite(s) through Phase I reactions, including oxidation, reduction, hydrolysis, etc., catalyzed by enzymes, such as CYP450 enzymes [62].

Phase II metabolism and reaction: Phase II metabolism is another part of drug metabolism that conjugates a compound with endogenous molecule and forms a larger, more water soluble metabolite, which is catalyzed by transferases enzymes [62].

Reactive site/bond/atom: the site/bond/atom whose properties are changed in a chemical reaction.

Site: a position within the molecule, could be a atom or a bond.

SoM**:** site of metabolism that describes where a reaction occurs in terms of atoms.

Xenobiotic compounds: chemical compounds that are not naturally produced or expected to be present within the organism

Appendix B

Supplemental Material

Table B.1: Hold-out results for the nine CYP450 enzymes compared with ADMET Predictor and the random classifier.

	1A2	2A6	2B6	2C8	2C9	2C19	2D6	2E1	3A4	All
Jaccard Score										WAvg*
CypBoM$_{\eta-\eta}$	0.463	0.577	0.474	0.471	0.500	0.655	0.689	0.600	0.485	0.546
ADMET †	0.405	0.519	0.333	0.278	0.242	0.414	0.563	0.296	0.475	0.434
Random ‡	0.066	0.079	0.067	0.042	0.065	0.065	0.049	0.204	0.060	0.063
MCC										WAvg*
CypBoM$_{\eta-\eta}$	0.605	0.708	0.615	0.623	0.653	0.776	0.806	0.659	0.630	0.681
ADMET †	0.544	0.654	0.461	0.410	0.359	0.563	0.705	0.328	0.619	0.571
Random ‡	0	0	0	0	0	0	0	0	0	0
AUROC										WAvg*
CypBoM$_{\eta-\eta}$	0.866	0.938	0.931	0.978	0.780	0.982	0.987	0.902	0.909	0.922
ADMET †	0.776	0.820	0.731	0.697	0.653	0.751	0.857	0.641	0.818	0.782
Random ‡	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500	0.500

†ADMET is the abbreviation for ADMET Predictor.

‡Random means the random classifier.

*WAvg means macro weighted average value.