Development of Computational Approaches for the Prediction of Regioselectivity and the Likely Products of Xenobiotic Metabolism

Cumulative Dissertation
with the aim of achieving the degree
Doctor rerum naturalium (*Dr. rer. nat.*)

at the Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics
Universität Hamburg

submitted by

Christina de Bruyn Kops

November 2020

Gutachter/innen:
Prof. Dr. Johannes Kirchmair
Prof. Dr. Gerhard Wolber

Tag der Disputation:
12. Februar 2021

Acknowledgements

First and foremost, I would like to thank my supervisor, Prof. Dr. Johannes Kirchmair, for the chance to do the research for this thesis and for his guidance, ideas and advice over the past few years. I really appreciate his enthusiasm and the inspiration he provided.

I am grateful to the entire ACM/COMP3D group, both current and former members, for the interesting discussions about science and fun conversations about other topics. I very much appreciate the warm welcome I received from Nils and the inclusivity and support as I began my research project. I would also like to thank Prof. Dr. Matthias Rarey and the AMD group for welcoming me into their group meetings and events, especially in the beginning when the ACM group was so small. Further, thanks to Gerd and Jörn not only for the IT support but also for all the socializing over lunch.

My gratitude goes as well to all my collaborators, both internal and external, especially Martin Šícho whose work has been so integral to my research. Thanks also to Anya, Conrad, Nils, and Anke for the chance to take enjoyable breaks from my thesis topic to explore and contribute to their interesting research topics.

Additionally, many thanks to Neann, though we did not have the chance to collaborate directly. I really appreciate our conversations about career goals and life, as well as the social support during the lockdown.

For the funding of my final year of work on this dissertation, I would like to thank the Zentrale Ausschuss für Nachwuchsförderung of Universität Hamburg for granting me an *Abschlussstipendium*.

I would also like to thank Prof. Dr. Gerhard Wolber for reviewing this thesis.

In addition, I am grateful to Neann Mathai, Jiayi Kong, and Conrad Stork for each proofreading parts of this dissertation.

Finally, many thanks to my family for their support throughout my entire life and educational journey. In particular, a very special thanks to my sister Alisa for the incredible emotional support, encouragement, and life advice over the past few years.

Abstract

Metabolism plays an important role in the effect that foreign compounds have on the human body. The consideration of xenobiotic metabolism is therefore vital for the development of safe and effective drugs, as well as other compounds such as cosmetics and agrochemicals. Computational approaches can aid in this endeavor by predicting how xenobiotics will be metabolized.

This dissertation focuses on the development of computational methods to predict metabolically labile atom positions, also known as sites of metabolism (SoMs), and to predict the chemical structures of the metabolites.

In the context of SoM prediction, this dissertation presents an exploration of the usefulness of molecular alignment as a basis for predicting cytochrome P450 (CYP) enzyme regioselectivity. This alignment-based approach was evaluated by considering the proximity of the SoMs of the aligned molecules in terms of three-dimensional (3D) distance. It was then possible to further improve the performance of the alignment-based approach by adding a reactivity prediction component to the method.

For metabolite structure prediction, the use of predicted SoMs as a way to address the common problem of an overabundance of predicted metabolites was investigated. Initially, this exploration centered on CYP-mediated metabolism, for which a set of reaction rules was developed based on the scientific literature on CYP-mediated metabolism. Two separate approaches to incorporating SoM prediction were examined: using the predicted SoMs as a hard filter to reduce the number of predicted metabolites and using the predicted SoM probabilities in order to score the predicted metabolites. A scoring approach that combined the predicted SoM probabilities with a binary distinction between common and uncommon CYP-mediated reaction types was found to be effective at ranking the predicted metabolites. These methods were implemented in a tool called GLORY.

Following the development of GLORY, the concept of using predicted SoM probabilities to score the predicted metabolites was further developed and extended to cover both phase I and phase II metabolism. The extended version of the method,

called GLORYx, required additional reaction rules to cover non-CYP metabolic reactions as well as an investigation of the effect of different SoM prediction models on GLORYx's ability to rank the predicted metabolites. An interesting finding was that this ranking was very difficult using the general phase II SoM prediction model, making it necessary to use individual reaction type-specific phase II SoM models for scoring the predicted phase II metabolites.

Both GLORY and GLORYx have been made publicly available for academic research, via a web server and as an open-source software package. The datasets compiled for the development and the validation of the methods developed over the course of this dissertation have also been made publicly available. These datasets include but are not limited to manually curated test datasets of CYP as well as phase I and phase II metabolite data.

Zusammenfassung

Metabolismus spielt eine wichtige Rolle in der Wirkung körperfremder Stoffe auf den menschlichen Organismus. Die Berücksichtigung des xenobiotischen Metabolismus ist daher von entscheidender Bedeutung für die Entwicklung sicherer und wirksamer Arzneimittel sowie anderer Chemikalien, wie zum Beispiel Kosmetika und Pestizide. Computerbasierte Methoden können hierbei durch Vorhersagen des Metabolismus von Xenobiotika helfen.

Diese Dissertation beschäftigt sich mit der Entwicklung computerbasierter Methoden zur Vorhersage metabolisch labiler Atompositionen (SoMs) in Molekülen sowie zur Vorhersage der chemischen Strukturen von Metaboliten.

Im Kontext der SoM-Vorhersage wird in dieser Dissertation das Potential molekularer Überlagerungs-Methoden für die Vorhersage der Regioselektivität von Cytochrom P450 (CYP) Enzymen untersucht. Der auf molekulkare Übereinanderlagerungen basierte Ansatz wurde mittels Distanzberechnung der SoMs übereinandergelagerter Molekülpaare im dreidimensionalen Raum ausgewertet. Aufbauend auf den Erkenntnissen aus dieser Evaluierung konnte die Performanz des computerbasierten Ansatzes durch die Implementierung einer Reaktivitätskomponente weiter verbessert werden.

Für die Vorhersage von Metabolitstrukturen wurde die Verwendung vorhergesagter SoMs als eine Strategie zur Senkung der generell hohen Anzahl vorhergesagter Metabolite untersucht. Anfangs konzentrierte sich diese Untersuchung auf die Vorhersage des CYP-mediierten Metabolismus, für welche ein Set von Reaktionsregeln aus der wissenschaftlichen Literatur abgeleitet wurde. Konkret wurden zwei unterschiedliche Herangehensweisen untersucht. Im ersten dieser beiden Ansätze wurden die vorhergesagten SoMs als harter Filter für die Reduktion der vorhergesagten Metabolite verwendet. Im zweiten Ansatz wurden die berechneten Wahrscheinlichkeiten für vorhergesagte SoMs in Kombination mit einer binären Klassifikation von CYP-Reaktionstypen in “verbreitete” und “seltene” Typen verwendet, um eine Priorisierung der vorhergesagten Metabolite zu erreichen. Der letztere Ansatz hat sich dabei als besonders effektiv herausgestellt. Beide Methoden

wurden in der Software “GLORY” implementiert.

Aufbauend auf GLORY wurde dieser kombinierte Ansatz weiterentwickelt, um auch den gesamten Phase I und Phase II Metabolismus abzudecken. Die erweiterte Methode, genannt GLORYx, erforderte die Entwicklung eines stark erweiterten Satzes an Reaktionsregeln, um auch nicht-CYP-mediierte Reaktionen abzudecken. Weiters wurde auch eine Untersuchung der Performanz und Anwendbarkeit verschiedener SoM Vorhersagemodelle notwendig, um ihren Einfluss auf die Metabolit-Reihung verstehen und optimieren zu können. Im Zuge dieser Analysen stellte sich die Reihung von Phase-II-Metaboliten basierend auf den vorhergesagten SoM-Wahrscheinlichkeiten als besonders herausfordernd dar. Dies machte die Entwicklung und Anwendung individueller SoM-Modelle für die einzelnen Phase-II Reaktionstypen notwendig.

GLORY und GLORYx stehen der akademischen Forschung sowohl in Form eines Webservers als auch eines Open-Source Softwarepakets frei zur Verfügung. Zudem wurden auch die im Rahmen der Entwicklung dieser Methoden aufwendig zusammengestellten Datensätze publiziert.

Contents

Introduction

Humans are constantly surrounded by foreign compounds, now more than ever. These foreign compounds range from man-made to microorganism-made, and we are regularly exposed to them either on purpose or accidentally. These substances include drugs, cosmetics, food additives, household products, industrial products and byproducts, agrochemicals, mycotoxins, and other environmental chemicals. The body's process for dealing with this phenomenon is called xenobiotic metabolism, which refers to the chemical reactions taking place in the body that modify foreign compounds. The overall goal of the body in this regard is to remove the foreign compounds. Along the way, the compounds are transformed into different molecules, called metabolites.

When investigating the effects of a chemical compound on humans, an examination of the interactions of the compound itself with various components of the body does not provide a complete picture of the overall effect of the compound. Metabolism causes the formation of metabolites that have different physicochemical and pharmacological properties compared to the original molecule [1], and the metabolites often go on to interact with the body before being eliminated, for example prolonging the activity of a drug, being the primary cause of a drug's activity, or causing off-target effects [2–5]. These changes in activity, combined with the observation that metabolism is the main clearance pathway for the majority of all drugs [6], mean that metabolism affects the half-life, oral bioavailability, and therefore proper dosing of drugs [6].

Although metabolism generally promotes the elimination of foreign compounds from the body, there can also be unfortunate effects when the metabolites turn out to be toxic or reactive [2, 7]. Failure to recognize the effects of a drug's metabolism can be devastating; drugs such as nefazodone have been withdrawn from the market due to fatal toxicity caused by a metabolite [8, 9]. In fact, a positive correlation has been found between adverse drug reactions (ADRs) and the formation of reactive metabolites of said drugs [10]. In the case of hepatotoxicity in particular, metabolism

to reactive metabolites has been found to occur, either in vivo or in vitro, for most hepatotoxic drugs [11]. Problematic reactive metabolites are typically electrophiles that can bind covalently to proteins and nucleic acids, thereby disrupting cellular functions [12] and potentially leading to pathological effects such as carcinogenesis or immunological hypersensitivity [2, 13].

Even drugs that are commonly used today may have toxic metabolites. One such drug with a known toxic metabolite is acetaminophen, for which proper dosing is important in order to prevent the toxic metabolite from building up in the body [11]. In addition to dose-dependent toxicity, idiosyncratic toxicity is a problem that is often presumed or known to be caused by toxic or reactive metabolites [7]. It remains unclear exactly what fraction of drug toxicity actually involves metabolism, but it is clear that metabolism does play a role [14]. A further source of complexity is that genetic variations in the expression and function of metabolic enzymes can also lead to increased toxicity or off-target effects of a drug, as well as to lower efficacy, depending on whether the enzyme activity is decreased or increased [15, 16]. Metabolism can also play a role in drug-drug, drug-food, and drug-herbal interactions. These interactions can cause slowed metabolism due to competitive inhibition, such as the interaction of terfenadine and ketoconazole, or can increase the rate of metabolism of a drug, such as is the case for St. John's wort and its effect on cytochrome P450 (CYP) 3A4 [17]. Further, the metabolites themselves may cause inhibition-based drug-drug interactions (DDIs), though this risk is considered low [18]. All told, the effects of metabolism should be taken into account during the drug development process in order to strive towards safety and efficacy.

1.1 Enzymes and Reactions Involved in Xenobiotic Metabolism

Traditionally, xenobiotic metabolism is separated into two phases: phase I and phase II [19]. Phase I metabolism encompasses oxidation, reduction, and hydrolysis reactions that typically form or unmask polar functional groups and thus often result in a metabolite that is more polar than its parent molecule. The more polar the molecule, the more readily excretable it tends to be. Phase II metabolism also

generally improves the hydrophilicity and excretability of the compound, but it does so through conjugation reactions.

Though the phase I-phase II nomenclature did arise from the general observed order of metabolic transformations, the nomenclature does not mean that any given molecule must undergo a phase I reaction before it can undergo a phase II reaction [19]. Many drugs undergo metabolism via multiple metabolism pathways [20]. Overall though, the ratio of phase II reactions to phase I reactions increases over the course of a sequence of metabolic reactions [21].

An analysis of 1171 xenobiotic compounds and their 6767 metabolites by Testa et al. [21] found that phase I metabolism accounts for approximately 70% of all metabolites and phase II metabolism for approximately 30%. Broken down further, the percentages corresponded to 57% of metabolites formed by oxidation-reduction (redox), 10% by hydrolysis, and 33% by conjugation reactions [21].

Both metabolism phases are relevant to the difficulties associated with metabolism discussed above. Reactions from both phases can lead to reactive and/or toxic metabolites [22]. For DDIs as well, phase I and phase II enzymes can be involved, though the CYP enzyme superfamily is generally considered the main perpetrator [17]. Further, genetic variations affecting drug metabolism have been found in many different enzymes across both phase I and phase II metabolism [23–26]. The following subsections discuss each phase of metabolism and the enzyme families involved in more detail.

A xenobiotic that enters the human body may undergo multiple metabolic reactions catalyzed by different metabolizing enzymes, and each metabolite may undergo subsequent further metabolic reactions. The result of this process is referred to as the metabolic tree, and the successive metabolic reactions result in successive generations of metabolites. Interestingly, it has been found that reactive metabolites are nearly just as likely to be formed at any of the first three generations of metabolites [21]. The percentages of metabolites that are formed by conjugation reactions increase after the first generation, while the percentage of redox-formed metabolites decreases [21]. On the other hand, the percentage of metabolites formed by hydrolysis reactions does not change much throughout the first few generations of metabolites [21].

1.1.1 Phase I Metabolism

The most well-known phase I enzyme family is the CYP enzyme family, which metabolizes around 75% of the small molecule drugs that undergo enzymatic metabolism [27] and accounts for approximately 40% of all metabolites [21]. This enzyme family is described in more detail in the subsection below. Although CYPs generally receive much attention, there are a number of other enzymes that also play important roles in phase I metabolism.

Oxidation and reduction reactions are two of the three types of reactions that make up phase I metabolism. Some families of metabolic enzymes, including CYPs [28], are able to catalyze both oxidation and reduction reactions, so oxidation and reduction are combined here for the purpose of describing the relevant enzyme families. In addition to CYPs, these types of reactions are catalyzed by oxidases, dehydrogenases, reductases, and peroxidases.

Flavin-containing monooxygenases (FMOs) catalyze oxidation reactions that oxygenate compounds containing a nucleophilic heteroatom such as nitrogen, sulfur, and phosphorus [29]. There are five human FMOs, with FMO3 typically being considered the most important FMO isozyme in the human liver for drug metabolism [30].

Monoamine oxidases (MAOs) are flavin-containing oxidoreductases that catalyze the oxidative deamination and dehydrogenation of amines [29]. In humans, two MAO enzymes have been identified: MAO-A and MAO-B [29].

Aldehyde oxidase (AOX) and xanthine oxidase (XO) are molybdoflavoenzymes that can catalyze both oxidation and reduction reactions [29]. AOX acts primarily on aldehydes, catalyzing their oxidation to carboxylic acids, whereas XO acts mainly on purines and pyrimidines [29]. AOX is not limited to substrates containing an aldehyde functionality, however, and can also oxidize a variety of other substrates, including aromatic heterocycles [31]. In terms of reduction, AOX and XO can catalyze the reduction of S- and N- functional groups [29]. AOX has only one human isoform, namely AOX1 [32].

Two main types of dehydrogenases involved in phase I metabolism are alcohol dehydrogenases (ADHs) and aldehyde dehydrogenases (ALDHs). ADHs catalyze the reversible oxidation of alcohols to aldehydes or ketones, while ALDHs oxidize aldehydes to carboxylic acids [29].

Reductases include aldo-keto reductases (AKRs) and quinone reductases (QRs) [29]. AKRs reduce aldehydes and ketones to alcohols, whereas QRs reduce quinones to hydroquinones [29].

A literature analysis by Rendic and Guengerich published in 2015 found that CYPs are responsible for 95% of metabolic redox reactions, with this percentage at 96% specifically for drugs [33]. The other 4-5% of redox reactions were carried out by other oxidoreductase enzymes, of which this study considered only FMO, AKR and MAO.

Hydrolysis reactions are the third type of phase I metabolic reaction. These reactions are catalyzed by hydrolases, such as epoxide hydrolases, peptidases, amidases, and esterases. Epoxide hydrolases are a category of hydrolases that catalyze the hydration of epoxides to 1,2-diols [29]. An example of a peptidase is cathepsin B, which has been found to catalyze the hydrolysis of prodrugs and antibody-drug conjugates [32]. The majority of the hydrolysis occurring in the human liver and small intestine has been found to be catalyzed by carboxylesterase (CES) enzymes, in particular human carboxylesterase 1 (hCE1) and human carboxylesterase 2 (hCE2) [34]. CESs catalyze the hydrolysis of esters, thioesters, amides, and carbamates [35]. CESs play an especially vital role in the metabolism of many cardiovascular drugs, in particular the ester prodrugs that are hydrolyzed by CESs to the active metabolite [35].

The above-mentioned analysis by Testa et al. found that hydrolases account for 7% of all metabolites [21]. In addition, a study of the 22 prodrugs approved by the United States (U.S.) Food and Drug Administration (FDA) between 2006 and 2015 found that in 86.4% of cases, it was hydrolases that were responsible for the bioactivation of the prodrug [36].

CYPs

There has been much focus on CYPs in the context of drug development due to their prominence as drug-metabolizing enzymes, including for their well-known involvement in DDIs [37] and in the formation of reactive metabolites [21], which are of particular concern for ADRs. As mentioned at the beginning of this section, published analyses of the scientific literature on drugs and other xenobiotics have also found that CYPs are the most prolific enzyme family in terms of the percentage

of drugs metabolized and the number of metabolites formed. The above-mentioned 2012 study by Testa et al. considered the metabolites of over 1000 compounds, including drugs and other xenobiotics, and found that 40% of the metabolites were formed by CYPs [21]. A 2016 analysis by M. A. Cerny examined the metabolism of 125 of the 221 small-molecule drugs approved by the FDA between 2006 and 2015, excluding prodrugs, imaging agents, drugs administered neither orally nor intravenously, and drugs without available ADME data [36]. Of these drugs, 44.8% were found to be metabolized only by CYPs, with a further 16.8% metabolized by a combination of CYPs and non-CYPs. Only 20.8% of these drugs were found to be metabolized only by non-CYPs, as 17.6% had no major metabolites. This means that approximately 75% of the drugs that are cleared by metabolism are metabolized by CYPs. These results are in line with the earlier analysis by Williams et al. of the top 200 most prescribed drugs in the U.S. in 2002, which had found that approximately two-thirds of the drugs cleared by metabolism were metabolized by CYPs [20].

The 2012 Testa et al. study also analyzed which reaction types resulted in toxic and chemically reactive metabolites [21]. In total, 7% of the metabolites were found to be toxic and/or reactive. Of these, 41% are the result of oxidation to quinones, 15% are formed by sp^2- and *sp*-carbon oxidation, 9.5% by oxidation of sulfur atoms, 8.2% by oxidation of NH or NOH groups, and 6.3% by sp^3-carbon oxidation. Each of these reaction types can be catalyzed by CYPs [21]. In total, this amounts to 80% of toxic and/or reactive metabolites that are potentially formed by CYPs.

In addition to these potentially problematic reaction types, CYPs mediate a wide variety of reactions. Common CYP reaction types include hydroxylation, *N*-dealkylation, and heteroatom oxygenation [28]. More unusual reaction types consist of reductions, oxidative ester cleavage, and ring formation and expansions [28]. There is little, if any, variability in the ability of the different CYP isozymes to catalyze the various reaction types, because all CYPs share the same chemical mechanism [38]. Instead, the main variability in the different CYP isozymes is in the substrates they can bind.

Of the 57 known human CYP enzymes, around half are associated with xenobiotic metabolism [39] and only a handful are considered especially important for xenobiotic metabolism [27]. The CYP isozymes relevant for xenobiotic metabolism belong mainly to the CYP1, CYP2, and CYP3 subfamilies [40]. The others are

primarily involved in endogenous metabolism, though drug-metabolizing isozymes can also play a role in endogenous metabolism [39].

The CYP isozymes share a common overall fold and mechanism of action, with the active site containing a heme group that plays a direct role in the catalytic cycle [41]. At the same time, the binding pockets of the CYP isozymes vary widely in terms of their volume, shape, flexibility, and the substrates they can bind [41, 42]. The sizes of the active sites of the CYP isozymes involved in xenobiotic metabolism range from a volume of 260 $Å^2$ for CYP2B6 to 1438 $Å^2$ for CYP3A4 [40]. This variety in the active sites allows CYPs to bind and mediate reactions of highly diverse substrates.

The most relevant CYP isozymes for xenobiotic metabolism are thought to be 3A4, 3A5, 2D6, 2C8, 2C9, 2C19, 1A1, 2B6, and 2E1 [6], with 1A2, 2C9, 2C19, 2D6, and 3A4 considered to be of the greatest importance to human drug metabolism [17]. Most of the main CYP isozymes involved in xenobiotic metabolism are affected by genetic polymorphism and have known inducers and inhibitors [43]. Hence it is important to be aware of the roles of the different CYP isozymes during drug development in order to prevent unanticipated effects caused by polymorphisms or DDIs combined with too heavy a reliance on one CYP isozyme for clearance. The above-mentioned literature analysis by Rendic and Guengerich published in 2015 examined the contributions of different CYP isozymes to the metabolism of all chemicals, including a specific analysis for drugs [33]. There were 860 drugs considered in the analysis, including both in-development and marketed drugs, and approximately 4000 metabolic reactions. Specifically for drugs, Rendic and Guengerich found that CYP3A4 contributed the most to CYP-mediated metabolism, at 27%, followed by 2D6 at 13% and 2C9 at 10%. The next highest percentages belonged to 2C19 and 1A2, each with 9% [33].

1.1.2 Phase II Metabolism

Nearly 90% of all phase II metabolites are formed by five main enzyme families: UDP-glucuronosyltransferases (UGTs), glutathione S-transferases (GSTs), sulfotransferases (SULTs), methyltransferases (MTs), and N-acetyltransferases (NATs) [21]. All of these transferases carry out conjugation reactions: glucuronida-

tion, glutathione (GSH) conjugation, sulfonation, methylation, and acetylation, respectively. Other phase II reactions include amino acid conjugation, such as glycination, and phosphorylation [21]. Even more unusual phase II reactions have been attributed to enzymes such as ADP-ribosyltransferase, which is typically involved in other cellular processes [32].

Glucuronidation by the UGT enzyme family is generally considered the most important phase II metabolic pathway [44]. UGT-catalyzed glucuronidation reactions occur at a nucleophilic atom such as O, N or S, forming a β-D-glucuronide [45]. In humans, there are at least 22 different UGTs from four different families [45]. Isozymes 1A1, 1A3, 1A4, 1A6, 1A9, 2B7, and 2B15 are considered the most important UGTs for metabolizing drugs in the human liver, while several other UGTs (1A7, 1A8, and 1A10) are found primarily in the gastrointestinal tract [45]. A review by Williams et al. of the top 200 drugs prescribed in the United States in 2002 found that 14 of those drugs have glucuronidation listed as a clearance mechanism [20]. This same study by Williams et al. found that of the drugs metabolized by UGTs, UGT2B7 was the most commonly listed isozyme, accounting for 35% of glucuronidation of drugs, followed by UGT1A4 at 20% and UGT1A1 at 15% [20].

In addition to catalyzing GSH conjugation reactions, GSTs can catalyze other non-conjugation reactions ranging from reduction reactions to isomerization reactions [46]. In the context of phase II metabolism, only the GSH conjugation reactions are typically considered. There are two known GST superfamilies, one of which is involved in xenobiotic metabolism and the other of which is primarily involved in arachidonic acid metabolism [45]. Over 20 GSTs have been identified in humans [47]. Though GSH conjugation is generally a detoxification reaction, there are a number of cases in which the GSH conjugates are either themselves reactive or undergo further modifications that lead to reactive molecules [48].

SULTs catalyze sulfonate conjugation at oxygen, nitrogen and sulfur acceptor groups [45]. Four SULT families have been found in humans [45]. As is the case for other phase II enzyme families, SULTS can end up forming reactive metabolites despite generally being considered detoxifying metabolic enzymes [49].

NATs catalyze the N-acetylation of arylamines and arylhydrazines as well as O-acetylation of *N*-hydroxyarylamines [29]. The O-acetylation of *N*-hydroxyarylamines to acetoxy esters is a bioactivation reaction that creates reactive metabolites [29]. In humans, there are two NAT isozymes: NAT1 and NAT2 [45].

The two main MTs are thiopurine S-methyl transferase (TPMT) and catechol O-methyl transferase (COMT) [45]. TPMT is known to S-methylate aromatic heterocyclic sulfhydryl compounds, while COMT O-methylates catecholamines [45]. Other MTs include nicotinamide N-methyltransferase (NNMT) and thiol methyltransferase (TMT) [29]. In most cases, methylation reactions result in more hydrophobic metabolites rather than increasing hydrophilicity; however, there are exceptions such as N-methylation of pyridine [29].

Of these five enzyme families, UGTs play the largest role in xenobiotic metabolism. The 2012 study by Testa et al. found that UGTs account for 14% of all metabolites, followed by GSTs and SULTs at 5.5% and 4.7%, respectively [21]. NATs and MTs lead to only 1.5% and 2.1% of all metabolites, respectively, according to this same study [21].

1.2 Predicting Xenobiotic Metabolism

Being able to make predictions about how a particular compound could be metabolized offers several key advantages. Knowledge of the potential metabolic pathways can be used to design modifications to make a drug more metabolically stable and decrease the chance of its having reactive metabolites, or to attempt to avoid reliance on highly polymorphic metabolizing enzymes. Studying the metabolism of a compound of interest either in vivo or in vitro is, however, a laborious process that can be aided by computational techniques.

Metabolism of xenobiotics in humans is a highly complex system involving many metabolizing enzymes, transporters, and various tissues in the body. There are several main aspects that have been the focus of computational methods. In particular, computational methods can be used to predict the metabolically labile atom positions, or sites of metabolism (SoMs), which is also referred to as predicting the regioselectivity of xenobiotic-metabolizing enzymes. The prediction of the chemical structures of metabolites, i.e. the products of metabolic reactions, is another promising area for computational methods to address. Further, quantitative structure-activity relationship (QSAR) models have been developed to predict the interaction of compounds of interest with particular metabolic enzymes, such as

CYP isozymes, UGTs, and SULTS [42]. For CYPs in particular, predicting isozyme specificity is of relevance in order to attempt to avoid the reliance on polymorphic CYPs for the clearance of a drug or to aim at predicting and preventing DDIs [50]. As such, the prediction of CYP inhibition and regioselectivity is a popular goal [51]. With the aim of preventing DDIs, a number of approaches have been developed to predict the inhibition and induction of key CYP isozymes [52].

There are two general categories of computational techniques for predicting xenobiotic metabolism: structure-based and ligand-based. Structure-based approaches take protein structural information into account, in this case the structural information about the specific metabolizing enzyme(s) of interest, including for example the interactions between the ligand and the active site. Ligand-based approaches, on the other hand, do not consider the protein structure and instead use compounds whose metabolism is known in order to make predictions about the metabolism of other molecules. This dissertation focuses on the ligand-based prediction of SoMs and of the chemical structures of metabolites.

Computational methods to predict metabolism are usually based on experimental data in one way or another, at least to test the methods. The source of much of the experimental data upon which ligand-based methods for SoM and metabolite prediction are based is from in vitro or in vivo studies in humans and other mammals. The in vitro studies typically use liver microsomes, liver S9 fractions, or hepatocytes. The identification of the metabolites is typically done using mass spectrometry (MS), though nuclear magnetic resonance (NMR) may be used to elucidate the exact structures of metabolites if the amount of material is sufficient [42]. Metabolite identification (MetID) is not trivial [53], and in many cases the exact location of a functional group, for example a hydroxyl group added via a metabolic reaction, cannot be easily determined. This is another reason that computational prediction of metabolites can be useful: the predictions can be compared to the MS data to get a better idea of the exact structures of the experimentally found metabolites.

1.2.1 Regioselectivity Prediction

In general, successfully predicting SoMs is dependent upon taking into account two distinct but synergistic aspects of the atoms in the molecule(s) in question [54].

The first is reactivity, which represents the intrinsic ability of each atom to be involved in a reaction. The second aspect is accessibility, which refers to whether an atom is located at a position in the molecule that allows the atom to interact with the metabolizing enzyme, either from a steric perspective or in terms of the orientation of the molecule in the active site. In the case of CYPs, for example, SoMs must be accessible to the heme. Many successful SoM prediction methods take both reactivity and accessibility into account, at least implicitly [54].

There have been examples of structure-based methods to predict regioselectivity, such as docking a molecule of interest into the binding pocket of a metabolic enzyme, but such approaches are typically computationally expensive and therefore most useful for examining individual substrate-enzyme interactions [54]. In terms of ligand-based methods, the majority are based on techniques such quantum mechanics, machine learning, and data mining [54].

Machine learning approaches in particular are relatively popular and quite successful at predicting SoMs. Prominent machine learning-based SoM predictors include XenoSite [55–57], SOMP [58], MetScore [59, 60], and the FAME tools [A4, A8, 61]. For example, XenoSite, the successor of RS-Predictor [62], uses a variety of descriptors with neural networks to predict SoMs for CYP-mediated metabolic reactions [55]. The descriptors explored in XenoSite include quantum chemical, reactivity, fingerprint similarity, and molecule-level descriptors [55]. XenoSite has been extended to cover other enzyme families, such as UGTs [57], as well as to include models focused on specific reaction types, for example to predict sites of quinone formation [63]. The latest version is Rainbow XenoSite, which predicts reaction types along with the SoMs for five main phase I reaction classes: stable oxidation, unstable oxidation, dehydrogenation, hydrolysis, and reduction [56].

Our research group has developed the machine learning-based methods FAME 2 [A4] and FAME 3 [A8], which both use extra trees classifiers and two-dimensional (2D) circular descriptors. FAME 2 predicts SoMs for CYP-mediated metabolic reactions, and FAME 3 predicts SoMs for phase I and phase II metabolic reactions, including separate models for phase I and phase II prediction. Both tools achieved high performance on their test sets, both in terms of overall measures of SoMs and non-SoMs being correctly predicted as well as measures of the ranks of the predicted probabilities of atoms being SoMs.

In contrast to the above-mentioned machine learning methods, SMARTCyp [64–

67] is a fragment-based method that uses a combination of a reactivity descriptor and a 2D accessibility descriptor for each atom to predict CYP SoMs. The reactivity descriptor represents the activation energy required for a reaction to occur at the particular atom position, which is calculated by matching the atom to a lookup table of energies computed using density functional theory (DFT). The accessibility descriptor represents the distance of the atom from the 2D center of the molecule. SMARTCyp is primarily based on reactivity, with the accessibility component allowing slightly less reactive atoms to be given greater consideration if they are potentially highly accessible to the CYP active site. In a later version of SMARTCyp, a solvent-accessible surface area descriptor was added [65]. The latest version, SMARTCyp 3.0, offers a similarity feature that scores the matched substructure according to how well it matches the molecule fragment that was the basis for the DFT calculation [67].

The focus of many SoM prediction methods is on CYP-mediated metabolism or on general phase I or phase II metabolism. However, some methods have also been developed that focus on predicting the regioselectivity of certain specific non-CYP metabolic enzymes [68]. For example, XenoSite UGT predicts SoMs for UGT-mediated metabolism [57], MetaSite offers AOX [69] and FMO3 [70] regioselectivity prediction along with CYP regioselectivity prediction [71], and SOMP predicts SoMs for UGT- as well as CYP-mediated metabolism [58].

Most of the approaches discussed so far are based on machine learning or other descriptor-based methods. A different approach to SoM prediction was proposed by Sykes et al. [72]. In this study, the molecule flurbiprofen, in its CYP2C9-bound conformation, was used as a template for aligning 69 other CYP2C9 substrates. The authors found that 60% of the molecules were aligned such that their SoMs were within 3 Å of flurbiprofen's SoM. This was an interesting study because it attempted to use molecular alignment as a spatial method for predicting CYP2C9 regioselectivity. However, this study was limited in scope as it only considered one CYP isoform and used a relatively lenient cutoff of 3 Å to define a correct SoM prediction. For small molecules, a radius of 3 Å around a SoM could encompass the majority of the molecule.

Prior to work on this dissertation, the primary public dataset of SoM data was the dataset published by Zaretzki et al. in 2013 in their paper on XenoSite [55]. This dataset includes 680 molecules annotated with their SoMs for CYP-mediated

metabolic reactions. During the work on this dissertation, the MetaQSAR database was published [73]. This database contains over 1800 molecules with annotated SoMs for phase I and phase II metabolic reactions. The MetaQSAR data was used to develop the FAME 3 SoM prediction models. In the same year that the MetaQSAR was published, 2018, a test dataset of 31 compounds annotated with phase I and phase II SoMs was published and used to validate the MetScore method [60].

Many of the above-mentioned approaches to SoM prediction were trained on the Metabolite database [74], a commercial database of metabolic data that is considered to be quite comprehensive and has been commonly used to develop methods for SoM and metabolite prediction but is no longer available to the public. Notable exceptions are FAME 2 and FAME 3.

1.2.2 Metabolite Structure Prediction

Most methods for metabolite structure prediction to date use a rule-based approach. A rule-based approach means that metabolic reactions are encoded in a set of reaction rules, and the rules are used to generate the possible products of metabolic reactions. One problem with this approach is that the reaction rule sets typically generate massive numbers of possible metabolites, which is cumbersome from a practical standpoint [75]. Therefore, it is important for rule-based methods to score or rank their predictions, and many of the available tools for metabolite structure prediction do so. Ranking the predictions has been attempted in a few ways, such as using occurrence ratios based on a large dataset of metabolic reactions (e.g. SyGMa [76]), using neural networks to assign probabilities to each of the generated metabolites (e.g. by Wang et al. [77]), or incorporating the ranking of predicted SoMs (e.g. MetaTox [78, 79]).

Commercial methods for metabolite structure prediction include Meteor Nexus (Lhasa Ltd.) [80], TIMES (LMC) [81], MetabolExpert (CompuDrug Ltd.) [82], ADMET Predictor (SimulationsPlus) [83], StarDrop (Optibrium) [84], and MetaSite (Molecular Discovery) [71]. Further, a number of non-commercial methods have been published, including SyGMa [76], BioTransformer [85], MetaTox [78, 79], and RD-Metabolizer [86]. These non-commercial methods are discussed in more detail in the remainder of this section.

SyGMa [76] has a total of 145 reaction rules, subdivided into 118 phase I reaction rules and 27 phase II rules. The scoring approach used by SyGMa is based on empirical probability scores that were calculated based on occurrence ratios in the Metabolite database [74]. Unfortunately, this database has been discontinued. Considering both phase I and phase II metabolism, by using the entire reaction rule set, SyGMa successfully predicted 68% of the metabolites in a test set of 175 parent molecules. In terms of ranking ability, 45% of the metabolites in the test set were ranked among the top ten predicted metabolites for each corresponding parent molecule. Though SyGMa does not have a CYP-specific model, the performance on a set of 127 single-step CYP-mediated reactions was measured using all of the phase I reaction rules, resulting in correct prediction of 84% of the metabolites and the ranking of 66% of the known metabolites within the top three ranked predictions. SyGMa, including the reaction rules, is freely available as an open-source Python package.

BioTransformer [85] provides five metabolite prediction modules: human CYP, human phase II, gut microbial, environmental microbial, and human "Enzyme Commission (EC)"-based metabolism. BioTransformer has 163 CYP rules and 74 phase II rules. Note that non-CYP phase I metabolic reactions are not covered by either of these two modules for human xenobiotic metabolism. Though BioTransformer does not rank its predictions, it attempts to reduce the number of excessive predictions by using a separate set of constraints to define the types of molecules that the different reaction rules are allowed to act on. Further, for CYP-specific metabolite prediction, BioTransformer uses a CYP selectivity prediction model prior to application of the reaction rules. With this approach, BioTransformer was able to predict 80% of the metabolites in its test set of 40 pharmaceuticals and pesticides. Considering CYP-mediated metabolism on its own, BioTransformer was able to successfully predict 90% of the metabolites of a CYP test dataset of 60 parent molecules. BioTransformer, including the reaction rules, additional constraints, and the MetXBioDB database of metabolic reactions, is freely available as an open-source Java package.

MetaTox [78, 79] combines prediction of the reacting atoms [87, 88] with predictions of the reaction class to predict both phase I and phase II metabolites. The published evaluation of MetaTox includes a separate validation of the reaction class and the reacting atom predictions, including a rank-based analysis of each

prediction component. It is therefore unclear exactly how well MetaTox is able to predict the actual structures of the metabolites, both in terms of how many experimental metabolites are predicted and in terms of whether the generated metabolites are ranked in a meaningful way, to distinguish the more likely from the less likely.

RD-Metabolizer [86] predicts metabolites for phase I and phase II metabolic reactions using a set of reaction rules. In addition, RD-Metabolizer predicts SoMs using occurrence ratios, which were based at least in part on the currently unavailable Metabolite database. It appears that RD-Metabolizer only applies the reaction rules to the predicted reaction centers and/or ranks its metabolite predictions based on the predicted SoMs, though the exact methodology remains unclear. This method was able to predict 36% of known metabolites within the top three ranked predictions for a test set of 425 compounds. Several example reaction rules are provided in the publication, but the full set of reaction rules and RD-Metabolizer itself do not seem to be easily accessible.

In addition, two new publicly available methods for SoM prediction have been published since the publication of the relevant portions of this dissertation's research. These new tools are discussed in section 5.2, as they can be directly compared to the methods developed as part of this dissertation.

Overall, many available metabolite structure predictors are knowledge-based approaches derived from the Metabolite database [74], which has the disadvantage of not being available for further analysis or reproducibility by other researchers. A notable exception to this trend is BioTransformer, which was based on a newly collated, freely available database called MetXBioDB [85].

Aims

Xenobiotic metabolism is highly relevant to the research and development of chemical entities that come into contact with humans. From drugs to cosmetics to agrochemicals, the metabolism of these compounds plays a role in their safety and, for compounds intended for direct use by humans, in their efficacy. Computational approaches can be used to predict how these compounds will be metabolized, in particular by predicting the SoMs and the metabolites of these molecules. This dissertation focuses on SoM prediction and metabolite structure prediction and aims to answer the following questions.

Firstly, to what extent can an alignment-based method be successfully used to predict SoMs? To study this problem, SoM data for multiple CYP isozymes were used to construct an alignment-based approach that measured the proximity of the SoMs in the aligned molecules. A further component of this study was the combination of the alignment approach with a reactivity model, in order to take into account the other key element of SoM prediction and thereby achieve an improvement in performance.

Second, how can SoM prediction be combined with a reaction rule set to produce an effective metabolite structure prediction approach that avoids the problem of excessive numbers of putative false positive predictions? Most existing methods for metabolite structure prediction suffer from this problem, which makes it difficult to use them to guide, for example, drug development decisions. We therefore desired to either reduce the number of false positive predictions or to rank the predictions in a meaningful way, in order to assign the putative false positive predictions less importance in the list of predictions.

Third, can comparable or better performance compared to existing freely available metabolite structure prediction methods be achieved if we only use publicly available data and attempt to reduce bias caused by dependence on a particular dataset for the development of the underlying method? Reproducibility is a key objective in scientific research, and so is being able to build on former research and

previously developed methods. We aspired to develop an approach that is not only reproducible but can be expanded or improved in the future.

To answer the second and third questions, we initially focused on the extensively studied CYP family of metabolic enzymes. The comparative abundance of metabolism data for this enzyme family, as well as its generally acknowledged importance in drug development, made CYP metabolism a desirable starting point. After developing a method for metabolite structure prediction focused on CYP-mediated metabolism, we extended the scope of our work to include all of phase I and phase II metabolism. The overall aim of this portion of the dissertation was to develop computational tools for metabolite structure prediction that can predict metabolites formed by CYP-mediated metabolism as well as by phase I and phase II metabolism in general, and to make these tools available to the scientific community.

Methods

3.1 Alignment of Small Molecules

The alignment of small molecules to each other, which is often referred to as a shape-based approach, is a well-studied technique for ligand-based virtual screening [89–91]. The idea behind this approach is that molecules with similar shape, and similar chemical makeup in relation to their shape, are likely to have similar function in terms of whether they bind to a particular target. So, if we have one or more molecules with known values for a particular property of interest, such as activity against a particular target, and one or more molecules for which we would like to predict this property, we can use molecular alignment to make these predictions.

In this dissertation, the property of interest for the alignment-based method was the SoMs in the context of CYP-mediated metabolism. We surmised that if two molecules are similar in shape, i.e. align well to each other, then the molecules could be assumed to bind with similar orientations to the same CYP isozyme(s). This supposition would suggest that the two molecules could be assumed to have SoMs located close to each other spatially.

So, if we have a database of molecules with known SoMs, and we have a query molecule for which we would like to predict the SoMs, then we can align the query molecule to all the reference molecules in the database. Based on which reference molecule aligns best to the query molecule, we can use the known SoMs of the reference molecule to predict that the SoMs of the query molecule should be nearby. Note that in the context of virtual screening, the terminology of query and reference is often reversed. The reason for this reversal is that in virtual screening, the query is often a known ligand and the aim is to find actives in the database of molecules. For SoM prediction, however, we use the database of molecules as the reference and make predictions for the query molecule instead.

In this work, the method used for the alignment of small molecules is ROCS [92].

The basis of the alignment with ROCS is the molecular shape and a quality called color that represents the chemical properties of functional groups. The molecular shape is represented by atom-centered Gaussian functions [92]. To align two molecules, the shared volume between a query molecule and a reference molecule is maximized. Meanwhile, a color force field is used in order to attempt to achieve a maximum overlap of functional groups with similar properties, such as hydrogen bond donors and acceptors, hydrophobic groups, anions and cations, and rings [92].

In order to properly compare two molecules using an alignment-based method, it is necessary to generate conformers for one of the molecules. Because molecules can be very flexible, a meaningful alignment is only possible if different conformations are considered. It has previously been found, in the context of virtual screening, that it is not necessary for the reference molecule to be in its bioactive conformation and that it is not necessary to generate multiple conformers for both the query and reference molecules [93]. In this dissertation, the conformer generation was performed with OMEGA [94], which has been found to be a leading conformer ensemble generator in terms of replicating protein-bound conformations of molecules [A2].

3.2 Incorporating SoM Prediction into Metabolite Structure Prediction

In general, metabolite structure prediction requires a knowledge-based approach in which potential metabolic reactions are encoded in a set of reaction rules. Because of this approach, most available metabolite structure predictors generate huge, unwieldy numbers of metabolites [75].

With the aim of addressing this common problem of a plethora of predicted metabolites, we considered two distinct conceptual approaches for incorporating SoM prediction into the metabolite prediction workflow. First, we considered whether predicted SoMs could be used as a hard filter to determine where in the molecule to apply the reaction rules (Figure 3.1A). For example, if there is only one predicted SoM in the molecule, then the reaction rules are only applied at that one position. Reaction rules may involve more than one atom, and in that case the predicted SoM must match one of the atoms in the rule. Since there are far fewer

SoMs than heavy atoms in a molecule, this approach is designed to drastically reduce the total number of metabolites predicted.

Second, we considered how the predicted SoM probabilities could be used to score the predicted metabolites (Figure 3.1B). In this case, the reaction rules were applied everywhere in the molecule. Note that not every rule can be applied at every position in a molecule. For example, if a reaction rule requires a sulfur atom to be involved in the reaction, then the rule cannot be applied to a molecule that does not contain a sulfur atom. Overall, however, the set of rules was applied to all positions in the molecule, wherever each rule matched. This approach did not reduce the total number of predicted metabolites. Instead, we devised a scoring approach that took the maximum SoM probability of the heavy atoms involved in the reaction into account. The scoring approach is explained in detail in sections 4.2.1 and 4.2.2.

3.3 Encoding Metabolic Reactions

Metabolic reactions can be encoded using the SMIRKS reaction transform language [95]. The SMIRKS language combines components of SMILES (Simplified Molecular Input Line Entry System) notation and the SMARTS molecular pattern language in order to describe the changes to atoms and bonds. These changes are described by representing the reactant and product(s) with a pairwise mapping of the atoms on the reactant side to the atoms on the product side of the transformation. Not all atoms must be mapped, however, so some atoms may appear only on the reactant side or only on the product side of the transform (Figure 3.2).

SMIRKS describe generic transforms, which do not have to be proper chemical reactions, per se. For example, SMIRKS can be used to convert between different tautomeric forms of the same molecule [96]. SMIRKS can also be used to standardize functional groups, for which they have also been used in this work.

To describe a transformation, the SMIRKS need only encode the affected atoms in the reactant and product(s). These atoms are described by SMARTS patterns, and the SMIRKS transformation can be applied to any molecule that matches the SMARTS pattern on the reactant side of the transformation. The SMIRKS

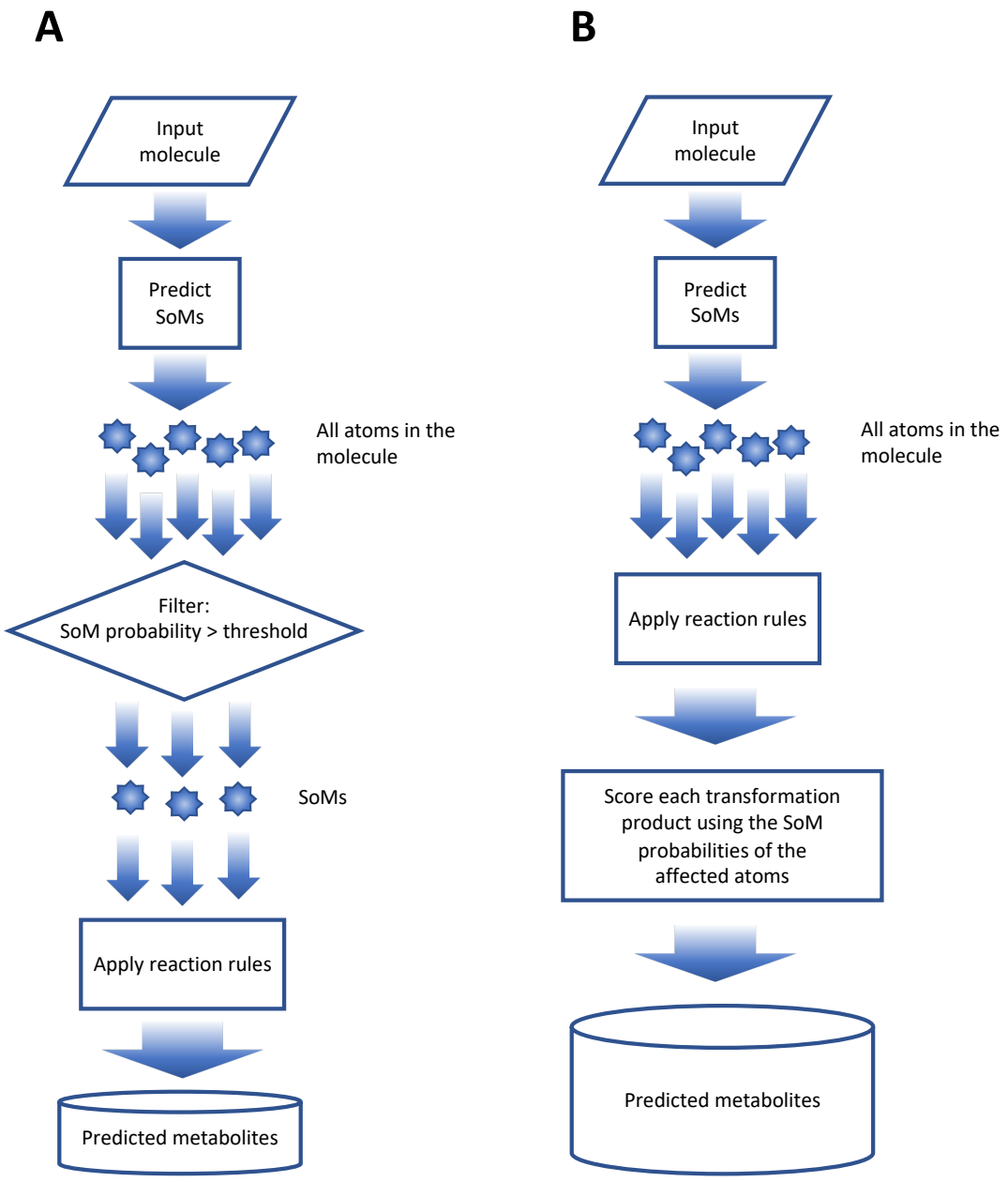


Figure 3.1: Workflows illustrating the two concepts for incorporating SoM prediction into metabolite prediction that were explored in this dissertation: A) using predicted SoMs as a hard filter, and B) using the predicted SoM probabilities for scoring the predicted metabolites.

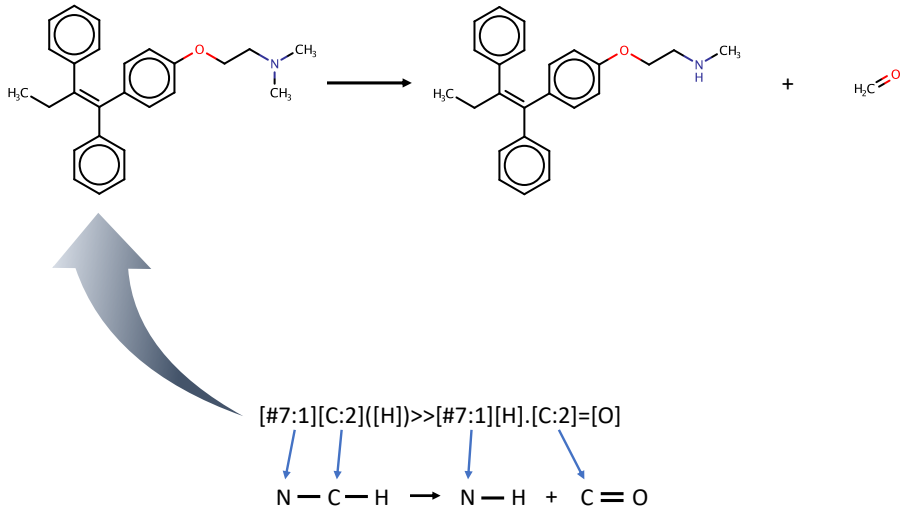


Figure 3.2: Illustration of how a SMIRKS string can be interpreted and applied, using the example of the N-dealkylation SMIRKS ("`[#7:1][C:2]([H])>>[#7:1][H].[C:2]=[O]`") applied to tamoxifen. The blue arrows indicate the mapped atoms, which are present in both the reactant and the product of the transformation.

language does not allow full SMARTS functionality, requiring for example that bonds be described as SMILES (i.e. not allowing bond queries).

There is a variant of SMIRKS called Reaction SMARTS, which allows more SMARTS functionality within the transform language. Reaction SMARTS are implemented in RDKit, an open-source cheminformatics Python library [97]. The term SMIRKS is sometimes used to mean Reaction SMARTS, as in the case of the metabolite structure predictor SyGMa [76], which is an open-source software library for Python that is based on RDKit. SMIRKS and Reaction SMARTS are not interchangeable from a software development perspective, however, as the language that can be used to encode reactions depends on the software library that is used to interpret the rules and carry out the transformations.

The software library used in this work to perform the transformations is Ambit SMIRKS [98], which is based on the Chemistry Development Kit (CDK) [99], a collection of open-source java libraries for cheminformatics. Ambit SMIRKS

is focused specifically on SMIRKS, as the name suggests, and cannot interpret Reaction SMARTS correctly. The SMIRKS language was therefore used in this work. In the third part of this work (section 4.2.2), in which the set of reaction rules from SyGMa were implemented in our developed program, it was necessary to translate the Reaction SMARTS to SMIRKS in order for the rules to be interpreted and applied correctly.

3.3.1 Considerations for Molecular Perception

The CDK software libraries are used to represent and process molecules in various ways. Prior to transformation with Ambit SMIRKS, the molecules must be protonated and the hydrogens must be explicitly represented in the virtual molecule. In addition, the aromaticity or lack thereof must be detected for each atom in the molecule. Aromaticity detection is necessary because SMIRKS may specify that a particular transformation may only occur at, for example, an aromatic carbon but not at an aliphatic carbon.

CDK offers four electron donation models for determining aromaticity [100]. Ambit SMIRKS includes an implementation of one of those models, the `cdk` electron donation model. It was determined during the course of this work, however, that this aromaticity model deviates substantially from the aromaticity interpretation by human beings with chemical knowledge. In addition, the `cdk` model's aromaticity detection differs substantially from RDKit's aromaticity perception, which was of vital relevance when implementing the set of reaction rules from SyGMa. Using the same set of reaction rules should result in the same products of the transformations, but instead there were initially large discrepancies. These discrepancies occurred in the case of rings or ring systems in which ring-adjacent atoms were involved in the aromatic system. In some cases, the discrepancies seemed to be due to the molecule being in keto form, when the eno form would be clearly recognized as aromatic. For example, Ambit SMIRKS combined with the `cdk` aromaticity model did not recognize 4-pyridone as aromatic, whereas this molecule's enol form (pyridin-4-ol) would be recognized as aromatic. Converting between tautomers is a complex problem [96], an in this case would be an unrewarding approach as it would not solve the aromaticity perception problem for some molecules, such

as 4-hydroxy-6-methylcoumarin (Figure 3.3). For a given aromatic methylation reaction rule to be applied to this molecule, the carbon must be recognized as aromatic but the oxygen must be protonated. In order for this and other rules to be applied correctly to aromatic molecules, it was necessary to consider other options for aromaticity detection.

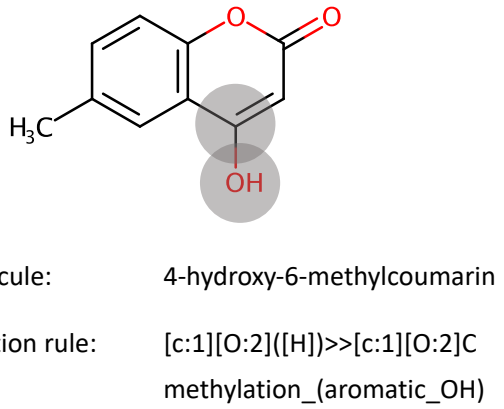


Figure 3.3: Example of a molecule containing atoms that should be recognized as aromatic but may or may not be, depending on the aromaticity model used by CDK. The atoms highlighted in gray are the atoms that should be matched by the shown reaction rule. This reaction rule is a phase II methylation rule sourced from SyGMa.

CDK offers several other aromaticity models. We found that the `daylight` electron donation model [100] could very closely approximate, according to the transformation products of the set of reaction rules, RDKit's model [97] as well as a reasonable human interpretation of chemistry. Hence the `daylight` electron donation model was used in this work (section 4.2.2) in order to ensure that the reaction rules sourced from SyGMa and translated manually into SMIRKS would predict nearly the same metabolites as SyGMa. One class of molecule was found (tetrazoles) for which GLORYx was able to predict products of aromatic reaction rules but SyGMa was not. Otherwise the `daylight` electron donation model enabled the exact same metabolites to be predicted as by SyGMa, as tested on a reference dataset containing 1420 parent molecules compiled from the DrugBank [101] and

MetXBioDB [85] databases.

3.4 Evaluating Performance

In order to measure performance, it is necessary to decide which definitions of success are relevant to the problem. In the context of this dissertation, there are two general types of performance metrics that are of interest. One is the overall performance, in terms of how many predictions are correct and how many are incorrect. The other is how early in the ranked list of predictions the correct predictions can be found. This second concept is often addressed by measuring the overall ranking performance or the early enrichment.

There are slightly different considerations depending on whether SoMs or metabolites are being predicted. SoM prediction is in principle a binary classification problem, in which atoms are classified either as SoMs or not SoMs. This means that each prediction falls into one of four categories: true positive (TP), true negative (TN), false positive (FP), or false negative (FN). In this work, however, the alignment-based approach to SoM prediction was not a classification method. Therefore, evaluation metrics were chosen that measured whether the top-ranked positive predictions are correct, i.e. whether the atoms predicted as most likely to be SoMs are actually SoMs.

Metabolite structure prediction is not a classification problem. Structures of potential metabolites are generated, but there are no negative predictions unless one considers every molecule that could possibly exist but was not predicted. There are therefore no TN predictions. There are, however, FN predictions: known metabolites that are not predicted.

Recall and precision are two metrics that provide an idea of the ratio of correct predictions to incorrect predictions. Because TN predictions are not considered in these metrics, precision and recall can be used to evaluate metabolite structure prediction methods. In the context of metabolite prediction, recall is the percentage of known metabolites that are predicted (Equation 3.1), also known as sensitivity. Precision is the percentage of predicted metabolites that correspond to known

metabolites (Equation 3.2).

$$Recall = \frac{TP}{TP + FN} \tag{3.1}$$

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

In order to measure how well the predictions are scored and ranked, for both SoM prediction and metabolite prediction, the receiver operating characteristic (ROC) curve can be used. The ROC curve is a visual metric whereby the true positive rate (sensitivity) is plotted against the false positive rate. Because the ROC curve describes the order of predictions within a given set of predictions, this metric can be used on non-classification problems such as metabolite structure prediction. ROC curves can be interpreted as follows (Figure 3.4). A ROC curve that is a diagonal line from the bottom left to the top right of the plot represents random performance (dashed line in Figure 3.4). Perfect ordering of the predictions is described by a ROC curve that follows a straight line from the lower left-hand corner to the upper left-hand corner, followed by a straight line from the upper left-hand corner to the upper right-hand corner (blue line in Figure 3.4). Conversely, an ordering that is the exact opposite of the correct one is represented by a ROC curve that progresses in a straight line from the lower left-hand corner to the lower right-hand corner, and therefrom to the upper right-hand corner (orange line in Figure 3.4).

A numerical metric that is based on a ROC curve is the area under the ROC curve (AUC). As the name indicates, the AUC is determined by calculating the area under the curve in a ROC plot. An AUC of 1 indicates perfect ordering of predictions, an AUC of 0.5 indicates random performance, and an AUC of 0 indicates that the performance was the exact opposite of the desired performance. The AUC is a useful numerical metric to represent the overall ranking performance. However, the same AUC value can represent differently shaped ROC curves, so an examination of the ROC curves themselves provides important additional information about early enrichment.

In order to specifically analyze early enrichment, the top-k and Boltzmann-Enhanced Discrimination of ROC (BEDROC) [102] metrics can be used. The top-k

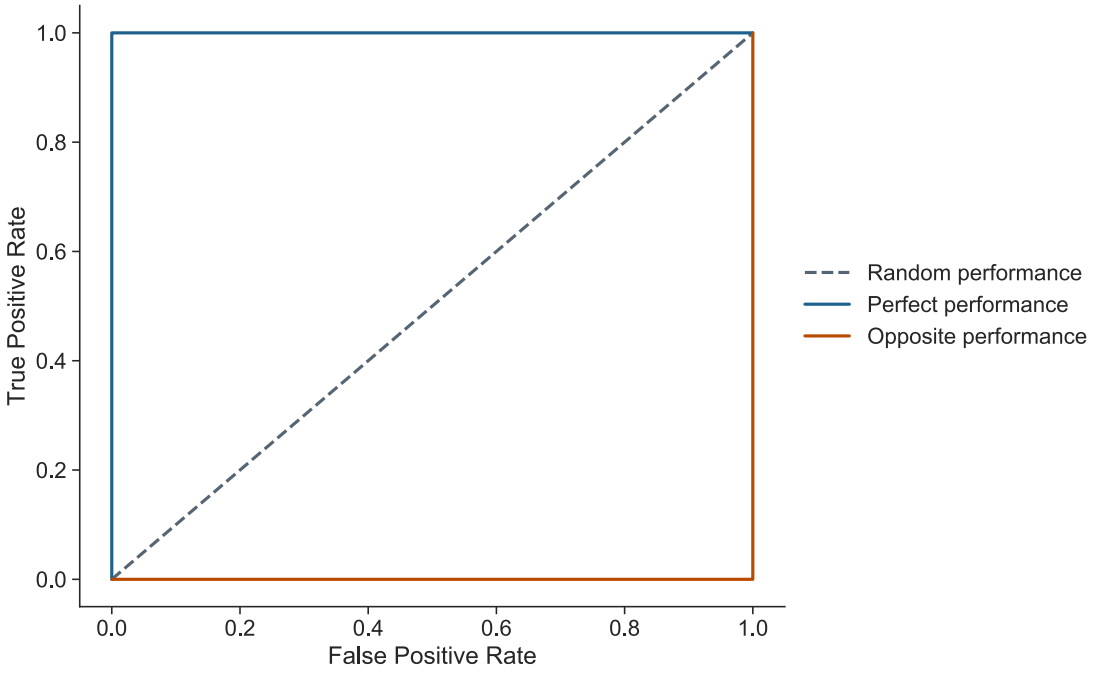


Figure 3.4: Illustration of three ROC curves representing random performance (diagonal dotted line), perfect ordering of the predictions (blue line), and the opposite of perfect ordering of the predictions (orange line).

metric measures the percentage of molecules for which a true prediction is within the top k ranked predictions. The top-k metric was used in this work to evaluate the performance of both the SoM prediction and metabolite prediction methods.

In this work, the BEDROC metric was used in the evaluation of the alignment-based SoM prediction approach (section 4.1). The BEDROC metric was developed to measure early enrichment in virtual screening and is commonly used for virtual screening [103], but in the case of this dissertation we have SoMs in place of actives and non-SoMs in place of inactives.

The BEDROC score is calculated as shown in Equation 3.3,

$$BEDROC = \frac{\sum_{i=1}^{n} e^{-\alpha r_i/N}}{R_a \frac{1-e^{-\alpha}}{e^{\alpha/N}-1}} \times \frac{R_a sinh(\alpha/2)}{cosh(\alpha/2) - cosh(\alpha/2 - \alpha R_a)} + \frac{1}{1-e^{\alpha(1-R_a)}} \quad (3.3)$$

where n is the number of actives (i.e. SoMs), N is the total number of compounds,

r_i is the rank of the ith active (i.e. SoM) in the ordered list, and R_a is the ratio of actives (i.e. SoMs) to the total number of compounds, meaning that $R_a = n/N$ [102]. The value of the BEDROC metric is always between 0 and 1, with higher values being better.

To calculate the BEDROC, a user-defined parameter α must be chosen. The value of α can be modulated to change the weighting of the top percentages of the ranked predictions, in effect representing how "early" the early enrichment being measured is. For $\alpha = 20$, 80% of the BEDROC score is accounted for by the top 8% of the ranked predictions. For $\alpha = 80.5$, the percentage of ranked predictions accounting for 80% of the score is only 2%. Hence the choice of α has a direct effect on the meaning of the BEDROC score, and BEDROC scores based on different α values cannot be compared [102].

Results

4.1 An Alignment-Based Approach to SoM Prediction

One way to endeavor to predict xenobiotic metabolism is to predict the metabolically labile atom positions. Of the many strategies for predicting SoMs that have previously been investigated, one approach that had been only briefly examined in the scientific literature was an alignment-based approach to predicting SoMs. In 2008, Sykes et al. used the CYP2C9-bound conformation of flurbiprofen as a template that they aligned, based on molecular shape and chemical features, to other CYP2C9 substrates [72]. The idea behind this approach is that the alignment might represent the relative orientations of the ligands to each other in terms of how they would bind in the binding pocket of the enzyme. This appeared to be the case, as 60% of the molecules were aligned such that their SoM was within 3 Å of the reference SoM in the template molecule [72]. These results suggested that such an easily interpretable alignment-based method could potentially be useful for SoM prediction; however, the study was limited in scope and in the depth of the analysis.

In the following study, the idea of an alignment-based approach to SoM prediction was further developed and expanded to include the most relevant CYP isozymes for humans. To this end, a previously published SoM dataset was modified to correct the stereochemistry of the molecules. Using this dataset, the approach was also analyzed in more detail, particularly in terms of key validation aspects such as early enrichment and false positives. As part of the evaluation, we examined how the performance of the approach was affected by the similarity of a query molecule to the reference molecules. We additionally examined the effect of adding an atom reactivity component to the method, which led to an improvement in performance.

[D1] Alignment-Based Prediction of Sites of Metabolism
Christina de Bruyn Kops, Nils-Ole Friedrich, and Johannes Kirchmair
Journal of Chemical Information and Modeling, 2017

Available at `https://doi.org/10.1021/acs.jcim.7b00165`.

Contribution:
C. de Bruyn Kops and J. Kirchmair conceptualized the research. C. de Bruyn Kops along with N.-O. Friedrich and J. Kirchmair designed the experiments. C. de Bruyn Kops conducted the computational method development and analysis. C. de Bruyn Kops wrote the manuscript, with contributions from J. Kirchmair. J. Kirchmair supervised this work.

The supplementary information for this article can be found in Appendix B.

JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

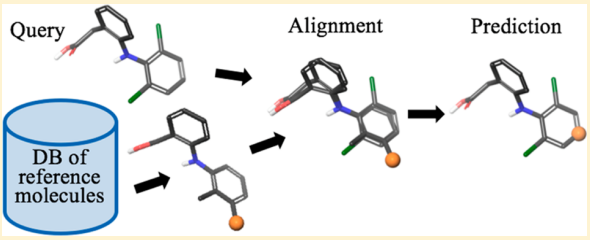
Alignment-Based Prediction of Sites of Metabolism

Christina de Bruyn Kops,[†] Nils-Ole Friedrich,[†] and Johannes Kirchmair*[,†]

[†]Faculty of Mathematics, Informatics and Natural Sciences, Department of Computer Science, Center for Bioinformatics, Universität Hamburg, Hamburg 20146, Germany

Supporting Information

ABSTRACT: Prediction of metabolically labile atom positions in a molecule (sites of metabolism) is a key component of the simulation of xenobiotic metabolism as a whole, providing crucial information for the development of safe and effective drugs. In 2008, an exploratory study was published in which sites of metabolism were derived based on molecular shape- and chemical feature-based alignment to a molecule whose site of metabolism (SoM) had been determined by experiments. We present a detailed analysis of the breadth of applicability of alignment-based SoM prediction, including transfer of the approach from a structure- to ligand-based method and extension of the applicability of the models from cytochrome P450 2C9 to all cytochrome P450 isozymes involved in drug metabolism. We evaluate the effect of molecular similarity of the query and reference molecules on the ability of this approach to accurately predict SoMs. In addition, we combine the alignment-based method with a leading chemical reactivity model to take reactivity into account. The combined model yielded superior performance in comparison to the alignment-based approach and the reactivity models with an average area under the receiver operating characteristic curve of 0.85 in cross-validation experiments. In particular, early enrichment was improved, as evidenced by higher BEDROC scores (mean BEDROC = 0.59 for α = 20.0, mean BEDROC = 0.73 for α = 80.5).



INTRODUCTION

Xenobiotic metabolism plays an important role in drug development, as the biological and physicochemical properties of the metabolites have a substantial effect on the successful utilization of a potential drug. Metabolism is one of the main clearance pathways for approximately 75% of existing drugs.[1] The toxicity of metabolites, as well as their potential activity or inactivity against the target protein, are important factors in the safety and efficacy of a drug.[2] For example, nearly 7% of known metabolites are confirmed toxic or contain a reactive group, whereas only around 3% are confirmed to maintain pharmacological activity.[3] From soft drugs, designed to have neither active nor toxic metabolites, to prodrugs, biotransformation of small molecules via metabolism plays a key role in the safety and efficacy of drugs.[2,4] An in-depth study of metabolism is therefore of vital importance for the drug discovery process.

Predicting metabolically labile atom positions, or sites of metabolism (SoMs), is a key component of the simulation of xenobiotic metabolism. Determining the SoMs makes it relatively straightforward, in most cases, to predict the chemical structures of the potential metabolites. Once these are known, an assessment of properties such as toxicity and reactivity can follow.[4,5]

Two key aspects that contribute to the location of a SoM in a molecule are the reactivity of the atom and its accessibility based on its position in the binding pocket, particularly accessibility from the catalytic center in enzymes. Cytochromes P450 (CYPs) are the most important class of enzymes for xenobiotic metabolism. Approximately 40% of all metabolites and 58% of first-generation metabolites are created by CYPs,[3] and this enzyme family is responsible for most drug−drug interactions and toxicity due to metabolism.[3,6] For CYP-mediated metabolism, reactivity is the major factor,[7] and many available SoM predictors focus solely on CYP substrates.

Current computational methods for SoM prediction range from knowledge-based approaches to machine learning, reactivity models, and structure-based approaches (mainly ligand docking).[4,5,8,9] These methods for SoM prediction can be generalized into three categories: knowledge-, structure-, and ligand-based. Knowledge-based systems use relatively simple rules derived from empirical knowledge accumulated by medicinal chemists over decades of research. Structure-based methods take a more detailed look at an interaction with a specific enzyme, e.g., using docking. This type of method allows for detailed examination of a specific protein and the binding of its substrates but requires high-quality protein structures and is limited by the conformational space covered by the crystal structures. In addition, the focus of structure-based methods is on the steric aspects, whereas any contributions from reactivity are largely ignored. Ligand-based approaches, on the other hand, do not require a crystal structure and therefore have broader applicability (e.g., multiple enzymes and coverage of a larger chemical space) but cannot provide information on binding or interaction of compounds with an enzyme. In

Received: March 21, 2017
Published: May 18, 2017

DOI: 10.1021/acs.jcim.7b00165
J. Chem. Inf. Model. 2017, 57, 1258−1264

general, the most accurate models are multicomponent models that make use of several different methods, e.g., docking combined with a reactivity model or a rule-based approach plus a reactivity model.[4,5,8] A typical metric for evaluating the success of these various methods at predicting SoMs is the percentage of molecules for which at least one known SoM is among the top 2 or 3 predicted atom positions in a molecule. For most existing methods, these percentages lie between 70 and 90%.[10]

In 2008, Sykes et al.[11] published an interesting study in which the potential of alignment-based methods for SoM prediction was probed. They extracted flurbiprofen in its CYP2C9-bound conformation from an X-ray structure and used it as a template for the alignment of 69 other known CYP2C9 substrates. It was assumed that molecular shape- and chemical feature-based alignment would allow the derivation of the correct orientation of the known substrates. Sykes et al. did indeed find that 60% of the molecules were aligned such that their site of metabolism was within 3 Å of the reference site of metabolism.

The study by Sykes et al. yielded encouraging results and suggested that alignment-based methods could have high potential for SoM prediction, in particular when used as one component of a more sophisticated model. However, this exploratory study is limited in some aspects. For example, it was based on a data set of just 70 molecules with all but one molecule annotated with only a single SoM corresponding to CYP2C9 (not considering topological symmetry). The approach was designed as a surrogate docking technique that uses the protein-bound ligand conformation as a reference, hence limiting it to the availability of such experimentally determined structures. Furthermore, the requirement for successful prediction, namely a distance of less than 3 Å between the aligned SoM in the reference molecule and the known SoM in the query molecule, is a lenient definition of what constitutes a good prediction, as a large portion of a typical druglike molecule (if not the entire molecule) can be placed within such a sphere (Figure 1). In addition, a limitation in terms of validation is that neither the existence nor the extent of false positive predictions was considered.

Inspired by the work of Sykes et al., we set out to fully explore the scope and limitations of alignment-based methods for SoM prediction. To examine the degree to which their applicability can be broadened, we investigate the effects of several aspects and extensions of the method. A shape- and chemical feature-based alignment between a query molecule and a set of reference molecules was used to rank the atoms in the query molecule in order of likelihood of being a SoM. By increasing the validation data set from a small data set based on a single CYP isozyme to a more comprehensive data set with SoM information for all CYP isozymes, we show that the applicability of this approach can be extended from CYP2C9 to encompass CYP metabolism of druglike molecules in general. In conjunction with using this larger data set, we adjusted the method from a structure-based to a purely ligand-based approach. We additionally examined the combination of a chemical reactivity model with the alignment-based approach, as chemical reactivity is of key importance to modeling SoMs.

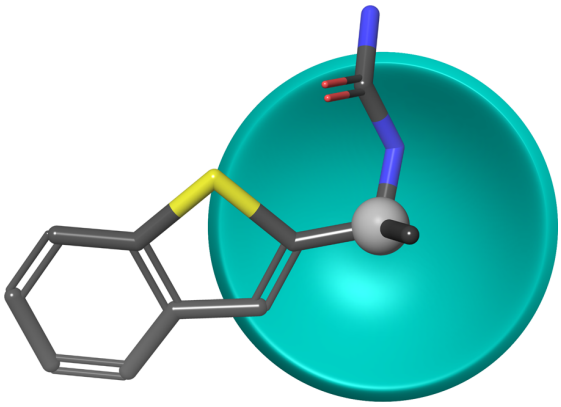


Figure 1. A sphere of 3 Å in radius (green hemisphere) placed around an atom (gray sphere) on a small druglike molecule can potentially cover a large portion of the molecule.

RESULTS

Approach of Sykes et al. Revisited. In a first experiment, we revisited the study of Sykes et al. and tested the performance of the ROCS-based approach for SoM prediction using the same experimental setup but more stringent criteria for analyzing the success of prediction. We aligned the 69 query molecules of the Sykes data set to the experimentally observed conformation of flurbiprofen bound to CYP2C9 using ROCS, as specified in the original study. We then determined the likelihood that a known SoM in each of the 69 query molecules would be within less than one bond length (approximated to 1 Å) of a SoM in flurbiprofen and extended this cutoff to 2 and 3 Å for comparison.

Sykes et al. reported that 60%, or 42 of 70 molecules (including flurbiprofen, which they aligned to itself in protein-bound conformation), had a known SoM aligned within 3 Å of the SoM in flurbiprofen.[11] We found that 55% of the molecules (38 of 69) had a known SoM within the same radius (Table 1).

Table 1. Percent of Molecules with a Known SoM within a 1, 2, or 3 Å Radius of a SoM in the Reference Molecule[a]

radius [Å]	Sykes Data Set [%]	Zaretzki Data Set [%]
1	23.2	54.3 ± 1.5
2	37.7	68.5 ± 1.9
3	55.1	78.7 ± 1.3

[a]For the Zaretzki Data Set, both the arithmetic mean and standard error are reported.

The results of both studies differ by only three molecules. Using a smaller radius resulted, as expected, in fewer correctly predicted molecules. Only 38% of molecules had a known SoM located within 2 Å of the flurbiprofen SoM.

Subsequently, using the same experimental setup described above, we ranked the atoms in each query molecule according to distance to a known SoM in the aligned reference molecule. With this ranking, we looked at the percentage of query molecules for which there is at least one known SoM in the top 1−3 atoms located closest to the reference molecule's SoM (top k metric) as well as the receiver operating characteristic (ROC) curve and the BEDROC metric. Only 23% of the molecules in the Sykes data set had a known SoM as the top-ranked atom, and only 39% of the molecules had a known SoM within the top 3 ranked atoms (Table 2). The area under the receiver operating characteristic curve (AUC) was 0.62 ± 0.038 (standard error; Figure 2a).

The BEDROC metric provides insight into early enrichment. An α of 20 means that 80% of the BEDROC score is accounted

DOI: 10.1021/acs.jcim.7b00165
J. Chem. Inf. Model. 2017, 57, 1258−1264

Table 2. Percentage of Molecules with a Known SoM in the Top-k Scored Atoms for Top 1−3 for Each Data Set[a]

	top 1 [%]	top 2 [%]	top 3 [%]
Sykes Data Set, alignment only	23.2	31.9	39.1
Zaretzki Data Set, alignment only	46.0 ± 1.8	57.7 ± 1.8	64.0 ± 2.1
Zaretzki Data Set, reactivity only	59.9 ± 1.4	73.4 ± 1.6	79.6 ± 0.9
Zaretzki Data Set, alignment and reactivity combined	53.8 ± 1.2	71.9 ± 1.4	80.3 ± 1.4

[a]For the Zaretzki Data Set, both the arithmetic mean and the standard error are reported.

for by the top 8% of the ranked data, whereas an α of 80.5 means that the top 2% of the data accounts for 80% of the score.[12,13] The BEDROC value obtained with an α of 20.0 was 0.28 compared to a BEDROC value of 0.22 for an α of 80.5 (Table 3). These low values for the BEDROC metric indicate relatively weak early enrichment, coinciding with the low top k values.

Toward an Alignment-Based Model for Predicting SoMs Related to CYP Metabolism. The study of Sykes et al. is focused on CYP2C9 metabolism and based on a data set of just 70 molecules with only a single known SoM assigned to most molecules. To study the potential of alignment-based approaches for SoM prediction in more detail and to extend the applicability from one isozyme to all relevant CYP isozymes, a larger data set was required. We therefore used a revised version (see Methods) of the Zaretzki data set of 680 molecules with SoMs from all CYP isozymes.[14] Unlike Sykes et al., we did not aim to use ROCS as a surrogate docking approach that would allow the prediction of the orientation of a ligand within the binding site. We therefore did not need to restrict ourselves to a CYP-bound ligand conformation for the reference molecule. This change enabled us to take advantage of the larger data set by modifying the method to allow for choosing the best reference molecule from a wide selection of molecules with known SoMs. This adjustment also facilitated automatic extension of the approach to all CYP isozymes, the idea being that the best-aligned reference molecule would have a similar binding profile for the set of CYP isozymes and also have correspondingly similar SoMs. If a query molecule is of similar shape to a reference molecule, then the molecules are assumed to bind with the same orientation to the same isozyme(s) and hence have SoMs located at similar atom positions. This concept is in direct contrast to the basis of the Sykes et al. study on a single isozyme that metabolizes all molecules in the data set. Contrasting schemes of the two methods are shown in Figure 3.

Once the best-aligned reference molecule had been selected, we ranked all atoms in the query molecule, as described above. In addition, the Zaretzki data set was analyzed with 10-fold cross-validation in which the data set was randomly split into 10 sets of query molecules, each containing 10% of the total data set. The remaining 90% of the molecules for each set were used as the corresponding reference set.

We again investigated the existence of a known SoM within 1, 2, and 3 Å of an aligned reference SoM. The percentage of molecules with a known SoM within 1 Å of a SoM of the best-aligned reference molecule was 54% (Table 1). When the radius was increased to 3 Å, an average of 79% of the query molecules had a known SoM within this distance of a reference SoM. This is a relatively low percentage when one considers that a sphere with radius 3 Å may cover a large portion of a druglike molecule (Figure 1) and that, if the reference molecule has more than one SoM, this effect can be even more striking. Using the adjusted method on the Zaretzki data set resulted in better SoM predictions according to the top k, AUC, and BEDROC metrics. For example, an average of 58% of molecules in the Zaretzki data set had a known SoM in the

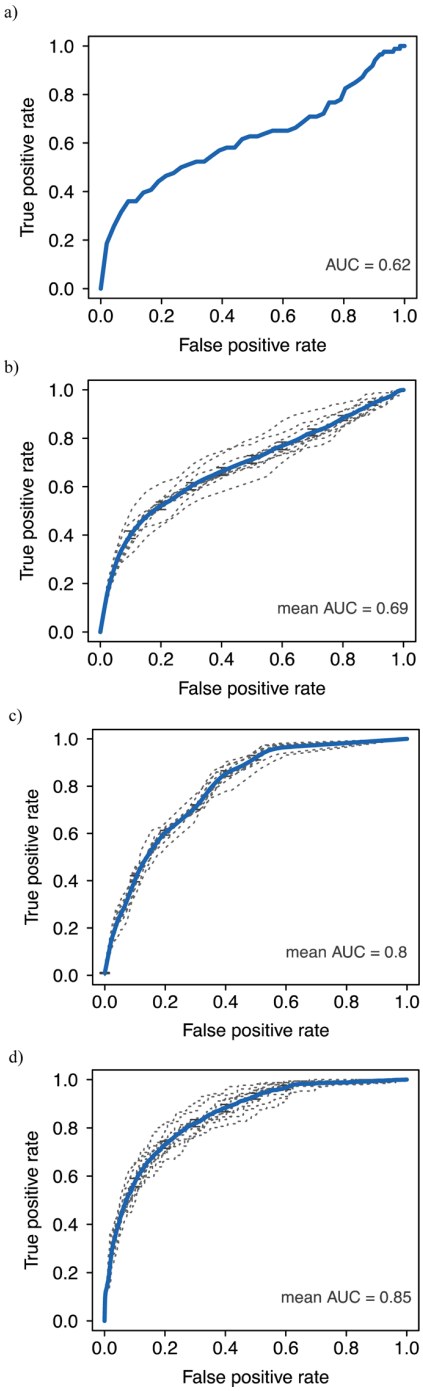


Figure 2. ROC curves for (a) alignment-based SoM prediction for the Sykes data set, (b) alignment-based SoM prediction for the Zaretzki data set, (c) reactivity-based SoM prediction for the Zaretzki data set, and (d) combination of alignment- and reactivity-based SoM prediction for the Zaretzki data set. Error bars represent standard error.

DOI: 10.1021/acs.jcim.7b00165
J. Chem. Inf. Model. 2017, 57, 1258−1264

Table 3. BEDROC for the Sykes and Zaretzki Datasets[a]

BEDROC α	Sykes data set, alignment only	Zaretzki data set, alignment only	Zaretzki data set, reactivity only	Zaretzki data set, alignment and reactivity combined
20.0	0.28	0.44 ± 0.01	0.44 ± 0.01	0.59 ± 0.01
80.5	0.22	0.50 ± 0.03	0.54 ± 0.02	0.73 ± 0.02

[a]For the Zaretzki data set, both the arithmetic mean and standard error are reported.

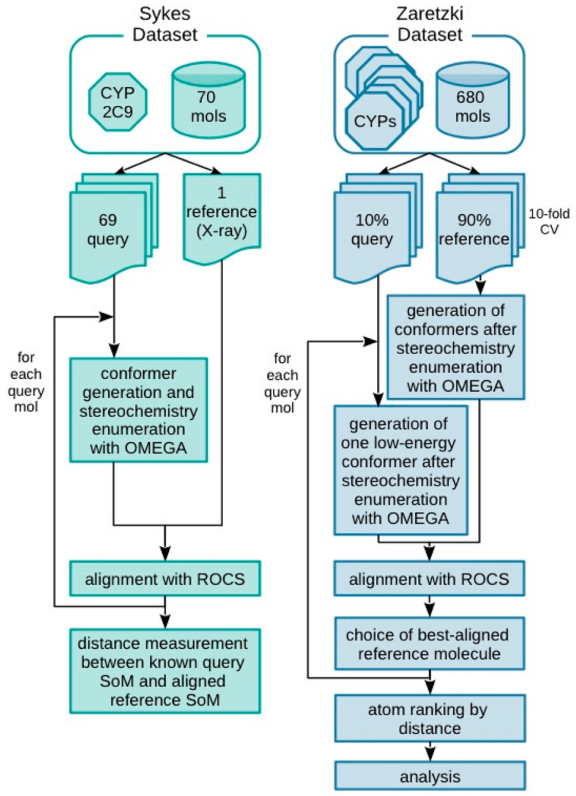


Figure 3. Workflow used by Sykes et al.[11] (left) compared to the workflow used in this study (right).

top 2 ranked atoms (Table 2). The average AUC was 0.69 (Figure 2b) with a standard error of 0.0078, and the average BEDROCs were 0.44 and 0.50 for α = 20.0 and α = 80.5, respectively (Table 3).

Note that differences in performance of the Sykes and our approach are the result of adjustments of the method and the use of a larger data set with a more comprehensive record of SoMs. Therefore, direct comparison of the performance of the two approaches should be avoided.

Molecular Similarity. The effect of molecular similarity on the quality of the alignment-based predictions was examined. Greater molecular similarity was found to correspond to improved SoM prediction for the alignment-based method. Both the percentage of molecules with a known SoM within 1, 2, or 3 Å of a reference SoM and the percentage of molecules within the top 1−3 ranked atoms are larger for more highly similar molecules. As an example illustrating this trend, the percentage of correctly predicted molecules according to the 1 Å criterion increases on average from 29%, for molecules with up to 0.25 similarity (calculated using the Tanimoto coefficient of ECFP6-like fingerprints) to the reference molecule, to 76% for molecules with similarity greater than 0.75 (Figure 4). Correspondingly, the percentage of molecules within the top

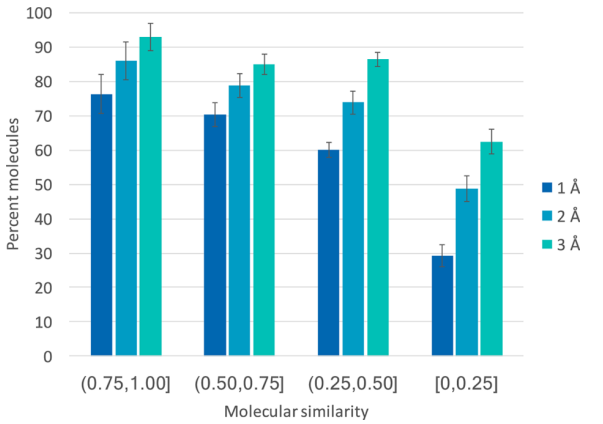


Figure 4. Percentage of molecules with a known SoM within 1, 2, and 3 Å of a reference SoM in the Zaretzki data set with percentage calculated separately for each bin of molecular similarity calculated using the Tanimoto coefficient of ECFP6-like fingerprints.

1−3 ranked atoms increases as the molecular similarity increases from 58% in the top 2 with similarity up to 0.25 to 88% in the top 2 with similarity greater than 0.75, as shown in Figure 5a.

Combination of an Alignment-Based Method with a Reactivity Model. Because steric accessibility of an atom is only one component of the likelihood of a reaction occurring at that position in a molecule, we also explored the possibility of increasing prediction accuracy by including a reactivity component in the model. SMARTCyp is a leading method for predicting SoMs related to CYP metabolism. The more recent versions of SMARTCyp contain steric accessibility components in addition to the reactivity component. Rydberg et al. showed that the reactivity component accounts for 86−97% of their model;[7] however, the SMARTCyp model with the latest accessibility descriptor was trained on a portion of the nonmodified version of the same data set we used in our study (the Zaretzki data set), whereas the reactivity descriptor in SMARTCyp is derived directly from DFT calculations.[7,15] We therefore extracted only the reactivity component from the SMARTCyp predictions in addition to using the standard SMARTCyp score for comparison. Because we found the results to be similar, the results with the standard SMARTCyp score are located in the Supporting Information.

The top k results for the reactivity component of SMARTCyp on its own (Table 2) are comparable to those previously reported (for CYP3A4) with the original version of SMARTCyp, which was not trained on the Zaretzki data set but contains an accessibility descriptor.[15] We combined this reactivity component with the alignment-based atom ranking to create new scores for all the atoms of all query molecules. The mean percentage of molecules with a known SoM included in the top 2 ranked molecules was higher for the combination of alignment and reactivity than for the alignment-only method, i.e., 72 and 58%, respectively (Table 2). Because the percentage of molecules for this same metric was 73% for the reactivity component only, the top k metric shows only a benefit of the combined approach compared to that of the alignment-based approach but not compared to the reactivity component on its

DOI: 10.1021/acs.jcim.7b00165
J. Chem. Inf. Model. 2017, 57, 1258−1264

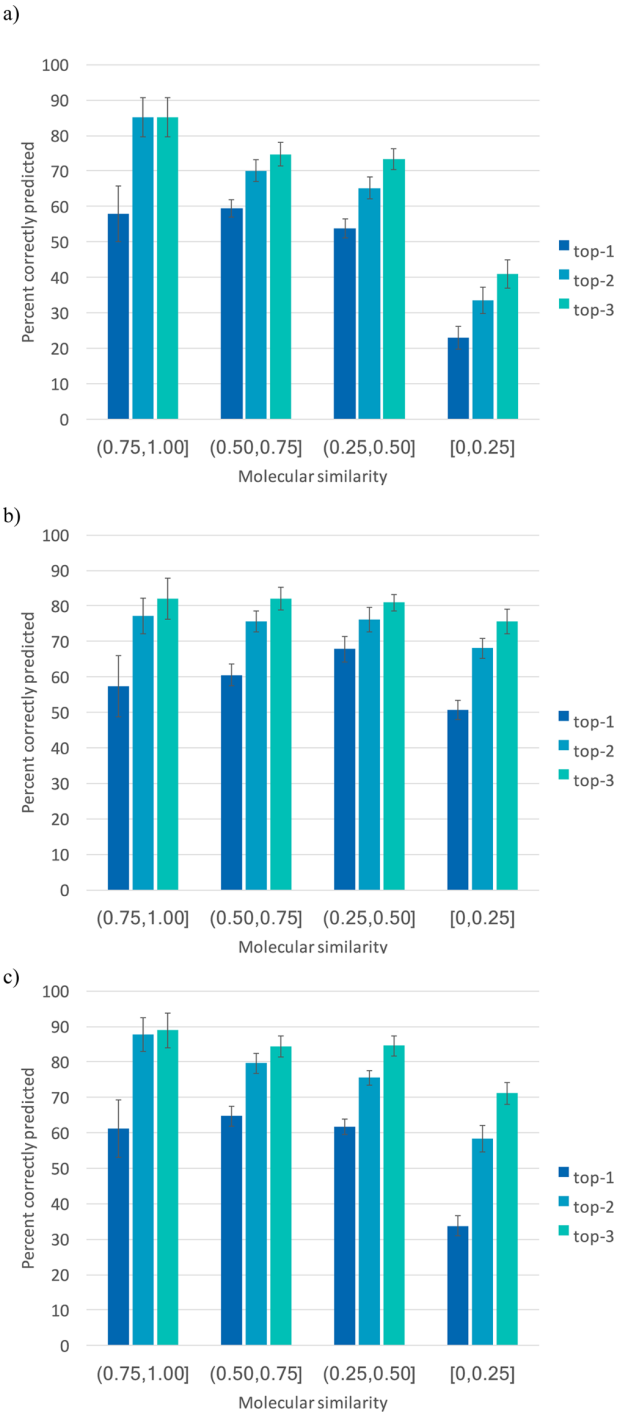


Figure 5. Top k analysis for different similarity ranges applied to (a) the alignment-based SoM prediction, (b) reactivity-based SoM prediction, and (c) the combination of alignment-based ranking and the reactivity descriptor for SoM prediction. The percentage is calculated separately for each bin of molecular similarity, which is calculated using the Tanimoto coefficient of ECFP6-like fingerprints.

own. More importantly, however, a larger difference in prediction according to the top k metric is seen when the results are broken down into categories of molecular similarity (calculated using the Tanimoto coefficient of ECFP6-like fingerprints; Table 2). As expected, there was no clear trend for the effect of molecular similarity on the top k results for the reactivity component of SMARTCyp on its own (Figure 5b) or even the total SMARTCyp score (Figure S1a). Combining the alignment-based ranking and the reactivity component results in a top 2 percentage of 88% for similarity greater than 0.75, as compared to 85% for alignment only and 77% for reactivity only for the same similarity range (Figure 5).

Analysis of the AUC and BEDROC scores (Figure 2, Table 3), both of which, unlike the top k metric, take false positives into account, shows a clear advantage of the combined approach over both models that rely on a single component. More specifically, the combined method had better early enrichment than the reactivity-based approach, reflected by a better BEDROC score at $\alpha = 20.0$ (BEDROC = 0.59 vs 0.54), and in particular, a higher BEDROC score at $\alpha = 80.5$ (BEDROC = 0.73 vs 0.54). The combined method also had a statistically significantly higher average AUC (AUC = 0.85 ± 0.0048 vs 0.80 ± 0.0051; $p = 4.0 \times 10^{-33}$). The AUC in particular reveals an even larger significant increase in performance for the combined approach compared to the alignment-based atom ranking (average AUC = 0.69 ± 0.0078; $p = 1.7 \times 10^{-139}$; mean BEDROC = 0.44 for $\alpha = 20.0$; mean BEDROC = 0.50 for $\alpha = 80.5$). For comparison, the AUC of the reactivity-based approach was also significantly higher than that of the alignment-based atom ranking ($p = 2.1 \times 10^{-33}$). All these results indicate substantially better early enrichment (and hence reduction of false-positive rates) of the combined approach. The same trend was observed for predictions using the full SMARTCyp score on its own and combined with the alignment-based atom ranking (Figures S1–S2, Tables S1–S3), including a significantly higher AUC with SMARTCyp than with alignment-based ranking ($p = 3.0 \times 10^{-45}$) and a significant improvement in AUC when using the combination of alignment with the full SMARTCyp score compared to both the SMARTCyp score on its own ($p = 5.0 \times 10^{-21}$) and the alignment-based prediction on its own ($p = 5.5 \times 10^{-145}$).

CONCLUSIONS

We have examined the scope and limitations of an alignment-based approach to SoM prediction. Starting from the CYP2C9 model introduced by Sykes et al.,[11] we widened coverage to include the CYP isozymes most important for metabolism in humans. Our analysis filled in the picture presented by Sykes et al. and provided information about early enrichment and false positives. We discovered that the alignment-based method can be used to predict SoMs to a considerable extent. On its own, the alignment-based approach to SoM prediction yields results comparable to those of rigid receptor docking.[16] Additional improvements to docking result in improved performance for more flexible docking models (e.g., in combination with molecular dynamics simulations techniques).[16,17] In this case, the advantage of an alignment-based method remains in that it is ligand-based as opposed to structure-based. The alignment-based approach does not have the accuracy and reach of other methods for SoM prediction such as machine learning and reactivity models, but it could be beneficial as a complementary component of SoM predictors, in particular in addition to reactivity. Reactivity is a key factor in an atom's involvement in a metabolism reaction, and small changes to a molecule can have a large impact on reactivity. For example, widely applied strategies for improving metabolic stability while maintaining target activity, such as fluorination, can have a substantial impact on the metabolic fate of compounds even though the

DOI: 10.1021/acs.jcim.7b00165
J. Chem. Inf. Model. 2017, 57, 1258–1264

molecular shape remains largely preserved.[18] Because the alignment-based ranking of atoms does not take atom reactivity into account, the combination of alignment with reactivity prediction makes sense on a conceptual level. We tested this combination using the reactivity descriptor from SMARTCyp and saw a substantial improvement in the predictions, in particular with respect to early enrichment and false-positive rates.

The quality of the alignment-based predictions is strongly affected by the availability of suitable reference molecules with sufficient similarity to the molecule in question. The alignment-based approach works well for structurally related molecules but less well for highly structurally different molecules. It is therefore expected that the implementation of a mechanism for toggling the alignment-based technique in two- or multi-component models will boost prediction accuracy. For example, the distance-based ranking that is the output of the alignment-based method could be used as a descriptor for machine learning models for SoM prediction. Furthermore, requiring no structural information, in contrast to the original study published by Sykes et al., has already removed a key constraint on the available data and results in a wider array of possible reference compounds to draw from. With the increasing availability of metabolism data, the value of alignment-based methods is expected to increase in the future.

METHODS

Data Set Preparation. *Sykes Data Set.* The structure of the reference molecule flurbiprofen was downloaded from the PDB in its crystallized conformation (PDB ID: 1R9O), as specified by ref 11. The structures of the 69 query molecules were downloaded as 2D SD files from the PubChem Compound database. The structure of kaempferide depicted in ref 11 was corrected. The SoMs were annotated manually according to the data published by Sykes et al.[11]

Zaretzki Data Set. The latest version of the Zaretzki data set[14] containing 680 molecules and SoM information from all CYP isozymes was used. Because the chirality flags were found to be unreliable, the chirality flags of all 298 chiral molecules were checked against the PubChem Compound database[19] and/or the primary literature and adjusted accordingly. Additionally, the topology of eight molecular structures was corrected, preserving correct SoM annotation. The modified version of the Zaretzki data set used in this study is available in the Supporting Information.

Workflow. The SoMs for all CYP isozymes were used for the Zaretzki data set. The Zaretzki data set was separated into query and reference molecules using 10-fold cross-validation with random fold selection by KFold from scikit-learn.[20,21] Each training set was used as a set of reference molecules, and each test set as the corresponding query molecules.

Because only heavy atoms are considered SoMs, any SoM annotated on a hydrogen atom was moved to the neighboring heavy atom. Topologically identical atoms, determined using the *CanonicalRankAtoms* function in the *Chem* module of RDKit,[22] were assigned the same status, i.e., if any of the atoms in a set of topologically identical atoms is a SoM, then all atoms in the set are considered SoMs.

For the Sykes data set, conformers were only generated for the query molecules, and the reference molecule was left in the fixed (i.e., experimentally determined, protein-bound) conformation. For the Zaretzki data set, only one conformer was generated for each query molecule, whereas the default number of conformers was generated for the reference molecules. Conformer generation was performed with OMEGA.[23,24] The flipper option in OMEGA was used to enumerate all stereoisomers for each undefined stereocenter. Default parameters were used for OMEGA with the exception of *canonOrder* to preserve the numbering of atoms from SoM annotations and, for the query molecules of the Zaretzki data set, *maxConfs* set to 1.

Each query molecule was aligned to all reference molecules. The alignment was performed with ROCS[25,26] using default parameters. Only the reference molecule with the highest TanimotoCombo score when aligned with the query molecule was used for SoM prediction. A single stereoisomer of each query molecule was chosen, if applicable, by selecting the stereoisomer with the highest alignment score out of all stereoisomers aligned with all reference molecules.

The heavy atoms in the query molecule were ranked according to distance to a known SoM of the chosen reference molecule. Reactivity prediction was performed with SMARTCyp (version 2.4.2)[7,15,27] for the query molecules using default parameters and the standard SMARTCyp model. To combine the distance ranking and the SMARTCyp score for each atom, both the alignment-based ranking and the SMARTCyp score were normalized and inverted so that the best value of each was 1. If an atom was predicted as reactive by SMARTCyp (energy not 999), the two values were then averaged to create a final prediction score. Otherwise, the prediction score was set to 0, the lowest possible combined score. The score returned by the current version of SMARTCyp contains two accessibility components in addition to the calculated reactivity. We used both the full SMARTCyp score with accessibility included and the reactivity component of the SMARTCyp score on its own. The same procedure to combine SMARTCyp prediction and distance ranking was carried out both using the full SMARTCyp score and only using the reactivity component of SMARTCyp.

Evaluation. ROC curves were calculated with the ROCR R package.[28,29] The average ROC curve for the cross-validation folds was determined using vertical averaging.

The statistical significance of differences in AUC was determined using the R package pROC with the DeLong test for correlated ROC curves.[30,31] The variance of the AUC, from which the standard error was derived, was also calculated with pROC using the DeLong method. For these calculations, all cross-validation runs were combined into one ROC curve, from which the total AUC was then calculated. The AUC calculated in this manner is equal to the average AUC for the cross-validation runs calculated in ROCR. All ROC curves calculated for the Zaretzki data set were considered to be correlated for this test.

Molecular similarity was calculated with ECFP6-like fingerprints and the Tanimoto coefficient using RDKit. The BEDROC[12] was calculated using the R package enrichvs.[32]

Hardware Setup. Calculations were performed on a Linux workstation running openSUSE 13.1 with an Intel i7 3.6 GHz processor and 32 GB of main memory and a Linux cluster running openSUSE 13.1 and equipped with Intel Xeon processors (2.2 to 2.7 GHz) and 126 GB of main memory.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jcim.7b00165.

DOI: 10.1021/acs.jcim.7b00165
J. Chem. Inf. Model. 2017, 57, 1258−1264

SMARTCyp and the combination of SMARTCyp with alignment-based prediction, top k results not considering molecular similarity, top k results considering molecular similarity, ROC curves, BEDROC scores, and standard error in the AUCs (PDF)

Text file containing SoM-annotated structures for the modified Zaretzki data set (ZIP)

AUTHOR INFORMATION

Corresponding Author

*E-mail: kirchmair@zbh.uni-hamburg.de. Tel.: +49 (0)40 42838 7303.

ORCID

Christina de Bruyn Kops: 0000-0001-8890-2137

Nils-Ole Friedrich: 0000-0002-8983-388X

Johannes Kirchmair: 0000-0003-2667-5877

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We thank OpenEye Scientific Software for providing licenses for ROCS and OMEGA via the Free Academic Licensing Program.

ABBREVIATIONS

SoM, site of metabolism; AUC, area under the receiver operating characteristic curve; CYP, cytochrome P450; ROC, receiver operating characteristic

REFERENCES

(1) Di, L. The Role of Drug Metabolizing Enzymes in Clearance. *Expert Opin. Drug Metab. Toxicol.* **2014**, *10*, 379−393.

(2) Testa, B. Metabolism in Drug Development. In *Drug Metabolism Prediction*; Kirchmair, J., Ed.; Methods and Principles in Medicinal Chemistry; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2014; pp 1−26.

(3) Testa, B.; Pedretti, A.; Vistoli, G. Reactions and Enzymes in the Metabolism of Drugs and Other Xenobiotics. *Drug Discovery Today* **2012**, *17*, 549−560.

(4) Kirchmair, J.; Göller, A. H.; Lang, D.; Kunze, J.; Testa, B.; Wilson, I. D.; Glen, R. C.; Schneider, G. Predicting Drug Metabolism: Experiment And/or Computation? *Nat. Rev. Drug Discovery* **2015**, *14*, 387−404.

(5) Olsen, L.; Oostenbrink, C.; Jørgensen, F. S. Prediction of Cytochrome P450 Mediated Metabolism. *Adv. Drug Delivery Rev.* **2015**, *86*, 61−71.

(6) Wienkers, L. C.; Heath, T. G. Predicting in Vivo Drug Interactions from in Vitro Drug Discovery Data. *Nat. Rev. Drug Discovery* **2005**, *4*, 825−833.

(7) Rydberg, P.; Rostkowski, M.; Gloriam, D. E.; Olsen, L. The Contribution of Atom Accessibility to Site of Metabolism Models for Cytochromes P450. *Mol. Pharmaceutics* **2013**, *10*, 1216−1223.

(8) Kirchmair, J.; Williamson, M. J.; Tyzack, J. D.; Tan, L.; Bond, P. J.; Bender, A.; Glen, R. C. Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms. *J. Chem. Inf. Model.* **2012**, *52*, 617−648.

(9) Oostenbrink, C. Structure-Based Methods for Predicting the Sites and Products of Metabolism. In *Drug Metabolism Prediction*; Kirchmair, J., Ed.; Methods and Principles in Medicinal Chemistry; Wiley-VCH Verlag GmbH & Co. KGaA: Weinheim, Germany, 2014; pp 243−264.

(10) Campagna-Slater, V.; Pottel, J.; Therrien, E.; Cantin, L.-D.; Moitessier, N. Development of a Computational Tool to Rival Experts in the Prediction of Sites of Metabolism of Xenobiotics by p450s. *J. Chem. Inf. Model.* **2012**, *52*, 2471−2483.

(11) Sykes, M. J.; McKinnon, R. A.; Miners, J. O. Prediction of Metabolism by Cytochrome P450 2C9: Alignment and Docking Studies of a Validated Database of Substrates. *J. Med. Chem.* **2008**, *51*, 780−791.

(12) Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.

(13) Chaput, L.; Martinez-Sanz, J.; Saettel, N.; Mouawad, L. Benchmark of Four Popular Virtual Screening Programs: Construction of the Active/decoy Dataset Remains a Major Determinant of Measured Performance. *J. Cheminf.* **2016**, *8*, 56.

(14) Zaretzki, J.; Matlock, M.; Swamidass, S. J. XenoSite: Accurately Predicting CYP-Mediated Sites of Metabolism with Neural Networks. *J. Chem. Inf. Model.* **2013**, *53*, 3373−3383.

(15) Rydberg, P.; Gloriam, D. E.; Zaretzki, J.; Breneman, C.; Olsen, L. SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. *ACS Med. Chem. Lett.* **2010**, *1*, 96−100.

(16) Li, J.; Cai, J.; Su, H.; Du, H.; Zhang, J.; Ding, S.; Liu, G.; Tang, Y.; Li, W. Effects of Protein Flexibility and Active Site Water Molecules on the Prediction of Sites of Metabolism for Cytochrome P450 2C19 Substrates. *Mol. BioSyst.* **2016**, *12*, 868−878.

(17) Moors, S. L. C.; Vos, A. M.; Cummings, M. D.; Van Vlijmen, H.; Ceulemans, A. Structure-Based Site of Metabolism Prediction for Cytochrome P450 2D6. *J. Med. Chem.* **2011**, *54*, 6098−6105.

(18) Murphy, C. D.; Sandford, G. Recent Advances in Fluorination Techniques and Their Anticipated Impact on Drug Metabolism and Toxicity. *Expert Opin. Drug Metab. Toxicol.* **2015**, *11*, 589−599.

(19) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; et al. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44*, D1202−D1213.

(20) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.

(21) *Scikit-learn*, version 0.18, 2016.

(22) *RDKit: Open-Source Cheminformatics*, version 2016.03.1, 2016.

(23) Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database. *J. Chem. Inf. Model.* **2010**, *50*, 572−584.

(24) *OMEGA*, version 2.5.1.4; OpenEye Scientific Software: Santa Fe, NM, 2011.

(25) Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools. *J. Med. Chem.* **2007**, *50*, 74−82.

(26) *ROCS*, version 3.2.1.4; OpenEye Scientific Software: Santa Fe, NM, 2015.

(27) Rydberg, P.; Gloriam, D. E.; Olsen, L. *SMARTCyp*, version 2.4.2, 2013.

(28) Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. ROCR: Visualizing Classifier Performance in R. *Bioinformatics* **2005**, *21*, 3940−3941.

(29) Sing, T.; Sander, O.; Beerenwinkel, N.; Lengauer, T. *ROCR*, version 1.0-7, 2015.

(30) Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Müller, M. pROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves. *BMC Bioinf.* **2011**, *12*, 77.

(31) Robin, X.; Turck, N.; Hainard, A.; Tiberti, N.; Lisacek, F.; Sanchez, J.-C.; Müller, M. *pROC*, version 1.7.2, 2014.

(32) Yabuuchi, H. *enrichvs*, version 0.0.5, 2011.

DOI: 10.1021/acs.jcim.7b00165
J. Chem. Inf. Model. 2017, 57, 1258−1264

4.2 Metabolite Structure Prediction

SoM prediction is one portion of the equation when it comes to predicting xenobiotic metabolism, but it many cases it is desirable to actually predict the structures of the metabolites. To this end, the following studies explore how SoM prediction can be used to improve metabolite structure prediction. The SoM prediction methods used in this portion of the thesis are not the alignment-based method described previously (section 4.1) but rather machine learning-based methods that were developed in our research group [A4, A8].

4.2.1 Predicting CYP Metabolites

A problem common to most of the previously existing methods for metabolite structure prediction is that they tend to generate huge numbers of false positive predictions, which can make finding the real metabolites difficult. The following study explored how SoM prediction could be incorporated into a metabolite structure prediction approach in order to address this problem. In this study, we focused exclusively on CYP-mediated metabolism and used FAME 2 [A4] to predict SoMs.

A major component of this study was the development of a set of reaction rules describing CYP-mediated metabolic reactions. The reaction rules were created manually based on an analysis of the scientific literature on CYP-mediated metabolism and were classified as either common or uncommon reaction types, according to the relevant literature.

To analyze the effects of incorporating SoM prediction, two concepts were implemented and evaluated. In the first approach, the predicted SoMs were used as a hard filter that determined where in the molecule the reaction rules were applied. In the second approach, there was no filter. Instead, the predicted SoM probabilities were used to score the predicted metabolites, along with consideration of the above-mentioned simple distinction between common and uncommon reaction types. We found that this scoring approach resulted in a meaningful ranking of the predicted metabolites.

For the development and evaluation of the method, a reference dataset of 848 parent molecules and their CYP metabolites was compiled from the DrugBank [101]

and MetXBioDB [85] databases. In addition, a test dataset was created and used in the final validation of the approach, including a comparison to two previously published open-source metabolite structure prediction tools. This test dataset was manually assembled from the scientific literature on the metabolism of drugs and other xenobiotics and contains 29 parent molecules and their 81 CYP-formed metabolites.

This study resulted in the development of GLORY, a novel tool for predicting metabolites formed by CYP-mediated reactions. GLORY is freely available as a web server.

[D2] GLORY: Generator of the Structures of Likely Cytochrome P450 Metabolites Based on Predicted Sites of Metabolism
Christina de Bruyn Kops, Conrad Stork, Martin Šícho, Nikolay Kochev, Daniel Svozil, Nina Jeliazkova, and Johannes Kirchmair
Frontiers in Chemistry, 2019

Available at `https://doi.org/10.3389/fchem.2019.00402`.

Contribution:
C. de Bruyn Kops, N. Kochev, N. Jeliazkova, and J. Kirchmair conceptualized the research. C. de Bruyn Kops developed the metabolite prediction method. C. de Bruyn Kops wrote the manuscript, with contributions from J. Kirchmair. C. Stork contributed to the development of the reaction rules. M. Šícho contributed to the implementation of SoM prediction. J. Kirchmair and D. Svozil supervised this work.

The following article was reprinted from:
de Bruyn Kops, C.; Stork, C.; Šícho, M.; Kochev, N.; Svozil, D.; Jeliazkova, N.; Kirchmair, J. GLORY: Generator of the Structures of Likely Cytochrome P450 Metabolites Based on Predicted Sites of Metabolism, *Front. Chem.* **2019**, *7*, 402.

The supplementary information for this article can be found in Appendix C.

frontiers in Chemistry

ORIGINAL RESEARCH
published: 12 June 2019
doi: 10.3389/fchem.2019.00402

GLORY: Generator of the Structures of Likely Cytochrome P450 Metabolites Based on Predicted Sites of Metabolism

Christina de Bruyn Kops[1†], *Conrad Stork*[1†], *Martin Šícho*[1,2†], *Nikolay Kochev*[3,4†], *Daniel Svozil*[2†], *Nina Jeliazkova*[3†] *and Johannes Kirchmair*[1,5,6*]

[1] Department of Computer Science, Center for Bioinformatics (ZBH), Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, Hamburg, Germany, [2] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, Prague, Czechia, [3] Ideaconsult Ltd., Sofia, Bulgaria, [4] Department of Analytical Chemistry and Computer Chemistry, University of Plovdiv, Plovdiv, Bulgaria, [5] Department of Chemistry, University of Bergen, Bergen, Norway, [6] Computational Biology Unit (CBU), University of Bergen, Bergen, Norway

Computational prediction of xenobiotic metabolism can provide valuable information to guide the development of drugs, cosmetics, agrochemicals, and other chemical entities. We have previously developed FAME 2, an effective tool for predicting sites of metabolism (SoMs). In this work, we focus on the prediction of the chemical structures of metabolites, in particular metabolites of xenobiotics. To this end, we have developed a new tool, GLORY, which combines SoM prediction with FAME 2 and a new collection of rules for metabolic reactions mediated by the cytochrome P450 enzyme family. GLORY has two modes: MaxEfficiency and MaxCoverage. For MaxEfficiency mode, the use of predicted SoMs to restrict the locations in the molecule at which the reaction rules could be applied was explored. For MaxCoverage mode, the predicted SoM probabilities were instead used to develop a new scoring approach for the predicted metabolites. With this scoring approach, GLORY achieves a recall of 0.83 and can predict at least one known metabolite within the top three ranked positions for 76% of the molecules of a new, manually curated test set. GLORY is freely available as a web server at https://acm.zbh.uni-hamburg.de/glory/, and the datasets and reaction rules are provided in the **Supplementary Material**.

Keywords: metabolism prediction, metabolite structure prediction, rule-based approach, sites of metabolism, xenobiotic metabolism, cytochrome P450, metabolites

OPEN ACCESS

Edited by:
Jose L. Medina-Franco, National Autonomous University of Mexico, Mexico

Reviewed by:
Anastasia Rudik, Institute of Biomedical Chemistry, Russian Academy of Medical Sciences (RAMS), Russia
Angelica Mazzolari, University of Milan, Italy

***Correspondence:**
Johannes Kirchmair
johannes.kirchmair@uib.no
orcid.org/0000-0003-2667-5877

†Christina de Bruyn Kops
orcid.org/0000-0001-8890-2137
Conrad Stork
orcid.org/0000-0002-5499-742X
Martin Šícho
orcid.org/0000-0002-8771-1731
Nikolay Kochev
orcid.org/0000-0001-6547-3675
Daniel Svozil
orcid.org/0000-0003-2577-5163
Nina Jeliazkova
orcid.org/0000-0002-4322-6179

Specialty section:
This article was submitted to Medicinal and Pharmaceutical Chemistry, a section of the journal Frontiers in Chemistry

Received: *28 March 2019*
Accepted: *17 May 2019*
Published: *12 June 2019*

Citation:
de Bruyn Kops C, Stork C, Šícho M, Kochev N, Svozil D, Jeliazkova N and Kirchmair J (2019) GLORY: Generator of the Structures of Likely Cytochrome P450 Metabolites Based on Predicted Sites of Metabolism. Front. Chem. 7:402. doi: 10.3389/fchem.2019.00402

INTRODUCTION

Metabolism is responsible for creating metabolites with different physicochemical and pharmacological properties compared to those of the original parent molecule. Xenobiotic metabolism in particular is directly relevant for humans, especially as it relates to, for example, the development of drugs, cosmetics, and agrochemicals. In fact, it is supposed that metabolism is the main clearance pathway for the vast majority of all xenobiotics (Kirchmair et al., 2015). However, metabolism can also result in pharmacologically active metabolites as well as toxic metabolites (Testa et al., 2012).

The cytochrome P450 (CYP) family of enzymes plays an important role in the metabolism of xenobiotics, especially in the formation of first-generation metabolites, of which roughly 60% are formed by CYPs (Testa et al., 2012). The importance of CYPs to drug discovery is clear from the observation that many drugs are metabolized by CYPs; common estimates range from 50% (Di, 2014) to 80% (Testa et al., 2012). A detailed meta-analysis of the metabolites of over 1,000 different xenobiotic substrates carried out by Testa et al., showed that 40% of all metabolites are formed by CYPs, including a substantial proportion of all toxic or highly reactive metabolites (Testa et al., 2012).

There are 57 known human CYP enzymes, the majority of which are primarily involved in endogenous metabolism. The CYP2 and CYP3 subfamilies are mainly responsible for metabolizing xenobiotics (Testa et al., 2012), and the key CYP isozymes for drug metabolism are CYP3A4, 3A5, 2D6, 2C8, 2C9, 2C19, 1A1, 2B6, and 2E1 (Di, 2014). Among the xenobiotic-metabolizing CYP isozymes, the binding pockets vary greatly; in some cases the binding pocket of a single isozyme is highly flexible and can accommodate a broad range of substrates with widely varying sizes (Kirchmair et al., 2015).

Computational methods can make a significant contribution to predicting xenobiotic metabolism, because they can be used to quickly make predictions that can focus the experimental aspects of the drug development process. Such a focusing effect is both cost-effective and time-effective (Kirchmair et al., 2015).

One relatively well-developed aspect of the computational prediction of xenobiotic metabolism is the identification of the metabolically labile atom positions, also known as sites of metabolism (SoMs) (Kirchmair et al., 2012). Being able to predict SoMs is important because knowing an atom position in a molecule at which a metabolizing reaction is likely to occur usually provides a chemist with a good idea of the ensuing metabolite structure. Besides a range of commercial offerings, several freely available tools, such as SMARTCyp (Olsen et al., 2019), SOMP (Rudik et al., 2015), Xenosite (Zaretzki et al., 2013), and FAME 2 (Šícho et al., 2017), are able to predict SoMs with high accuracy (Tyzack and Kirchmair, 2018). FAME 2, which is used in the present work for SoM prediction, is a machine learning-based tool developed recently in our group. The extra trees classifier models of FAME 2, which are based on a set of 2D circular descriptors, were developed specifically to predict SoMs of metabolic reactions catalyzed by the CYP family of enzymes in humans. FAME 2 is highly accurate, achieving, on an independent test set, a Matthews correlation coefficient of 0.57 and an area under the receiver operating characteristic curve (AUC) of 0.91.

In contrast to *in silico* SoM prediction, computational prediction of the structures of metabolites lags behind with respect to prediction accuracy. In general, existing methods for predicting metabolite structures for xenobiotics are dominated by rule-based approaches. There are a number of well-established commercial tools for metabolite structure prediction, including Meteor Nexus (Lhasa Ltd.), a rule-based metabolite prediction software (Marchant et al., 2008). Meteor Nexus offers three different reasoning methods to prioritize the plethora of generated metabolites. The current default reasoning method is SoM scoring, which compares the SoM identified by the reaction rule to experimental data in order to assign scores to the predicted metabolites[1]. Other rule-based computational tools include TIMES (LMC; Mekenyan et al., 2004), which uses a heuristic algorithm to generate possible metabolic maps, and MetabolExpert (CompuDrug; Darvas, 1987).

In addition to commercial metabolite structure prediction tools, there is an increasing number of freely available options. Again, many of the available options rely primarily on a set of reaction rules to generate structures of possible metabolites. One well-known approach that has been around for some time is SyGMa (Ridder and Wagener, 2008), which in this work is used as a reference method. SyGMa predicts metabolites using knowledge-based reaction rules, some of which were derived from common knowledge of metabolism reactions and some of which were developed using the Metabolite Database (MDL Metabolite Database, Elsevier, 2001), for a total of 144 reaction rules covering both phase I and phase II metabolism. The predicted metabolites are ranked by empirical probability scores calculated based on the fraction of predicted metabolites produced by the particular reaction rule that match reported metabolites in the database. Using all 144 phase I and phase II reaction rules in up to three successive reaction steps, SyGMa was able to predict 68% of all known metabolites in the test set. In terms of ranking, SyGMa ranked 45% of the known metabolites in the test set in the top 10. The authors additionally examined SyGMa's potential usefulness for predicting CYP-mediated metabolism by evaluating its performance on a set of 127 single-step CYP-mediated reactions. Using only the 118 phase I reaction rules, which include but are not specific to CYP-mediated reactions, SyMGa was able to predict 84% of all known CYP-formed metabolites and predict 66% of the known metabolites within the top three ranked predicted metabolites. However, the proprietary nature of the dataset that was used to derive SyGMa's reaction rules and validate the method, not to mention the current unavailability of the dataset, hinders the reproducibility of the results as well as further use of the models derived from the data.

A recent, free software designed to predict metabolites from multiple sources and enzyme families is BioTransformer (Djoumbou-Feunang et al., 2019), which in this work is used as the second reference method. BioTransformer is a comprehensive metabolite prediction tool that contains a CYP metabolite prediction module (in addition to four other metabolite prediction modules). BioTransformer predicts CYP-formed metabolites using a knowledge-based approach combined with built-in CYP selectivity prediction by CypReact (Tian et al., 2018), a machine learning-based tool, as a precursor to metabolite prediction. Aside from the initial CYP isoform-specificity prediction, the basis of BioTransformer's CYP450 metabolite prediction module is a rule-based method whose

Abbreviations: AUC, area under the receiver operating characteristic curve; CYP, cytochrome P450; ROC, receiver operating characteristic; SoM, site of metabolism.

[1]Meteor Reasoning Methodologies, Lhasa Limited, https://www.lhasalimited.org/products/meteor-reasoning-methodologies.htm

reaction rules are derived partly from the metabolic reactions in MetXBioDB (Djoumbou-Feunang et al., 2019), a freely available database of metabolism reactions that was established in the context of developing BioTransformer. In the current version of BioTransformer, the predicted metabolites are not ranked. BioTransformer also offers an option for identifying metabolites based on masses from mass spectrometry data. On a test dataset of 60 parent molecules with a total of 180 known metabolites, BioTransformer's CYP450 metabolite prediction module achieved a recall of 0.90 and a precision of 0.46.

Another freely available metabolite prediction tool is MetaTox (Rudik et al., 2017), which encompasses both phase I and phase II metabolism and combines the prediction of the reaction class and the reacting atom in order to predict metabolites. Additionally, the open-source software Toxtree (Patlewicz et al., 2008) contains a metabolism prediction module called "SMARTCyp—Cytochrome P450-Mediated Drug Metabolism" that predicts SoMs using SMARTCyp (Rydberg et al., 2010) and then applies a small set of reaction rules to the predicted SoMs in order to predict metabolites.

Common to all modern approaches for metabolite prediction is that they remain challenged by the combinatorial explosion of predictions, in particular when looking at several generations of metabolites (Judson, 2014). It is not unusual for metabolite structure predictors to produce several pages full of predicted metabolites, a fact which is often and not without reason criticized, particularly by experts in metabolism. The key to tackling this problem lies in the development of approaches for the accurate ranking of metabolites according to their relevance in terms of metabolic rates and biological properties. A number of methods attempt to get a handle on the immense number of predicted metabolites by ranking their predictions according to various approaches.

Another option, which has primarily been implemented in commercial tools to date, is to use SoM prediction as a preliminary step to reduce the number of generated metabolites. Commercial tools for metabolite prediction that incorporate SoM prediction include ADMET Predictor (SimulationsPlus)[2], which predicts SoMs and the corresponding metabolite structures for nine CYP isoforms, and StarDrop (Optibrium; Tyzack et al., 2016), whose "P450 metabolism" module predicts SoMs using quantum mechanical simulations and displays the structures of the metabolites corresponding to the predicted SoMs. In addition, META Ultra (MultiCASE Inc.; Klopman et al., 1994) predicts SoMs and metabolites, and MetaSite (Cruciani et al., 2005) was a SoM and CYP isoform selectivity prediction software that now also predicts metabolite structures[3].

Few freely available metabolite prediction methods combine information on predicted SoMs with a rule set. MetaTox predicts reaction classes and reacting atoms (i.e., SoMs, in principle) separately for each parent molecule, then combines the predictions to generate metabolites. The probability that the metabolite is formed is calculated based on the predicted probabilities of the reaction class and of the SoM predicted with the SOMP method (Rudik et al., 2015). However, the validation of MetaTox considers the performance of the reaction class prediction and the reacting atom prediction separately, without evaluating the prediction of the metabolite structures themselves, and it is unclear how exactly the reaction class and reacting atom predictions are combined to generate a metabolite structure (Rudik et al., 2017). On the other hand, it is clear that SoM prediction is used directly as a prefilter before applying reaction rules in the SMARTCyp Toxtree module. However, a validation of this method has not been published.

In terms of the availability of rule sets for metabolite structure prediction, there are a few existing freely available collections of reaction rules described in an easily accessible, computer-readable format such as SMIRKS[4], a reaction transform language within the Daylight system. One source of CYP reaction rules is the SMARTCyp Toxtree module, which uses 16 reaction rules and makes the SMIRKS freely available as part of the source code. A larger selection of reaction rules is provided in the freely available SyGMA Python package. The reaction rules are clearly separated into phase I and phase II rules; however, there is no indication of which of the 118 phase I reaction rules specifically describe CYP-mediated reactions. In addition, these rules were derived from a proprietary and no longer distributed dataset. BioTransformer offers a large number of CYP-specific biotransformation rules in SMIRKS format as well as additional constraint(s) for each rule as part of its Reaction Knowledgebase.

In this work, we present a multipronged approach to the prediction of metabolites formed by the CYP enzyme family in humans. In reference to FAME, we name this approach GLORY. One fundamental aspect of GLORY is a new, easily interpretable rule base for CYP metabolism that was developed solely from the scientific literature and basic chemistry knowledge, without relying on any dataset of metabolic reactions. In addition, we have examined the effect of using SoM prediction as a preliminary filter for the positions at which reaction rules are allowed to be applied and also as part of a new approach to ranking the predicted metabolites. GLORY therefore has two modes: MaxCoverage, which focuses solely on recall, and MaxEfficiency, which focuses more on precision. Further, we have validated GLORY on a new, high quality, manually curated dataset that is provided in the **Supplementary Material**.

RESULTS AND DISCUSSION

Two key aspects are at the core of GLORY, which aims to predict metabolites within the context of human, CYP-mediated metabolism: reaction rules and predicted SoMs. In terms of the rule-based aspect, GLORY uses reaction rules to convert parent molecules into their possible metabolites. To this end, we developed a collection of rules based entirely on the scientific literature to ensure that the rule set was not biased by any particular metabolism dataset. The information on the CYP-mediated reactions from the literature was combined with

[2]ADMET Predictor Metabolism Module, SimulationsPlus, https://www.simulations-plus.com/software/admetpredictor/metabolism/

[3]MetaSite, https://www.moldiscovery.com/software/metasite/

[4]SMIRKS—A Reaction Transform Language, Daylight, http://www.daylight.com/dayhtml/doc/theory/theory.smirks.html

basic chemistry knowledge to develop SMIRKS to describe each reaction type. In some cases, such as for O-dearylation, multiple SMIRKS were required for a single reaction type, resulting in a total of 73 SMIRKS for the 61 reaction types present in our collection (**Supplementary Table 1**). We additionally use a simple binary distinction between common and uncommon reaction types, which were thoroughly discussed and distinguished from each other in Guengerich (2001), and which distinction we were able to extrapolate to the CYP-mediated reactions found elsewhere in the literature (see Methods for details). We do not use occurrence ratios calculated based on a given dataset in order to rank the predicted metabolites, due to the limited size, quality, and accessibility of existing datasets. Out of our collection of 61 CYP reaction types, 22 have been designated as common.

The second key aspect of GLORY is its use of the SoM probabilities predicted by FAME 2 for each heavy atom in a molecule to (i) reduce the false-positive prediction rate while maintaining an acceptable recovery rate and (ii) augment the ranking of predicted metabolites. In order to reduce the false-positive prediction rate, the possibility of utilizing a hard cutoff based on SoM probabilities was explored. This cutoff was used to determine at which atom positions the rules were allowed to be applied. In the context of GLORY, we have called this approach, in which SoM prediction is used as a preliminary filter, MaxEfficiency mode. In contrast, we designate the approach in which SoM probabilities are used for ranking metabolites derived for all positions in a molecule regardless of SoM probability the MaxCoverage mode. The difference in workflow between the two modes is illustrated in **Figure 1**.

Datasets

To choose a SoM probability cutoff for the MaxEfficiency mode and develop a priority score to rank predicted metabolites, a large reference dataset was generated by combining the CYP metabolism data extracted from DrugBank (Wishart et al., 2018) and MetXBioDB. MetXBioDB is a recently published database of metabolic reactions, whose substrates are mainly comprised of xenobiotics and also include a few sterol lipids and mammalian primary metabolites, and whose reaction data came from the scientific literature as well as publicly available databases (Djoumbou-Feunang et al., 2019). In addition, a manually curated, high-quality dataset was compiled from the scientific literature for the validation of GLORY. This test dataset contains 29 parent molecules and a total of 81 metabolites, resulting in 2.79 metabolites per parent molecule on average. Importantly, any parent compounds that are in the test dataset were removed from the reference dataset before any analysis occurred. In total, the reference dataset contains 848 parent molecules and a total of 1,588 metabolites, for an average of 1.87 metabolites per parent molecule. Predictions could be made for 847 of 848 molecules in the reference dataset (one molecule could not be processed successfully with FAME 2; see Methods for details).

MaxEfficiency Mode: Selection of a Cutoff for Metabolite Structure Generation Based on SoM Probability

In order to determine the effect of a SoM prediction-based prefilter on predicting preferably only the most relevant metabolites and reducing the number of false positive predictions, we tried several different cutoffs for the SoM probability that must be achieved by at least one atom involved in the reaction (as defined by the reaction's SMIRKS). For each heavy atom in a molecule, FAME 2 reports a probability between 0 and 1, corresponding to the fraction of trees of the extra trees classifier that predict that a particular atom is a SoM. The decision threshold in FAME 2 for whether or not an atom is considered likely enough to be a SoM to be designated as such was determined by the trained model to be 0.4 (Šícho et al., 2017).

We examined the effect of different SoM probability cutoffs using the reference dataset and selected the cutoff to be used

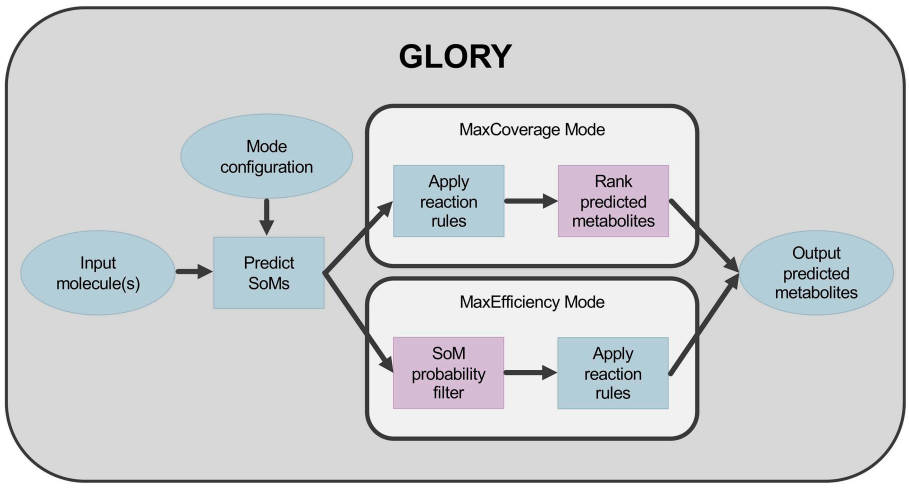


FIGURE 1 | Workflow for GLORY indicating the difference between MaxCoverage mode and MaxEfficiency mode.

in MaxEfficiency mode based on these results. In particular, we inspected the effect of the SoM probability cutoffs on precision and recall, which are defined as follows:

$$Recall = TP / (TP + FN)$$
$$Precision = TP / (TP + FP)$$

where TP is the number of true positive predictions, FP is the number of (putative) false positive predictions, and FN is the number of false negative predictions. In other words, recall measures the portion of known metabolites that were reproduced by the method and precision measures the fraction of all predicted metabolites that are represented in the dataset.

Here it is worth noting that the number of false positives, and the designation of a prediction as false positive, is especially dependent on the dataset that is being used for comparison. Many metabolites that are formed in humans have not yet been discovered, or their structures have not yet been exactly elucidated. Since even the highest-quality dataset is limited by the available experimental data, the reality is that the distinction between a real false positive prediction and the true positive prediction of an as yet unknown or unconformed metabolite may not be possible. Nevertheless, with this caveat, we evaluate our method based on the available data, including the putative false positives.

The purpose of the MaxEfficiency mode is to use the SoM probability cutoff to predict metabolites with increased precision compared to no cutoff (i.e., MaxCoverage mode). At the same time, however, we did not want to sacrifice too much in terms of recall, as it is still important to predict a molecule's actual metabolites even while reducing the number of putative false positive predictions.

For the purpose of metabolite prediction, we found that using FAME 2's decision threshold of 0.4 as the cutoff for SoM probability resulted in a relatively low recall of 0.65 (especially when compared to the recall of 0.83 achieved in MaxCoverage mode, as will be discussed later in this work). Hence, despite the increased precision afforded by a cutoff of 0.4, it was determined that this cutoff too greatly reduced the achieved recall. We therefore additionally tested lower SoM probability cutoffs (**Table 1**). Observing the trade-off between precision and recall with cutoffs ranging from 0.4 to 0.1 and comparing them to MaxCoverage mode, we determined that a SoM probability cutoff of 0.2, which resulted in a precision of 0.19 and a still-high recall of 0.75, offered the best compromise. A SoM probability cutoff of 0.2 for MaxEfficiency mode was therefore fixed based on the results shown in this section. Note that although all of the precision values shown in **Table 1** are quite low, the precision of GLORY using a SoM probability cutoff is comparable to the precision of existing methods for metabolite structure prediction (see below for the results on the test dataset).

TABLE 1 | Effect of different SoM probability cutoffs on precision and recall over the entire reference dataset.

SoM Probability Cutoff[a]	0.4	0.3	0.2	0.1	None
Precision	0.24	0.22	0.19	0.13	0.07
Recall	0.65	0.71	0.75	0.80	0.83

[a]*Note that 0.4 is the default decision threshold in FAME 2, a cutoff of none corresponds to MaxCoverage mode, and a cutoff of 0.2 was chosen for MaxEfficiency mode.*

Development of a Priority Score to Rank Predicted Metabolites for MaxCoverage Mode

In order to rank the predicted metabolites for a particular molecule, we developed a priority score for each predicted metabolite based on the SoM probability of the atoms involved in the transformation and whether the reaction type is common or not. Specifically, the SoM probability calculated by FAME 2 for all atoms in the parent molecule that are involved in a reaction as defined by the SMIRKS is considered, and the maximum SoM probability among these atoms is then incorporated into the score, as illustrated in **Figure 2**. The priority score was calculated using a simple formula:

$$score_{\text{predictedmetabolite}} = P \times F$$

where P is the maximum SoM probability out of the atoms in the parent molecule that were matched by the applied transformation and F is the factor according to whether the reaction type was designated as common or uncommon. In case the same predicted metabolite resulted from multiple transformations, the maximum priority score over all transformations leading to that prediction was used. A higher priority score is intended to indicate a higher likelihood of the prediction being true. For all uncommon reaction types, $F = 1$. The factor F for common reaction types affects the early enrichment of the predictions. Specifically, the early enrichment improves when common reaction types are given more weight in the score than uncommon reaction types, i.e. $F_{common} > 1$ (**Figure 3**). Based on an analysis of the receiver operating characteristic (ROC) curves and area under the ROC curves (AUC) for varying F_{common}, shown in **Figure 3**, a factor of 5, resulting in an AUC of 0.90, was chosen. All subsequent results based on ranking the predicted metabolites therefore used $F_{common} = 5$ in the calculation of the priority score, and the priority score can therefore range from 0 to 5.

Comparison of Performance on a New, Manually Curated Test Set

The performance of the MaxEfficiency and MaxCoverage modes of GLORY was evaluated on the curated test set of 29 parent molecules with a total of 81 metabolites. This evaluation includes a comparison with BioTransformer and SyGMa as well as an analysis of how well the scoring and ranking aspects of the different approaches work. Specifically, we employed the CYP450 module of BioTransformer and the phase I metabolism reactions of SyGMa (SyGMa does not feature a dedicated module for CYP metabolism, but phase I metabolism is carried out to a significant extent by CYP enzymes) for the comparison.

Analysis of MaxEfficiency Mode

GLORY's MaxEfficiency mode was designed to address the problem of low precision caused by a high number of putative

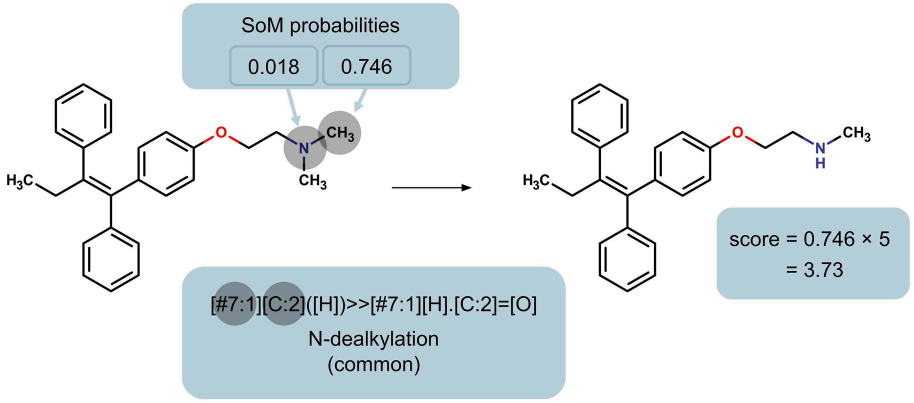


FIGURE 2 | Illustration of the determination of the maximum SoM probability of all heavy atoms in the parent molecule that are matched by the reaction rule, using the N-dealkylation reaction rule (common reaction type; factor $F = 5$) as an example. This maximum probability is used to calculate the priority score of the product.

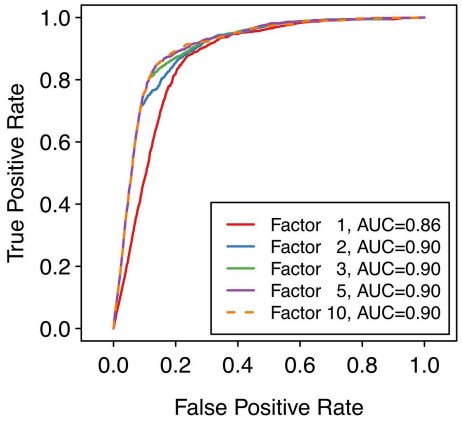


FIGURE 3 | Receiver operating characteristic (ROC) curves over the entire reference dataset of 848 compounds with 1,588 known metabolites, with varying values of the factor used for common reaction types when calculating the priority score for each metabolite. Note that a factor of 1 means that only the SoM probability (i.e., the maximum SoM probability for all atoms that are matched by the SMIRKS) affects the priority score of the predicted metabolite, regardless of the reaction type. Note also that a ROC curve can be calculated despite there being no "true negative" predictions overall (all predicted metabolites are "positive" predictions). To generate the ROC curve, the false positive rate (FPR) is calculated at each score threshold. At each point, predictions with scores below the threshold are considered "negative" predictions and predictions with scores above the threshold are considered "positive" predictions. Hence the number of "true negative" predictions and therefore the FPR can be calculated for each point of the ROC curve.

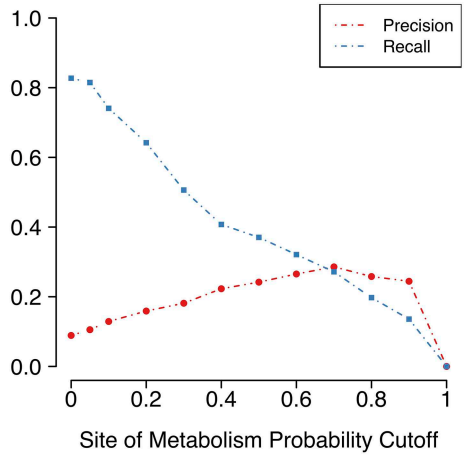


FIGURE 4 | Precision (portion of predictions that are true positives) and recall (portion of known metabolites that are predicted) vary according to the cutoff for FAME 2's predicted SoM probability. A SoM probability cutoff of 0.4 corresponds to the decision threshold used in FAME 2. The SoM probability cutoff chosen for the MaxEfficiency mode of GLORY was 0.2.

false positive metabolite predictions. This general problem of an excess of predictions is well-documented for metabolite prediction tools (Judson, 2014). However, as mentioned above, it is important to note that the designation of predictions as false positive is particularly dataset-dependent.

As described previously, the MaxEfficiency mode uses a cutoff based on the SoM probabilities that FAME 2 predicts for each heavy atom in order to restrict the locations in the molecule at which the reaction rules are allowed to be applied. This SoM probability cutoff was set to 0.2 based on the analysis on the reference dataset; however, we also examine the effect of different SoM probability cutoffs using the high-quality test dataset in order to get a more complete picture of how much can be gained by a cutoff-based approach.

As expected, using SoM predictions to confine the application of reaction rules to certain positions does involve a trade-off between precision and recall (**Figure 4**). Recall measures the portion of known metabolites that the method was able to reproduce, and precision measures the fraction of all predicted metabolites that are actually known metabolites (see previous section for definitions). The larger the SoM probability required to be present among the atoms involved in the transformation, the lower the recall but the higher the precision as measured across the entire test dataset. In addition, the larger the SoM probability cutoff, the more parent molecules there are for which

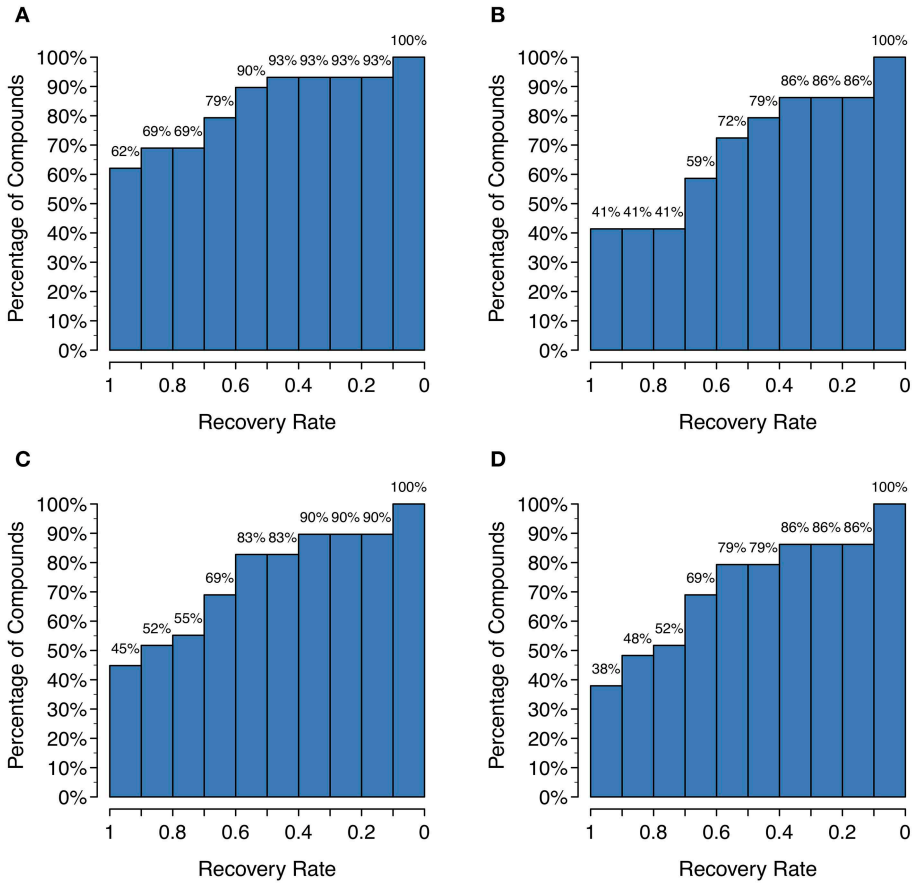


FIGURE 5 | Histograms of the recovery rate of known metabolites broken down by parent compound: **(A)** GLORY in MaxCoverage mode, **(B)** GLORY in MaxEfficiency mode, **(C)** SyGMa, **(D)** BioTransformer. For example, a recovery rate of 0.5 indicates that for x% of all parent molecules, at least half of all recorded metabolites from the test dataset were predicted.

no metabolite predictions can be made. Without any such cutoff and even up to a SoM probability cutoff of 0.2, metabolites can be predicted for all parent molecules in the test dataset. However, with a SoM probability cutoff of 0.3, no metabolites are predicted for two parent molecules, and this number increases to three for a cutoff of 0.4 (**Supplementary Table 2**). The number of molecules affected is small in this case, yet is approximately 10% of the size of the test dataset. Overall, as the cutoff increases, the total number of predicted metabolites decreases drastically (**Supplementary Table 2**).

Unfortunately, as **Figure 4** shows, there is a large decrease in recall for a small increase in precision when using SoM probability cutoffs of 0.1 or greater. Looking more closely at the recovery rates per parent molecule, we see that GLORY's MaxEfficiency mode (using the selected cutoff of 0.2 as described above) can predict at least half of the known metabolites for 72% of the parent molecules in the test dataset, as opposed to 83% for SyGMa and 79% for BioTransformer (**Figure 5**). GLORY's MaxEfficiency mode can predict all known metabolites for 41% of the parent molecules in the test dataset, as opposed to 45% for SyGMa and 38% for BioTransformer. On the other hand, the number of putative false positives per parent molecule is brought to within the same range as was measured for SyGMa and BioTransformer (**Figure 6**). Using MaxEfficiency mode, most parent molecules have fewer than 10 putative false positives, which is also the case for BioTransformer but not quite the case for SyGMa (however, as mentioned above, SyGMa's rule base also includes rules for non-CYP-mediated phase I reactions).

Based on these results, it appears that using FAME 2's predicted SoM probabilities as a hard cutoff for metabolite prediction may not be sufficient for many use cases. However, the SoM predictions are useful for more than just as a hard cutoff, namely to rank the predicted metabolites, as will be shown in the next section.

Comparison of MaxCoverage Mode to SyGMa and BioTransformer

Neither SyGMa nor BioTransformer uses regioselectivity prediction as a prefilter before applying reaction rules. The same is true of MaxCoverage mode, which only uses SoM prediction in order to score and rank the predicted metabolites. Hence, we compared SyGMa and BioTransformer to GLORY's

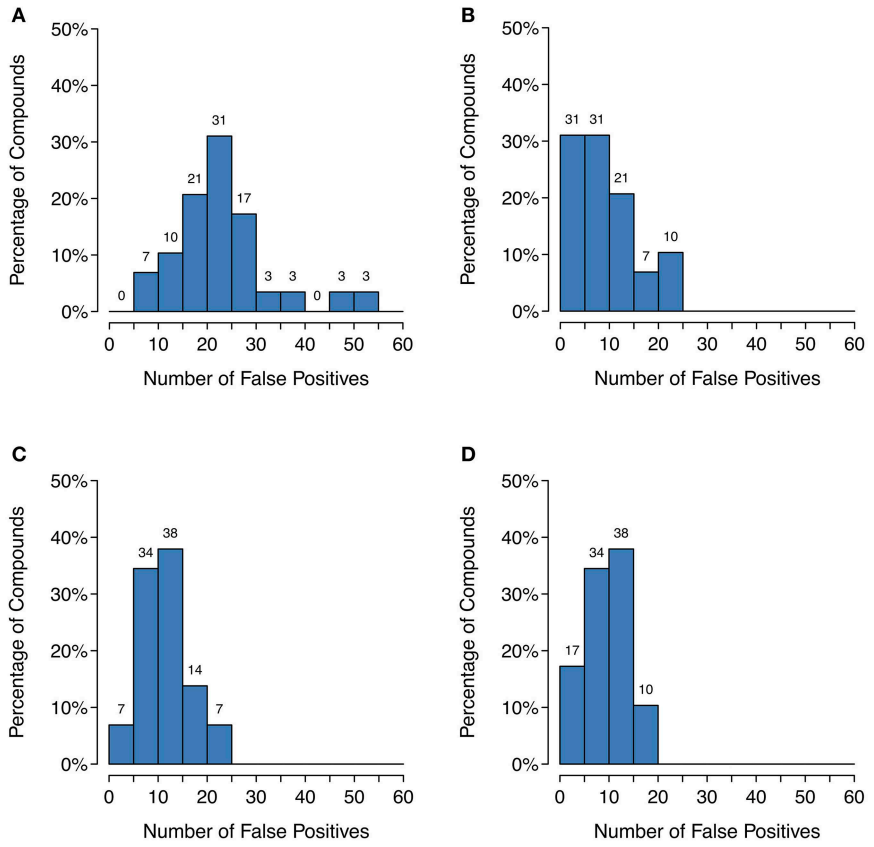


FIGURE 6 | Histograms of the number of putative false positive predictions: **(A)** GLORY in MaxCoverage mode, **(B)** GLORY in MaxEfficiency mode, **(C)** SyGMa, **(D)** BioTransformer. These histograms use right-closed intervals.

MaxCoverage mode in terms of recall, precision, and ability to rank the predicted metabolites.

A high recall is important for any use case of a metabolite structure predictor, but even more so for applications in which it is of utmost importance to not miss any physically existing metabolites, such as, for example, when attempting to identify metabolites based on MS data. GLORY's MaxCoverage mode performs well in terms of recall, with a recall of 0.83 compared to 0.74 and 0.72 for SyGMa and BioTransformer, respectively, across the entire test dataset (**Table 2**). A closer look at recall broken down to the level of the recovery rate of known metabolites for each parent molecule shows that GLORY is able to predict all known metabolites for 62% of the parent molecules, whereas SyGMa and BioTransformer achieve only 45% and 38%, respectively, in this regard (**Figure 5**). The number of parent molecules for which GLORY is able to predict at least half of the known metabolites is 90%, compared to 83% for SyGMa and 79% for BioTransformer (**Figure 5**).

Precision can be a useful metric for measuring how well a method is able to keep the number of putative false positive predictions under control. Precision was low across the board for metabolite prediction on the test dataset, with BioTransformer reaching the highest precision of the three tools at 0.17. SyGMa was close behind at 0.15, and GLORY's MaxCoverage mode lagged further behind at a precision of only 0.08 (**Table 2**). Again breaking this down to a slightly more detailed overview, we see that BioTransformer and SyGMa both always produce fewer than 25 putative false positives per parent molecule and, for the majority of parent molecules, fewer than 15 putative false positives or even, in the case of BioTransformer, fewer than 10 (**Figure 6**). GLORY in MaxCoverage mode, on the other hand, often produces so many predictions per parent molecule that there are up to 53 putative false positives per parent molecule in the test dataset and on average a relatively high number of putative false positive predictions compared to the other two tools (**Figure 6**).

In the case of the low precision observed for SyGMa, it is important to note that SyGMa's rule set is not specific to CYP-mediated metabolism but rather covers phase I metabolism in general. This could indicate that SyGMa might achieve higher precision if only the CYP-specific rules were used.

BioTransformer's CYP450 prediction module, which has the highest precision of all three methods, uses isoform prediction as a preliminary filter. Only the relevant reactions for the predicted metabolizing CYP isoform(s) are applied to the parent molecule, which could contribute to the observed precision.

Although the precision of MaxCoverage mode (as well as SyGMa and BioTransformer) was found to be low and high rates

TABLE 2 | Evaluation results for SyGMa, BioTransformer, and GLORY's MaxCoverage and MaxEfficiency modes on the manually curated test dataset.

	SyGMa	BioTransformer	GLORY, MaxCoverage mode	GLORY, MaxEfficiency mode[b]
Precision	0.15	0.17	0.08	0.16
Recall	0.74	0.72	0.83	0.64
Total number of predicted metabolites	406	344	793	327
Number of successfully predicted reported metabolites[a]	60	58	67	52
Top-1	0%	N/A	68.97%	68.97%
Top-2	48.28%	N/A	72.41%	72.41%
Top-3	68.97%	N/A	75.86%	75.86%

[a] *The total number of reported metabolites in the dataset was 81.*

[b] *The SoM probability cutoff used for MaxEfficiency mode is 0.2, chosen based on the results of the analysis on the reference dataset. Data on the performance of MaxEfficiency mode with different SoM probability cutoffs are reported in* ***Supplementary Table 2****.*

of false positive predictions are problematic in general, in the case of metabolite structure predictors a low precision is only problematic if there is no way to distinguish between the true positive and putative false positive predicted metabolites. This distinction can be achieved with a well-working ranking of the predicted metabolites, which circumvents the need to reduce the total number of predicted metabolites. Hence it is important that a metabolite prediction tool can rank the predicted metabolites in terms of likelihood of occurrence.

GLORY scores its predicted metabolites based partly on the maximum SoM probability of all the atoms involved in the reaction and also takes the type of reaction into account (see above for a more detailed description of the priority score). SyGMa uses empirical probability scores calculated based on the percentage of all predictions for each reaction rule that are found in the training dataset. SyGMa's scoring system thereby relies entirely on the discontinued Metabolite dataset. The scores generated by GLORY or by SyGMa can be used to rank the predicted metabolites for a given parent compound in terms of their likelihood of occurring. The current version of BioTransformer, on the other hand, does not score or rank its predictions.

We compared the ranking capability of GLORY's MaxCoverage mode with that of SyGMa. SyGMa was able to predict a known metabolite within the top three ranked positions for 69% of the parent molecules in the test dataset, whereas GLORY's MaxCoverage mode predicted a known metabolite within the top three predictions for 76% of the parent molecules (**Table 2**).

To look at the overall quality of the scoring as well as the ranking ability of SyGMa compared to GLORY, we generated ROC curves for each method using the score of each predicted metabolite as well as the rank of each predicted metabolite for a given molecule. The rank-based analysis corresponds better to the actual use case, in which it is desired to prioritize the predicted metabolites for a particular parent molecule, as opposed to over an entire dataset [note that SyGMa was originally only evaluated in terms of ranking per parent molecule (Ridder and Wagener, 2008)]. However, we additionally used the score-based ROC curve to visualize the performance of GLORY's priority score across the whole test dataset. To better allow for comparison of the ROC curves, false negatives were included in the ROC curves and thereby in the calculated AUCs by adding those molecules to the set of data points and artificially assigning them a score of 0 or rank of 1,000, as applicable, for the purpose of this evaluation.

Though the AUC values are low, due in part to the inclusion of false negative data points in the ROC curves, the ROC curves show a much better earlier enrichment for GLORY than for SyGMa (**Figure 7**). SyGMa does not rank a known metabolite in the best-ranked position for any parent molecule in the test dataset (**Table 2**), which is reflected in the ROC curve. This decent early enrichment with GLORY, which is corroborated by the top-3 value, is a highly encouraging result indicating that the most likely predictions are closer to the top of the ranked list than the putative false positive predictions are.

One possible explanation for why SyGMa performs poorly in terms of scoring could be that its scoring scheme was derived from occurrence ratios in the Metabolite database and therefore optimized to predict the metabolites in that particular dataset. Although the Metabolite database was large, the authors of SyGMa report that the database was nevertheless biased toward compounds with one known metabolite and postulate that many of the metabolite profiles were incomplete (Ridder and Wagener, 2008). Our manually curated test dataset consists of parent molecules with metabolites that have been published since 2014, while SyGMa was developed using the 2001 version of Metabolite, so we assume that the overlap, if any, between SyGMa's training dataset and our test dataset is low. Without access to the dataset that was used to develop SyGMa's scoring methodology, it remains unclear how well the types of the reactions that lead to the metabolites in the test dataset were represented in their training dataset. Related to that, an additional downside of SyGMa's approach of basing their scoring approach on a database of metabolic reactions is that, since reaction rules can only be included if the database contains enough examples of a specific reaction type to calculate

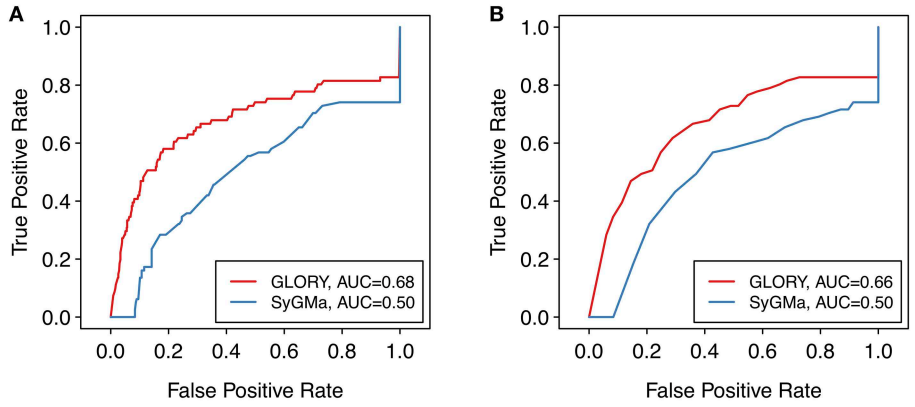


FIGURE 7 | ROC curves over the entire test dataset comparing the **(A)** scoring and **(B)** ranking approaches of SyGMa to GLORY's MaxCoverage mode. For a better comparison of the two methods, false negatives were included in the ROC curve by assigning those data points a score of 0 or rank of 1,000, as applicable.

a probability score, more unusual reaction types or reaction types that are for some reason not well enough represented in the database may be missing from SyGMa's rule base (Ridder and Wagener, 2008).

There are several other differences in methodology between GLORY and SyGMa that could contribute to the difference in performance. Firstly, SyGMa does not specifically predict CYP-mediated metabolism but rather phase I metabolism in general, meaning that it could predict other phase I metabolites that are simply not present in the test dataset because they are not formed by CYPs. Second, in the current Python package implementation that was used for this validation, SyGMa does not appear to require its predicted metabolites to have a certain minimum size. Unlike GLORY, which does not output a potential metabolite if it has fewer than three heavy atoms, SyGMa predicts a handful of metabolites (across the whole test dataset) with only one or two heavy atoms.

Computation Time

The run time for GLORY was measured on a workstation equipped with eight Intel(R) Core(TM) i7-4790 CPUs, 32 GB of main memory, and a Linux operating system. For the test dataset, the total run time (using eight cores) was 4.6 min in MaxCoverage mode and 4.3 min in MaxEfficiency mode (averaged over three runs). On average, the computation time per molecule required to predict metabolites was 10.9 s for MaxCoverage mode and 10.3 for MaxEfficiency mode (averaged over three runs).

METHODS

Development of a Collection of Transformations

A collection of transformations, defined by SMIRKS and representing reaction types, was assembled based on known CYP-mediated reactions found in the literature (see **Supplementary Material** for details). The SMIRKS were defined to be as general as possible while being restricted to reasonable reaction chemistry, as indicated by the literature and common chemical knowledge. Therefore, if a reaction was found in the literature but it was not clear how the reaction would apply to other molecules besides the provided example, the reaction was excluded from the collection. This was the case for most reactions involving large ring systems as well as ring fusions and ring contractions. Specifically, the following types of reactions were excluded from our collection: reactions that appeared to be singleton reactions, reactions involving more than two fused rings that are not part of a steroid backbone, ring fusions, ring contractions, reactions in which the substrate or product is a radical, and reactions specifically indicated to have been found only in the case of plant CYP isozymes.

A few of the SMIRKS used to describe the transformations were taken from the Toxtree SMARTCyp module[5]. Most of the SMIRKS, however, were newly developed specifically for GLORY. When developing the SMIRKS expressions, care was taken to include as few atoms as possible in the explicit mapping, since SoM probabilities were considered for all atoms in the mapping.

Each reaction type was designated as either "common" or "uncommon." Whenever possible, this label was assigned according to the reaction's classification by Guengerich in his 2001 review of CYP-mediated reactions (Guengerich, 2001), which explicitly divided the reactions into these two categories. If the reaction type was not described in that publication, a "common" or "uncommon" label was chosen based on extrapolation (on the basis of empirical similarity to reaction types present in the publication).

Our collection of CYP reaction rules consists of 61 reaction types. In some cases, multiple transformations were required to describe the same reaction type, leading to a total of 73 transformations in the collection of defined reactions. A full list of the reaction types and their SMIRKS can be found in **Supplementary Table 1**.

Metabolite Prediction Program

Predicting the structures of the metabolites involves applying the reaction rules at all relevant positions. The relevant positions

[5]Toxtree Module: SMARTCyp—Cytochrome P450-Mediated Drug Metabolism, http://toxtree.sourceforge.net/smartcyp.html

are determined by the reaction rule itself and, in the case of the MaxEfficiency mode, by the SoM probability predicted for each heavy atom. In MaxCoverage mode, the SoM probabilities are also used to score the predicted metabolites.

SoM Prediction With FAME 2

The SoM predictions were carried out using the FAME 2 software (Šícho et al., 2017), which included preprocessing of the molecules. The circCDK_ATF_6 trained model, which had the best average performance during the independent test set validation in Šícho et al. (2017), was used for the SoM prediction within GLORY.

Application of Transformations

The transformations of parent molecules into predicted metabolites based on the defined SMIRKS strings were performed using Ambit-SMIRKS [Kochev et al., 2018; Ambit-SMARTS Java Library, version 3.1.0. http://ambit.sourceforge.net/smirks.html (accessed Oct 4, 2017)]. Some transformations may result in multiple products. Products that contain fewer than three heavy atoms are not included in the set of predicted metabolites generated by GLORY.

When SoM prediction is used as a preliminary filter, a transformation rule is only applied at a particular location in the parent molecule if one of the heavy atoms involved is predicted to be a SoM with a probability over a certain threshold (see Results for more information on this threshold).

Scoring of Predicted Metabolites

The scoring of the predicted metabolites was based on SoM probability predictions and whether the reaction type was designated as common or uncommon. Each atom in the parent molecule was assigned a likelihood of being a SoM by FAME 2. When applying the transformations defined by SMIRKS, Ambit-SMIRKS maps the reactant portion of the defined transformation to any matching set of atoms in the parent molecule. Within this mapping, the maximum SoM probability was calculated and used to score the predicted metabolite that resulted from this particular transformation and mapping.

For each predicted metabolite, the priority score is calculated by multiplying the maximum SoM probability within the mapping with a factor F depending on whether the reaction type was classified as "common" or "uncommon." Priority scores for the predicted metabolites therefore range from 0 to F_{common}. The higher the score, the more likely the predicted metabolite is considered to be. See Results for further details on the selection of values for F.

If multiple transformations of a given parent molecule lead to the same metabolite structure, the priority score is calculated separately in each case and the highest score is retained. To calculate top-k values and rank-based ROC curves, it was necessary to rank the predicted metabolites for each parent molecule based on their priority scores. If different metabolites of the same parent compound have the same priority score, then they receive the same rank. In the case of a tie, one or more rank numbers, according to the number of tied predictions, following the tied rank are skipped. For example, if the highest score is 2.5 and two predicted metabolites both have this score, then both of these metabolites are assigned a rank of 1, no predicted metabolite is assigned a rank of 2, and the predicted metabolite(s) with the next highest score are assigned the rank of 3.

Program Output

The predicted metabolites are provided as an SD file with the following information for each predicted metabolite: rank (out of all predicted metabolites for a particular parent molecule), priority score, reaction name, and the InChI, SMILES, and ID of the parent molecule. If multiple transformations led to the same product, the highest priority score and the corresponding reaction name are reported. If the input consists of multiple molecules, the ID of a parent molecule is set to the molecule's position in the ordered list of input molecules (i.e., its position in the input file).

Creation of the Reference Dataset

The reference dataset was made by combining the CYP metabolism data from DrugBank and MetXBioDB. The total size of the combined reference dataset, not including any metabolism information for any of the parent molecules contained in the manually curated test dataset, is 848 parent molecules and 1588 metabolites (an average of 1.87 metabolites per parent molecule).

DrugBank Dataset

The DrugBank database (DrugBank, version 5.1.2. https://www.drugbank.ca/ [accessed Jan 14, 2019]) was downloaded from the website. In addition to the database in XML format, the structures of all of the molecules, both parents and metabolites, were downloaded in SD format from the website (drug group "All" for the parent molecules).

Any parent or metabolite molecule without an available structure was ignored. One parent compound (DrugBank ID: DB09327) was ignored because its SMILES had two components of which the main component could not be unambiguously identified. All available generations of metabolism reactions were considered, as long as the reaction was annotated as mediated by one or more CYP isozymes. The enzymes for the reactions listed in DrugBank do not have any apparent species information, so all were assumed to be human and thereby relevant for this dataset.

For all CYP-mediated reactions, the reactant was considered to be the parent molecule and the product was considered to be a first-generation metabolite of that particular parent molecule. Any metabolite with the same InChI, ignoring stereochemistry information, as its parent molecule was removed from the set of metabolites for that parent molecule. Only those parent molecules with at least one valid metabolite were included in the final dataset.

Finally, the six parent molecules that are also present in the manually curated test dataset were removed from the DrugBank dataset prior to any evaluation, along with their corresponding metabolism information. These parent compounds were bupropion, ticlopidine, imipramine, ifosfamide, bosentan, and olanzapine.

After preprocessing, including removal of the overlap with the manually curated test dataset, the DrugBank dataset contained

364 parent molecules and 702 metabolites in total, with an average of 1.93 metabolites per parent molecule in the dataset.

MetXBioDB Dataset

The human, CYP-mediated reactions were extracted from the MetXBioDB dataset (MetXBioDB, version 1.0. https://bitbucket.org/djoumbou/biotransformerjar/src/master/database/ [accessed Jan 11, 2019]). As the only structural information provided in the MetXBioDB is in the form of InChIs and InChIKeys, any substrate or product without a reported InChI could not be considered. A lacking InChI was only the case for one out of 1468 CYP-mediated, human reactions in MetXBioDB.

Stereochemistry information was removed by generating InChIs without a stereochemistry layer, resulting in 751 CYP, human parent compounds in total. Of these, 259 are also present in the DrugBank dataset. For these overlapping parent compounds, 512 of 569 DrugBank metabolites are also in MetXBioDB, and MetXBioDB has an additional 93 metabolites for these overlapping parent compounds.

Eight parent compounds (olanzapine, bupropion, metoclopramide, bosentan, imipramine, ticlopidine, ifosfamide, and atomoxetine) from the manually curated test dataset were also present in the MetXBioDB dataset, only two of which (metoclopramide and atomoxetine) were not also present in the DrugBank dataset. These parent compounds and the corresponding metabolism data were removed from the MetXBioDB dataset.

After preprocessing, including removal of the overlap with the manually curated test dataset, the MetXBioDB dataset contained 743 parent molecules and 1385 metabolites in total, with an average of 1.86 metabolites per parent molecule in the dataset.

Merger of the DrugBank and MetXBioDB Datasets

The DrugBank dataset and the MetXBioDB dataset were combined to form the reference dataset via a straightforward consolidation of the parent and metabolite information. All molecule comparisons occurred using InChIs generated without stereochemistry information. For any parent molecule that was present in both the DrugBank and the MetXBioDB datasets, which was the case for 259 parent molecules, the sets of metabolites from both datasets were combined, disregarding stereochemistry, to yield the final set of metabolites for that parent molecule in the reference dataset.

Creation of the Manually Curated Test Dataset

A new dataset for testing GLORY was manually assembled from the scientific literature. The data were extracted from publications on metabolism that were found in two journals: *Xenobiotica* and *Drug Metabolism and Disposition*. The time frame considered was from January 2014 to June 2018 for *Xenobiotica* and from January 2014 to June 2017 for *Drug Metabolism and Disposition*.

Publications were chosen and the metabolism information they contain included in the dataset if the following criteria were fulfilled:

1. The publication must contain a figure that depicts the metabolism scheme and includes the chemical structures of the parent compound and the first-generation metabolites.
2. The metabolism data must have been experimentally determined from a human source (i.e., either humans, human cells, or recombinant human CYP enzymes). If some but not all of the data were from humans, any non-human metabolites in the metabolism scheme were excluded from the dataset.
3. For at least 75% of all of the first-generation human metabolites depicted in the metabolism scheme (note that any metabolite that is depicted as merely being an intermediate is not considered), the following two criteria must be satisfied. First, the identity of the enzyme(s) responsible for the formation of the metabolite must be known. For this purpose, it is sufficient to know whether or not this metabolite is formed by CYPs. Second, the exact chemical structure, including the connectivity of all atoms, of the metabolite must be known. There is one exception to this rule: If the metabolite is known to not be CYP-formed, then the exact structure is not relevant and the metabolite is counted anyway.

Based on these criteria, 29 metabolism schemes containing at least one human, CYP-formed first-generation metabolite with a fully defined structure were found and included in the dataset. For these 29 parent molecules, there are 81 metabolites in total that fulfill the criteria (first-generation, human, CYP, fully defined structure) for inclusion in the dataset. Note that only first-generation metabolites are included in the dataset. Note also that intermediates, as depicted in the metabolism scheme, are not included in the dataset. Instead, the first non-intermediate metabolite in the pathway is used.

The SMILES for the metabolites were generated using ChemSpider (ChemSpider. http://www.chemspider.com/ [accessed Feb 13, 2019]). Consistency of stereochemistry information between parents and their metabolites was maintained.

Validation of Metabolite Structure Predictors

Predicted metabolites were compared to known metabolites from the reference and test datasets using their InChIs. The InChIs used for this comparison were generated without stereochemistry information using CDK (Willighagen et al., 2017; Chemistry Development Kit, version 2.0. https://cdk.github.io/ [accessed Nov 3, 2017]).

During the validation, a predicted aldehyde metabolite was considered equivalent to the corresponding carboxylic acid, because there is evidence that some percentage of an aldehyde metabolite acts as an intermediate that is further oxidized to a carboxylic acid without leaving the CYP enzyme active site (Bell-Parikh and Guengerich, 1999).

In the case of one parent molecule in the reference dataset, no predictions could be made because the parent molecule contains boron. FAME 2 is unable to make predictions for molecules containing boron because no boron-containing molecules were present in the dataset used to train the model.

The SyGMa predictions were carried out in Python using the SyGMa Python package (SyGMa, version 1.1.0), and RDKit (RDKit: Open-Source Cheminformatics, version 2017_03_01, 2017). Only the phase I reaction rule set was used and one reaction cycle was applied.

The BioTransformer predictions were performed using the CYP450 mode of the BioTransformer (BioTransformer, version 1.0.8. https://bitbucket.org/djoumbou/biotransformerjar/src/master/ [accessed Feb 5, 2019]) command line tool. BioTransformer was run individually for each parent compound using single SMILES input.

The ROC curves were generated using the ROCR R package (Sing et al., 2005; ROCR, version 1.0-7, 2015). When false negative data points were added to the curve, these data points were assigned a score of 0 or a rank of 1,000, respectively, depending on whether the ROC curve represented scores or ranks.

CONCLUSIONS

We have developed GLORY, a new tool for predicting the structures of human metabolites formed by CYPs. GLORY incorporates two key ideas: a literature-based collection of CYP-mediated reaction rules and SoM prediction, which was used particularly auspiciously to develop a new scoring approach for the predicted metabolites.

For GLORY, we developed a new collection of 73 reaction rules, describing 61 reaction types, for CYP-mediated metabolism. In developing this collection, we prioritized the reproducibility of our rule set and therefore based the rules on the scientific literature rather than on any dataset. In addition to the rules themselves, each reaction type was designated as either common or uncommon, again based on the scientific literature rather than on any dataset.

In addition, we have devised a priority score for predicted metabolites based on predicted SoM probabilities and the simple, literature-based distinction between common and uncommon reaction types. Hence neither our rule set nor our scoring approach is directly based on any dataset of metabolic reactions, setting our approach apart from other tools, for example SyGMa, which uses reaction rules and occurrence ratios derived from a proprietary dataset, and BioTransformer, whose rules were to some extent based on a freely available dataset.

GLORY has two modes: MaxEfficiency, which uses SoM prediction as a prefilter for the positions in a molecule at which reactions are allowed to occur, and MaxCoverage, which does not use a prefilter and instead focuses on high recall and an accurate ranking of the predicted metabolites. Using SoM prediction as a preliminary filter, i.e., in MaxEfficiency mode, does not work as well as might be expected in terms of reducing the number of putative false positive predictions while still keeping a high rate of recovery of reported metabolites. However, by developing a priority score for the predicted metabolites using SoM prediction combined with a simple binary distinction between common and uncommon reaction types, we are able to rank the metabolites predicted by MaxCoverage mode to the extent that GLORY can predict at least one known metabolite within the top three ranked positions for 76% of the molecules in the independent test set while achieving a recall of 0.83. GLORY's MaxCoverage mode outperforms both SyGMa and BioTransformer in terms of recall and outperforms SyGMa in terms of ranking (BioTransformer does not currently rank its metabolite predictions). One use case for the MaxCoverage mode could be, for example, identifying metabolites from mass spectrometry data.

Along with the collection of reaction rules, we provide a new, manually curated test dataset for free use as a benchmark dataset. In addition, GLORY is freely available as a web server at https://acm.zbh.uni-hamburg.de/glory/.

Importantly, the concept of GLORY is such that it can be extended to predict metabolites formed by enzymes not belonging to the CYP family. The enzymes that this approach can be expanded to is limited, in principle, only by the extent of the available data and the coverage of the relevant metabolic reactions by SoM prediction tools.

DATA AVAILABILITY

Publicly available datasets were analyzed in this study. This data can be found here: https://bitbucket.org/djoumbou/biotransformerjar/ and https://www.drugbank.ca/.

AUTHOR'S NOTE

The GLORY web service is available at the following address: https://acm.zbh.uni-hamburg.de/glory/.

AUTHOR CONTRIBUTIONS

CdBK and JK: conceptualization; CdBK, CS, and JK: methodology; CdBK, CS, MŠ, NK, and NJ: software development; CdBK: validation; JK and DS: resources; CdBK: data curation; all authors: writing—original draft preparation; CdBK: visualization; DS, NJ, and JK: supervision; JK: project administration; DS and JK: funding acquisition.

FUNDING

CS and JK are supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)–project number KI 2085/1-1. DFG is a German research funding organization supporting research in science, engineering and the humanities. JK is also supported by the Bergen Research Foundation (BFS) [BFS2017TMT01]. BFS gives grants toward research and research supporting activities at the University of Bergen (UiB) and Haukeland University Hospital (HUS), and other Norwegian research institutions that cooperate with institutions in Bergen. The foundation also gives grants to support research at UiB and HUS at the interface between basic research and clinical research. MŠ and DS are supported

by the Ministry of Education, Youth and Sports (MEYS)—project numbers MSMT No 20-SVV/2017, LM2015063 and RVO: 68378050-KAV-NPUI. MEYS is responsible for public administration in education, for developing educational, youth and sport policies and international cooperation in these fields. It is also the central administrative office responsible for research and development and one of the main public research coordinating and funding bodies of the Czech Republic.

ACKNOWLEDGMENTS

We thank Craig Knox, Michael Wilson, and David Wishart from the University of Alberta, AB, Canada and OMx Personal Health Analytics Inc., AB, Canada for sending us the structures of the metabolites in the DrugBank database prior to their being made freely available online.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fchem.2019.00402/full#supplementary-material

Supplementary Tables | Reaction rules and additional evaluation results (DOCX).

Supplementary Data Sheet 1 | Test dataset containing SMILES for parent molecules and metabolites as well as publication references (CSV).

Supplementary Data Sheet 2 | Reference dataset containing SMILES (with stereochemistry information), InChI (without stereochemistry information), and DrugBank and MetXBioDB identifiers for parent molecules and metabolites (JSON).

REFERENCES

Bell-Parikh, L. C., and Guengerich, F. P. (1999). Kinetics of cytochrome P450 2E1-catalyzed oxidation of ethanol to acetic acid via acetaldehyde. *J. Biol. Chem.* 274, 23833–23840. doi: 10.1074/jbc.274.34.23833

Cruciani, G., Carosati, E., De Boeck, B., Ethirajulu, K., Mackie, C., Howe, T., et al. (2005). MetaSite: understanding metabolism in human cytochromes from the perspective of the chemist. *J. Med. Chem.* 48, 6970–6979. doi: 10.1021/jm050529c

Darvas, F. (1987). "Metabolexpert: an expert system for predicting metabolism of substances," in *QSAR in Environmental Toxicology - II*, ed K. L. E. Kaiser, (Dordrecht: Springer), 71–81.

Di, L. (2014). The role of drug metabolizing enzymes in clearance. *Expert Opin Drug Metabol. Toxicol.* 10, 379–393. doi: 10.1517/17425255.2014.876006

Djoumbou-Feunang, Y., Fiamoncini, J., Gil-de-la-Fuente, A., Greiner, R., Manach, C., and Wishart, D. S. (2019). BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminform.* 11:2. doi: 10.1186/s13321-018-0324-5

Guengerich, F. P. (2001). Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. *Chem. Res. Toxicol.* 14, 611–650. doi: 10.1021/tx0002583

Judson, P. N. (2014). "Knowledge-based approaches for predicting the sites and products of metabolism," in *Drug Metabolism Prediction*, ed J. Kirchmair, Methods and Principles in Medicinal Chemistry (Weinheim: Wiley-VCH), 293–318.

Kirchmair, J., Göller, A. H., Lang, D., Kunze, J., Testa, B., and Wilson, I. D. (2015). Predicting drug metabolism: experiment and/or computation? *Nat. Rev. Drug Disc.* 14, 387–404. doi: 10.1038/nrd4581

Kirchmair, J., Williamson, M. J., Tyzack, J. D., Tan, L., Bond, P. J., Bender, A., et al. (2012). Computational prediction of metabolism: sites, products, S. A. R., P450 enzyme dynamics, and mechanisms. *J . Chem. Inform. Model.* 52, 617–48. doi: 10.1021/ci200542m

Klopman, G., Dimayuga, M., and Talafous, J. (1994). META. 1. a program for the evaluation of metabolic transformation of chemicals. *J. Chem. Inform. Comp. Sci.* 34, 1320–1325.

Kochev, N., Avramova, S., and Jeliazkova, N. (2018). Ambit-SMIRKS: a software module for reaction representation, reaction search and structure transformation. *J. Cheminform.* 10:42. doi: 10.1186/s13321-018-0295-6

Marchant, C. A., Briggs, K. A., and Long, A. (2008). *In silico* tools for sharing data and knowledge on toxicity and metabolism: Derek for Windows, Meteor, and Vitic. *Toxicol. Mechanisms Methods.* 18, 177–187. doi: 10.1080/15376510701857320

Mekenyan, O. G., Dimitrov, S. D., Pavlov, T. S., and Veith, G. D. (2004). A systematic approach to simulating metabolism in computational toxicology. I. the TIMES heuristic modelling framework. *Curr. Pharmaceutical Design.* 10, 1273–1293. doi: 10.2174/1381612043452596

Olsen, L., Montefiori, M., Tran, K. P., and Jørgensen, F. S. (2019). SMARTCyp 3.0: Enhanced cytochrome P450 site-of-metabolism prediction server. *Bioinformatics.* doi: 10.1093/bioinformatics/btz037. [Epub ahead of print].

Patlewicz, G., Jeliazkova, N., Safford, R., J., Worth, A., P., and Aleksiev, B. (2008). An evaluation of the implementation of the Cramer classification scheme in the Toxtree software. *SAR QSAR Environ. Res.* 19, 495–524. doi: 10.1080/10629360802083871

Ridder, L., and Wagener, M. (2008). SyGMa: combining expert knowledge and empirical scoring in the prediction of metabolites. *Chem. Med. Chem.* 3, 821–832. doi: 10.1002/cmdc.200700312

Rudik, A., Dmitriev, A., Lagunin, A., Filimonov, D., and Poroikov, V. (2015). SOMP: web server for *in silico* prediction of sites of metabolism for drug-like compounds. *Bioinformatics* 31, 2046–2048. doi: 10.1093/bioinformatics/btv087

Rudik, A. V., Bezhentsev, V. M., Dmitriev, A. V., Druzhilovskiy, D. S., Lagunin, A. A., Filimonov, D. A., et al. (2017). MetaTox: web application for predicting structure and toxicity of xenobiotics' metabolites. *J. Chem. Inform. Model.* 57, 638–642. doi: 10.1021/acs.jcim.6b00662

Rydberg, P., Gloriam, D. E., Zaretzki, J., Breneman, C., and Olsen, L. (2010). SMARTCyp: a 2D method for prediction of cytochrome p450-mediated drug metabolism. *ACS Med. Chem. Lett.* 1, 96–100. doi: 10.1021/ml100016x

Šícho, M., de Bruyn Kops, C., Stork, C., Svozil, D., and Kirchmair, J. (2017). FAME 2: simple and effective machine learning model of cytochrome P450 regioselectivity. *J. Chem. Inform. Model.* 57, 1832–1846. doi: 10.1021/acs.jcim.7b00250

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics* 21, 3940–3941. doi: 10.1093/bioinformatics/bti623

Testa, B., Pedretti, A., and Vistoli, G. (2012). Reactions and enzymes in the metabolism of drugs and other xenobiotics. *Drug Disc. Today* 17, 549–560. doi: 10.1016/j.drudis.2012.01.017

Tian, S., Djoumbou-Feunang, Y., Greiner, R., and Wishart, D. S. (2018). CypReact: a software tool for *in silico* reactant prediction for human cytochrome P450 enzymes. *J. Chem. Inform. Model.* 58, 1282–1291. doi: 10.1021/acs.jcim.8b00035

Tyzack, J. D., Hunt, P. A., and Segall, M. D. (2016). Predicting regioselectivity and lability of cytochrome p450 metabolism using quantum mechanical simulations. *J. Chem. Inform. Model.* 56, 2180–2193. doi: 10.1021/acs.jcim.6b00233

Tyzack, J. D., and Kirchmair, J. (2018). Computational methods and tools to predict cytochrome P450 metabolism for drug discovery. *Chem. Biol. Drug Design* 93, 377–386. doi: 10.1111/cbdd.13445

Willighagen, E. L., Mayfield, J. W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., et al. (2017). The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminform.* 9:33. doi: 10.1186/s13321-017-0220-4

Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., et al. (2018). DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 46, D1074–D1082. doi: 10.1093/nar/gkx1037

Zaretzki, J., Matlock, M., and Swamidass, S. J. (2013). XenoSite: accurately predicting CYP-mediated sites of metabolism with neural networks. *J . Chem. Inform. Model.* 53, 3373–3383. doi: 10.1021/ci400518g

Conflict of Interest Statement: NJ is a founder and co-owner of Ideaconsult Ltd. and has been the technical manager of the company since 2009. NK works for Ideaconsult Ltd. on a part-time basis.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

4.2.2 Predicting Phase I and Phase II Metabolites

The approach to metabolite structure prediction developed for CYPs (see section 4.2.1) explored how SoM prediction could be incorporated to improve a metabolite structure predictor's ability to prioritize the predictions and enable focusing on the most likely ones. We concluded that the most effective approach was to use the predicted SoM probabilities not as a hard filter but as a component of a score for each predicted metabolite, which enabled a meaningful ranking of the predictions.

The following publication describes the extension of this method to cover both phase I and phase II metabolism, incorporating SoM prediction by the previously published FAME 3 models [A8] to score the predicted metabolites. This new development required additional reaction rules covering non-CYP phase I metabolic reactions as well as phase II reactions. These new reaction rules were developed by implementing the reaction rules from SyGMa [76], an existing open-source metabolite prediction tool, as well as manually developing reaction rules for GSH conjugation based on the scientific literature.

In addition, we compiled a new reference dataset from phase I and phase II metabolite data in the DrugBank [101] and MetXBioDB [85] databases in order to analyze the ability of our new method to predict both phase I and phase II metabolites. With this reference dataset of 1420 parent molecules and their phase I and phase II metabolites, we were able to analyze the performance separately for phase I and phase II, which allowed us to see that although FAME 3's phase I SoM model's predictions led to a reasonable ranking of the phase I metabolites, ranking the predicted phase II metabolites based on the predictions made by FAME 3's phase II SoM model was difficult. This discovery led to the use of newly developed FAME 3 models for individual phase II reaction classes, which greatly improved the ranking of the predicted phase II metabolites.

The result of this work was GLORYx, a novel tool for predicting phase I and phase II metabolites of xenobiotics. For the final evaluation of GLORYx, we developed a new, manually curated test dataset of 37 best-selling drugs and their phase I and phase II metabolites from the scientific literature. GLORYx is freely available as a web server and is open source.

[D3] GLORYx: Prediction of the Metabolites Resulting from Phase 1 and Phase 2 Biotransformations of Xenobiotics

Christina de Bruyn Kops, Martin Šícho, Angelica Mazzolari, and Johannes Kirchmair

Chemical Research in Toxicology, 2020

Available at `https://doi.org/10.1021/acs.chemrestox.0c00224`.

Contribution:
C. de Bruyn Kops and J. Kirchmair conceptualized the research. C. de Bruyn Kops developed the metabolite prediction method. C. de Bruyn Kops wrote the manuscript, with contributions from J. Kirchmair. M. Šícho developed the machine learning models. A. Mazzolari contributed the dataset for the development of the machine learning models. J. Kirchmair supervised this work.

The following article was reprinted from:
de Bruyn Kops, C.; Šícho, M.; Mazzolari, A.; Kirchmair, J. GLORYx: Prediction of the Metabolites Resulting from Phase 1 and Phase 2 Biotransformations of Xenobiotics, *Chem. Res. Toxicol.* **2020**, `DOI 10.1021/acs.chemrestox.0c00224`.

The supplementary information for this article can be found in Appendix D.

Chemical Research in Toxicology

pubs.acs.org/crt

Article

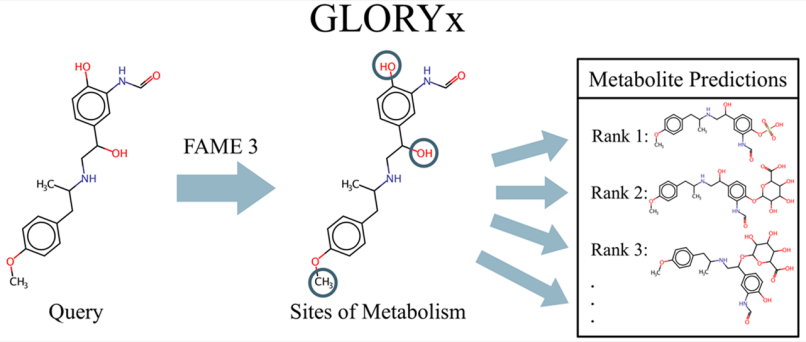
GLORYx: Prediction of the Metabolites Resulting from Phase 1 and Phase 2 Biotransformations of Xenobiotics

Christina de Bruyn Kops, Martin Šícho, Angelica Mazzolari, and Johannes Kirchmair*

Cite This: https://dx.doi.org/10.1021/acs.chemrestox.0c00224

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



ABSTRACT: Predicting the structures of metabolites formed in humans can provide advantageous insights for the development of drugs and other compounds. Here we present GLORYx, which integrates machine learning-based site of metabolism (SoM) prediction with reaction rule sets to predict and rank the structures of metabolites that could potentially be formed by phase 1 and/or phase 2 metabolism. GLORYx extends the approach from our previously developed tool GLORY, which predicted metabolite structures for cytochrome P450-mediated metabolism only. A robust approach to ranking the predicted metabolites is attained by using the SoM probabilities predicted by the FAME 3 machine learning models to score the predicted metabolites. On a manually curated test data set containing both phase 1 and phase 2 metabolites, GLORYx achieves a recall of 77% and an area under the receiver operating characteristic curve (AUC) of 0.79. Separate analysis of performance on a large amount of freely available phase 1 and phase 2 metabolite data indicates that achieving a meaningful ranking of predicted metabolites is more difficult for phase 2 than for phase 1 metabolites. GLORYx is freely available as a web server at https://nerdd.zbh.uni-hamburg.de/ and is also provided as a software package upon request. The data sets as well as all the reaction rules from this work are also made freely available.

INTRODUCTION

Metabolism has a large impact on the safety and efficacy of the xenobiotics that enter the human body, from drugs to cosmetics and agrochemicals, because metabolic reactions can change these compounds into metabolites with different physicochemical and pharmacological properties.[1] Computational approaches can be useful for predicting how drugs and other xenobiotics will be metabolized in humans, allowing, for example, the focusing of the drug development process on the most promising compounds in order to save time and reduce costs.

Human xenobiotic metabolism is generally separated into two phases, phase 1 and phase 2, based on the type of reaction (note that the nomenclature does not indicate that a phase 1 reaction must occur before a phase 2 reaction can take place). Phase 1 metabolism consists of oxidation, reduction, and hydrolysis reactions that generally result in increased polarity of the metabolite compared to the parent molecule by creating or unmasking polar functional groups. The main enzyme family responsible for phase 1 metabolism is the cytochrome P450 (CYP) enzyme family, which is responsible for the formation of approximately 60% of first-generation metabolites but only for approximately 40% of metabolites overall (all percentages presented here are based on the current version[2] of the MetaQSAR database,[3] which contains over 4000 parent molecules, including drugs and other xenobiotics, along with their first-, second-, and third-generation metabolites produced by mammalian metabolic enzymes in vitro and/or in vivo). CYPs are also the cause of a large portion of toxic metabolites and drug−drug interactions.[4] Of the non-CYP phase 1 enzymes, the ones with the highest impact on metabolite formation are hydrolases and flavin-containing monooxyge-

Special Issue: Computational Toxicology

Received: June 3, 2020
Published: August 10, 2020

nases (FMOs), which account for approximately 9% and approximately 4% of all metabolites, respectively.

Phase 2 metabolism consists of conjugation reactions that, like phase 1 metabolic reactions, tend to modify compounds to more excretable forms. In all, phase 2 metabolism accounts for approximately 30% percent of all metabolites.[2] The enzymes responsible for phase 2 drug metabolism belong primarily to five enzyme families: UDP-glucuronosyltransferases (UGTs), glutathione *S*-transferases (GSTs), sulfotransferases (SULTs), *N*-acetyltransferases (NATs), and methyltransferases (MTs).[5] These five enzyme families are responsible for nearly 90% of all phase 2 metabolites.[4]

Computational prediction of xenobiotic metabolism encompasses several aspects, including the prediction of the metabolically labile atom positions in molecules, which are known as sites of metabolism (SoM), and prediction of metabolite structures.[1,6] Although SoM prediction can provide valuable information and often allows the structure of the resulting metabolite to be inferred by a chemist, it is also of interest to directly predict the metabolites themselves. There are many freely available and commercial methods for metabolite structure prediction, though most focus exclusively on CYP-mediated metabolism. Commercial tools that offer comprehensive prediction of metabolites for both phase 1 and phase 2 metabolism include Meteor Nexus (Lhasa Ltd.),[7] TIMES (LMC),[8] and MetabolExpert (CompuDrug Ltd.).[9]

In terms of freely available metabolite structure predictors, a popular and relatively long-lived, open-source tool is SyGMa.[10] SyGMa generates and ranks metabolites based on reaction rules and their occurrence ratios derived from the Metabolite database.[11] The current version of SyGMa predicts phase 1 and phase 2 metabolites using a set of 145 knowledge-based reaction rules (118 phase 1 rules and 27 phase 2 rules). Using the combined set of reaction rules, SyGMa was able to predict 68% of the metabolites in a test set consisting of 175 parent compounds and 385 reactions (taken from a later release of the Metabolite database compared to the training data).[10] The predictor was able to rank 45% of the metabolites in the test set within the top 10 predictions for their corresponding parent molecules. Unfortunately, the Metabolite database, which was used to develop the reaction rules and the occurrence ratios for the scoring, has been discontinued.

Another freely available tool for metabolite prediction is BioTransformer,[12] an open-source, comprehensive program that predicts metabolite structures for human CYP and phase 2 metabolism as well as gut microbial, environmental microbial, and human "enzyme commission (EC)-based" metabolism. BioTransformer has 163 CYP rules and 74 phase 2 rules as well as additional constraints regarding molecule types that various rules are allowed to be applied to. Using a combination of the CYP, phase 2, and EC-based modules (408 rules), BioTransformer achieved a recall of 88% on its test set of 40 pharmaceuticals and pesticides, with a precision of 0.49 (188 true positive predictions and 198 putative false positive predictions). BioTransformer does not currently rank its predictions.

A further freely available tool, MetaTox,[13,14] predicts metabolites by separately predicting the probability that each potential reaction class is relevant to the given molecule and also predicting the probability of a reaction occurring at each possible reaction center given each possible reaction class. The probability that the resulting metabolite is formed is calculated by combining both of these probabilities, which can then be used to rank the predictions. The reaction types include both phase 1 and phase 2 reaction types, though it is unclear how many reaction rules there are in total. During leave-one-out cross-validation, MetaTox obtained invariant accuracy prediction (IAP, a metric related to area under the receiver operating characteristic curve (AUC)) values between 0.79 and 0.95 for the prediction of the correct reaction class and IAP values between 0.86 and 0.99 for the prediction of the reacting atoms for each of the biotransformation classes.[14]

We recently reported on the development of a tool called GLORY that predicts the structures of metabolites formed by the CYP enzyme family.[15] GLORY includes a new set of reaction rules for CYP-mediated metabolism, whereby common reaction types are distinguished from more unusual reactions. Importantly, GLORY explored how SoM prediction could be effectively employed within the context of metabolite structure prediction. We were able to demonstrate that using the predicted SoM probabilities for each atom in a molecule to score the predicted metabolites, resulting from reactions taking place at those atom positions, led to a meaningful ranking of the predictions.

The software for SoM prediction that was used in GLORY was FAME 2,[16] a machine learning-based SoM prediction program that uses extremely randomized trees classifiers combined with two-dimensional (2D) circular descriptors to predict SoMs for CYP-mediated metabolism. Since the development of GLORY, a successor to FAME 2 has become available. FAME 3[17] continues to use the concept of extra trees classifiers and 2D circular descriptors developed in FAME 2 and applies this approach to generate comprehensive SoM prediction models for both phase 1 and phase 2 metabolism.

There are several other metabolite prediction tools, both commercial (e.g., ADMET Predictor[18] from SimulationsPlus, StarDrop[19] from Optibrium, and MetaSite from Molecular Discovery[20]) and freely available (e.g., MetaTox[13,14]), that incorporate SoM prediction into their metabolite prediction approaches. These tools either focus solely on CYP-mediated metabolism, have not been published, or, as described above for MetaTox, have been evaluated in such a way that makes it difficult to determine how well the metabolite structures themselves were predicted. Thus, although the concept of combining SoM prediction with reaction rules for comprehensive metabolite prediction has been applied in various ways, a systematic analysis of the performance for both phase 1 and phase 2 metabolism has not yet been published.

We have extended the approach developed in GLORY to create a new tool, called GLORYx, that combines SoM prediction with a set of reaction rules to predict metabolites for both phase 1 and phase 2 metabolism. GLORYx employs FAME 3 for SoM prediction, the results of which are used to score and rank the predicted metabolites. Compared to GLORY, GLORYx requires more reaction rules in order to cover non-CYP phase 1 metabolic reactions as well as phase 2 metabolic reactions. GLORYx is freely available via a web server at https://nerdd.zbh.uni-hamburg.de/.

METHODS

Reference Data Set. A reference data set of compound-metabolite pairs was compiled from the freely available metabolism data in the DrugBank (drug group "All")[21,22] and MetXBioDB[23] databases to serve as a basis for evaluation of the method during the development of GLORYx. For each metabolic reaction in either database, the reactant was considered to be the parent molecule, and

https://dx.doi.org/10.1021/acs.chemrestox.0c00224
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

the product was considered to be the metabolite. The reference data set is therefore in the format of a map of each parent molecule to its first-generation metabolites, regardless of whether the parent molecule is itself the metabolite of another molecule.

The extraction of the data from DrugBank and MetXBioDB is consistent with the method used in GLORY (see ref 15). The differences in the preprocessing of the data (i.e., assigning a phase and removing the minor component of salts; see below) arise from considering both phase 1 and phase 2 metabolism rather than just CYP metabolism.

The preprocessing procedure is as follows:

(1) Structural information for both the parent and the metabolite was required in order for a reaction to be included. For DrugBank, the structures are provided in SD file format. In MetXBioDB, only InChIs and InChIKeys are provided, so the InChI was used to generate the structure. Note that stereochemistry information was ignored for parent molecules as well as metabolites, so stereoisomers were thereby condensed.
(2) The multicomponent parent molecules in the DrugBank database had to be handled (no multicomponent parent molecules were found in MetXBioDB). The minor component of each salt was removed (e.g., K^+, Ca^{2+}). There was one multicomponent compound (DrugBank ID: DB09327) in which the main component could not be automatically determined, so this compound was excluded from the reference data set. Note that multicomponent metabolites, on the other hand, are simply separated into their individual components and each is considered a separate metabolite.
(3) Any metabolite that contained only one heavy atom (six cases consisting of metal ions, SeH_2, and a water molecule; DrugBank only) or had the same InChI, ignoring stereochemistry, as its parent molecule, was excluded.
(4) The metabolites were classified as either phase 1 or phase 2 metabolites, according to the enzyme or biotransformation type annotation (see subsection below for details). If a metabolite could not be assigned a phase, the metabolite was ignored.
(5) Parent molecules with no remaining valid metabolites after applying the above criteria were removed.
(6) The metabolism data corresponding to all parent molecules that overlap with a manually curated test data set (described below) were removed from the reference data set. The removal of the overlap with the test data set affected 15 parent molecules from DrugBank and 9 from MetXBioDB.

The DrugBank and MetXBioDB data were combined in a straightforward manner using InChIs generated without stereochemistry information to compare molecules. If the same parent molecule was present in both DrugBank and MetXBioDB, then the metabolites from both sources were combined, disregarding stereochemistry, into one set.

Assigning a Metabolism Phase to Metabolites in the Data Set. To enable separate evaluation for phase 1 and phase 2 metabolite prediction, we assigned each metabolite in the reference data set to a phase based on the relevant information in DrugBank and MetXBioDB. This allowed the creation of two distinct subsets of the reference data set. The phase 1 and phase 2 subsets of the reference data set represent only phase 1 and phase 2 biotransformations, respectively. If a parent compound has no relevant metabolites for the given phase, then it was excluded from the corresponding subset of the data set.

For the DrugBank data, the metabolites were assigned to a metabolism phase based on the enzyme annotation of the reaction. Some enzymes were omitted completely because they are not enzymes typically associated with human xenobiotic metabolism (e.g., hemoglobin, serum albumin, and lyases). See Table S1 in the Supporting Information for a list of all enzymes that were excluded. This criterion resulted in the exclusion of only 17 metabolites from 11 parent compounds.

For the MetXBioDB data, the appropriate phase for each metabolite was determined based on the "Biotransformation type" annotation in the database. The reactions annotated "Human Phase 1" or "Human Phase 2" were classified as phase 1 or phase 2, respectively.

Manually Curated Test Data Set. The test data set was manually assembled from the scientific literature. We wanted to include all known metabolites of the parent compounds (i.e., all metabolites which have been experimentally observed and reported in the scientific literature), so we chose to structure the data as metabolic trees, including all generations of metabolites that were found in the literature.

The selection of parent molecules for the test data set was based on the top 100 best-selling drugs from 2018.[24,25] For all the small-molecule drugs within these 100 drugs that are made up of only the atoms H, C, N, S, O, F, Cl, Br, I, and P, we searched the scientific literature for relevant metabolism information, specifically the structures of human metabolites and preferably a scheme depicting the metabolic tree (see below for more detail). For the listed pharmaceutical products that are a combination of two or more named drugs (e.g., Mavyret is composed of glecaprevir and pibrentasvir), a separate literature search was undertaken for each drug component. For sources of metabolite information, we considered all scientific journal publications that could be found online with Google.

The basic criteria for inclusion in the data set were as follows:

(1) The metabolites must be clearly indicated to be found in humans (either in vivo or in vitro using human hepatocytes, human liver microsomes, or human liver S9 fractions).
(2) Structures of metabolites must be provided. In cases in which a metabolism scheme is not shown, it must be clear, based on chemistry knowledge, that the depicted metabolites are not metabolites of each other, that is, that the depicted metabolites are all first-generation metabolites of the parent drug.
(3) Only fully defined metabolite structures (i.e., the exact position of the added functional group is shown) are included in the data set. The branches of the metabolic tree are followed, and the metabolites included and annotated with the corresponding generation, until a not-fully defined structure is reached. Any further metabolites derived from such a not-fully defined structure are ignored. The maximum metabolite generation included in the data set was generation five, which occurred for only two parent molecules.
(4) Intermediates designated as such in the scheme are not included in the data set.
(5) Some metabolites could not be considered first-generation (based on chemistry knowledge and additional information from the text of the publication), even though the visual scheme indicated that their precursors were only intermediates. Such cases were also removed.
(6) Fatty acid conjugation was not considered.
(7) In the case of one prodrug (abiraterone acetate), we used the drug itself (abiraterone) as the parent molecule in the data set because it had more (first-generation) metabolites shown in the scheme.

The data set was assembled by extracting the SMILES for the parent compounds from the ChEMBL Database[26,27] by looking up each drug name. These structures were manually verified for correctness before proceeding. The SMILES for the metabolites were generated with MarvinSketch[28] by modifying the parent molecules to create the metabolites and saving them in SMILES format. Metabolite stereoisomers were combined, resulting in a structure with unspecified stereochemistry at the relevant stereocenter.

Data Set Structure. The final data set contains 37 parent molecules and is provided as a JSON file (see Notes). There are 136 first-generation metabolites in total.

The JSON file is structured to represent the metabolic trees, which include multiple generations of metabolites, whenever relevant,

https://dx.doi.org/10.1021/acs.chemrestox.0c00224
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

following the procedure explained above. For each parent compound, the DOI or PMID of the reference paper(s) is provided, along with the drug name, SMILES, and metabolites. For each metabolite, the name it was given in the publication is provided for reference (this name is often something like "M1") along with the metabolism generation number and the SMILES. Due to the JSON file format, it is always clear for the second, third, and subsequent generations of metabolites which first-generation metabolites were their precursor, and so on.

No distinction between phase 1 and phase 2 metabolites is made, and enzyme annotations are not included, as this information was only rarely provided in the original literature used to assemble the data set.

Analysis of the Metabolite Data from MetXBioDB and DrugBank. The data from MetXBioDB and DrugBank were considered separately. The data were extracted and preprocessed from each source as described in the Reference Data Set subsection, except that the parent molecules that overlap with the test data set were not removed. For the analysis described here, only the properties of the parent molecules were considered.

Calculations of molecular weight and log *P* in order to plot the distributions were performed using RDKit.[29] The molecular weight calculated was the average molecular weight including hydrogens. One molecule in MetXBioDB was not considered a valid molecule by RDKit (explicit valence greater than permitted) and was therefore excluded from all analysis described in this section.

Principal component analysis (PCA) was performed with scikit-learn[30,31] using 44 physicochemical descriptors calculated with the Molecular Operating Environment (MOE).[32] A full list of the descriptors, as well as a brief description of each, can be found in Table S2.

Reaction Rules. The metabolic reaction rules used in GLORYx are encoded as SMIRKS.[33] Three sets of reaction rules were used: (1) all of SyGMa's reaction rules, which include both phase 1 and phase 2 rules; (2) GLORY's reaction rules, covering only CYP metabolism; and (3) a newly developed set of GSH conjugation rules to augment SyGMa's phase 2 rules, which are missing reactions of this type. The reaction rules from GLORY were used unchanged.

Implementation of SyGMa Reaction Rules. Because the so-called SMIRKS provided in SyGMa's open-source python package[34] are actually in the format of RDKit's reaction SMARTS, it was necessary to convert them to proper SMIRKS in order to implement them in our software. This conversion was performed manually, with care being taken to preserve the chemical meaning of the reaction.

In one case, namely that of oxidative deamination, additional SMIRKS strings were necessary to achieve the same result with the SMIRKS that SyGMa achieved with its reaction SMARTS. The reason is that double bonds in an aromatic ring are not automatically shifted during transformation in GLORYx. We therefore added two additional SMIRKS in order to explicitly shift the double bonds for 6-rings and 5-rings. Any invalid products generated by the SMIRKS for this reaction are ignored because the molecule validity checker in GLORYx discards transformation products with a carbon atom of invalid atom type (in this case, a valence >4 due to incorrect bond placement).

Development of Reaction Rules for Glutathione Conjugation. The scientific literature indicates that glutathione (GSH) conjugation by the GST enzyme family occurs mainly at the following functional groups: epoxides, α,β-unsaturated carbonyls, quinones, nucleophilic substitution (aliphatic and aromatic), isocyanates (and isothiocyanates), and nitriles.[35−40] The SMIRKS for these cases were developed based on the reaction descriptions and example reactions present in the referenced literature.

Metabolite Prediction Program GLORYx. GLORYx applies the reaction rules to all appropriate positions in the molecule, determined by where each reaction rule SMIRKS matches, if it matches at all. Within the program, SoMs are predicted with FAME 3,[17] and the predicted SoM probabilities are used to score and rank the predicted metabolites. The software is written in Java and uses CDK version 2.0.[41,42]

GLORYx performs an initial preprocessing step for all input molecules to check that the input molecule can be successfully parsed by CDK, does not have multiple components, and contains no element types other than C, N, S, O, H, F, Cl, Br, I, P, B, and Si (FAME 3's allowed element types; note that FAME 3 does not make predictions for B and Si due to a lack of training data, and for this reason the test set was chosen to not include any molecules with a B or Si atom). If any of these checks fail, no predictions are made for the input molecule. Further preprocessing steps that occur within the context of SoM prediction and the application of the reaction rules are described in the following subsections.

SoM Prediction. The SoM prediction in GLORYx is performed using FAME 3.[17] FAME 3 was trained on the SoM data from the MetaQSAR database[3,43] and offers three SoM prediction models: the P1 model predicts SoMs corresponding to phase 1 metabolic reactions, the P2 model predicts SoMs corresponding to phase 2 metabolic reactions, and the P1 + P2 model predicts SoMs corresponding to both metabolism phases.

The FAME 3 code includes preprocessing of the input molecules, involving the standardization of nitro groups, aromaticity detection, and automatic addition of hydrogens if the hydrogens of the input molecule are not explicitly specified. Because the SoM prediction step comes before the application of the reaction rules within the GLORYx program, the standardization of the molecules described here remains in place for the subsequent transformation step described below.

FAME 3 uses circular descriptors that incorporate 15 basic 2D CDK descriptors and circular atom-type fingerprints (see ref 17 for details). During the development of FAME 3, the effect of the bond depth of the circular descriptors was examined, and a bond depth of five was chosen as the default bond depth for the descriptors. GLORYx uses the default models with a descriptor depth of five.

In order to improve GLORYx's ability to rank its predictions of phase 2 metabolites (see Results for details), we used previously unpublished reaction class-specific individual phase 2 SoM models from FAME 3. These models were created using the identical modeling procedure described in the FAME 3 paper (see ref 17), but with each model trained on only a subset of the data. The SoM data from the reaction classes in the MetaQSAR database[43] corresponding approximately to the five main enzyme families of phase 2 xenobiotic metabolism (UGTs, GSTs, SULTs, MTs, and NATs) were selected, and a separate model was created for each subset. The reaction class types and the number of molecules used to train the models are described in Table S3. Note that the data from two classes of glucuronidation reactions in MetaQSAR were combined to create a single glucuronidation model.

For the reaction class-specific phase 2 SoM models, GLORYx again uses the models with a descriptor depth of five, for consistency.

Transformation of Molecules According to Reaction Rules. The reaction rules were applied using Ambit-SMIRKS.[44,45] As for GLORY, any product containing fewer than three heavy atoms is not included in the set of predicted metabolites.

In order to apply the reaction rules correctly, that is, to achieve the same predicted metabolites as SyGMa while using the same rules, it was necessary to use an aromaticity model that could recognize aromaticity in rings with exocyclic heteroatoms. To achieve this, we chose an aromaticity model in CDK that uses the *ElectronDonation.daylight()* electron donation model. In order to allow for better ring recognition in molecules with more than three rings, we set the cycles portion of the aromaticity model to *Cycles.or(Cycles.all(), Cycles.relevant())*, indicating that all cycles would be used whenever possible and the "relevant" cycles would be used in cases in which the molecule contained too many cycles for all of them to be considered. This new aromaticity model is applied directly before reaction rule-based transformation using all reaction rules, not just the rules sourced from SyGMa, and does not affect the aromaticity recognition used for SoM prediction with FAME 3.

There is one noticeable remaining discrepancy related to aromaticity, as determined by comparison on the reference data set, between GLORYx's predictions and SyGMa's predictions for the "same" reaction rules. Tetrazoles appear to be recognized as aromatic

https://dx.doi.org/10.1021/acs.chemrestox.0c00224
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

by GLORYx but not by SyGMa, as indicated by an aromatic glucuronidation reaction being successfully applied by GLORYx but not by SyGMa. This discrepancy affects only three parent molecules in the phase 2 subset of the reference data set.

Scoring. Each predicted metabolite is assigned a priority score in order to rank the predictions. The priority score has two components. The first component is the predicted SoM probability from the FAME 3 model used to make the prediction. The maximum SoM probability among the atoms in the mapping of the reaction rule's SMIRKs onto the parent molecule is used.

The second component is a reaction rule weighting factor based on a simple designation of either "common" or "uncommon" for each reaction type. This designation, which we previously introduced for CYP reactions in ref 15, was based primarily on a detailed review of CYP-mediated reaction types that described both common and uncommon types of reactions.[46] In this work, we use the common−uncommon designation more loosely, as a simple differentiation in reaction type prioritization that allows a binary weighting of the reaction rules. A weighting factor corresponding to the common−uncommon classification is multiplied with the maximum SoM probability mentioned above in order to calculate the priority score for the predicted metabolite. In GLORYx, a weighting factor of 1 is used for reaction rules designated "common", and a weighting factor of 0.2 is used for reaction rules designated "uncommon". These weighting factors thereby maintain the same ratio of 5:1 as described previously in ref 15 but are scaled such that the final priority score more reflects a probability-like concept, with values ranging from 0 to 1.

The final priority score of a predicted metabolite is thereby the product of the maximum SoM probability and the weighting factor corresponding to the priority level, common or uncommon, of the reaction type.

The final assignment of a priority level to the reaction rules was determined rationally. The priority levels of the CYP metabolism-based rules from GLORY were not changed. All of the phase 1 rules from SyGMa were designated uncommon, which does not affect the higher priority given to common CYP-mediated reaction types in the case of duplicate reaction types in the SyGMa and GLORY rule sets. The phase 2 rules corresponding to the five main phase 2 enzyme families were designated common, while the others (glycination, phosphorylation, and dephosphorylation) were designated uncommon.

Validation. Predicted metabolites were compared to the known metabolites from either the reference data set or the test data set using InChIs that were generated without stereochemistry information.

Special Consideration for CYP Reactions. Spontaneous oxidation from an aldehyde to a carboxylic acid was considered during the evaluation process, as in GLORY (see ref 15), but only for predicted metabolites that were the product of a phase 1 reaction rule. It was intended that this consideration only apply to CYP reactions, but SyGMa's phase 1 reaction rules do not distinguish between CYP and non-CYP, so this step was applied to all phase 1 products. Note that this applies only to the validation and does not affect the predicted metabolites that are provided to the users of GLORYx.

Comparison to SyGMa. The comparison of GLORYx to SyGMa was performed using SyGMa[34] with RDKit.[29] One change to the standard usage of SyGMa was required, in the case where both phase 1 and phase 2 metabolite predictions were desired. When SyGMa is run with a single metabolism Scenario object specifying both phase 1 and phase 2, the rule sets for the phases are applied sequentially, that is, the first rule set listed (phase 1) is applied first, and then the second rule set (phase 2) is applied to the parent compound as well as the products of the first rule set. This behavior corresponds to a different research question than the one posed in our evaluation, so SyGMa was instead run twice for each molecule in the test set, once using only the phase 1 rules and then separately using only the phase 2 rules. The predictions from both runs were combined.

In addition, any predicted metabolite with the same InChI as the parent compound was ignored, and, for the sake of comparison, a filter to remove all predicted metabolites with fewer than three heavy atoms was added (SyGMa's built-in percentage-based size filter was turned off). Implicit hydrogens were also added to SyGMa's output SMILES before generating the InChIs for comparison with GLORYx's predictions.

RESULTS

The concept of GLORYx is that SoMs, or rather the probability of each heavy atom being a SoM, are predicted with FAME 3, and, building on these predictions, a set of reaction rules is applied in order to generate the structures of predicted metabolites for both phase 1 and phase 2 metabolism. We have previously determined, for our earlier CYP-focused metabolite prediction tool GLORY, that using the predicted SoM probabilities as a hard cutoff to determine whether or not to apply a reaction rule at a given position is not a particularly effective approach, except if the goal is to simply reduce the number of predictions.[15] Instead, we found that using the predicted SoM probabilities to score and rank the predicted metabolites enabled a reasonable ranking of the predicted metabolites while retaining a high recall of known metabolites. Therefore, we again use the predicted SoM probabilities to rank the metabolites predicted by GLORYx. For GLORYx, we also have the capability of using a different FAME 3 SoM prediction model depending on which phase of metabolism is being predicted.

GLORYx was developed and analyzed using a large reference data set containing metabolism data from DrugBank and MetXBioDB. This reference data set was used to examine phase 1 and phase 2 metabolism separately to make sure each phase could be handled satisfactorily on its own as well as to determine how to best combine predictions for both phases. The final validation of GLORYx was subsequently performed on a manually curated test data set.

Analysis of the Approach Using a Large Reference Data Set. A reference data set for the development of the GLORYx method was created by combining the freely available metabolism data from DrugBank and MetXBioDB (see Methods for details). Considering both phase 1 and phase 2 metabolism, and using the data preparation process described in Methods, we collected metabolite data for 560 parent molecules from DrugBank and 1188 parent molecules from MetXBioDB. Of these parent molecules, 310 are identical, not considering stereochemistry, meaning there are 1438 parent molecules total from both sources combined. The metabolites for the overlapping parent molecules were consolidated when forming the reference data set. Within this overlap, 555 of 868 metabolites were present in both data sets. Of the rest, 135 were from DrugBank and 178 from MetXBioDB.

It is relevant to mention here that DrugBank does not contain species annotations for the metabolism data, while MetXBioDB specifies "Human Phase 1" and "Human Phase 2" metabolic reactions. Neither data source includes annotations regarding whether any given metabolite data were collected in an in vivo or an in vitro study.

Beyond noting the amount of overlap between the two data sources, we wanted to examine the chemical space covered by each, in terms of the parent molecules. To the best of our knowledge, such an analysis has not yet been done for MetXBioDB. For DrugBank, an analysis focused specifically on the compounds for which there is metabolite data has also not yet been undertaken. When performing this analysis, we retained the overlapping parent molecules in both data sets.

https://dx.doi.org/10.1021/acs.chemrestox.0c00224
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

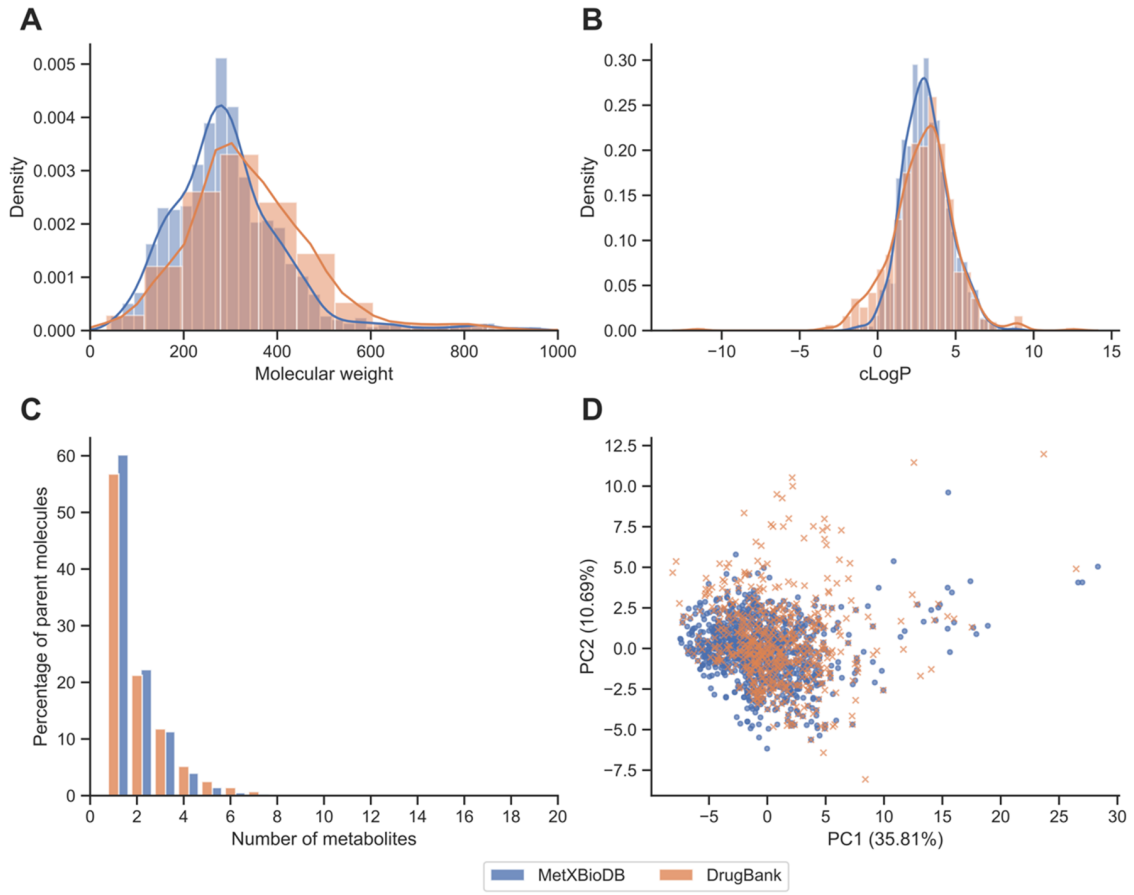


Figure 1. Comparison of the metabolite data from MetXBioDB and DrugBank, in terms of parent molecules. (A) Distribution of molecular weight. (B) Distribution of clog *P*. (C) Histogram of the number of metabolites per parent molecule in terms of percentage of parent molecules. (D) Comparison of the chemical space of the parent molecules from MetXBioDB and DrugBank using PCA calculated using 44 physicochemical descriptors. The percentage of the total variance explained by each of the first two principal components is included in the axis labels.

In terms of molecule size, we observe a narrower distribution among the parent molecules of MetXBioDB than among those of DrugBank, as seen in Figure 1A for molecular weight. In addition, we noted a shift in the distributions, whereby DrugBank has a median molecular weight of 322 while MetXBioDB has a median of only 282. The mean values are not compared due to the presence of an outlier with a molecular weight of 4114 Da (semaglutide) in the DrugBank data. For calculated log *P* (clog *P*) values as well, a narrower distribution is observed for MetXBioDB (Figure 1B). However, for clog *P* the median values of the two distributions are very similar, at 3.04 for DrugBank and 3.05 for MetXBioDB.

In the context of metabolite prediction, it is especially interesting to compare the ratio of parent molecules and metabolites recorded in a data set as this ratio can give an indication of the comprehensiveness of the metabolism data (metabolism data are generally incomplete; more metabolites are typically known for compounds of high relevance, in particular approved drugs). In the case of the DrugBank and MetXBioDB data, the distributions of the number of metabolites per parent molecule are quite similar (Figure 1C). In both cases, the majority of parent molecules have only one known metabolite. At the same time, over 40% of the parent molecules from each data source have multiple metabolites.

Finally, to achieve a visual comparison that takes into account multiple physicochemical properties of the parent molecules, we performed principal component analysis (PCA) on each set of parent molecules using 44 physicochemical descriptors (Figure 1D; see Methods for details). From the PCA we see that there is a large amount of overlap between the two data sets, which is unsurprising given that most of the molecules in the DrugBank data set are also included in the MetXBioDB data set. However, we also see that there are portions of the chemical space populated by parent molecules from DrugBank but not from MetXBioDB, which is consistent with the results from the comparison of the distributions of molecular weight and clog *P*. Inspection of the PCA loading plot (Figure S1) shows that molecule size and polarity seem to play a large role in the variance in the PCA plot. In particular, molecule size seems to influence the first principal component, while polarity seems to influence the second principal component. Interestingly, the five data points (two from DrugBank, three from MetXBioDB) in the far right portion of the PCA plot correspond to the five largest molecules included in the calculation, all of which have a molecular weight between 1000 and 1300 Da (the outlier with a molecular weight of over 4000 Da was not included in the PCA). These five molecules consist of five macrocyclic peptides (including cyclosporine) and one nonmacrocyclic peptide (angiotensin II).

https://dx.doi.org/10.1021/acs.chemrestox.0c00224
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

Whereas the above chemical space analysis included all valid metabolite data from DrugBank and MetXBioDB, a further data preprocessing step was performed for the formation of the final reference data set used for the evaluation of the metabolite structure prediction approach. All metabolism data corresponding to parent molecules contained in the test set were removed from the reference data set. This removal resulted in a final reference data set containing 1420 parent molecules and a total of 2453 metabolites.

The reference data set was further separated into two subsets, corresponding to phase 1 and phase 2 metabolism. The phase 1 subset contains 944 parent molecules and 1763 metabolites, and the phase 2 subset contains 582 parent molecules and their 690 metabolites (Table 1). Most of the phase 1 metabolites are CYP metabolites, and most of the phase 2 metabolites are UGT metabolites (Table 1). Note that some of the phase 2 metabolites do not correspond to any of the listed enzyme families, just as some of the phase 1 metabolites are not formed by CYPs.

Table 1. Composition of the Reference Data Set[a] in Terms of Metabolism Phase and Enzyme Family

	number of metabolites[b]	number of parent molecules
phase 1, all	1763	944
CYP	1640	–
phase 2, all	690	582
UGT	480	–
SULT	92	–
GST	46	–
NAT	34	–
MT	17	–
phase 1 + phase 2	2453	1420

[a]The reference data set was created by combining the DrugBank and MetXBioDB metabolism data and removing the data for all parent molecules contained in the test set. [b]Note that the total numbers of phase 1 and phase 2 metabolites do not equal the sum of the metabolites from the listed enzyme families, because not all metabolites in the data set correspond to these main enzyme families.

The two separate subsets of the reference data set were used to analyze the performance of GLORYx for phase 1 and phase 2 individually, because there are slightly different considerations for each metabolism phase. In addition, the entire reference data set was used to analyze the combined prediction of both phase 1 and phase 2 metabolites.

Note that GLORYx is unable to process two parent molecules in the phase 1 subset of the reference data set and one parent molecule in the phase 2 subset. Both of the phase 1 parent molecules contain a Se atom, which FAME 3 cannot handle (partial charges cannot be calculated; see Methods for a list of allowed element types). Because no SoM predictions can be made, no metabolites are predicted. The parent molecule in the phase 2 subset is unable to be processed because it contains a nitrogen atom with a state that FAME 3 does not recognize. This is the case regardless of which FAME 3 model is used.

Phase 1 Metabolism. The fundamental concept of our approach to predicting metabolites is to integrate machine learning-based SoM prediction in order to score the predicted metabolites. Therefore, the first thing we wanted to know is how GLORYx's SoM probability-based scoring approach compares to the scoring approach used by the state-of-the-art, open source, comprehensive metabolite prediction tool SyGMa.

To compare the scoring approaches, GLORYx was initially implemented using only the phase 1 reaction rules sourced from SyGMa. The phase 1-specific FAME 3 SoM prediction model (model P1) was used to predict SoMs. The predicted metabolites were scored using the maximum SoM probability predicted among all heavy atoms in the mapping onto the parent molecule of the reaction rule that led to the particular predicted metabolite. In this case, the score was therefore equal to this SoM probability; no weighting based on reaction type was used. SyGMa, on the other hand, ranks its predictions based on probability scores that are calculated using the occurrence ratios of each reaction rule in the Metabolite database. Each of SyGMa's predicted metabolites is assigned a probability score corresponding to the reaction rule that formed the predicted metabolite.

Given the same reaction rules, SyGMa with its reaction probability score-based ranking performed slightly better than our SoM probability-based ranking, with an AUC of 0.76 compared to 0.73, respectively, as shown in Figure 2A. This result is reasonable if we suppose that the Metabolite database, which was used to calculate the occurrence ratios for SyGMa's reaction types, was so exhaustive even in its 2001 version (the version used to develop SyGMa) that it contained most of the contents of the current versions of DrugBank and MetXBioDB. This supposition is consistent with the observation in 2013 by Kirchmair et al. that nearly all of the Approved Drugs in DrugBank at the time (1341 out of 1391) were found in the 2011 version of the Metabolite database as top-level substrates.[47] Unfortunately, without access to the Metabolite database, which is currently unavailable, we are unable to perform a comparison ourselves. Nevertheless, it appears that GLORYx achieves a comparable ranking performance to SyGMa.

Phase 1 Metabolism: Combination of Reaction Rules from SyGMa and GLORY. For GLORY, we had developed a set of reaction rules specific to CYP-mediated metabolism.[15] These reaction rules were manually created based on the scientific literature on CYP-mediated reaction types and mechanisms, and each reaction rule received a designation of either common or uncommon reaction type, also according to the literature. SyGMa's phase 1 reaction rules are not separated into CYP and non-CYP rules, so it was of interest to determine whether adding these CYP-specific rules to the phase 1 rules sourced from SyGMa would result in any gains in performance for GLORYx.

When combining the rule sets, the overlap of the rules from the two different sources is handled in a straightforward manner. Duplicate metabolite predictions are combined by retaining the highest priority score. The addition of the CYP reaction rules from GLORY resulted in a substantial jump in recall (portion of known metabolites that were successfully predicted, also known as sensitivity) from 0.72 to 0.84 (Table 2). The precision (portion of predictions that match known metabolites), on the other hand, was halved, as the number of total predicted metabolites more than doubled, from over 10,000 to nearly 25,000. Note that only a fraction of the metabolites generated by organisms is experimentally observed and reported in the scientific literature and databases, for a number of reasons (e.g., lack of chemical stability, low concentrations of metabolites, limitations of the in vitro system, research interest focused on a specific metabolic

https://dx.doi.org/10.1021/acs.chemrestox.0c00224
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

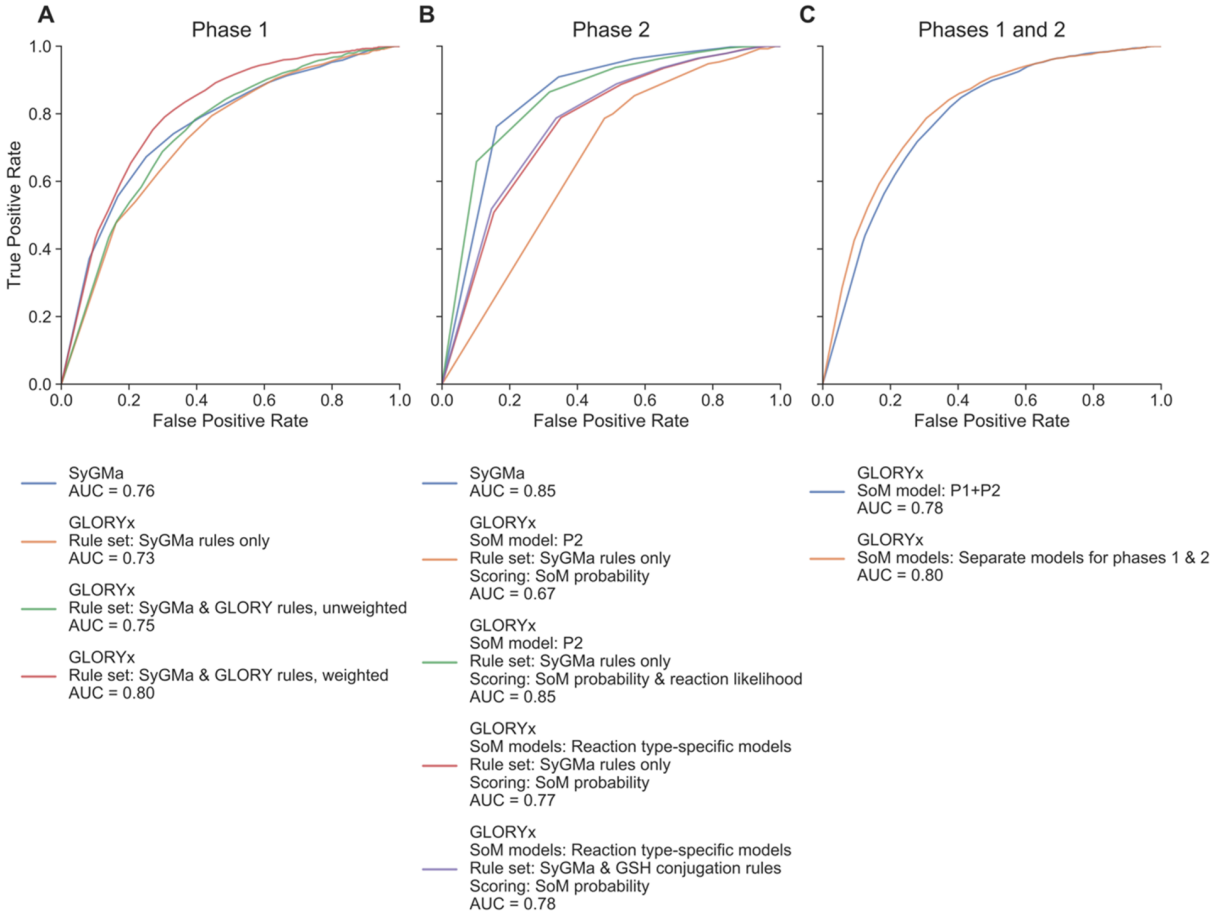


Figure 2. Rank-based ROC curves for the evaluation of metabolite prediction performance on the reference data set. The ranks are calculated based on the priority scores of the predicted metabolites for each parent molecule. (A) Comparison of GLORYx, which scores its predicted metabolites based on predicted SoM probabilities, to SyGMa, which uses reaction probability scores, for phase 1 metabolite prediction. Weighted rules refer to the weighting of the SoM probability-based score based on whether the reaction type is designated common or uncommon. (B) Comparison of the ranking performance of GLORYx with different scoring approaches and rule sets as well as direct comparison to SyGMa's performance, for phase 2 metabolite prediction. The scoring approach that is based on both SoM probability and reaction probability is achieved by a simple multiplication of the two components. (C) Comparison of the ranking performance of GLORYx for combined prediction of metabolites for phases 1 and 2 metabolism, using different SoM prediction approaches to score the predicted metabolites. In both cases, the score is based on predicted SoM probability with weighting according to reaction type, and the rule set is made up of the final phase 1 rule set (SyGMa and GLORY rules) and final phase 2 rule set (SyGMa and GSH conjugation rules).

Table 2. Performance of GLORYx on Predicting Phase 1 Metabolites

	GLORYx using reaction rules from both SyGMa and GLORY	GLORYx using reaction rules from SyGMa only
recall	0.84	0.72
precision	0.060	0.12
total number of predictions	24,906	10,550
number of true positive predictions	1487	1262
AUC (rank-based)	0.80	0.73

enzyme or reaction or metabolite). Therefore, any predicted metabolites that are not "known" should more correctly be considered as putative false positive predictions. Nevertheless, the number of predicted metabolites is enormous, so it is crucial that metabolite prediction methods are able to rank their predictions in a meaningful way.

To examine the ranking performance of GLORYx using the combined rule set, we first used only the SoM probability to score and rank the predicted metabolites, as described above. This nonweighted scoring approach resulted in an AUC of 0.75 (Figure 2A), which was close to SyGMa's AUC of 0.76. Note that even though the sets of predicted metabolites are different in this case, the ranking ability of each approach can still be compared using the ROC curves and AUC. We then applied the concept of weighting reaction rules that we first developed for GLORY, namely applying a simple common vs uncommon distinction between reaction types and generating the priority score for a predicted metabolite by multiplying the SoM probability by a factor corresponding to whether or not the reaction that led to that particular predicted metabolite was designated common (see ref 15 for details). The common–uncommon designations of the reaction rules from GLORY were used unchanged. Then we simply designated all of SyGMa's phase 1 reaction rules as uncommon, based on the following logic: The CYP enzyme family is the most prevalent enzyme family involved in phase 1 metabolism,[4] SyGMa's

https://dx.doi.org/10.1021/acs.chemrestox.0c00224
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

phase 1 reaction rules contain rules for both CYP- and non-CYP-mediated reactions, and our process of combining duplicate predictions by keeping the highest score ensures that any CYP rules from SyGMa that are also "common" rules from GLORY will be in effect scored appropriately as being "common". The result of this weighting of the rule sets was a jump in AUC to 0.80 (Figure 2A).

A similar trend in AUCs for GLORYx in terms of the weighting approach is observed when the ROC curves are calculated based on score rather than rank (Figure S2A). This means that predicted metabolites are compared across different parent molecules in the reference data set in terms of their priority scores. Here, it is important to note that the original publication of SyGMa implied that its score was only intended to be used to compare likelihoods of predicted metabolites of the same parent molecule, and the evaluation in that publication only considered the ranking per parent molecule.[10] This consideration should be kept in mind when viewing all score-based ROC curves for SyGMa throughout this manuscript, which are included for the sake of completeness, especially since the score-based ROC curves for GLORYx tend to yield a higher AUC than the rank-based curves, yet the opposite is true for SyGMa (Figure S2 and Figure 2).

It is also relevant to note that the phase 1 subset of our reference data set is heavily biased toward CYP-mediated metabolism, with over 90% of the metabolites in the data set being CYP metabolites (Table 1). Although CYPs are widely considered the most relevant enzyme family for phase 1 human xenobiotic metabolism, the available data are perhaps even more skewed toward CYP data than would be realistic in humans. Due to the composition of this phase 1 reference data set, it is reasonable that the addition of the CYP-specific rules from GLORY leads to improved performance.

Phase 2 Metabolism. For phase 2 metabolite structure prediction, we again examined the question of how scoring the predicted metabolites based on the SoM probability predicted by FAME 3 compares to SyGMa's scoring approach. Similarly to the phase 1 protocol, the initial comparison was carried out using only the phase 2 reaction rules from SyGMa, along with the general phase 2 SoM prediction model from FAME 3 (model P2), and scoring the predicted metabolites using only the SoM probability predicted by the SoM model. This comparison showed a large difference in ranking performance between SyGMa and our approach (Figure 2B). SyGMa achieved an AUC of 0.85, while our approach, which used the SoM probabilities predicted by the FAME 3 P2 model to rank the predicted metabolites, achieved an AUC of only 0.67.

It therefore appears that SoM probabilities are a surprisingly poor indication of the likelihood of phase 2 metabolism occurring. We know, however, that FAME 3 predicts SoMs corresponding to phase 2 metabolic reactions very well (AUC of 0.97 on a holdout data set consisting of 157 randomly selected compounds with a total of 3476 annotated atoms).[17] The reason for this discrepancy is that multiple predicted metabolites, corresponding to different reaction types, receive the same score because they correspond to the same predicted SoM. Phase 2 metabolic reactions are more specific in terms of functional groups at which they can occur than, for example, CYP-mediated reactions, which makes it easier to predict SoMs but more difficult to predict which reaction type would be more likely to actually occur at a given location. To illustrate this point, consider the case of a hydroxyl group. A hydroxyl group that is a phase 2 SoM could be glucuronidated, sulfated, methylated, or phosphorylated. Another difficult case would be an amine group, which, if it is a phase 2 SoM, could be glucuronidated or N-acetylated. These observations combined with the poor ranking performance indicate that, so far, GLORYx struggles to discriminate between phase 2 reaction types.

In light of this observation and to further investigate the relationship between the predictive capabilities of SoM probabilities and reaction probabilities, we attempted to combine the two scores, since in theory both the SoM and the likelihood of a particular reaction rule compared to other reaction rules that could be applied at a given location are both relevant to the likelihood of the predicted metabolite. We tried two combination approaches: multiplying the reaction probability with the SoM probability and calculating a weighted average. Despite trying various weights (Table S4), a combination score was unable to do better than SyGMa's reaction probability-based scoring approach alone at ranking the predictions. In addition, by varying the weights, it became clear that the more highly the predicted SoM probability was weighted compared to the reaction probability, the worse the ranking performance was (Table S4). The weighted average score combination, using weights up to 10:1, achieved a maximum rank-based AUC of 0.83 (Table S4), whereas multiplying the SoM probability by the reaction probability resulted in a rank-based AUC of 0.85 (Table S4, Figure 2B), which is the same as for SyGMa's reaction probability score alone (however, the shape of the ROC curve is slightly different). These results indicated that the SoM probabilities predicted in this way could not compete with SyGMa's reaction probability scores when it comes to ranking performance.

SyGMa's good ranking performance was to be expected, for the same reasons discussed in the above section on phase 1 metabolism regarding the use of the Metabolite database to develop SyGMa's reaction probability scores. Meanwhile, the poor showing by the SoM probability scoring approach indicates that reactivity is not sufficient to discriminate between the different types of phase 2 reactions, especially not when compared to the data-derived likelihoods of each reaction type. We therefore examined how we could use SoM prediction to achieve a distinction between different reaction types without resorting to precomputed occurrence ratios for the reaction rules.

Phase 2 Metabolism: Reaction Type-Specific SoM Prediction Models. In order to attempt to better predict which reaction type would be more likely at a given SoM, without using SyGMa's reaction probabilities, we developed FAME 3 reaction type-specific SoM prediction models that roughly correspond to the five main phase 2 enzyme families: UGTs, GSTs, SULTs, MTs, and NATs. These models were created using the same training protocol as for the previously published FAME 3 models. Each model was trained on only a subset of the FAME 3 data set, whereby the subsets were selected based on the reaction class annotation in the MetaQSAR database. The reaction classes and the number of molecules used to train each model are provided in Table S3. The 10-fold cross-validation performance of these models was high across the board, with average AUCs all above 0.95 and the average percentage of molecules in which a correct SoM was predicted among the top two atom positions with the highest SoM probabilities (top 2 metric) all above 0.87 (Table 3), despite the relatively small number of molecules used for

https://dx.doi.org/10.1021/acs.chemrestox.0c00224
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

Table 3. Average SoM Prediction Performance of the FAME 3 Reaction Class-Specific Models During Cross-Validation

reaction class	average top 2	average AUC
glucuronidations and glycosylations	0.957	0.988
GSH and RSH[a] conjugations	0.874	0.950
sulfonations	0.966	0.992
methylations	0.877	0.968
acetylations and acylations	0.956	0.992

[a]RSH = protein thiol.

training in each case. Note, however, that these models are trained on atoms, not molecules, so the number of training instances (although not entirely independent from each other) is much larger than the number of molecules.

Because these reaction type-specific SoM models were each trained on only a subset of the molecules that were used to train the general phase 2 SoM model, not all atom types (i.e., Sybyl atom types) were represented in the training data for each individual model, which can then not make predictions for molecules containing these unrepresented atom types. Therefore, these individual reaction type-specific SoM models were used to overrule the predicted SoM probabilities from the general P2 model for the molecules to which they apply rather than as a complete substitute for the general model.

There are a few phase 2 reaction rules in SyGMa that do not correspond to any of the five main phase 2 enzyme families. These rules are simply designated "uncommon", while all other phase 2 reaction rules are designated "common", and the general P2 SoM prediction model is always used to score the products of these uncommon reaction rules.

The general P2 model is also used to score the predicted metabolites corresponding to the individual reaction type-specific SoM models that can not make predictions for a given input molecule. For example, if the SoM model for sulfonation reactions could not make predictions, the predicted metabolites resulting from sulfonation reaction rules are scored using the predicted SoM probabilities from the general P2 model. An illustration of the workflow for predicting phase 2 metabolites using the reaction type-specific SoM models for scoring is shown in Figure 3.

In this way, the same metabolites are predicted as if only the general P2 model was used, but the reaction type-specific scoring approach results in different ranks of the metabolites and, perhaps most importantly, a drastic reduction in the number of tied ranks for predicted metabolites of a single parent molecule.

Using the individual reaction type-specific phase 2 SoM models to score the predicted metabolites resulted in a large improvement in the ranking, with an AUC of 0.77 compared to an AUC of 0.67 using the general P2 model (Figure 2B) and only the reaction rules sourced from SyGMa for comparison. Similarly, the score-based AUC increased from 0.66 to 0.79 upon implementation of the reaction type-specific SoM models (Figure S2). Unfortunately, even using the reaction type-specific SoM prediction models resulted in a ranking performance that was worse than SyGMa's (AUC of 0.77 compared to 0.85, respectively). However, as discussed for phase 1 metabolism above, SyGMa's approach has the advantage of having derived its scoring approach directly based on, in effect, all available metabolism data from a comprehensive but not freely available database. Meanwhile, the difficulty of using SoM prediction for phase 2 metabolism appears to be that there are relatively few potential SoMs, but that the atom environments may not be specific enough to differentiate between different types of reactions.

Based on these results, we therefore use the individual reaction type-specific phase 2 SoM models to score the phase 2 metabolites predicted in all subsequent sections of this manuscript.

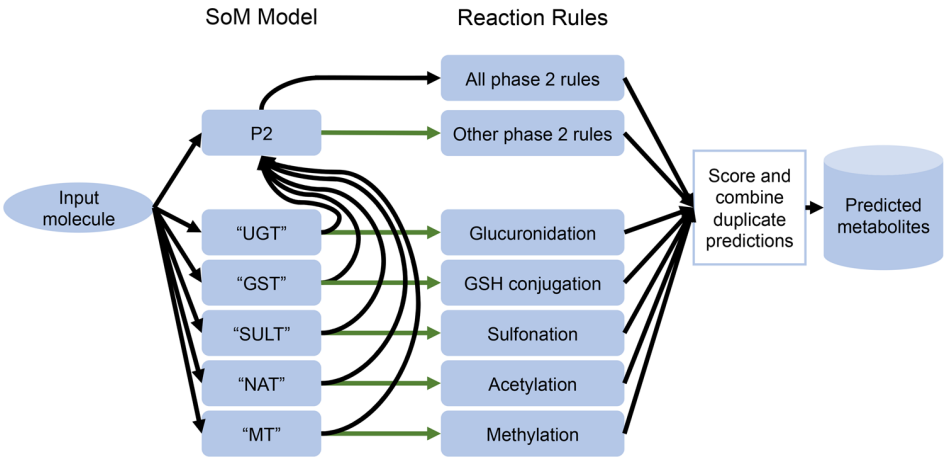


Figure 3. Workflow of phase 2 metabolite prediction using reaction type-specific SoM models to score and rank the predicted metabolites. The reaction type-specific SoM models ("UGT", "GST", "SULT", "NAT", "MT") are used instead of the general phase 2 SoM model (P2) to score the products of the relevant reactions for all molecules in which all of the reaction type-specific models are able to make a prediction. The green arrows indicate the molecules that were predicted successfully by the relevant reaction type-specific SoM model. If one or more of the reaction type-specific models cannot make predictions for a given molecule, then that molecule additionally follows the path of the black arrows, followed by a deduplication of predictions. The "UGT" model covers glucuronidation and glycosylation reactions, the "GST" model covers GSH and RSH conjugations, the "SULT" model covers sulfonations, the "NAT" model covers acetylations and acylations, and the "MT" model covers methylation reactions. The "other phase 2 rules" refer to the rules that are neither glucuronidation, GSH conjugation, sulfonation, acetylation, or methylation rules.

https://dx.doi.org/10.1021/acs.chemrestox.0c00224
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

Phase 2 Metabolism: Addition of GSH Conjugation Reaction Rules. We found that the reaction rules sourced from SyGMa do not contain any GSH conjugation reactions, which correspond to the GST enzyme family, one of the five main enzyme families for phase 2 xenobiotic metabolism. We therefore developed a set of GSH conjugation reaction rules based on descriptions of GSH conjugation metabolic reactions in the scientific literature. This resulted in nine new reaction rules.

When only the reaction rules sourced from SyGMa were implemented, GLORYx achieved a recall of 0.78 and a precision of 0.21 (Table 4). When the new GSH conjugation reaction rules were added, GLORYx achieved a recall of 0.80 with the same precision, because only 141 more metabolites were predicted in total (Table 4). Though we do not see a very large improvement in performance on the reference data set after adding these GSH conjugation reaction rules to GLORYx, we believe that this addition is actually meaningful for the purpose of metabolite structure prediction in the real world, because GSTs are actually the second most relevant phase 2 enzyme family for xenobiotic metabolism in terms of number of metabolites formed.[4]

Table 4. Performance of GLORYx on Predicting Phase 2 Metabolites

	GLORYx using SyGMa rules only	GLORYx using SyGMa rules plus GSH conjugation rules
recall	0.78	0.80
precision	0.21	0.21
total number of predictions	2509	2650
number of true positive predictions	539	555
AUC (rank-based)	0.77	0.78

As was expected based on the relatively low number of GST-mediated metabolites in the reference data set (Table 1), the ranking performance remained similar upon the addition of GSH conjugation reaction rules (AUC of 0.77 and 0.78, respectively; Figure 2B). This comparable performance seems to suggest that the products of the new GSH conjugation reaction rules are scored in a meaningful way based on the corresponding reaction type-specific SoM prediction model.

Combined Phase 1 and Phase 2. A general use case of predicting "all" possible metabolites at once was also considered. FAME 3 provides one model, "P1 + P2", that predicts all SoMs from both phases of metabolism. For this use case, we therefore examined whether it makes sense to use the P1 + P2 FAME 3 model's predictions to score the predicted metabolites or to use the separate models, as determined separately for phase 1 and phase 2 (see sections Phase 1 Metabolism and Phase 2 Metabolism), and combine the predictions. The predicted metabolite structures are the same in both cases; what changes is their scores, since those are based on the predicted SoM probabilities.

Using separate SoM prediction models for the two phases did provide a slight advantage in terms of the ranking performance, with an improvement in AUC from 0.78 to 0.80 compared to using the P1 + P2 SoM model, as shown in Figure 2C. An improvement of the same amount is seen in the AUCs of the score-based ROC curves (AUC increased from 0.79 to 0.81; Figure S2C). Although this advantage appears small at first glance, it is important to recall the composition of the reference data set. This data set contains more than twice as much phase 1 data as phase 2 data, in terms of number of known metabolites, which may cause the benefit of using separate SoM prediction models for the two phases to be underrepresented by this analysis. Based on these considerations along with the ROC curves, we conclude that the multimodel approach should be used for optimal performance, and we use this approach in the validation on the test data set (see section Performance on a Manually Curated Test Data Set).

Performance on a Manually Curated Test Data Set. The performance of the final version of GLORYx was evaluated on a manually curated test data set consisting of 37 parent molecules that were among the top 100 best-selling drugs in 2018. For these parent molecules, the data set contains a total of 136 first-generation metabolites, which equates to an average of 3.7 known metabolites per parent molecule. This test data set does not contain enzyme or metabolism phase annotations, so the evaluation was carried out from the perspective of predicting all possible metabolites, from both phase 1 and phase 2 metabolism.

GLORYx was able to predict 77% of the known metabolites in the test data set, which is higher than SyGMa's recall of 68% (Table 5). In conjunction with this higher recall, GLORYx had a lower precision than SyGMa (0.061 compared to 0.12, respectively), which is unsurprising given that GLORYx contains many more reaction rules than SyGMa due to the addition of the CYP metabolism rules from GLORY and the new GSH conjugation rules. The total number of metabolites predicted by GLORYx was nearly double the number predicted by SyGMa. However, SyGMa's precision of 0.12 was also very low, due to a relatively large number of predictions (800 total). Another potential contribution to the low precision of both tools is that experimentally determined metabolites whose structures have not been fully defined were not included in the test data set. It is possible that this aspect of the data set has an effect on the number of false positive predictions, which would have an effect on the precision as well.

Table 5. Performance of GLORYx and SyGMa on the Test Data Set of 37 Parent Compounds and Their 136 Metabolites

	GLORYx	SyGMa
recall	0.77	0.68
precision	0.061	0.12
total number of predictions	1724	800
number of true positives (out of 136)	105	93
AUC (rank-based)	0.79	0.74

The relatively large number of predictions made by both SyGMa and GLORYx is a general problem that is shared by all available metabolite structure prediction approaches.[48] This phenomenon clearly underlines the need to have a meaningful way to rank the predicted metabolites. In our case in particular, neither SyGMa nor GLORYx has sufficiently high precision to be used without ranking the predicted metabolites.

In terms of the ability to rank the predicted metabolites, GLORYx showed better performance than SyGMa, as indicated by the ROC curves shown in Figure 4. The AUC of the rank-based ROC curve was 0.79, compared to 0.74 for

https://dx.doi.org/10.1021/acs.chemrestox.0c00224
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

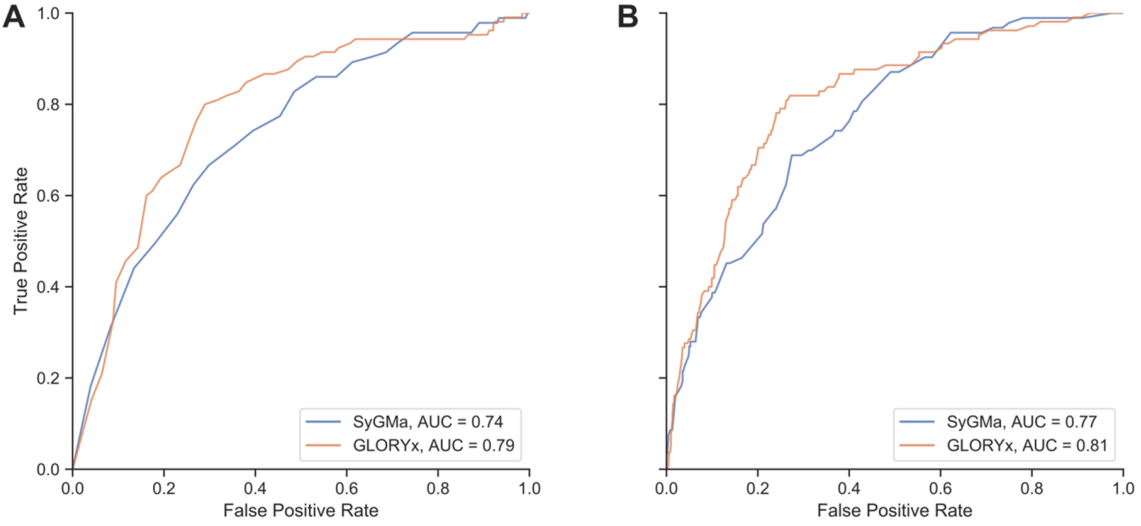


Figure 4. ROC curves for GLORYx and SyGMa representing ranking performance on the test set based on the (A) ranks and (B) scores of the predicted metabolites.

SyGMa. In addition, GLORYx's priority score seems to be a meaningful score in and of itself, not just for ranking the predictions for individual parent molecules separately, because the ROC curve and AUC were actually slightly better using the score than they were using the rank (0.81 compared to 0.79, respectively; Figure 4). For SyGMa, the AUC of the score-based approach was also higher than that of the rank-based approach (0.77 compared to 0.74); however, it is important to note that SyGMa's score was most likely not intended to be used to compare predicted metabolites from different parent compounds (see section Phase 1 Metabolism: Combination of Reaction Rules from SyGMa and GLORY).

To get an idea of the variability in the ranking performance on the test data set, we calculated the AUCs while systematically removing one parent molecule at a time from the data set. This resulted in 37 different AUCs for each tool and AUC type (rank-based or score-based), which are plotted in Figure S3. From this analysis, we observed a similar amount of variability in the AUCs between the two tools with the different metrics (score based and rank based). In all cases, the median AUCs from this analysis were within 0.01 of the AUCs reported in Figure 4. From the outliers observed in Figure S3, we learn that two to three molecules are particularly challenging for each tool. In the case of GLORYx, the two outliers correspond to having excluded everolimus (the furthest outlier) and budesonide for both the score-based and rank-based AUCs. For SyGMa, the furthest outlier is also caused by the exclusion of everolimus, while the second-furthest outlier corresponds to having excluded darunavir. Everolimus is a macrocycle with 12 known metabolites in the test data set, while budesonide and darunavir each have 6 known metabolites.

Overall, a clear difference in performance is observed between the two tools, with GLORYx outperforming SyGMa in both cases. The improvement in ranking performance seems to indicate that combining predicted SoM probabilities with reaction rules to score the predicted metabolites, whereby the SoM model and the reaction rules correspond to the same type(s) of reactions, provides very valuable information. This approach also has the benefit of not relying on reaction rule occurrence ratios based on existing metabolism data to score and rank the predictions. Our reference data set was used to measure performance during development but was not used to develop reaction rules or calculate occurrence ratios. This difference could potentially make GLORYx more flexible with regard to never-before-seen input molecules.

CONCLUSION

GLORYx is a new tool for predicting the structures of metabolites formed by both phase 1 and phase 2 metabolic reactions in humans. The tool utilizes FAME 3 to predict, for all atom positions in a molecule, the likelihood of a biotransformation to take place at this position and, based on these predictions, applies a set of reaction rules to generate and rank likely metabolites.

In conjunction with a high recall of known metabolites (77% on the test data set), GLORYx ranked the predicted metabolites with an AUC of 0.79 on the manually curated test data set. This recall and ranking performance is better than we observed for the established, freely available tool SyGMa on the same data set. However, when considering only phase 2 metabolite prediction, SyGMa's ranking performance was better than that of GLORYx.

We have observed that it is difficult to predict phase 2 metabolites, that is, difficult to rank the predicted metabolites in a meaningful way, based on predicted SoM probabilities despite high performance of the SoM prediction models themselves. We have concluded that the cause of this difficulty is that reactivity is an insufficient metric for determining which type of conjugation reaction would be more likely to occur at a particular atom position. We were able to mitigate this problem substantially by using individual reaction type-specific SoM prediction models corresponding roughly to the five main phase 2 enzyme families.

During each run of GLORYx, the algorithm generates and ranks one generation of metabolites based on the parent compound(s) provided. Users may of course provide (predicted) metabolites as input to GLORYx, hence enabling multigeneration metabolite prediction.

Given the scarcity of the available high-quality data on small-molecule metabolism, it is difficult to provide a robust definition of the applicability domain of GLORYx. However,

https://dx.doi.org/10.1021/acs.chemrestox.0c00224
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

we know from thorough analyses of FAME 3 that the metabolic properties of the atoms in a molecule are first and foremost determined by the proximate atom environment, and these environments are much more redundant across the chemical space than the overall (global) structure of molecules. Considering also that the reaction rules implemented in GLORYx are based on only a few connected atoms, GLORYx is expected to provide reliable results for a wide range of synthetic compounds and natural products alike.

GLORYx is freely available as a web server at https://nerdd.zbh.uni-hamburg.de/ and is also provided as a software package upon request. Note that GLORYx should be considered an extension of GLORY rather than a replacement. Hence both tools are available on the Web site, to enable users to choose between CYP-specific metabolite structure prediction with GLORY and comprehensive phase 1 and phase 2 metabolite structure prediction with GLORYx.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemrestox.0c00224.

Additional tables and figures (PDF)

■ AUTHOR INFORMATION

Corresponding Author

Johannes Kirchmair − *Department of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria; Center for Bioinformatics (ZBH), Department of Informatics, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, 20146 Hamburg, Germany;* orcid.org/0000-0003-2667-5877; Phone: +43 1-4277-55104; Email: johannes.kirchmair@univie.ac.at

Authors

Christina de Bruyn Kops − *Center for Bioinformatics (ZBH), Department of Informatics, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, 20146 Hamburg, Germany;* orcid.org/0000-0001-8890-2137

Martin Šícho − *CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, 166 28 Prague 6, Czech Republic;* orcid.org/0000-0002-8771-1731

Angelica Mazzolari − *Facoltà di Scienze del Farmaco, Dipartimento di Scienze Farmaceutiche "Pietro Pratesi", Università degli Studi di Milano, I-20133 Milan, Italy;* orcid.org/0000-0003-1352-1094

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemrestox.0c00224

Funding

C.d.B.K. is funded by a Ph.D. completion scholarship from the Universität Hamburg. M.S. was supported from the Ministry of Education, Youth and Sports of the Czech Republic, project number LM2018130.

Notes

The authors declare no competing financial interest.

The reference and test data sets as well as all the reaction rules from this work are freely available at https://github.com/christinadebruynkops/GLORYx.

■ ACKNOWLEDGMENTS

We would like to thank Nina Jeliazkova and Nikolay Kochev from Idea Consult, Ltd. of Sofia, Bulgaria for discussions about using Ambit-SMIRKS for metabolite structure prediction. We also thank Conrad Stork from the Center of Bioinformatics (ZBH) of the Universität Hamburg for technical support with the Web service. We further acknowledge the authors of BioTransformer for their efforts in assembling the MetXBioDB, which we used as one of the data sources for the compilation of our reference dataset.

■ ABBREVIATIONS

AUC, area under the receiver operating characteristic curve; clog *P*, calculated log *P*; CYP, cytochrome P450; EC, Enzyme Commission; GSH, glutathione; GST, glutathione *S*-transferase; MT, methyltransferase; NAT, *N*-acetyltransferase; ROC, receiver operating characteristic; SoM, site of metabolism; SULT, sulfotransferase; UGT, UDP-glucuronosyltransferase

■ REFERENCES

(1) Kirchmair, J., Göller, A. H., Lang, D., Kunze, J., Testa, B., Wilson, I. D., Glen, R. C., and Schneider, G. (2015) Predicting Drug Metabolism: Experiment and/or Computation? *Nat. Rev. Drug Discovery 14* (6), 387−404.

(2) Pedretti, A., Mazzolari, A., Vistoli, G., and Testa, B. (2020) *MetaQSAR Database, version MetaQSAR_05_2020.*

(3) Pedretti, A., Mazzolari, A., Vistoli, G., and Testa, B. (2018) MetaQSAR: An Integrated Database Engine to Manage and Analyze Metabolic Data. *J. Med. Chem. 61* (3), 1019−1030.

(4) Testa, B., Pedretti, A., and Vistoli, G. (2012) Reactions and Enzymes in the Metabolism of Drugs and Other Xenobiotics. *Drug Discovery Today 17* (11−12), 549−560.

(5) Jancova, P., Anzenbacher, P., and Anzenbacherova, E. (2010) Phase II Drug Metabolizing Enzymes. *Biomed. Pap. 154* (2), 103−116.

(6) Kirchmair, J., Williamson, M. J., Tyzack, J. D., Tan, L., Bond, P. J., Bender, A., and Glen, R. C. (2012) Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms. *J. Chem. Inf. Model. 52* (3), 617−648.

(7) Marchant, C. A., Briggs, K. A., and Long, A. (2008) In Silico Tools for Sharing Data and Knowledge on Toxicity and Metabolism: Derek for Windows, Meteor, and Vitic. *Toxicol. Mech. Methods 18* (2−3), 177−187.

(8) Mekenyan, O. G., Dimitrov, S. D., Pavlov, T. S., and Veith, G. D. (2004) A Systematic Approach to Simulating Metabolism in Computational Toxicology. I. The TIMES Heuristic Modelling Framework. *Curr. Pharm. Des. 10* (11), 1273−1293.

(9) Darvas, F. (1987) Metabolexpert: An Expert System for Predicting Metabolism of Substances. In *QSAR in Environmental Toxicology - II* (Kaiser, K. L., Ed.), pp 71−81, Springer, Dordrecht.

(10) Ridder, L., and Wagener, M. (2008) SyGMa: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites. *ChemMedChem 3* (5), 821−832.

(11) (2001) *MDL Metabolite Database*, Elsevier, Amsterdam.

(12) Djoumbou-Feunang, Y., Fiamoncini, J., Gil-de-la-Fuente, A., Greiner, R., Manach, C., and Wishart, D. S. (2019) BioTransformer: A Comprehensive Computational Tool for Small Molecule Metabolism Prediction and Metabolite Identification. *J. Cheminf. 11* (1), 2.

(13) Rudik, A. V., Bezhentsev, V. M., Dmitriev, A. V., Druzhilovskiy, D. S., Lagunin, A. A., Filimonov, D. A., and Poroikov, V. V. (2017) MetaTox: Web Application for Predicting Structure and Toxicity of Xenobiotics' Metabolites. *J. Chem. Inf. Model. 57* (4), 638−642.

(14) Rudik, A., Bezhentsev, V., Dmitriev, A., Lagunin, A., Filimonov, D., and Poroikov, V. (2019) Metatox - Web Application for

Generation of Metabolic Pathways and Toxicity Estimation. *J. Bioinf. Comput. Biol.* *17* (1), 1940001.
(15) de Bruyn Kops, C., Stork, C., Šícho, M., Kochev, N., Svozil, D., Jeliazkova, N., and Kirchmair, J. (2019) GLORY: Generator of the Structures of Likely Cytochrome P450 Metabolites Based on Predicted Sites of Metabolism. *Front. Chem.* *7*, 402.
(16) Šícho, M., de Bruyn Kops, C., Stork, C., Svozil, D., and Kirchmair, J. (2017) FAME 2: Simple and Effective Machine Learning Model of Cytochrome P450 Regioselectivity. *J. Chem. Inf. Model.* *57* (8), 1832−1846.
(17) Šícho, M., Stork, C., Mazzolari, A., de Bruyn Kops, C., Pedretti, A., Testa, B., Vistoli, G., Svozil, D., and Kirchmair, J. (2019) FAME 3: Predicting the Sites of Metabolism in Synthetic Compounds and Natural Products for Phase 1 and Phase 2 Metabolic Enzymes. *J. Chem. Inf. Model.* *59* (8), 3400−3412.
(18) *ADMET Predictor Metabolism Module, SimulationsPlus*. Simulations-Plus.com/software/admetpredictor/metabolism/ (accessed 2020/05/07).
(19) Tyzack, J. D., Hunt, P. A., and Segall, M. D. (2016) Predicting Regioselectivity and Lability of Cytochrome P450 Metabolism Using Quantum Mechanical Simulations. *J. Chem. Inf. Model.* *56* (11), 2180−2193.
(20) Cruciani, G., Carosati, E., De Boeck, B., Ethirajulu, K., Mackie, C., Howe, T., and Vianello, R. (2005) MetaSite: Understanding Metabolism in Human Cytochromes from the Perspective of the Chemist. *J. Med. Chem.* *48* (22), 6970−6979.
(21) *DrugBank*, version 5.1.4. https://www.drugbank.ca/ (accessed 2019/09/26).
(22) Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C., and Wilson, M. (2018) DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* *46* (D1), D1074−D1082.
(23) *MetXBioDB*, version 1.0, https://bitbucket.org/djoumbou/biotransformer/src/master/ (accessed 2019/01/11).
(24) *Top 200 Brand Name Drugs by Retail Sales in 2018*, The Njardarson Group at the University of Arizona, Tucson, AZ.https://njardarson.lab.arizona.edu/content/top-Pharmaceuticals-Poster (accessed 2019/09/12).
(25) McGrath, N. A., Brichacek, M., and Njardarson, J. T. (2010) A Graphical Journey of Innovative Organic Architectures That Have Improved Our Lives. *J. Chem. Educ.* *87* (12), 1348−1349.
(26) Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I., and Leach, A. R. (2017) The ChEMBL Database in 2017. *Nucleic Acids Res.* *45* (D1), D945−D954.
(27) *ChEMBL Database*. https://www.ebi.ac.uk/chembl/ (accessed 2020/02/21).
(28) (2017) *MarvinSketch*, version 17.1.9, ChemAxon, Budapest, Hungary. http://www.chemaxon.com (accessed 2017/10/01).
(29) (2018) *RDKit: Open-Source Cheminformatics*, version 2018.09.1.
(30) Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* *12*, 2825−2830.
(31) *Scikit-Learn: Machine Learning in Python*, version 0.20.1.
(32) (2018) *Molecular Operating Environment (MOE)*, version 2018.01, Chemical Computing Group, Montreal, QC, Canada.
(33) *SMIRKS - A Reaction Transform Language*, Daylight, Laguna Niguel, CA. https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html (accessed 2020/05/08).
(34) *SyGMa*, version 1.1.0.
(35) Allocati, N., Masulli, M., Di Ilio, C., and Federici, L. (2018) Glutathione Transferases: Substrates, Inhibitors and pro-Drugs in Cancer and Neurodegenerative Diseases. *Oncogenesis* *7* (1), 8.
(36) van Bladeren, P. J. (2000) Glutathione Conjugation as a Bioactivation Reaction. *Chem.-Biol. Interact.* *129* (1−2), 61−76.
(37) Awasthi, Y. C. (2006) *Toxicology of Glutathione Transferases*, CRC Press, Boca Raton, FL.
(38) Armstrong, R. N. (1991) Glutathione S-Transferases: Reaction Mechanism, Structure, and Function. *Chem. Res. Toxicol.* *4* (2), 131−140.
(39) Zhang, Y., den Braver-Sewradj, S. P., den Braver, M. W., Hiemstra, S., Vermeulen, N. P. E., van de Water, B., Commandeur, J. N. M., and Vos, J. C. (2018) Glutathione S-Transferase P1 Protects Against Amodiaquine Quinoneimines-Induced Cytotoxicity but Does Not Prevent Activation of Endoplasmic Reticulum Stress in HepG2 Cells. *Front. Pharmacol.* *9*, 388.
(40) MacFaul, P. A., Morley, A. D., and Crawford, J. J. (2009) A Simple in Vitro Assay for Assessing the Reactivity of Nitrile Containing Compounds. *Bioorg. Med. Chem. Lett.* *19* (4), 1136−1138.
(41) Willighagen, E. L., Mayfield, J. W., Alvarsson, J., Berg, A., Carlsson, L., Jeliazkova, N., Kuhn, S., Pluskal, T., Rojas-Chertó, M., Spjuth, O., Torrance, G., Evelo, C. T., Guha, R., and Steinbeck, C. (2017) The Chemistry Development Kit (CDK) v2.0: Atom Typing, Depiction, Molecular Formulas, and Substructure Searching. *J. Cheminf.* *9* (1), 33.
(42) *Chemistry Development Kit*, version 2.0.
(43) Pedretti, A., Mazzolari, A., Vistoli, G., and Testa, B. (2018) *MetaQSAR Database* (snapshot from March 15, 2018).
(44) *Ambit-SMARTS Java Library*, version 3.1.0.
(45) Kochev, N., Avramova, S., and Jeliazkova, N. (2018) Ambit-SMIRKS: A Software Module for Reaction Representation, Reaction Search and Structure Transformation. *J. Cheminf.* *10* (1), 42.
(46) Guengerich, F. P. (2001) Common and Uncommon Cytochrome P450 Reactions Related to Metabolism and Chemical Toxicity. *Chem. Res. Toxicol.* *14* (6), 611−650.
(47) Kirchmair, J., Howlett, A., Peironcely, J. E., Murrell, D. S., Williamson, M. J., Adams, S. E., Hankemeier, T., van Buren, L., Duchateau, G., Klaffke, W., and Glen, R. C. (2013) How Do Metabolites Differ from Their Parent Molecules and How Are They Excreted? *J. Chem. Inf. Model.* *53* (2), 354−367.
(48) Judson, P. N. (2014) Knowledge-Based Approaches for Predicting the Sites and Products of Metabolism. In *Drug Metabolism Prediction*; Kirchmair, J., Ed.; Wiley-VCH, Weinheim, Germany; pp 293−318.

https://dx.doi.org/10.1021/acs.chemrestox.0c00224
Chem. Res. Toxicol. XXXX, XXX, XXX−XXX

Discussion

Xenobiotic metabolism plays a huge role in the safety and efficacy of all the various chemical entities that are so important for our 21st-century lifestyle. It is therefore important to study and predict xenobiotic metabolism of drugs, cosmetics, agrochemicals, and other chemical concoctions. SoM prediction and metabolite structure prediction are two aspects of xenobiotic metabolism that can be predicted computationally.

5.1 Exploring Alignment-Based SoM Prediction

Reactivity and accessibility are the two key components of SoM prediction, as discussed in section 1.2.1. The alignment-based approach to SoM prediction explored in this dissertation (section 4.1) is in effect a ligand-based approach to modeling accessibility. The reference molecule with the known SoM provides information about accessibility, following the assumption that the alignment to the reference molecule indicates the orientation with which the query molecule would bind to the metabolizing enzyme.

Using molecular alignment and a public CYP SoM dataset of 680 molecules with newly corrected stereochemistry, a known SoM was predicted among the top three ranked atoms (top-3) for 64% of the molecules on average. In terms of other ranking-based metrics, an average AUC of 0.69 and an average BEDROC ($\alpha = 80.5$) of 0.50 were achieved.

In contrast, the machine-learning based method FAME 2 [A4], which used the same dataset as the alignment-based method but with an 80:20 training-testing split, achieved top-3 values above 90% on the independent test set (i.e. 20% of the total dataset). Even though the other performance metrics were different for each of the two approaches and cannot be compared, it is clear that FAME 2 and the alignment-only approach are not in the same league in terms of performance.

Nevertheless, in the current age of popular black-box machine learning methods, our alignment-based approach could offer refreshing clarity and insight into the reason for a particular prediction. Although machine learning methods such as FAME 2 perform better at the task of predicting SoMs, the particular insights that this alignment-based approach provides are related to the potential orientation of molecules in the binding cite of the metabolizing enzyme (CYPs in this case), which could be valuable information. However, because the performance of the alignment approach on its own was not satisfactory, we also considered the other key component of SoM prediction: reactivity.

In order to predict reactivity, we extracted the reactivity component from the predictions made by SMARTCyp [64]. In addition to the reactivity component derived from DFT calculations, SMARTCyp contains an atom accessibility component as well. For our purposes, however, we wanted to use only the reactivity component of the SMARTCyp predictions, based on the assumption that the alignment implicitly takes accessibility into account. This reactivity-only prediction approach resulted in a top-3 value of 79.6% on average, an AUC of 0.80 on average, and a BEDROC ($\alpha = 80.5$) of 0.54 on average. Unsurprisingly, reactivity on its own was shown to be a better predictor of SoMs than alignment on its own. Combining reactivity and alignment resulted in a comparable average top-3 value of 80.3%, an improved AUC of 0.85 on average, and an improved BEDROC ($\alpha = 80.5$) of 0.73 on average. So, based on the ROC curve, AUC, and BEDROC metrics, it appears that adding the alignment component to the reactivity component resulted in better early enrichment and overall ranking of the heavy atoms in the molecules in terms of their likelihood of being SoMs. Interestingly, this combination method resulted in a slightly better performance than the full SMARTCyp as well, in which the reactivity was combined with an atom accessibility component. SMARTCyp achieved a top-3 of 82.6%, a BEDROC ($\alpha = 80.5$) of 0.65, and an AUC of 0.82 on average. Of the measured metrics, only top-k did not show a benefit of the alignment-reactivity approach compared to SMARTCyp.

A further finding of this study was that the alignment-based approach performed best on query molecules for which there was a similar molecule in the reference database. The molecular similarity was measured using fingerprints and was therefore a 2D similarity. We found that performance was quite poor for query molecules with high dissimilarity to all of the reference molecules. This finding

indicates a limitation of the alignment-based approach, which was not shared by the reactivity-only approach. The combination of alignment and reactivity was affected less by molecular similarity than the alignment-only approach, but still appeared to be affected.

Machine learning approaches that use, for example, atom environments to predict SoMs are more able to avoid this problem because the same atom environments may be found across highly dissimilar molecules [A8]. In this respect, machine learning-based methods for SoM prediction are able to be more robust than the alignment-based approach. This robustness seems to be the case, at least for FAME 2 [A4] and FAME 3 [A8], because of the use of atom environments as descriptors.

Due to the limitations of the alignment-based approach to SoM prediction in terms of performance and robustness, this method's primary potential for further use would be in combination with another approach. One possibility would be to devise alignment-based descriptors to use in the development of a machine learning model, combined with other descriptors. Such an approach would require some sort of set of reference molecules to which both the training and test set molecules would be aligned. The SoMs of the top several best-aligned reference molecules could then potentially be used as descriptors, in combination with the alignment score for each of these reference molecules. This approach would lend itself more to making predictions for a specific metabolic enzyme or enzyme family, as the alignment descriptors would depend on the availability of relevant reference molecules. It would be interesting to see whether alignment-based descriptors could lead to improved performance of a machine learning model.

Finally, this work has provided an additional benefit. This study used the Zaretzki dataset [55], which is one of the most comprehensive public datasets for CYP SoMs. However, this dataset was lacking correct stereochemistry information and therefore had to be revised in order to use it to validate the alignment-based SoM prediction approach. The revised dataset was provided to the public in the supporting information for [D1]. The hope is that the revised version of this dataset, with the stereochemistry included, may be useful to other scientists for further development and validation of methods for predicting SoMs.

5.2 Metabolite Structure Prediction

Metabolite prediction is an interesting problem, because although much is known about metabolic reactions that can occur in humans, the number of potential reactions that could occur for any given molecule is in most cases much higher than the number of metabolites that are actually experimentally observed for that molecule. This is the difficulty facing developers of tools for predicting metabolites. In knowledge-based approaches, which take a set of reaction rules and apply them to a given input molecule, the number of predicted metabolites can easily be cumbersomely large.

Yet the concept of SoMs still applies. In theory, the locations in the molecule at which the metabolic reactions will occur depend on the reactivity and accessibility of the atoms. Therefore, it seemed reasonable to attempt to predict the SoMs and subsequently only apply the reaction rules at the SoMs. When we attempted to reduce the putative false positive predictions in this manner, however, we found that this approach doubled the precision but decreased the recall. On a test set of CYP metabolites, the recall decreased by 0.19, from 0.83 to 0.64, compared to not limiting the application of the reaction rules. The recall of 0.64 was lower than the recalls of SyGMa and BioTransformer, which were 0.74 and 0.72 on the same dataset, respectively. An important caveat here is that SyGMa does not predict only CYP metabolites but rather phase I metabolites in general, which should be kept in mind for comparisons to SyGMa. Meanwhile, the precision was comparable to that of SyGMa and BioTransformer, at 0.16 compared to 0.15 and 0.17, respectively, all of which are low precisions. These results occurred after trying different thresholds for SoM cutoffs and attempting to balance the improvement in precision with the loss of recall.

Using predicted SoMs as a hard filter in this way therefore seems to not be worth the cost of decreased recall. The main use case of metabolite structure prediction methods would be to predict all of the metabolites that can be experimentally determined but to predict as few additional metabolites as possible. At the same time, we know that the SoM prediction method that we used, FAME 2, had high performance on an independent SoM test set and seems to be one of the leading SoM prediction methods in terms of performance. It is therefore unlikely that the

particular SoM prediction method we used was the problem.

Because this first approach to using SoM prediction to reduce the total number of predicted metabolites was not satisfactory in terms of recall, we subsequently developed a new scoring function to use the predicted SoMs to provide direction towards the most likely predicted metabolites. We believe that having a meaningful way to score and rank the predicted metabolites is of utmost importance in general, since the precision even of the hard-filter approach, as well as of SyGMa and BioTransformer, is extremely low.

When developing the scoring approach, we wanted to focus on reproducibility and ease of comprehension. Other metabolite prediction tools such as SyGMa use occurrence ratios derived from a dataset, and in SyGMa's case the dataset is no longer available, which limits the possibilities for reproducing and developing the method further. We would like to enable other researchers to continue to use and develop our methods, and we also wanted to attempt to reduce bias based on any one dataset as much as possible. Of course, it is not possible to eliminate bias completely, as we are always limited by the existing experimental data. However, we wanted to see if we could avoid using occurrence ratios to score the predicted metabolites and still achieve a meaningful ranking of the predictions. We were able to accomplish this by paying attention during the development of the reaction rules to whether each reaction type is a common CYP reaction or a more unusual one. These two categories of CYP reactions were clearly delineated in a detailed review by F. P. Guengerich published in 2001 [28], from which we took the nomenclature for our binary distinction between "common" and "uncommon" reaction types. Some reaction rules were developed based on additional CYP-mediated reactions described in other publications, and for these reaction rules the proper category was determined according to literature-based knowledge of CYP-mediated metabolism. The final reaction rule set developed from combing the scientific literature consisted of 61 reaction types described by 73 reaction rules.

We then devised a scoring function that weighted common reactions more highly than uncommon reactions. After using a reference dataset to compare the scoring and ranking performance using different weights, we determined that a 5:1 common:uncommon weighting scheme worked best. During evaluation on the test set, the rank-based ROC curve showed better early enrichment than the ROC curve for SyGMa, and the AUC was higher for our approach as well, at 0.66 compared to

0.50, respectively. These ROC curves included false negative predictions for each method, i.e. known metabolites that were not predicted, in order to attempt to take into account the difference in missed known metabolites between our method and SyGMa. However, this choice resulted in lower AUC values than would typically be expected and is therefore mentioned here as a caveat when reporting the AUC values. In our subsequent work on phase I and phase II metabolite prediction, we decided to reverse this decision.

At this point, the overall result was the CYP metabolite prediction tool GLORY, which offers two modes to reflect the two different approaches to incorporating SoM prediction. The approach using SoMs as a hard filter was called MaxEfficiency mode, and the other mode was called MaxCoverage because there was no filter to reduce the number of predicted metabolites. In both modes, the SoM predictions were used, along with the weighting scheme to distinguish between common and uncommon reaction types, to score and rank the predicted metabolites.

Subsequently, GLORYx was developed to predict phase I and phase II metabolites, using the principles we discovered during the development of GLORY as a starting point. For GLORYx, SoMs were used only to score and rank the predicted metabolites, not as a hard filter. We additionally kept the 5:1 weight ratio between common and uncommon reaction types.

For GLORYx, we used separate FAME 3 SoM prediction models to score predicted metabolites resulting from phase I and phase II reaction rules. We initially implemented the reaction rules from SyGMa, which are separated into phase I and phase II rules. Then, since the predicted metabolites were the same as for SyGMa, with only a few exceptions due to molecule processing and aromaticity perception, we could directly compare our scoring approach to SyGMa's. We initially made no distinction between common or uncommon reaction types, so all reaction rules received the same weight. This means that the score was based only on the maximum SoM probability of the atoms involved in the transformation. Using a reference dataset separated into a phase I subset and a phase II subset, we found that our scoring approach was comparable to SyGMa's for phase I metabolite prediction, with AUCs of 0.73 and 0.76, respectively. For phase II, however, our scoring approach led to a rank-based AUC of 0.67, whereas SyGMa achieved an AUC of 0.85. This is not only a substantial difference, but an AUC of 0.67 generally indicates relatively low performance.

Further investigation into the difficulties of scoring the phase II metabolite predictions revealed that our scoring approach, using the maximum predicted SoM probability of an atom involved in the transformation, was not able to distinguish between more likely and less likely predicted metabolites, despite the high performance of FAME 3's phase II SoM model. We concluded that this discrepancy probably occurs because the functional groups involved in conjugation reactions are similar across different reaction types, and knowing the probability that a conjugation reaction occurs at a particular location does not aid in determining which conjugation reaction is more likely at this location. To solve this problem, we used reaction type-specific phase II SoM models that were newly developed for this work using the FAME 3 framework. An individual SoM model was developed corresponding approximately to each of the five main phase II enzyme families, and the general phase II SoM model was used for reaction types not corresponding to these five main enzyme families. This refinement to the method meant that any given heavy atom in a molecule was assigned separate SoM probabilities for each type of conjugation reaction. This drastically reduced the number of predictions that received the same score. The ranking performance also improved as a result, with a change in AUC from 0.67 to 0.77 on the reference dataset.

After this initial evaluation of the scoring approach by direct comparison to SyGMa using the same reaction rules, we added the CYP reaction rules from GLORY to the phase I rule set, and we manually created additional reaction rules for GSH conjugation. We then implemented a weighting component to the score, applying the same common-uncommon distinction that we developed for GLORY. For the reaction rules from GLORY, we kept the original common-uncommon annotation. The phase I rules sourced from SyGMa were designated "uncommon" and the phase II rules from the five main phase II enzyme families were designed "common" while the other phase II rules were labeled "uncommon". This categorization scheme resulted in a noticeable improvement over an unweighted scoring function (i.e. only using SoM probabilities for the score) for phase I, as measured on the reference dataset. For phase II, there were only a few uncommon reaction types and a relatively small number of metabolites in the dataset corresponding to these reaction types, so the weighting was perhaps not as necessary as for phase I in terms of its effect on performance.

Overall, both GLORY and GLORYx have competitive recall compared to existing

freely available metabolite prediction tools, based on the test sets that were manually curated from the scientific literature for this work. GLORY, in MaxCoverage mode, achieves a recall of 0.83 on a test set of CYP metabolites, compared to 0.74 and 0.72 for SyGMa and BioTransformer, respectively, and GLORYx achieves a recall of 0.77 on a test set of phase I and phase II metabolites, compared to 0.68 for SyGMa.

Both GLORY and GLORYx have low precision, which is lower than that of SyGMa and BioTransformer. However, the precision of the other tools is also very low, which emphasizes the need to have a meaningful way to rank the predicted metabolites. In this regard, both GLORY and GLORYx appear to be advantageous. Both GLORY and GLORYx outperformed SyGMa on ranking the predicted metabolites of the test sets. However, as discussed above, GLORYx lags behind SyGMa in terms of the ability to rank predicted phase II metabolites.

One of the aims of this dissertation was to use publicly available metabolite data to develop new approaches to metabolite structure prediction. This was the case for both GLORY and GLORYx, and the performance of both tools was compared to a method (SyGMa) developed based on a proprietary, currently unavailable dataset that is widely considered the most comprehensive metabolite dataset and hence has been used to develop many methods.

Further, we compared the metabolite data from the MetXBioDB [85] with the metabolite data from the DrugBank database [101]. This comparison considered both phase I and phase II metabolism, combined. Though there is a lot of overlap between these two data sources in terms of the parent molecules and metabolites, and the MetXBioDB is larger, the DrugBank data does cover additional areas of the chemical space not covered by the MetXBioDB parent molecules.

One of the difficulties of working with publicly available data is the limitations inherent in the experimental studies which published the data. For metabolite data in particular, many studies only present the experimentally determined structures of the main metabolites, or only the metabolites the researchers are interested in. It is also common for such studies to provide structures that are not completely determined. In many cases, the MS data used to determine the structures is insufficient to determine the exact location of the added functional group, for example. These circumstances currently limit the amount of data available for the evaluation of computational approaches for metabolite structure prediction.

5.2.1 Ongoing Developments and Future Research Directions for Metabolite Structure Prediction

A key part of this work on metabolite structure prediction was the development of a reference dataset and a test dataset for each study, resulting in separate datasets for CYP metabolism and for comprehensive phase I and phase II metabolism. Since the publication of GLORYx, there have already been two new publications of methods for metabolite structure prediction that have used one of the test datasets created as part of this dissertation, allowing direct comparison of the performance of the new methods to GLORY and GLORYx.

The Swamidass group has recently published two related methods for metabolite prediction. The first is the Metabolic Forest [104], which contains 24 reaction rules aimed at predicting phase I metabolites, but which nevertheless include conjugation reaction types. Most of the 24 reaction rules are composed of multiple sub-rules, with approximately 60 Reaction SMARTS in total. The Metabolic Forest has been implemented in XenoNet [105], a metabolic network predictor that aims to infer the metabolic pathways connecting a substrate and a target metabolite. The SoM prediction scores assigned by Rainbow XenoSite [56] are used to score each step in the metabolic pathway. When XenoNet is applied only to a parent molecule and predicts only one metabolism step, the normalized Rainbow XenoSite scores are used to rank the predicted metabolites for the parent molecule. The authors of XenoNet compared their tool to GLORY as well as to the reference methods from [D2] using our CYP metabolite test dataset. Prior to this comparison, they used the reference dataset published in [D2] to define thresholds for each reaction type to filter out putative false positive predictions. The authors found that XenoNet outperformed GLORY in terms of recall, predicting 89% of metabolites in the test dataset compared to 83% for GLORY in MaxCoverage mode. Despite its relatively small number of reaction rules and reaction type-specific score-based filtering procedure, XenoNet predicts 1179 metabolites for the test set of 29 parent molecules, which is nearly 400 more metabolites than GLORY predicts in MaxCoverage mode and around 800 more than predicted by SyGMa or BioTransformer. This finding could indicate that XenoNet's reaction rules are less specific than GLORY's, which could also explain the higher recall. In terms of ranking, the ROC curves seem to indicate

that XenoNet performs comparably to GLORY. The top-3 metric indicates that XenoNet performs slightly better, at 79% of molecules having a known metabolite predicted within the top three ranked positions, compared to GLORY's 76%. It is interesting to compare XenoNet to GLORY, even though XenoNet was originally designed to solve a different problem, because XenoNet also uses SoM prediction in order to prioritize the predicted metabolites. However, a downside of XenoNet is that it was developed based on the discontinued Metabolite database [74].

The other recent addition to the field worth discussing here is a publication by Litsa et al. [106] on a new rule-free approach to phase I and II metabolite structure prediction called MetaTrans. Litsa et al. reframed metabolite prediction as a sequence translation problem, using end-to-end learning with neural networks based only on SMILES of parent molecules and metabolites. MetaTrans was pre-trained on general chemical reactions, then tuned on human metabolic reactions sourced from public databases. Litsa et al. compared MetaTrans to GLORYx and SyGMa using a test dataset based on our CYP test dataset from [D2], to which they added additional parent molecules and metabolites from DrugBank in order to create a test dataset covering both phase I and phase II metabolism. Considering the top twenty predicted metabolites for each parent molecule, GLORYx had a recall of 74% whereas MetaTrans had a recall of 65%. Considering only the top five predicted metabolites for each parent molecule, however, GLORYx had a recall of 30% while MetaTrans had a recall of 43%. These results indicate that MetaTrans may do a better job of prioritizing its predictions, though it does not rank them. Instead, the algorithm provides a fixed number of predicted metabolites for each input molecule, depending on the user-defined "beam size" parameter. Overall, the highest possible recall reported for MetaTrans was only 68%.

It is exciting that the field of metabolite structure prediction is continuing to advance at such a rapid pace, and that the test datasets developed as part of this dissertation can contribute. Further advancement of the field may depend to some extent on the availability of more high-quality metabolite data. For example, more SoM data could be used to build more specific models. Perhaps reaction type-specific SoM models for phase I metabolism could also lead to improvements in distinguishing between the likelihoods of different predicted metabolites. It would be interesting to see if such a data-driven approach would perform better at ranking the predictions instead of or in addition to the simple distinction between common

and uncommon reaction types.

A further question that could be addressed is the prediction of multiple generations of metabolites, which is a capability of several existing tools such as SyGMa. However, multigenerational prediction leads to even larger numbers of predicted metabolites compared to predicting only one generation. Both GLORY and GLORYx predict only one generation of metabolites at a time, though it is possible to apply each tool sequentially in order to generate predicted metabolites of the predicted metabolites and thereby effectively predict multiple generations of metabolites. A potential future direction for work on this topic could be to examine how the problem of prioritizing the predicted metabolites, such as by ranking them, could be addressed for multigenerational predictions, and whether predicted SoM probabilities can help with this task.

Conclusions

Xenobiotic metabolism presents an ongoing challenge for the development of safe drugs, agrochemicals, and cosmetics. This dissertation aimed at addressing this challenge by exploring new methods for SoM and metabolite prediction.

First, the possibility of using molecular alignment to predict SoMs was investigated. An alignment-based approach had first been introduced by Sykes et al. in 2008 [72] but was accompanied by a limited analysis. In this dissertation, the alignment-based concept was further developed and expanded from a single CYP isozyme and one reference molecule in its bioactive conformation to an approach including all major CYP isozymes and a set of potential reference molecules. The analysis was broadened as well, to include several metrics for early enrichment and ranking-based performance. Since the performance of the alignment-based method on its own was found to be at an impractical level, a reactivity component was combined with the alignment. The combination of reactivity and alignment led to a boost in performance compared to either component on its own.

Further, two tools for metabolite structure prediction were developed. This research aimed at using SoM prediction to address the general problem of predicting more metabolites than a user can easily deal with. Initially, the development of the metabolite prediction approach focused on CYP-mediated metabolism. In this context, two distinct approaches for incorporating SoM prediction were investigated. One was to use the predicted SoMs as a hard filter to determine where in a molecule the reaction rules would be applied. The other approach was to employ the predicted SoM probabilities as part of a score for each predicted metabolite. This scoring approach was found to be more successful and useful than the hard-filter approach. The resulting tool was GLORY.

Subsequently, the focus of the metabolite prediction approach was expanded to cover all of phase I and phase II metabolism, resulting in the development of GLORYx. In this case, SoM prediction was only used to score the predictions, based on the conclusions of the previous work. Extending the approach to phase

I and phase II was not trivial. For phase II in particular, it proved tricky to use SoMs to determine which metabolites were more likely to occur experimentally. We determined that this is because phase II reactions are more likely to occur at similar functional groups, to the extent that multiple reaction types can occur at the same functional group. To address this problem, additional SoM prediction models were required. These additional models were reaction type-specific, which allowed each atom to have a different SoM probability depending on which reaction type was being considered. Using these new SoM prediction models, we were able to improve the ranking performance for phase II metabolism.

For both GLORY and GLORYx, rule sets of metabolic reactions were developed. For GLORY, a CYP reaction rule set was developed based on the scientific literature. For GLORYx, the phase I and phase II rule sets from SyGMa were translated into SMIRKS and combined with the CYP rule set from GLORY as well as a new GSH conjugation rule set developed based on the scientific literature.

Over the course of this dissertation, two new test sets for metabolite structure prediction were manually created, one for CYP-mediated metabolism and one for phase I and phase II metabolism. In addition, a public dataset for SoM prediction was modified to include proper stereochemistry and two reference datasets for metabolite structure prediction, one for CYP-mediated and one for phase I and phase II metabolism, were assembled by combining data from two public sources. The hope is that these datasets can be used in the future development of new tools and the further improvement of existing tools for metabolite structure prediction.

Bibliography

[1] Kirchmair, J.; Howlett, A.; Peironcely, J. E.; Murrell, D. S.; Williamson, M. J.; Adams, S. E.; Hankemeier, T.; van Buren, L.; Duchateau, G.; Klaffke, W.; Glen, R. C. How Do Metabolites Differ from Their Parent Molecules and How Are They Excreted?, *J. Chem. Inf. Model.* **2013**, *53*, 354–367.

[2] Attia, S. M. Deleterious effects of reactive metabolites, *Oxid. Med. Cell. Longev.* **2010**, *3*, 238–253.

[3] Rautio, J.; Meanwell, N. A.; Di, L.; Hageman, M. J. The expanding role of prodrugs in contemporary drug design and development, *Nat. Rev. Drug Discov.* **2018**, *17*, 559–587.

[4] Whitebread, S.; Dumotier, B.; Armstrong, D.; Fekete, A.; Chen, S.; Hartmann, A.; Muller, P. Y.; Urban, L. Secondary pharmacology: screening and interpretation of off-target activities - focus on translation, *Drug Discov. Today* **2016**, *21*, 1232–1242.

[5] Fura, A.; Shu, Y.-Z.; Zhu, M.; Hanson, R. L.; Roongta, V.; Humphreys, W. G. Discovering Drugs through Biological Transformation: Role of Pharmacologically Active Metabolites in Drug Discovery, *J. Med. Chem.* **2004**, *47*, 4339–4351.

[6] Di, L. The role of drug metabolizing enzymes in clearance, *Expert Opin Drug Metab Toxicol* **2014**, *10*, 379–393.

[7] Eno, M. R.; Cameron, M. D. Gauging reactive metabolites in drug-induced toxicity., *Curr. Med. Chem.* **2015**, *22*, 465–89.

[8] Walgren, J. L.; Mitchell, M. D.; Thompson, D. C. Role of Metabolism in Drug-Induced Idiosyncratic Hepatotoxicity, *Crit. Rev. Toxicol.* **2005**, *35*, 325–361.

[9] Kalgutkar, A. S.; Vaz, A. D. N.; Lame, M. E.; Henne, K. R.; Soglia, J.; Zhao, S. X.; Abramov, Y. A.; Lombardo, F.; Collin, C.; Hendsch, Z. S.; Hop, C. E. C. A. Bioactivation of the Nontricyclic Antidepressant Nefazodone to a Reactive Quinone-Imine Species in Human Liver Microsomes and Recombinant Cytochrome P450 3A4, *Drug Metab Dispos* **2005**, *33*, 243.

[10] Humphreys, W. G.; Unger, S. E. Safety Assessment of Drug Metabolites: Characterization of Chemically Stable Metabolites, *Chem. Res. Toxicol.* **2006**, *19*, 1564–1569.

[11] Baillie, T. A.; Rettie, A. E. Role of biotransformation in drug-induced toxicity: influence of intra- and inter-species differences in drug metabolism, *Drug Metab Pharmacokinet* **2011**, *26*, 15–29.

[12] Tang, W.; Lu, A. Y. H. Metabolic bioactivation and drug-related adverse effects: current status and future directions from a pharmaceutical research perspective, *Drug Metab. Rev.* **2010**, *42*, 225–249.

[13] Liebler, D. C.; Guengerich, F. P. Elucidating mechanisms of drug-induced toxicity, *Nat. Rev. Drug Discov.* **2005**, *4*, 410–420.

[14] Guengerich, F. P. Cytochrome P450s and other enzymes in drug metabolism and toxicity, *AAPS J.* **2006**, *8*, E101–E111.

[15] Pinto, N.; Dolan, M. E. Clinically relevant genetic variations in drug metabolizing enzymes, *Curr. Drug Metab.* **2011**, *12*, 487–497.

[16] Zhou, S.-F.; Liu, J.-P.; Chowbay, B. Polymorphism of human cytochrome P450 enzymes and its clinical impact, *Drug Metab. Rev.* **2009**, *41*, 89–295.

[17] Guengerich, F. P. A history of the roles of cytochrome P450 enzymes in the toxicity of drugs, *Toxicol. Res.* **2020**, DOI 10.1007/s43188-020-00056-z.

[18] Yu, H.; Balani, S. K.; Chen, W.; Cui, D.; He, L.; Humphreys, W. G.; Mao, J.; Lai, W. G.; Lee, A. J.; Lim, H.-K.; MacLauchlin, C.; Prakash, C.; Surapaneni, S.; Tse, S.; Upthagrove, A.; Walsky, R. L.; Wen, B.; Zeng, Z. Contribution of Metabolites to P450 Inhibition-Based Drug-Drug Interactions: Scholarship from the Drug Metabolism Leadership Group of the Innovation and Quality Consortium Metabolite Group, *Drug Metab Dispos* **2015**, *43*, 620.

[19] Josephy, P. D.; Peter Guengerich, F.; Miners, J. O. "Phase I and Phase II" Drug Metabolism: Terminology that we Should Phase Out?, *Drug Metab. Rev.* **2005**, *37*, 575–580.

[20] Williams, J. A.; Hyland, R.; Jones, B. C.; Smith, D. A.; Hurst, S.; Goosen, T. C.; Peterkin, V.; Koup, J. R.; Ball, S. E. Drug-Drug Interactions for UDP-Glucuronosyltransferase Substrates: A Pharmacokinetic Explanation for Typically Observed Low Exposure (AUC_I/AUC) Ratios, *Drug Metab Dispos* **2004**, *32*, 1201.

[21] Testa, B.; Pedretti, A.; Vistoli, G. Reactions and enzymes in the metabolism of drugs and other xenobiotics, *Drug Discov. Today* **2012**, *17*, 549–560.

[22] Macherey, A.-C.; Dansette, P. M.; Wermuth, C. G.; Aldous, D.; Raboisson, P.; Rognan, D., Chapter 25 - Biotransformations Leading to Toxic Metabolites: Chemical Aspects in *The Practice of Medicinal Chemistry (Fourth Edition)*, Academic Press, San Diego, **2015**, pp. 585–614.

[23] Amacher, D. E. Reactive intermediates and the pathogenesis of adverse drug reactions: the toxicology perspective, *Curr. Drug Metab.* **2006**, *7*, 219–29.

[24] Wormhoudt, L. W.; Commandeur, J. N. M.; Vermeulen, N. P. E. Genetic Polymorphisms of Human N-Acetyltransferase, Cytochrome P450, Glutathione-S-Transferase, and Epoxide Hydrolase Enzymes: Relevance to Xenobiotic Metabolism and Toxicity, *Crit. Rev. Toxicol.* **1999**, *29*, 59–124.

[25] Phillips, I. R.; Shephard, E. A. Flavin-containing monooxygenase 3 (FMO3): genetic variants and their consequences for drug metabolism and disease, *Xenobiotica* **2020**, *50*, 19–33.

[26] Garattini, E.; Terao, M. The role of aldehyde oxidase in drug metabolism, *Expert Opin Drug Metab Toxicol* **2012**, *8*, 487–503.

[27] Guengerich, F. P., Human Cytochrome P450 Enzymes in *Cytochrome P450, Structure, Mechanism, and Biochemistry*, (Ed.: Ortiz de Montellano, P. R.), Springer International Publishing, **2015**, pp. 523–785.

[28] Guengerich, F. P. Common and Uncommon Cytochrome P450 Reactions Related to Metabolism and Chemical Toxicity, *Chem. Res. Toxicol.* **2001**, *14*, 611–650.

[29] Penner, N.; Woodward, C.; Prakash, C., Appendix: Drug Metabolizing Enzymes and Biotransformation Reactions in *ADME-Enabling Technologies in Drug Design and Development*, (Eds.: Zhang, D.; Surapaneni, S.), John Wiley & Sons, Ltd, **2012**, pp. 545–565.

[30] Cashman, J. R. Some distinctions between flavin-containing and cytochrome P450 monooxygenases., *Biochem. Biophys. Res. Commun* **2005**, *338*, 599–604.

[31] Garattini, E.; Terao, M. Aldehyde oxidase and its importance in novel drug discovery: present and future challenges, *Expert Opin Drug Discov* **2013**, *8*, 641–654.

[32] Fan, P. W.; Zhang, D.; Halladay, J. S.; Driscoll, J. P.; Khojasteh, S. C. Going Beyond Common Drug Metabolizing Enzymes: Case Studies of Biotransformation Involving Aldehyde Oxidase, γ-Glutamyl Transpeptidase, Cathepsin B, Flavin-Containing Monooxygenase, and ADP-Ribosyltransferase, *Drug Metab Dispos* **2016**, *44*, 1253.

[33] Rendic, S.; Guengerich, F. P. Survey of Human Oxidoreductases and Cytochrome P450 Enzymes Involved in the Metabolism of Xenobiotic and Natural Chemicals, *Chem. Res. Toxicol.* **2015**, *28*, 38–42.

[34] Imai, T. Human Carboxylesterase Isozymes: Catalytic Properties and Rational Drug Design, *Drug Metab Pharmacokinet* **2006**, *21*, 173–185.

[35] Laizure, S. C.; Herring, V.; Hu, Z.; Witbrodt, K.; Parker, R. B. The role of human carboxylesterases in drug metabolism: have we overlooked their importance?, *Pharmacotherapy* **2013**, *33*, 210–222.

[36] Cerny, M. A. Non-P450 Metabolism of FDA-Approved Drugs: 2006-2015, *Drug Metab Dispos* **2016**, *44*, 1246–1252.

[37] Ballard, P.; Brassil, P.; Bui, K. H.; Dolgos, H.; Petersson, C.; Tunek, A.; Webborn, P. J. H.; Hill, R. G.; Rang, H. P., Metabolism and pharmacokinetic optimization strategies in drug discovery in *Drug Discovery and Development (Second Edition)*, Churchill Livingstone, **2013**, Chapter 10, pp. 135–155.

[38] Guengerich, F. P., Mechanisms of Cytochrome P450-Mediated Reactions in *Encyclopedia of Drug Metabolism and Interactions*, American Cancer Society, **2012**, pp. 1–23.

[39] Lewis, D. F. V. 57 varieties: the human cytochromes P450, *Pharmacogenomics* **2004**, *5*, 305–318.

[40] Lewis, D. F. V.; Ito, Y. Human CYPs involved in drug metabolism: structures, substrates and binding affinities, *Expert Opin. Drug Metab. Toxicol.* **2010**, *6*, 661–674.

[41] Denisov, I. G.; Makris, T. M.; Sligar, S. G.; Schlichting, I. Structure and Chemistry of Cytochrome P450, *Chem. Rev.* **2005**, *105*, 2253–2278.

[42] Kirchmair, J.; Göller, A. H.; Lang, D.; Kunze, J.; Testa, B.; Wilson, I. D.; Glen, R. C.; Schneider, G. Predicting drug metabolism: experiment and/or computation?, *Nat Rev Drug Discov* **2015**, *14*, 387–404.

[43] Zanger, U. M.; Schwab, M. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation, *Pharmacol. Ther.* **2013**, *138*, 103–141.

[44] Ge, S.; Tu, Y.; Hu, M. Challenges and Opportunities with Predicting In Vivo Phase II Metabolism via Glucuronidation From In Vitro Data, *Curr. Pharmacol. Rep.* **2016**, *2*, 326–338.

[45] Jancova, P.; Anzenbacher, P.; Anzenbacherova, E. Phase II Drug Metabolizing Enzymes, *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub.* **2010**, *154*, 103–116.

[46] Allocati, N.; Masulli, M.; Di Ilio, C.; Federici, L. Glutathione transferases: substrates, inihibitors and pro-drugs in cancer and neurodegenerative diseases, *Oncogenesis* **2018**, *7*, 8.

[47] Omiecinski, C. J.; Vanden Heuvel, J. P.; Perdew, G. H.; Peters, J. M. Xenobiotic Metabolism, Disposition, and Regulation by Receptors: From Biochemical Phenomenon to Predictors of Major Toxicities, *Toxicol Sci* **2011**, *120*, S49–S75.

[48] van Bladeren, P. J. Glutathione conjugation as a bioactivation reaction, *Chem. Biol. Interact.* **2000**, *129*, 61–76.

[49] Glatt, H. Sulfotransferases in the bioactivation of xenobiotics, *Chem. Biol. Interact.* **2000**, *129*, 141–170.

[50] Kirchmair, J.; Williamson, M. J.; Tyzack, J. D.; Tan, L.; Bond, P. J.; Bender, A.; Glen, R. C. Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms, *J. Chem. Inf. Model.* **2012**, *52*, 617–648.

[51] Raunio, H.; Kuusisto, M.; Juvonen, R. O.; Pentikäinen, O. T. Modeling of interactions between xenobiotics and cytochrome P450 (CYP) enzymes, *Front. Pharmacol.* **2015**, *6*, 123.

[52] Kato, H. Computational prediction of cytochrome P450 inhibition and induction, *Drug Metab Pharmacokinet* **2020**, *35*, 30–44.

[53] Zhu, M.; Zhang, H.; Humphreys, W. G. Drug metabolite profiling and identification by high-resolution mass spectrometry, *J. Biol. Chem.* **2011**, *286*, 25419–25425.

[54] Tyzack, J. D.; Kirchmair, J. Computational methods and tools to predict cytochrome P450 metabolism for drug discovery, *Chem Biol Drug Des* **2019**, *93*, 377–386.

[55] Zaretzki, J.; Matlock, M.; Swamidass, S. J. XenoSite: Accurately Predicting CYP-Mediated Sites of Metabolism with Neural Networks, *J. Chem. Inf. Model.* **2013**, *53*, 3373–3383.

[56] Dang, N. L.; Matlock, M. K.; Hughes, T. B.; Swamidass, S. J. The Metabolic Rainbow: Deep Learning Phase I Metabolism in Five Colors, *J. Chem. Inf. Model.* **2020**, *60*, 1146–1164.

[57] Dang, N. L.; Hughes, T. B.; Krishnamurthy, V.; Swamidass, S. J. A simple model predicts UGT-mediated metabolism, *Bioinformatics* **2016**, *32*, 3183–3189.

[58] Rudik, A.; Dmitriev, A.; Lagunin, A.; Filimonov, D.; Poroikov, V. SOMP: web server for in silico prediction of sites of metabolism for drug-like compounds, *Bioinformatics* **2015**, *31*, 2046–2048.

[59] Finkelmann, A. R.; Göller, A. H.; Schneider, G. Site of Metabolism Prediction Based on ab initio Derived Atom Representations, *ChemMedChem* **2017**, *12*, 606–612.

[60] Finkelmann, A. R.; Goldmann, D.; Schneider, G.; Göller, A. H. MetScore: Site of Metabolism Prediction Beyond Cytochrome P450 Enzymes, *ChemMedChem* **2018**, *13*, 2281–2289.

[61] Kirchmair, J.; Williamson, M. J.; Afzal, A. M.; Tyzack, J. D.; Choy, A. P. K.; Howlett, A.; Rydberg, P.; Glen, R. C. FAst MEtabolizer (FAME): A rapid and accurate predictor of sites of metabolism in multiple species by endogenous enzymes, *J. Chem. Inf. Model.* **2013**, *53*, 2896–2907.

[62] Zaretzki, J.; Bergeron, C.; Rydberg, P.; Huang, T.-w.; Bennett, K. P.; Breneman, C. M. RS-Predictor: A New Tool for Predicting Sites of Cytochrome P450-Mediated Metabolism Applied to CYP 3A4, *J. Chem. Inf. Model.* **2011**, *51*, 1667–1689.

[63] Hughes, T. B.; Swamidass, S. J. Deep Learning to Predict the Formation of Quinone Species in Drug Metabolism, *Chem. Res. Toxicol.* **2017**, *30*, 642–656.

[64] Rydberg, P.; Gloriam, D. E.; Zaretzki, J.; Breneman, C.; Olsen, L. SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism, *ACS Med. Chem. Lett.* **2010**, *1*, 96–100.

[65] Rydberg, P.; Rostkowski, M.; Gloriam, D. E.; Olsen, L. The Contribution of Atom Accessibility to Site of Metabolism Models for Cytochromes P450, *Mol. Pharmaceutics* **2013**, *10*, 1216–1223.

[66] Rydberg, P.; Gloriam, D. E.; Olsen, L. The SMARTCyp cytochrome P450 metabolism prediction server, *Bioinformatics* **2010**, *26*, 2988–2989.

[67] Olsen, L.; Montefiori, M.; Tran, K. P.; Jørgensen, F. S. SMARTCyp 3.0: enhanced cytochrome P450 site-of-metabolism prediction server, *Bioinformatics* **2019**, *35*, 3174–3175.

[68] Dixit, V. A.; Lal, L. A.; Agrawal, S. R. Recent advances in the prediction of non-CYP450-mediated drug metabolism, *WIREs Comput Mol Sci* **2017**, *7*, e1323.

[69] Cruciani, G.; Milani, N.; Benedetti, P.; Lepri, S.; Cesarini, L.; Baroni, M.; Spyrakis, F.; Tortorella, S.; Mosconi, E.; Goracci, L. From Experiments to a Fast Easy-to-Use Computational Methodology to Predict Human Aldehyde Oxidase Selectivity and Metabolic Reactions, *J. Med. Chem.* **2018**, *61*, 360–371.

[70] Cruciani, G.; Valeri, A.; Goracci, L.; Pellegrino, R. M.; Buonerba, F.; Baroni, M. Flavin Monooxygenase Metabolism: Why Medicinal Chemists Should Matter, *J. Med. Chem.* **2014**, *57*, 6183–6196.

[71] Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: Understanding Metabolism in Human Cytochromes from the Perspective of the Chemist, *J. Med. Chem.* **2005**, *48*, 6970–6979.

[72] Sykes, M. J.; McKinnon, R. A.; Miners, J. O. Prediction of Metabolism by Cytochrome P450 2C9: Alignment and Docking Studies of a Validated Database of Substrates, *J. Med. Chem.* **2008**, *51*, 780–791.

[73] Pedretti, A.; Mazzolari, A.; Vistoli, G.; Testa, B. MetaQSAR: An Integrated Database Engine to Manage and Analyze Metabolic Data, *J. Med. Chem.* **2018**, *61*, 1019–1030.

[74] Elsevier. *MDL Metabolite Database*, **2001**.

[75] Judson, P. N., Knowledge-Based Approaches for Predicting the Sites and Products of Metabolism in *Drug Metabolism Prediction*, (Ed.: Kirchmair, J.), Wiley-VCH, Weinheim, Germany, **2014**, pp. 293–318.

[76] Ridder, L.; Wagener, M. SyGMa: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites, *ChemMedChem* **2008**, *3*, 821–832.

[77] Wang, D.; Liu, W.; Shen, Z.; Jiang, L.; Wang, J.; Li, S.; Li, H. Deep Learning Based Drug Metabolites Prediction, *Front. Pharmacol.* **2020**, *10*, 1586.

[78] Rudik, A. V.; Bezhentsev, V. M.; Dmitriev, A. V.; Druzhilovskiy, D. S.; Lagunin, A. A.; Filimonov, D. A.; Poroikov, V. V. MetaTox: Web Application for Predicting Structure and Toxicity of Xenobiotics' Metabolites, *J. Chem. Inf. Model.* **2017**, *57*, 638–642.

[79] Rudik, A.; Bezhentsev, V.; Dmitriev, A.; Lagunin, A.; Filimonov, D.; Poroikov, V. Metatox - Web application for generation of metabolic pathways and toxicity estimation, *J. Bioinform. Comput. Biol.* **2019**, *17*, 1940001.

[80] Marchant, C. A.; Briggs, K. A.; Long, A. In Silico Tools for Sharing Data and Knowledge on Toxicity and Metabolism: Derek for Windows, Meteor, and Vitic, *Toxicol. Mech. Methods* **2008**, *18*, 177–187.

[81] Mekenyan, O. G.; Dimitrov, S. D.; Pavlov, T. S.; Veith, G. D. A systematic approach to simulating metabolism in computational toxicology. I. The TIMES heuristic modelling framework., *Curr. Pharm. Des.* **2004**, *10*, 1273–93.

[82] Darvas, F., Metabolexpert: An Expert System for Predicting Metabolism of Substances in *QSAR in Environmental Toxicology - II*, (Ed.: Kaiser, K.), Springer, Dordrecht, **1987**.

[83] ADMET Predictor Metabolism Module, SimulationsPlus, SimulationsPlus.com/software/admetpredictor/metabolism/.

[84] StarDrop, Optibrium, https://www.optibrium.com/stardrop/stardrop-p450-models.php.

[85] Djoumbou-Feunang, Y.; Fiamoncini, J.; Gil-de-la-Fuente, A.; Greiner, R.; Manach, C.; Wishart, D. S. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification, *J. Cheminformatics* **2019**, *11*, 2.

[86] Meng, J.; Li, S.; Liu, X.; Zheng, M.; Li, H. RD-Metabolizer: an integrated and reaction types extensive approach to predict metabolic sites and metabolites of drug-like molecules, *Chem. Cent. J.* **2017**, *11*, 65.

[87] Rudik, A. V.; Dmitriev, A. V.; Lagunin, A. A.; Filimonov, D. A.; Poroikov, V. V. Metabolism Site Prediction Based on Xenobiotic Structural Formulas and PASS Prediction Algorithm, *J. Chem. Inf. Model.* **2014**, *54*, 498–507.

[88] Rudik, A. V.; Dmitriev, A. V.; Lagunin, A. A.; Filimonov, D. A.; Poroikov, V. V. Prediction of reacting atoms for the major biotransformation reactions of organic xenobiotics, *J. Cheminformatics* **2016**, *8*, 68.

[89] McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J.-F.; Cornell, W. D. Comparison of Topological, Shape, and Docking Methods in Virtual Screening, *J. Chem. Inf. Model.* **2007**, *47*, 1504–1519.

[90] Finn, P. W.; Morris, G. M. Shape-based similarity searching in chemical databases, *WIREs Computational Molecular Science* **2013**, *3*, 226–241.

[91] Singh, N.; Chaput, L.; Villoutreix, B. O. Virtual screening web servers: designing chemical probes and drug candidates in the cyberspace, *Briefings in Bioinformatics* **2020**, DOI `10.1093/bib/bbaa034`.

[92] Hawkins, P. C. D.; Skillman, A. G.; Nicholls, A. Comparison of Shape-Matching and Docking as Virtual Screening Tools, *J. Med. Chem.* **2007**, *50*, 74–82.

[93] Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How To Optimize Shape-Based Virtual Screening: Choosing the Right Query and Including Chemical Information, *J. Chem. Inf. Model.* **2009**, *49*, 678–692.

[94] Hawkins, P. C. D.; Skillman, A. G.; Warren, G. L.; Ellingson, B. A.; Stahl, M. T. Conformer Generation with OMEGA: Algorithm and Validation Using High Quality Structures from the Protein Databank and Cambridge Structural Database, *J. Chem. Inf. Model.* **2010**, *50*, 572–584.

[95] SMIRKS - A Reaction Transform Language, Daylight Chemical Information Systems, Inc, `https://www.daylight.com/dayhtml/doc/theory/theory.smirks.html` (visited on 09/17/2020).

[96] Dhaked, D. K.; Ihlenfeldt, W.-D.; Patel, H.; Delannée, V.; Nicklaus, M. C. Toward a Comprehensive Treatment of Tautomerism in Chemoinformatics Including in InChI V2, *J. Chem. Inf. Model.* **2020**, *60*, 1253–1275.

[97] RDKit: Open-Source Cheminformatics, `http://www.rdkit.org`.

[98] Kochev, N.; Avramova, S.; Jeliazkova, N. Ambit-SMIRKS: a software module for reaction representation, reaction search and structure transformation, *J. Cheminformatics* **2018**, *10*, 42.

[99] Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Chertó, M.; Spjuth, O.; Torrance, G.; Evelo, C. T.; Guha, R.; Steinbeck, C. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching, *J. Cheminformatics* **2017**, *9*, 33.

[100] ElectronDonation, `http://cdk.github.io/cdk/2.2/docs/api/org/openscience/cdk/aromaticity/ElectronDonation.html` (visited on 11/06/2020).

[101] Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: a major update to the DrugBank database for 2018, *Nucleic Acids Res* **2018**, *46*, D1074–D1082.

[102] Truchon, J.-F.; Bayly, C. I. Evaluating Virtual Screening Methods: Good and Bad Metrics for the "Early Recognition" Problem, *J. Chem. Inf. Model.* **2007**, *47*, 488–508.

[103] Zhao, W.; Hevener, K. E.; White, S. W.; Lee, R. E.; Boyett, J. M. A statistical framework to evaluate virtual screening, *BMC Bioinformatics* **2009**, *10*, 225.

[104] Hughes, T. B.; Dang, N. L.; Kumar, A.; Flynn, N. R.; Swamidass, S. J. Metabolic Forest: Predicting the Diverse Structures of Drug Metabolites, *J. Chem. Inf. Model.* **2020**, *60*, 4702–4716.

[105] Flynn, N. R.; Dang, N. L.; Ward, M. D.; Swamidass, S. J. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation, *J. Chem. Inf. Model.* **2020**, *60*, 3431–3449.

[106] Litsa, E. E.; Das, P.; Kavraki, L. E. Prediction of drug metabolites using neural machine translation, *Chem. Sci.* **2020**, DOI `10.1039/D0SC02639E`.

Bibliography of this Dissertation's Publications

This Dissertation's Publications

[D1] de Bruyn Kops, C.; Friedrich, N.-O.; Kirchmair, J. Alignment-Based Prediction of Sites of Metabolism, *J. Chem. Inf. Model.* **2017**, *57*, 1258–1264.

[D2] de Bruyn Kops, C.; Stork, C.; Šícho, M.; Kochev, N.; Svozil, D.; Jeliazkova, N.; Kirchmair, J. GLORY: Generator of the Structures of Likely Cytochrome P450 Metabolites Based on Predicted Sites of Metabolism, *Front. Chem.* **2019**, *7*, 402.

[D3] de Bruyn Kops, C.; Šícho, M.; Mazzolari, A.; Kirchmair, J. GLORYx: Prediction of the Metabolites Resulting from Phase 1 and Phase 2 Biotransformations of Xenobiotics, *Chem. Res. Toxicol.* **2020**, DOI 10.1021/acs.chemrestox.0c00224.

Additional Publications

[A1] Chen, Y.; de Bruyn Kops, C.; Kirchmair, J. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products, *J. Chem. Inf. Model.* **2017**, *57*, 2099–2111.

[A2] Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators, *J. Chem. Inf. Model.* **2017**, *57*, 2719–2728.

[A3] Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators, *J. Chem. Inf. Model.* **2017**, *57*, 529–539.

[A4] Šícho, M.; de Bruyn Kops, C.; Stork, C.; Svozil, D.; Kirchmair, J. FAME 2: Simple and Effective Machine Learning Model of Cytochrome P450 Regioselectivity, *J. Chem. Inf. Model.* **2017**, *57*, 1832–1846.

[A5] Stork, C.; Wagner, J.; Friedrich, N.-O.; de Bruyn Kops, C.; Šícho, M.; Kirchmair, J. Hit Dexter: A Machine-Learning Model for the Prediction of Frequent Hitters, *ChemMedChem* **2018**, *13*, 564–571.

[A6] Chen, Y.; de Bruyn Kops, C.; Kirchmair, J., Resources for Chemical, Biological, and Structural Data on Natural Products in *Progress in the Chemistry of Organic Natural Products 110: Cheminformatics in Natural Product Research*, (Eds.: Kinghorn, A. D.; Falk, H.; Gibbons, S.; Kobayashi, J.; Asakawa, Y.; Liu, J.-K.), Springer International Publishing, Cham, **2019**, pp. 37–71.

[A7] Ehm, P. A.; Lange, F.; Hentschel, C.; Jepsen, A.; Glück, M.; Nelson, N.; Bettin, B.; de Bruyn Kops, C.; Kirchmair, J.; Nalaskowski, M.; Jücker, M. Analysis of the FLVR motif of SHIP1 and its importance for the protein stability of SH2 containing signaling proteins, *Cellular Signalling* **2019**, *63*, 109380.

[A8] Šícho, M.; Stork, C.; Mazzolari, A.; de Bruyn Kops, C.; Pedretti, A.; Testa, B.; Vistoli, G.; Svozil, D.; Kirchmair, J. FAME 3: Predicting the Sites of Metabolism in Synthetic Compounds and Natural Products for Phase 1 and Phase 2 Metabolic Enzymes, *J. Chem. Inf. Model.* **2019**, *59*, 3400–3412.

[A9] Fan, N.; Bauer, C. A.; Stork, C.; de Bruyn Kops, C.; Kirchmair, J. ALADDIN: Docking Approach Augmented by Machine Learning for Protein Structure Selection Yields Superior Virtual Screening Performance, *Molecular Informatics* **2020**, *39*, 1900103.

[A10] Stork, C.; Embruch, G.; Šícho, M.; de Bruyn Kops, C.; Chen, Y.; Svozil, D.; Kirchmair, J. NERDD: a web portal providing access to in silico tools for drug discovery, *Bioinformatics* **2020**, *36*, 1291–1292.

[A11] Wilm, A.; Norinder, U.; Agea, M. I.; de Bruyn Kops, C.; Stork, C.; Kühnl, J.; Kirchmair, J. Skin Doctor CP: Conformal prediction of the skin sensitization potential of small organic molecules, *Chem. Res. Toxicol.*, in review.

List of Abbreviations

2D	Two-dimensional
3D	Three-dimensional
ADH	Alcohol dehydrogenase
ADP	Adenosine diphosphate
ADR	Adverse drug reaction
AKR	Aldo-keto reductase
ALDH	Aldehyde dehydrogenase
AOX	Aldehyde oxidase
AUC	Area under the receiver operating characteristic curve
BEDROC	Boltzmann-Enhanced Discrimination of ROC
CDK	Chemistry Development Kit
CES	Carboxylesterase
COMT	Catechol O-methyl transferase
CYP	Cytochrome P450
DDI	Drug-drug interaction
DFT	Density functional theory
EC	Enzyme Commission
FDA	Food and Drug Administration
FMO	Flavin-containing monooxygenase
FN	False negative
FP	False positive
GSH	Glutathione
GST	Glutathione S-transferase
hCE1	Human carboxylesterase 1
hCE2	Human carboxylesterase 2
MAO	Monoamine oxidase
MetID	Metabolite identification
MCC	Matthews correlation coefficient

MS	Mass spectrometry
MT	Methyltransferase
NAT	N-acetyltransferase
NMR	Nuclear magnetic resonance
NNMT	Nicotinamide N-methyltransferase
QR	Quinone reductase
QSAR	Quantitative structure-activity relationship
Redox	Oxidation-reduction
ROC	Receiver operating characteristic
SoM	Site of metabolism
SULT	Sulfotransferase
TMT	thiol methyltransferase
TN	True negative
TP	True positive
TPMT	Thiopurine S-methyl transferase
UDP	Uridine diphosphate
UGT	UDP-glucuronosyltransferase
U.S.	United States
XO	Xanthine oxidase

Appendices

Scientific Contributions

The following sections list the scientific contributions made by the author throughout the course of this cumulative dissertation.

A.1 Publications

This section lists the author's publications in peer-reviewed scientific journals and a book chapter.

This Dissertation's Publications

[D1] de Bruyn Kops, C.; Friedrich, N.-O.; Kirchmair, J. Alignment-Based Prediction of Sites of Metabolism, *J. Chem. Inf. Model.* **2017**, *57*, 1258–1264.

C. de Bruyn Kops and J. Kirchmair conceptualized the research. C. de Bruyn Kops along with N.-O. Friedrich and J. Kirchmair designed the experiments. C. de Bruyn Kops conducted the computational method development and analysis. C. de Bruyn Kops wrote the manuscript, with contributions from J. Kirchmair. J. Kirchmair supervised this work.

[D2] de Bruyn Kops, C.; Stork, C.; Šícho, M.; Kochev, N.; Svozil, D.; Jeliazkova, N.; Kirchmair, J. GLORY: Generator of the Structures of Likely Cytochrome P450 Metabolites Based on Predicted Sites of Metabolism, *Front. Chem.* **2019**, *7*, 402.

C. de Bruyn Kops, N. Kochev, N. Jeliazkova, and J. Kirchmair conceptualized the research. C. de Bruyn Kops developed the metabolite prediction method. C. de Bruyn Kops wrote the manuscript, with contributions from J. Kirchmair. C. Stork contributed to the development of the reaction rules. M. Šícho contributed to the implementation of SoM prediction. J. Kirchmair and D. Svozil supervised this work.

[D3] de Bruyn Kops, C.; Šícho, M.; Mazzolari, A.; Kirchmair, J. GLORYx: Prediction of the Metabolites Resulting from Phase 1 and Phase 2 Biotransformations of Xenobiotics, *Chem. Res. Toxicol.* **2020**, DOI 10.1021/acs.chemrestox.0c00224.

C. de Bruyn Kops and J. Kirchmair conceptualized the research. C. de Bruyn Kops developed the metabolite prediction method. C. de Bruyn Kops wrote the manuscript, with contributions from J. Kirchmair. M. Šícho developed the machine learning models. A. Mazzolari contributed the dataset for the development of the machine learning models. J. Kirchmair supervised this work.

Additional Publications

[A1] Chen, Y.; de Bruyn Kops, C.; Kirchmair, J. Data Resources for the Computer-Guided Discovery of Bioactive Natural Products, *J. Chem. Inf. Model.* **2017**, *57*, 2099–2111.

C. de Bruyn Kops and Y. Chen contributed equally to this work. C. de Bruyn Kops wrote the manuscript, with contributions from J. Kirchmair and Y. Chen. Y. Chen collected the databases and performed the analysis. C. de Bruyn Kops contributed to the analysis in the section Coverage and Reach of Chemical Structures Deposited in Natural Product Libraries.

[A2] Friedrich, N.-O.; de Bruyn Kops, C.; Flachsenberg, F.; Sommer, K.; Rarey, M.; Kirchmair, J. Benchmarking Commercial Conformer Ensemble Generators, *J. Chem. Inf. Model.* **2017**, *57*, 2719–2728.

C. de Bruyn Kops developed the concept for the statistical analysis of the performance tests, verified the results and contributed to the manuscript.

[A3] Friedrich, N.-O.; Meyder, A.; de Bruyn Kops, C.; Sommer, K.; Flachsenberg, F.; Rarey, M.; Kirchmair, J. High-Quality Dataset of Protein-Bound Ligand Conformations and Its Application to Benchmarking Conformer Ensemble Generators, *J. Chem. Inf. Model.* **2017**, *57*, 529–539.

C. de Bruyn Kops verified the results of the statistical analysis of the performance tests and contributed to the manuscript.

[A4] Šícho, M.; de Bruyn Kops, C.; Stork, C.; Svozil, D.; Kirchmair, J. FAME 2: Simple and Effective Machine Learning Model of Cytochrome P450 Regioselectivity, *J. Chem. Inf. Model.* **2017**, *57*, 1832–1846.

C. de Bruyn Kops contributed to the conceptualization of the research.

[A5] Stork, C.; Wagner, J.; Friedrich, N.-O.; de Bruyn Kops, C.; Šícho, M.; Kirchmair, J. Hit Dexter: A Machine-Learning Model for the Prediction of Frequent Hitters, *ChemMedChem* **2018**, *13*, 564–571.

C. de Bruyn Kops contributed to the design of the experiments and contributed to the manuscript.

[A6] Chen, Y.; de Bruyn Kops, C.; Kirchmair, J., Resources for Chemical, Biological, and Structural Data on Natural Products in *Progress in the Chemistry of Organic Natural Products 110: Cheminformatics in Natural Product Research*, (Eds.: Kinghorn, A. D.; Falk, H.; Gibbons, S.; Kobayashi, J.; Asakawa, Y.; Liu, J.-K.), Springer International Publishing, Cham, **2019**, pp. 37–71.

C. de Bruyn Kops, along with Y. Chen and J. Kirchmair, wrote the book chapter.

[A7] Ehm, P. A.; Lange, F.; Hentschel, C.; Jepsen, A.; Glück, M.; Nelson, N.; Bettin, B.; de Bruyn Kops, C.; Kirchmair, J.; Nalaskowski, M.; Jücker, M. Analysis of the FLVR motif of SHIP1 and its importance for the protein stability of SH2 containing signaling proteins, *Cellular Signalling* **2019**, *63*, 109380.

C. de Bruyn Kops, along with J. Kirchmair, performed the molecular modeling and wrote the corresponding portion of the manuscript.

[A8] Šícho, M.; Stork, C.; Mazzolari, A.; de Bruyn Kops, C.; Pedretti, A.; Testa, B.; Vistoli, G.; Svozil, D.; Kirchmair, J. FAME 3: Predicting the Sites of Metabolism in Synthetic Compounds and Natural Products for Phase 1 and Phase 2 Metabolic Enzymes, *J. Chem. Inf. Model.* **2019**, *59*, 3400–3412.

C. de Bruyn Kops contributed to the conceptualization of the research and to improvements in the FAME 3 program.

[A9] Fan, N.; Bauer, C. A.; Stork, C.; de Bruyn Kops, C.; Kirchmair, J. ALADDIN: Docking Approach Augmented by Machine Learning for Protein Structure Selection Yields Superior Virtual Screening Performance, *Molecular Informatics* **2020**, *39*, 1900103.

C. de Bruyn Kops contributed to the conceptualization of the research.

[A10] Stork, C.; Embruch, G.; Šícho, M.; de Bruyn Kops, C.; Chen, Y.; Svozil, D.; Kirchmair, J. NERDD: a web portal providing access to in silico tools for drug discovery, *Bioinformatics* **2020**, *36*, 1291–1292.

C. de Bruyn Kops contributed to the overall design of the web portal, developed the tool GLORY, and designed the results page used by GLORY and FAME 3.

[A11] Wilm, A.; Norinder, U.; Agea, M. I.; de Bruyn Kops, C.; Stork, C.; Kühnl, J.; Kirchmair, J. Skin Doctor CP: Conformal prediction of the skin sensitization potential of small organic molecules, *Chem. Res. Toxicol.*, in review.

C. de Bruyn Kops contributed to the design of the experiments and contributed to the manuscript.

A.2 Conference Contributions

This section lists the author's oral and poster presentations at national and international conferences.

Oral Presentations

[O1] On the Prediction of Xenobiotic Metabolism, *Workweek Hamburg-Groningen*, chemistry symposium for members of Universität Hamburg and the University of Groningen at Universität Hamburg, Sep 29, 2016, Hamburg, Germany.

[O2] Novel Computational Approaches for Early Drug Discovery, *4th Annual G20 World Brain Mapping Summit/Neuroscience-20*, July 5, 2017, Hamburg, Germany.

[O3] Metabolite Structure Prediction Focused on the Cytochrome P450 Enzyme Family, *33rd Molecular Modeling Workshop (MMWS)*, Apr 9, 2019, Erlangen, Germany.

- Lecture award, 3rd place.

[O4] Predicting Sites of Metabolism for Both Phase I and Phase II Metabolism Using Machine Learning Models, *Computer Aided Drug Design Gordon Research Seminar (CADD GRS)*, July 14, 2019, West Dover, VT, USA.

Poster Presentations

[P1] de Bruyn Kops, C.; Mazzolari, A.; Pedretti, A.; Testa, B.; Vistoli, G.; Kirchmair, J. Prediction of Regioselectivity in the Metabolism of Drugs based on Molecular Shape Similarity, *German Conference on Cheminformatics (GCC)*, Nov 6-8, 2016, Fulda, Germany.

[P2] de Bruyn Kops, C.; Friedrich, N.-O.; Kirchmair, J. Alignment-Based Method for the Prediction of Sites of Metabolism of Xenobiotics, *31st Molecular Modelling Workshop (MMWS)*, Mar 27-29, 2017, Erlangen, Germany.

[P3] de Bruyn Kops, C.; Friedrich, N.-O.; Kirchmair, J. Prediction of Xenobiotic Sites of Metabolism: Exploring an Alignment-Based Approach, *Vienna Summer School on Drug Design*, Sep 17-22, 2017, Vienna, Austria.

[P4] de Bruyn Kops, C.; Stork, C.; Kochev, N.; Jeliazkova, N.; Kirchmair, J. Generating Structures of Likely Metabolites Based on Cytochrome P450 Regioselectivity, *32nd Molecular Modelling Workshop (MMWS)*, Mar 12-14, 2018, Erlangen, Germany.

[P5] de Bruyn Kops, C.; Stork, C.; Kochev, N.; Jeliazkova, N.; Kirchmair, J. Metabolite Structure Prediction Benefits from Cytochrome P450 Regioselectivity Prediction, *11th International Conference on Chemical Structures (ICCS)*, May 27-31, 2018, Noordwijkerhout, Netherlands.

[P6] de Bruyn Kops, C.; Stork, C.; Kochev, N.; Jeliazkova, N.; Kirchmair, J. On Cytochrome P450 Regioselectivity Prediction as a Stepping Stone for the Generation of Structures of Likely Metabolites, *EuroQSAR*, Sep 16-20, 2018, Thessaloniki, Greece.

[P7] Šícho, M.; de Bruyn Kops, C.; Stork, C.; Mazzolari, A.; Pedretti, A.; Testa, B.; Vistoli, G.; Svozil, D.; Kirchmair, J. Predicting Sites of Metabolism for Both Phase I and Phase II Metabolism Using Machine Learning Models, *Computer Aided Drug Design Gordon Research Conference (CADD GRC)*, July 13-19, 2019, West Dover, VT, USA.

Supporting Information for [D1]

This appendix contains the supporting information for the publication:

de Bruyn Kops, C.; Friedrich, N.-O.; Kirchmair, J. Alignment-Based Prediction of Sites of Metabolism, *J. Chem. Inf. Model.* **2017**, *57*, 1258–1264.

Supporting Information

Alignment-Based Prediction of Sites of Metabolism

Christina de Bruyn Kops,[1] *Nils-Ole Friedrich,*[1] *Johannes Kirchmair*[1]*

[1] Universität Hamburg, Faculty of Mathematics, Informatics and Natural Sciences, Department of Computer Science, Center for Bioinformatics, Hamburg, 20146, Germany
*J. Kirchmair. E-mail: kirchmair@zbh.uni-hamburg.de. Tel.: +49 (0)40 42838 7303.

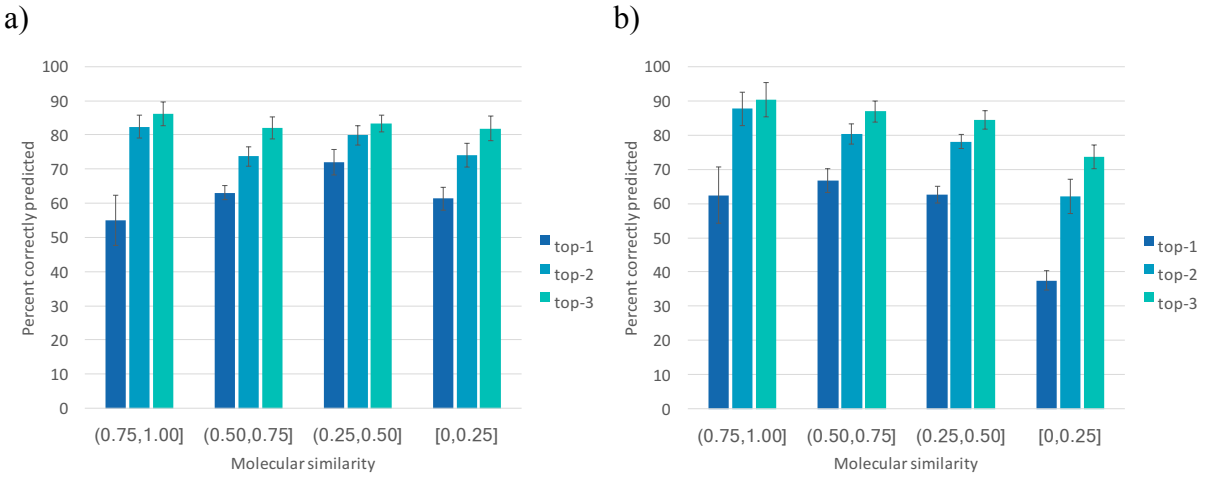


Figure S1. Top-*k* analysis for different similarity ranges applied to a) SoM prediction with the SMARTCyp standard model and b) the combined model, taking into account alignment-based ranking and the SMARTCyp standard model score. Error bars represent standard error. Molecular similarity was calculated using the Tanimoto coefficient of ECFP6-like fingerprints.

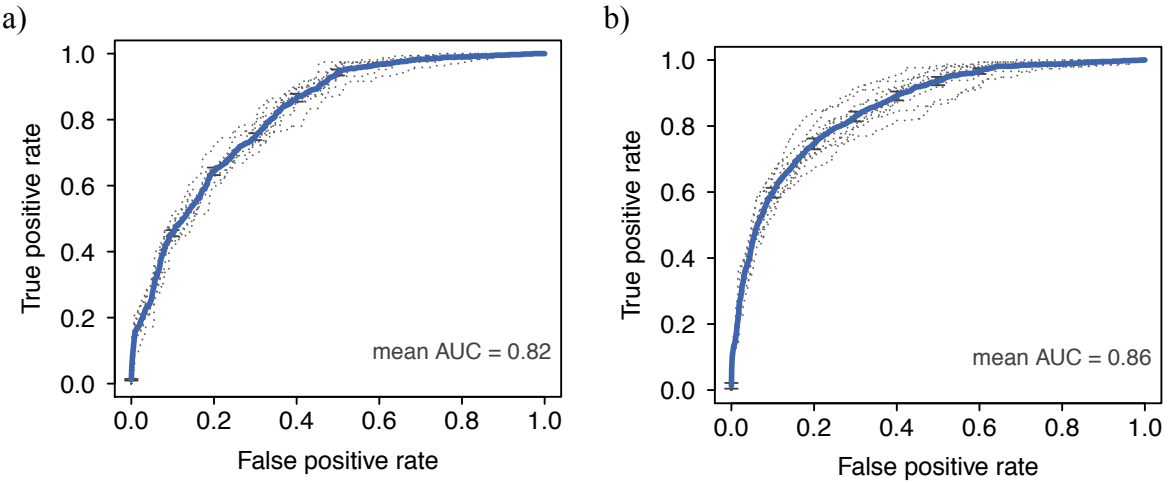


Figure S2. ROC curves for SMARTCyp. a) SoM prediction with the SMARTCyp standard model for the modified Zaretzki dataset; b) combination of alignment-based and SMARTCyp standard model SoM prediction for the modified Zaretzki dataset. Error bars represent standard error.

Table S1. Percentage of Molecules with a Known SoM in the top-*k* Scored Atoms, for top-1, top-2, and top-3 for SMARTCyp and Combined SMARTCyp and Alignment-Based Ranking Predictions using the Zaretzki Dataset.[a]

	top-1 [%]	**top-2 [%]**	**top-3 [%]**
SMARTCyp	64.9 ± 1.4	76.5 ± 0.7	82.6 ± 1.0
Alignment and SMARTCyp combined	56.2 ± 1.2	74.0 ± 1.5	82.1 ± 1.4

[a] Both the arithmetic mean and standard error are reported.

Table S2. BEDROC for SMARTCyp and Combined SMARTCyp and Alignment-Based Ranking Predictions using the Zaretzki Dataset.[a]

BEDROC α	**SMARTCyp**	**Alignment and SMARTCyp combined**
20.0	0.49 ± 0.01	0.60 ± 0.01
80.5	0.65 ± 0.02	0.75 ± 0.02

[a] Both the arithmetic mean and standard error are reported.

Table S3. AUC and Standard Error of AUC for SMARTCyp and Combined SMARTCyp and Alignment-Based Ranking Predictions using the Zaretzki Dataset.

	AUC	**Standard error**
SMARTCyp	0.82	0.0049
Alignment and SMARTCyp combined	0.86	0.0047

Supporting Information for [D2]

This appendix contains the supporting information for the publication:

de Bruyn Kops, C.; Stork, C.; Šícho, M.; Kochev, N.; Svozil, D.; Jeliazkova, N.; Kirchmair, J. GLORY: Generator of the Structures of Likely Cytochrome P450 Metabolites Based on Predicted Sites of Metabolism, *Front. Chem.* **2019**, *7*, 402.

Supplementary Material

Supplementary Table 1: Reaction Rules for GLORY, Including the SMIRKS, Sources, and Common vs Uncommon Label.

Reaction Name	Source[a]	Common/ Uncommon	SMIRKS
aliphatic hydroxylation	(1)	common	"[C;X4:1][H:2]>>[C:1][O][H:2]"
aliphatic hydroxylation with allylic rearrangement	(2)	uncommon‡	"[C;!$(C(=C)CC=C);X3:1]=[C;X3:2][C;!$(C(C=C)C=C);X4:3]([H])>>[C:1](O)-[C:2]=[C:3]"
			"[C;X3:1]=[C;X3:2][C;$(C(C=C)C=C)X4:3]([H])([H])>>[C:1](O)-[C:2]=[C:3]"
aromatic hydroxylation	(1)	common	"[c:1][H:2]>>[c:1][O][H:2]"
aromatic hydroxylation with NIH shift	(1, 3)	common	"[c:1]([H:5])[c;$(c1c([H])c([H])[c;H0]c([H])c([H])1):2][CH3,Br,Cl:3]>>[c:1]([*:3])[c:2][O][H:5]"
aromatic hydroxylation of pyrazolone	other[b]	common‡	"[#6;$([#6]1[#6](=O)[#7][#7][#6]:,=1),$([#6]1:,=[#6][#6](=O)[#7][#7]1):1][H:2]>>[#6:1][O][H:2]"
hydroxylation of cyclopropane	(2)	uncommon‡	"[C:1]1[C:2][C:3]1[C:4]([H])>>[C:1](O)[C:2][C:3]=[C:4]"
amine hydroxylation	(1)	common	"[N:1]([H:3])[#6:2]>>[N:1]([O][H:3])[#6:2]"
N-dealkylation	(1)	common	"[#7:1][C:2]([H])>>[#7:1][H].[C:2]=[O]"
N-dealkylation of piperazine	(4)	common‡	"[*;!#1:1][N;X3:2]1[C:3][C:4][N;X3:5][CH2][CH2]1>>[*:1][N:2][C:3][C:4][N:5]"
N-dealkylation of morpholine	(4)	common‡	[N;X3;$(N1CCOCC1):1][CH2;$(C1NCCOC1)][CH2;$(C1OCCNC1)][O;$(O1CCNCC1):4]>>[N:1].[O:4]
S-dealkylation	(1)	common	"[#16:1][C:2]([H])>>[#16:1]([H]).[C:2]=[O]"
O-dealkylation of methylenedioxyphenyl	(5)	common‡	"[O$(O1c2ccccc2OC1):1][C:2]([H])([H])[O$(O1c2ccccc2OC1):3]>>[O:1]([H]).[C:2](=O)[O-].[O:3]([H])"
S-oxidation	(1)	common	"[#16:1]>>[#16:1](=[O])"
N-oxidation	(1)	common	"[#7;X3,X2;H0:1][#6:2]>>[#7+:1]([O-])[#6:2]"
P-oxidation	(1)	common	"[#15;X3:1]>>[#15;X4:1]=[O]"
aldehyde oxidation to carboxylic acid	(1)	common	"[C:1]([H])=[O:2]>>[C:1](O)=[O:2]"

aldehyde oxidation to olefin plus HCO2H	(1)	uncommon	"[C:1]([H])[C:2]([H])[C:3]([H])=[O:4]>>[C:1]=[C:2].[C:3](O)=[O:4]"
olefin oxidation	(1)	common	“[C:1]([C:3])([C:4])=[C:2]([C:5])([C:6])>>[C:1](=O)([C:3])[C:2]([C:4])([C:5])([C:6])”
acetylene oxidation	(1)	common	“[#6:3][C:1]#[C:2][#6:4]>>[#6:3][C:1]([#6:4])=[C:2](=O)”
			“[#6:3][C:1]#[C:2][#6:4]>>[#6:3][C:1]([#6:4])[C:2](=O)[O]”
oxidation of indole	(1)	uncommon	"[c;$(c1cc2ccccc2n1),$(c1c2ccccc2nc1):1]([H])=,:[c:2]-,:[n:3]>>[C:1](=[O])-[C:2]-[N:3]"
alcohol oxidation	(1)	common	"[C:1]([H])[O:2][H]>>[C:1]=[O:2]"
oxidation of 4-substituted phenol to quinone	(6)	uncommon‡	“[c:1]1([O:7][H])[c:2][c:3][c;X3:4]([!C:8])[c:5][c:6]1>>[C:1]1(=[O:7])-[C:2]=[C:3]-[C;X3:4](=O)-[C:5]=[C:6]-1.[!C:8]”
oxidation of 4-substituted anisole to quinone	(6)	uncommon‡	“[c:1]1([O:7][C:9])[c:2][c:3][c;X3:4]([!C:8])[c:5][c:6]1>>[C:1]1(=[O:7])-[C:2]=[C:3]-[C;X3:4](=O)-[C:5]=[C:6]-1.[!C:8].[C:9]”
oxidation of 4-substituted phenol to quinone if the substituent is not a leaving group	(6)	uncommon‡	“[c:1]1([O:7][H])[c:2][c:3][c;X3:4]([C:8])[c:5][c:6]1>>[C:1]1(=[O:7])-[C:2]=[C:3]-[C:4]([C:8])(O)-[C:5]=[C:6]-1”
oxidation of 4-substituted phenol to quinone imine (dehydrogenation)	(7)	uncommon‡	"[c:1]1([O:7][H])[c:2][c:3][c;X3:4]([N:8][H])[c:5][c:6]1>>[C:1]1(=[O:7])-[C:2]=[C:3]-[C;X3:4](=[N:8])-[C:5]=[C:6]-1"
oxidation of 4-substituted anisole to quinone if the substituent is not a leaving group	(6)	uncommon‡	“[c:1]1([O:7][C:9])[c:2][c:3][c;X3:4]([C:8])[c:5][c:6]1>>[C:1]1(=[O:7])-[C:2]=[C:3]-[C:4]([C:8])(O)-[C:5]=[C:6]-1.[C:9]”
oxidation of 1,4-dihydropyrridines	(1)	common	"[N;X3:1]1([H])[#6:2]=[#6:3][#6;X4:4]([H])[#6:5]=[#6:6]1>>[n;H0:1]1=[#6:2][#6:3]=[#6:4][#6:5]=[#6:6]1"
aliphatic epoxidation	(1)	common	"[C:1]=[C:2]>>[C:1]1[C:2][O]1"
N-dearylation	(1)	uncommon	“[c;R1:1]1[c;R1:2][c;R1:3][c;R1:4][c;R1:5][c;R1:6]1[N:7][c:8]>>[C:1]1=[C:2]-[C:3]-[C:4]=[C:5]-[C:6]1=[O].[c:8][NH2:7]”
O-dearylation	(1)	uncommon	“[c;R1:1]1[c;R1:2][c;R1:3][c;R1:4][c;R1:5][c;R1:6]1[O:7][c:8]>>[C:1]1=[C:2]-[C:3]-[C:4]=[C:5]-[C:6]1=[O].[c:8][O:7]”
			“[c;R1:1]1[c;R1:2][c;R1:3][c;R1:4][c;R1:5][c;R1:6]1[O:7][c;R1:8]2[c;R1:9][c;R1:10][c;R1:11][c;R1:12][c;R1:13]2>>[C:1]1=[C:2]-[C:3]-[C:4]=[C:5]-[C:6]1=[O:7].[C:8]2(=[O])-[C:9]=[C:10]-[C:11]-[C:12]=[C:13]-2”
			“[c;R1:1]1[c;R1:2][c;R1:3]([O:20][H])[c;R1:4][c;R1:5][c;R1:6]1[O:7][c:8]>>[C:1]1=[C:2]-[C:3](=[O:20])-[C:4]=[C:5]-[C:6]1=[O].[c:8][O:7]”
			“[c;R1:1]1[c;R1:2][c;R1:3]([O:20][H])[c;R1:4][c;R1:5][c;R1:6]1[O:7][c;R1:8]2[c;R1:9][c;R1:10][c;R1:11][c;R1:12][c;R1:13]2>>[C:1]1=[C:2]-[C:3](=[O:20])-[C:4]=[C:5]-[C:6]1=[O:7].[C:8]2(=[O])-[C:9]=[C:10]-[C:11]-[C:12]=[C:13]-2”

deformylation	(1)	uncommon	"[C:1]([H])-[C:2]([H])-[C:3]=[O:4]>>[C:1]=[C:2].[C:3]=[O:4]"
oxidative desulfuration of phosphor	(8)	common‡	"[*:1][P:2](=S)([*:3])[*:4]>>[*:1][P:2](=O)([*:3])[*:4]"
desulfuration of carbon	(8)	uncommon‡	“[*:1][C:2](=S)[*:3]>>[*:1][C:2](=O)[*:3]”
reduction of N-oxide	(1)	uncommon	“[#7+;X4:1]([O-])>>[#7;X3:1]”
reduction of RNOR	(1)	uncommon	“[#8;$([#8][#6]):1][#7:2]:,=[#6:3]>>[#8:1]([H]).[#7:2]([H])([H])-[#6:3]([H])”
			“[#8;$([#8][#6]):1][#7:2][#6;!X4:3]>>[#8:1]([H]).[#7:2].[#6:3]=[O]”
			“[#8;$([#8][#6]):1][#7;$([#7][#6]):2]>>[#8:1]([H]).[#7:2]([H])”
reduction of nitro group	(1)	uncommon	"[N;X3:1](=O)=[O]>>[NH2:1]"
			"[N+;X3:1](=O)[O-]>>[NH2:1]"
reduction of C- or N-nitroso compound	(1)	uncommon	“[C,N:1][N;X2:2](=O)>>[C,N:1][N;H2:2]”
azo reduction	(1)	uncommon	“[#6:1][N:2]=[N:3][#6:4]>>[#6:1][NH2:2].[NH2:3][#6:4]”
hydrazine reduction	(1)	uncommon	“[NX3:1]-[NX3:2]>>[N:1]([H]).[N:2]([H])”
alkyl oxidative dehalogenation	(1)	common	“[C:1]([H])[F,Cl,Br:2]>>[C:1]=[O].[F,Cl,Br:2]”
benzyl oxidative dehalogenation	(1, 9)	uncommon	“[c;$([c;!H]1ccccc1),$(c1[c;!H]cccc1),$(c1c[c;!H]ccc1),$(c1cc[c;!H]cc1),$(c1ccc[c;!H]c1),$(c1cccc[c;!H]1):1][F,Cl,Br,I:2]>>[c:1][O].[F,Cl,Br,I:2]”
reductive dehalogenation	(1)	uncommon	“[C:1]([F,Cl,Br:3])[C:2]([F,Cl,Br:4])>>[C:1]=[C:2].[*:3].[*:4]”
alkyl dehydrogenation	(1)	uncommon	“[C:1]([H])-[C:2]([H])>>[C:1]=[C:2]”
dehydrogenation of N-C bond	(8)	uncommon‡	“[N;X3:1]([H])[C;!H3:2][H]>>[N:1]=[C:2]”
			"[#7;X3:1]([H]):,-[#6;!H3:2]([H]):,=[#6:3]-[C:4]([H])>>[#7:1]=[#6:2]-[#6:3]=[C:4]"
			“[N;X3:1][C:2][H]>>[N+:1]=[C:2]”
oxidative ether cleavage (O-dealkylation)	(1)	common	"[#6:1][O:2][C:3]([H])>>[#6:1][O:2].[C;X3:3](=O)"
oxidative ester cleavage	(1)	uncommon	"[C$(C(O)([#6])=O):2][O:3][C:4][H]>>[C:2][O:3].[C:4]=[O]"
monothiophosphate ester cleavage	(8)	uncommon‡	“[S:1]=[P$(P(O)(O)=S):2][O:3][#6:4]>>[S:1]=[P:2][O:3].[#6:4][O]”
			“[S:1]=[P$(P(O)(O)=S):2][O:3][#6:4]>>[S:1].[O]=[P:2][O:3].[#6:4][O]”
phosphoester cleavage	(1)	uncommon	“[O:1]=[P$(P(O)(O)=O):2][O:3][#6:4]>>[O:1]=[P:2][O:3].[#6:4][O]”

carbamate cleavage	(8)	uncommon‡	“[#7:1][C;$([C](O)=O):2][O:3][C:4]>>[#7:1][H].[C:2]=[O:3].[C:4][O]”
carbamide cleavage	(2)	uncommon‡	“[N:1][C;$([C](N)(N)=O):2][N:3]>>[N:1][C:2].[N:3]”
oxidation of N-nitrosamine	(1)	uncommon	“[N$(N(C)C):1]([C:3][H])[N$(N(N)=O):2]>>[N:1]([H])([H]).[N:2]([O-]).[C:3](=O)”
scission of unsaturated fatty acid peroxides	(1, 2)	uncommon	"[C:1]([H])=[C:2]-[C:3]=[C:4]-[C:5]-[O:6]([O])>>[C:1]=[C:2][C:3](O)[C:4]1-[C:5]-[O:6]1"
			"[C:1]([H])=[C:2]-[C:3]=[C:4]-[C:5]-[O:6]([O])>>[C:1]([O])[C:2]=[C:3][C:4]1-[C:5]-[O:6]1"
dehydration of an aldoxime to a nitrile	(1)	uncommon	“[C:1]([H])=[N:2][O]([H])>>[C:1]#[N:2]”
cyclization to 6-membered lactone	(10)	uncommon	“[C:1]([H])([OH])[#6:2][#6:3][#6:4][C;$(C=O):5][O:6][*:7]>>[C:1]1[#6:2][#6:3][#6:4][C;$(C=O):5][O:6]1.[*:7]”
cyclization to 5-membered lactone	(1)	uncommon	“[C:1]([H])[c:2][c:3][C;$(C=O):4][O:5][C,#1:6]>>[C:1]1[c:2][c:3][C:4][O:5]1.[*:6]”
cyclization to 6-membered NCN ring	(2)	uncommon	“[N;X3:1]([H])~[*:2]~[*:3]~[*:4]~[N:5]-[C:6]([H])>>[N:1]1~[*:2]~[*:3]~[*:4]~[N:5]-[C:6]1”
cyclization to 5-membered NCN ring	(2)	uncommon	“[N;X3:1]([H])~[*:2]~[*:3]~[N:5]-[C:6]([H])>>[N:1]1~[*:2]~[*:3]~[N:5]-[C:6]1”
cyclization to furan	(1)	uncommon	“[O:1]=[C;R1:2][C;R1:3]=[C:4][C:5]([H])[H]>>[O:1]1[C:2]=[C:3][C:4]=[C:5]1”
cyclobutamine expansion	(1)	uncommon	"[C:1]1-[C:2]-[C:3]-[C$(C1(C)CCC1):4]1[N:5]([H])>>[C:1]1-[C:2]-[C:3]-[C:4]=[N+:5]1"
oxidation of spiro[2,5]oxane	(11)	uncommon	"[C:1]1[C:2]2([C:3][C:4]2)[C:5]([H])[C:6][C:7][C:8]1>>[C:1]1[C:2]2(O)[C:3][C:4][C:5]2[C:6][C:7][C:8]1"
D-homoannulation of 17 alpha-ethinyl steroids	(1)	uncommon	"[C$([#6R1]~1~[#6R1]~[#6R1]~[#6R2]~2~[#6R2]~1~[#6R1]~[#6R1]~[#6R2]~3~[#6R2]~2~[#6R1]~[#6R1]~[#6R2]~4~[#6R1]~[#6R1]~[#6R1]~[#6R1]~[#6R2]~3~4):1]1([O:6][H])([C:7]#[C:8])[C:2][C:3][C:4][C:5]1>>[C:1]1(=[O:6])[C:7](=[C:8](O))[C:2][C:3][C:4][C:5]1"

[a] Note that the source(s) provided for each reaction type is not an exhaustive list. Many reaction types were found in multiple publications. When listing the source in this table, priority was given to the 2001 review by FP Guengerich (1) (because that is where the common/uncommon designation came from) and a second source only provided if it provided additional information used in the development of the SMIRKS. If the reaction was not found in reference 1, then the most general source that was applicable to the development of the SMIRKS was provided.

[b] Special case of aromatic hydroxylation. This additional reaction type was needed because pyrazolone is not recognized as aromatic by CDK and Ambit SMIRKS.

‡ This designation was based on extrapolation, as this reaction type was not included in the 2001 review by FP Guengerich (1).

Supplementary Table 2: Evaluation Results for GLORY in MaxEfficiency Mode with Varying Site of Metabolism (SoM) Probability Cutoffs on the Manually Curated Test Dataset.

SoM Probability Cutoff	**0.4**	**0.3**	**0.2**	**0.1**
Precision	0.22	0.18	0.16	0.13
Recall	0.41	0.51	0.64	0.74
Total number of predicted metabolites	148	226	327	465
Number of successfully predicted reported metabolites[a]	33	41	52	60
Number of molecules for which no metabolites could be predicted	3	2	0	0
Top-1	65.52 %[b]	65.52 %	68.97 %	68.97 %
Top-2	65.52 %[b]	65.52 %	72.41 %	72.41 %
Top-3	65.52 %[b]	68.97 %	75.86 %	75.86 %

[a] The total number of reported metabolites in the dataset was 81.

[b] Note: If it existed, the best rank of the a known metabolite was always 1 for the SoM probability cutoff of 0.4. No known metabolite was predicted for nearly half of the parent molecules.

References

1. FP Guengerich. Common and uncommon cytochrome P450 reactions related to metabolism and chemical toxicity. *Chem Res Toxicol* 14, 611–650 (2001)

2. PR Ortiz de Montellano; SD Nelson. Rearrangement reactions catalyzed by cytochrome P450s. *Arch Biochem Biophys* 507, 95–110 (2011)

3. PG Wislocki; GT Miwa; AYH Lu. Reactions Catalyzed by the Cytochrome P-450 System. In: Enzymatic Basis of Detoxication. WB Jakoby, ed. , Elsevier (1980)

4. J Bolleddula; K DeMent; JP Driscoll; P Worboys; PJ Brassil; DL Bourdet. Biotransformation and bioactivation reactions of alicyclic amines in drug molecules. *Drug Metab Rev* 46, 379–419 (2014)

5. M Murray. Toxicological actions of plant-derived and anthropogenic methylenedioxyphenyl-substituted chemicals in mammals and insects. *J Toxicol Environ Health B Crit Rev* 15, 365–395 (2012)

6. PR Ortiz de Montellano. Substrate Oxidation by Cytochrome P450 Enzymes. In: Cytochrome P450: Structure, Mechanism, and Biochemistry. PR Ortiz de Montellano, ed. , Springer (2015)

7. B Testa; SD Krämer. The Biochemistry of Drug Metabolism – An Introduction Part 2. Redox Reactions and Their Enzymes. *Chem Biodivers* 4, 257–405 (2007)

8. A Parkinson. Biotransformation of Xenobiotics. In: Casarett and Doull's Toxicology: The Basic Science of Poisons. CD Klaassen, ed. , McGraw-Hill (2007)

9. IM Rietjens; C den Besten; RP Hanzlik; PJ van Bladeren. Cytochrome P450-catalyzed oxidation of halobenzene derivatives. *Chem Res Toxicol* 10, 629–635 (1997)

10. S-F Zhou; Z-W Zhou; L-P Yang; J-P Cai. Substrates, inducers, inhibitors and structure-activity relationships of human Cytochrome P450 2C9 and implications in drug development. *Curr Med Chem* 16, 3480–3675 (2009)

11. K Auclair; Z Hu; DM Little; PRO de Montellano; JT Groves. Revisiting the Mechanism of P450 Enzymes with the Radical Clocks Norcarane and Spiro[2,5]octane. *J Am Chem Soc* 124, 6020–6027 (2002)

Supporting Information for [D3]

This appendix contains the supporting information for the publication:

de Bruyn Kops, C.; Šícho, M.; Mazzolari, A.; Kirchmair, J. GLORYx: Prediction of the Metabolites Resulting from Phase 1 and Phase 2 Biotransformations of Xenobiotics, *Chem. Res. Toxicol.* **2020**, DOI 10.1021/acs.chemrestox.0c00224.

Supporting Information

GLORYx: Prediction of the Metabolites Resulting from Phase 1 and Phase 2 Biotransformations of Xenobiotics

*Christina de Bruyn Kops[1], Martin Šícho[2], Angelica Mazzolari[3], Johannes Kirchmair[1,4]**

[1] Center for Bioinformatics (ZBH), Department of Informatics, Faculty of Mathematics, Informatics and Natural Sciences, Universität Hamburg, 20146 Hamburg, Germany

[2] CZ-OPENSCREEN: National Infrastructure for Chemical Biology, Department of Informatics and Chemistry, Faculty of Chemical Technology, University of Chemistry and Technology Prague, 166 28 Prague 6, Czech Republic

[3] Facoltà di Scienze del Farmaco, Dipartimento di Scienze Farmaceutiche "Pietro Pratesi", Università degli Studi di Milano, I-20133 Milan, Italy

[4] Department of Pharmaceutical Chemistry, Faculty of Life Sciences, University of Vienna, 1090 Vienna, Austria

* Correspondence: johannes.kirchmair@univie.ac.at; Tel.: +43 1-4277-55104

TABLES

Table S1. Enzymes Excluded From Consideration When Extracting Relevant Metabolites From DrugBank.

Enzyme
Cocaine esterase
Thymidine phosphorylase
Serum albumin
Ribulose-phosphate 3-epimerase
UDP-galactose 4-epimerase
cGMP-specific 3'5'-cyclic phosphodiesterase
Dihydropyrimidinase-related protein 2
Aromatic-L-amino-acid decarboxylase
Elongation of very long chain fatty acids protein 4
Elongation of very long chain fatty acids protein 5
Hemoglobin subunit beta
Hemoglobin subunit alpha
Selenocysteine lyase
Lysosomal protective protein
Enoyl-CoA hydratase mitochondrial
NADPH--cytochrome P450 reductase
Cytochrome b

Table S2. Descriptors Used for Principal Component Analysis.

Name	Description (1)
a_acc	Hydrogen bond acceptor atom count
a_acid	Acidic atom count
a_aro	Aromatic atom count
a_base	Basic atom count
a_don	Hydrogen bond donor atom count
a_heavy	Heavy atom count
a_hyd	Hydrophobic atom count
a_nB	Boron atom count
a_nBr	Bromine atom count
a_nC	Carbon atom count
a_nCl	Chlorine atom count
a_nF	Fluorine atom count
a_nH	Hydrogen atom count
a_nI	Iodine atom count
a_nN	Nitrogen atom count
a_nO	Oxygen atom count
a_nP	Phosphorus atom count
a_nS	Sulfur atom count
b_ar	Number of aromatic bonds
b_count	Number of bonds
b_double	Number of double bonds
b_rotN	Number of rotatable bonds
b_rotR	Fraction of rotatable bonds[a]
b_single	Number of single bonds
b_triple	Number of triple bonds
chiral	Number of chiral centers

FCharge	Total charge of the molecule
logP(o/w)	Log of the octanol/water partition coefficient
logS	Log of the aqueous solubility (mol/L)
mr	Molecular refractivity
PC+	Total positive partial charge
PC-	Total negative partial charge
rings	Number of rings
TPSA	Polar surface area ($Å^2$)
vdw_area	Area of van der Waals surface ($Å^2$)
vdw_vol	van der Waals volume ($Å^3$)
vsa_acc	Approximation of the sum of VDW[b] surface areas ($Å^2$) of pure hydrogen bond acceptors[c]
vsa_acid	Approximation of the sum of VDW surface areas of acidic atoms ($Å^2$)
vsa_base	Approximation of the sum of VDW surface areas of basic atoms ($Å^2$)
vsa_don	Approximation of the sum of VDW surface areas of pure hydrogen bond donors[d]
vsa_hyd	Approximation of the sum of VDW surface areas of hydrophobic atoms ($Å^2$)
vsa_other	Approximation of the sum of VDW surface areas ($Å^2$) of atoms typed as "other"
vsa_pol	Approximation of the sum of VDW surface areas ($Å^2$) of polar atoms
Weight	Molecular weight

[a] b_rotN divided by the number of bonds between heavy atoms
[b] VDW = van der Waals
[c] Not counting acidic atoms and atoms that are both hydrogen bond donors and acceptors
[d] Not counting basic atoms and atoms that are both hydrogen bond donors and acceptors

Table S3. Number of Molecules Used to Train the FAME 3 Reaction Type-Specific SoM Prediction Models.

Reaction class	Number of molecules	ClassID(s) from MetaQSAR
Glucuronidations & glycosylations	440 + 153 = 593	14, 15
GSH & RSH[a] conjugations	243	17
Sulfonations	148	16
Methylations	94	20
Acetylations & acylations	83	18

[a]RSH = protein thiol

Table S4. Ranking Performance of Phase 2 Metabolite Prediction using the Reaction Rules from SyGMa and Various Formulas for Combining the Predicted SoM Probabilities[a] with SyGMa's Reaction Probabilities.

Score equation[b]	AUC of rank-based ROC curve
S x R	0.85
(S + R) / 2	0.82
(2S + R) / 3	0.81
(3S + R) / 4	0.80
(5S + R) / 6	0.80
(10S + R) / 11	0.80
(S + 2R) / 3	0.82
(S + 3R) / 4	0.82
(S + 5R) / 6	0.82
(S + 10R) / 11	0.83

[a] The SoM probabilities were predicted with FAME 3 model P2
[b] S = SoM probability, R = reaction probability

FIGURES

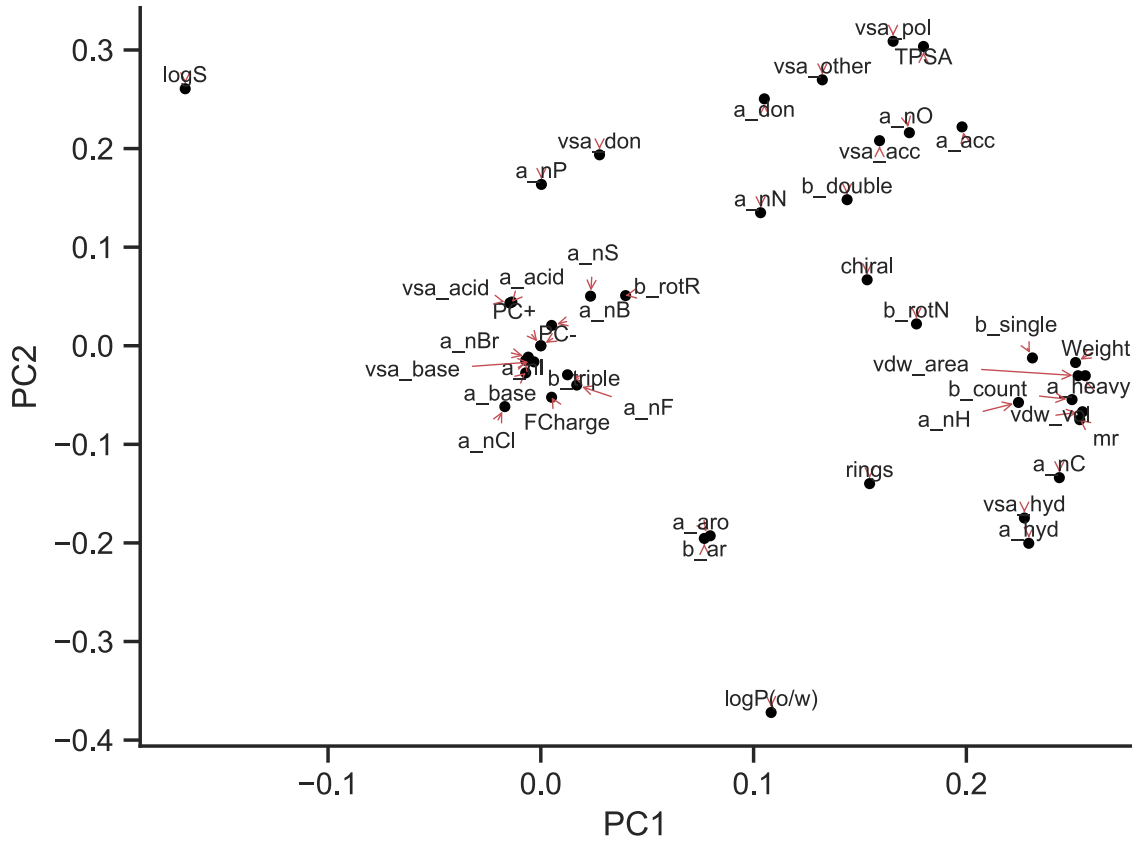


Figure S1. PCA loading plot for the PCA plot shown in Figure 1D. The PCA compares parent molecules from DrugBank and MetXBioDB using 44 physicochemical descriptors (Table S2). The percentage of the total variance explained by each of the first two principal components is 35.81% for PC1 and 10.69% for PC2.

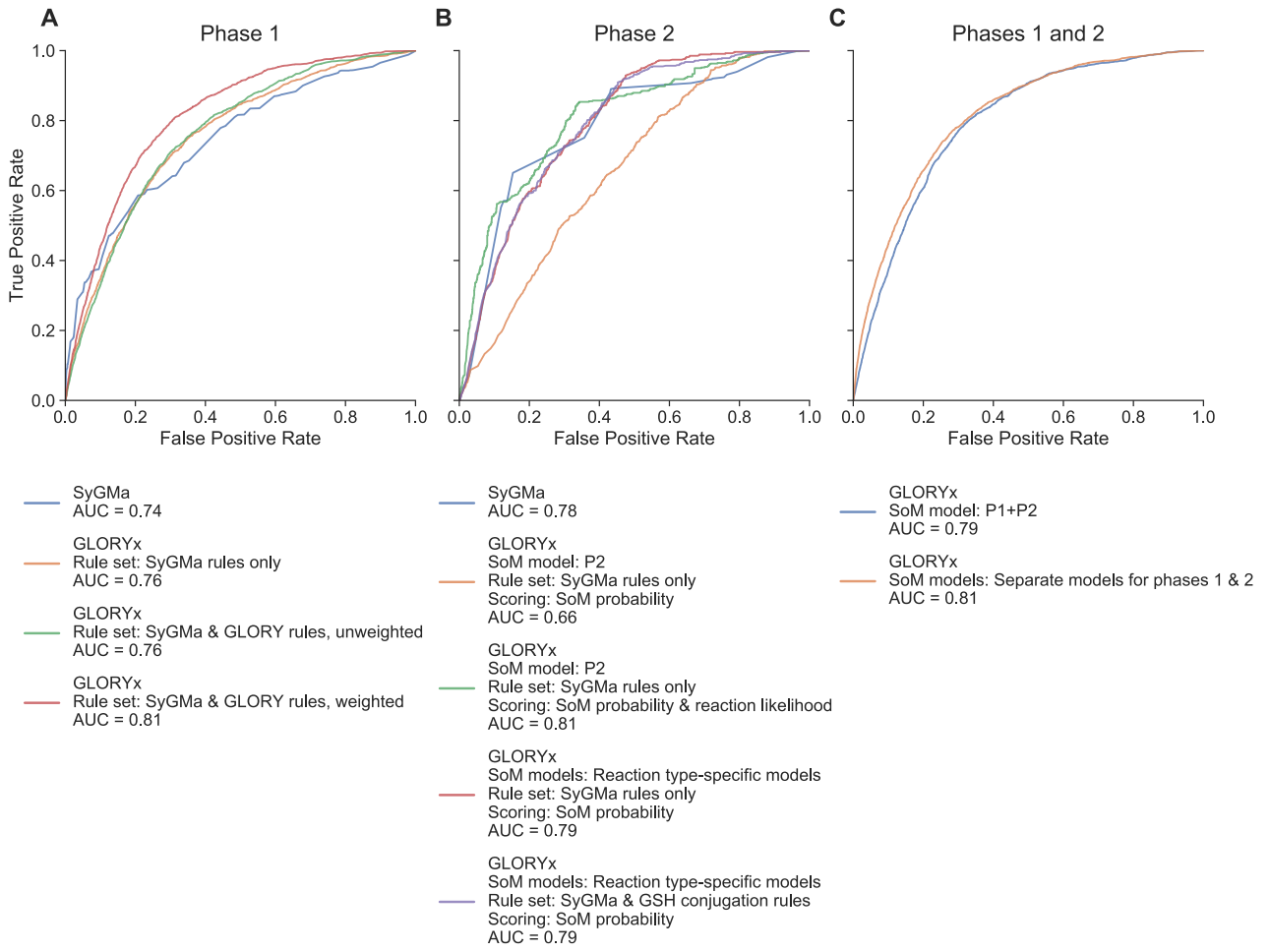


Figure S2. Score-based ROC curves for the evaluation of metabolite prediction performance on the reference dataset. (A) Comparison of GLORYx, which scores its predicted metabolites based on predicted SoM probability, to SyGMa, which uses reaction probability-based scoring, for phase 1 metabolite prediction. Weighted rules refer to the weighting of the SoM probability-based score based on whether the reaction type is designated common or uncommon. (B) Comparison of the ranking performance of GLORYx with different scoring approaches and rule sets, as well as a direct comparison to SyGMa's performance, for phase 2 metabolite prediction. The scoring approach that is based on both SoM probability and reaction probability is achieved by a simple multiplication of the two components. (C) Comparison of the ranking performance of GLORYx for combined prediction of metabolites for phases 1 and 2 metabolism, using different SoM prediction approaches to score the predicted metabolites. The predicted metabolites are scored based on predicted SoM probability. The rule set in both cases is the same and is made up of the final phase 1 rule set (SyGMa and GLORY rules) and final phase 2 rule set (SyGMa and GSH conjugation rules). Note that the score-based ROC curves for SyGMa should be viewed cautiously because SyGMa's scoring approach was only intended to compare scores among predicted metabolites of the same parent molecule (i.e. a rank-based comparison).

Figure S3. Variability in the ranking performance of SyGMa and GLORYx on the test set based on the rank and the score of the predicted metabolites. The data points were calculated by systematically removing one parent molecule from the test set at a time and calculating the AUC from the remaining predictions. There are therefore 37 AUC data points for each combination of tool and AUC type, corresponding to the size of the test set.

REFERENCES

(1) Chemical Computing Group ULC. *MOE User Guide, MOE 2018.01*. Chemical Computing Group ULC: Montreal, Canada, 2018.

Eidesstattliche Erklärung

Hiermit erkläre ich an Eides statt, dass ich die vorliegende Dissertationsschrift selbst verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

I hereby declare upon oath that I have written the present dissertation independently and have not used further resources and aids than those stated.

Hamburg, den 13. November 2020

Christina de Bruyn Kops