Journal of Medicinal Chemistry

pubs.acs.org/jmc

Article

Predicting Regioselectivity of AO, CYP, FMO, and UGT Metabolism Using Quantum Mechanical Simulations and Machine Learning

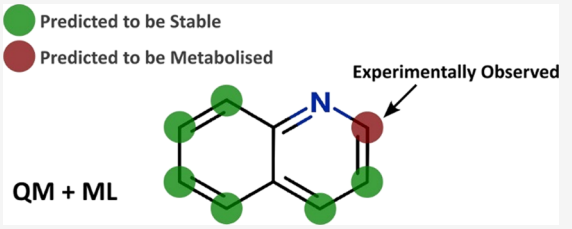
Mario Öeren,* Peter J. Walton, James Suri, David J. Ponting, Peter A. Hunt, and Matthew D. Segall

Cite This: https://doi.org/10.1021/acs.jmedchem.2c01303

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Unexpected metabolism in modification and conjugation phases can lead to the failure of many late-stage drug candidates or even withdrawal of approved drugs. Thus, it is critical to predict the sites of metabolism (SoM) for enzymes, which interact with drug-like molecules, in the early stages of the research. This study presents methods for predicting the isoform-specific metabolism for human AOs, FMOs, and UGTs and general CYP metabolism for preclinical species. The models use semi-empirical quantum mechanical simulations, validated using experimentally obtained data and DFT calculations, to estimate the reactivity of each SoM in the context of the whole molecule. Ligand-based models, trained and tested using high-quality regioselectivity data, combine the reactivity of the potential SoM with the orientation and steric effects of the binding pockets of the different enzyme isoforms. The resulting models achieve κ values of up to 0.94 and AUC of up to 0.92.



INTRODUCTION

The characterization of xenobiotic metabolism using *in silico* methods enables chemists to predict sites of metabolism (SoM) of potential drug candidates, agrochemicals, nutritional supplements, and cosmetics. Therefore, optimizing the structure of new chemical entities can be more cost-effective, and toxic metabolites can be identified early in the project.[1,2] Historically, predictive models have targeted the metabolism by human isoforms of the Cytochrome P450 (CYP) family of enzymes due to their irrefutable importance in the metabolism of drug-like compounds in the modification phase (Phase I).[3] However, studies on how to predict metabolism for other modification phase enzymes, such as aldehyde oxidases (AO)[4,5] and flavin-containing monooxygenases (FMO),[6] and conjugation phase (Phase II) enzymes, such as uridine 5′-diphospho-glucuronosyltransferases (UGT),[7−11] are increasing in prevalence.

There are many reasons why chemists are interested in expanding their portfolio of predictive models beyond CYPs. For example, introducing azaheterocyclic rings into compounds decreases their lability toward CYP metabolism but increases the likelihood of oxidation by AOs. The rapid clearance of molecules by AOs (not predicted by CYP-only modeling) has caused the discontinuation of multiple projects during clinical trials.[12−14] Similarly, the role of FMOs has been underestimated—the chemical space of its substrates overlaps with that of CYPs and metabolism by FMOs has sometimes been falsely attributed to CYPs. Predicting the sites of metabolism by FMOs would help chemists tailor compounds to be metabolized by multiple enzyme families, thereby avoiding drug−drug interactions, and detect potential toxic metabolites such as sulfenic and sulfinic acids, and S-oxides and S,S-dioxides of thiocarbonyls.[15,16] Finally, UGTs are the major enzymes contributing to the conjugation phase; approximately 15% of known drugs are glucuronidated.[17] Predicting metabolism by UGTs helps researchers to avoid the inactivation of potential drug candidates and detect the formation of potentially toxic acyl glucuronides.[18]

Despite the success of human CYP models, tests on animals are still conducted regularly. Testing the metabolism of potential drugs in animal models is primarily for toxicology studies. As each animal's metabolism is unique, the human metabolism cannot be replicated precisely by a single preclinical species, leading to the criterion that these trials must be conducted in at least two mammalian species (one rodent and one non-rodent). *In silico* modeling of the metabolism of preclinical species could aid in ensuring that the preclinical trials produce the likely human metabolites, using the model as an indicator of the best preclinical species. As well as the ethical benefits of this modeling approach, trials would be quicker and less expensive.

This study aims to build models that predict the SoM for various isoforms of AOs, FMOs, and UGTs found in humans. In addition, the study aims to expand the existing CYP SoM prediction models to preclinical species. The following subsections give a brief overview of the enzymes—their

Received: August 9, 2022

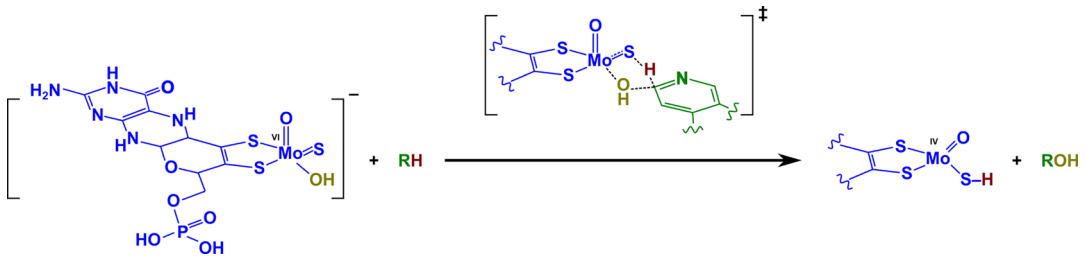


Figure 1. The transition state for oxidation by AO.

substrate space and reaction types—and the available data for building and validating models. Following this, we summarize the spectra of available modeling methods, give an overview of the existing models for the enzymes above and provide a rationale to train new models based on the reactivity-accessibility approach.

Aldehyde Oxidases. The existence of AOs in the liver was predicted as early as 1936.[19] However, the first time they were isolated was in 1940 by Gordon et al.[20] AOs were initially observed to react with aldehydes, hence the name, but they are also known to be responsible for catalyzing the oxidation of aromatic heterocycles[21] and iminium ions.[22−25] It is intriguing that AOs, which are considered to contribute to the modification phase, have also been observed to catalyze the reduction[26] of various molecules, e.g., nitro-compounds. However, with a few exceptions,[27] the reductive metabolism occurs at lower oxygen concentrations and is thought to play a role in human physiology (sensing low oxygen tensions).[24] In 2015, Sodhi et al. reported an additional metabolic activity mediated by AOs—amide hydrolysis.[28] It should be noted that the prevalent chemical reaction of AOs is considered to be oxidation and the majority of the known substrates are azaheterocycles.[25] Thus, this study concentrates on AO oxidation, and reactions such as reduction and hydrolysis fall out of the scope of the present work.

AOs belong to the molybdo-flavoenzyme family of enzymes and require the Molybdenum-cofactor (MoCo) alongside flavin adenine dinucleotide (FAD) and iron−sulfur clusters to catalyze the aforementioned reactions.[24,29] We present the detailed catalytic cycle of AOs in the Supporting Information; here, we concentrate on the oxidation step, which is understood to be the rate-limiting step of catalysis (the detailed descriptions for the catalytic cycles for CYP, FMO, and UGT can be found from our previous publications.[1,30] The MoCo structure varies between molybdoenzymes,[29] and in the case of AOs the molybdenum atom is surrounded by bidentate molybdopterin, double-bonded oxygen, and sulfur atoms and a hydroxide ion. The currently accepted hypothesis, suggested by Skibo et al.,[31] states that after the substrate is bound to the active site, the hydroxide ion of MoCo makes a nucleophilic attack on the carbon atom of the substrate, while the proton and two electrons (from the carbon atom) are transferred to the sulfur atom of MoCo. Computational studies using density functional theory (DFT) by Montefiori et al. and Alfaro et al. have confirmed the proposed concerted reaction.[14,32] The described transition state is depicted in Figure 1.

AOs can be found in certain prokaryotes and most eukaryotes, including mice, rats, rabbits, dogs, rhesus monkeys, chimpanzees, and humans. Unlike CYPs, the AO family does not have many isoforms; mice and rats have the largest number of isoforms—four, and humans have only one (orthologous to the Aox1 found in mice). The single isoform for humans is found in the liver, respiratory, digestive, urogenital, and endocrine tissues, with the majority in the liver. It is contained in the cytosol of the cells.[29]

Prediction of AO-mediated reactions has become an important avenue in drug development. Structural motifs such as azaheterocycles, in which carbon atoms are prevalent SoM for AOs, are common in drug-like molecules. In addition, researchers are actively trying to reduce the CYP-mediated metabolism, which gives rise to the increased prevalence of other routes of metabolism. There are several examples where AO metabolism has terminated a drug discovery program due to high metabolic clearance (e.g., carbazeran,[33] BIBX1382[34]) or toxicity (e.g., JNJ-38877605[35]).[13]

The first attempt to predict the SoM by AOs was by Torres et al., who assessed the relative energy values of a simplified tetrahedral intermediate structure for all potential SoM. The method was very successful (considering it did not take into account the protein structure) and had an accuracy of 93%. The drawback of the method was its slow execution time since it depended on the DFT method and the set of compounds for testing was relatively small—27 compounds.[4] The results were later used by Jones and Korzekwa to predict clearance for drugs and drug candidates metabolized by AOs[36] and Xu et al., who built a decision tree model based on the stability of the intermediate structure and an additional steric descriptor.[37] Montefiori et al. expanded the work from using relative energy values from the tetrahedral intermediate to calculating the activation energy value (E_a) using a simplified MoCo. While the activation energy was excellent in identifying the site of metabolism, only six substrates were tested. They also tried various other proxy descriptors (e.g., stability of the product, ESP charges) for the E_a and found out that they were as good but considerably faster to calculate.[14] Montefiori et al. subsequently expanded the study to a more extensive data set (78 compounds) and used various aforementioned proxy descriptors to build classification models. The resulting models had receiver operating characteristic area under curve (ROC-AUC) values of up to 0.96 and κ values of up to 0.89.[5] A notable experimental and computational study was performed by Lepri et al., who acquired or synthesized over 270 compounds to study the oxidation of azaheterocycles and hydrolysis of amides by AOs.[12] The study yielded guidelines for recognizing carbon atoms labile to AO metabolism and agreed with the work of Montefiori et al.[14] that the most positively charged carbon within an azaheterocycle is the potential site of metabolism.

Cytochromes P450s. Quantitively, the CYP enzymes are the most important family for the metabolism of xenobiotics. These enzymes contribute to the modification phase and are responsible for the metabolism of 75−90% of hepatically cleared drugs in humans.[3,38,39] The catalytic action of CYPs is predominantly that of a monooxygenase (*C*-hydroxylation,

https://doi.org/10.1021/acs.jmedchem.2c01303
J. Med. Chem. XXXX, XXX, XXX−XXX

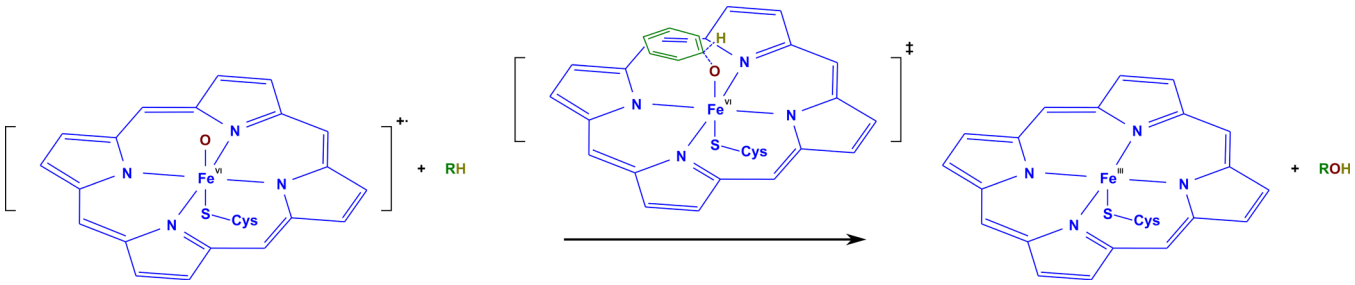


Figure 2. The transition state for oxidation by CYP.

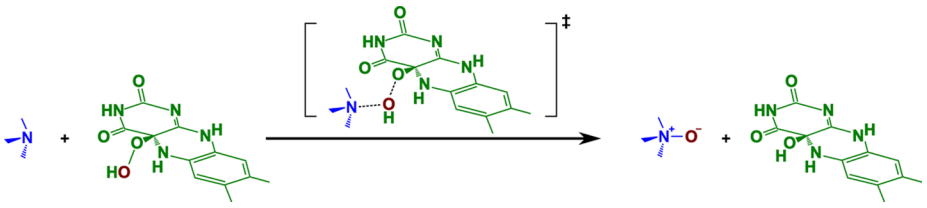


Figure 3. The transition state for oxidation by FMO.

heteroatom oxygenation, dealkylation) but also includes epoxide formation and aromatic dehalogenation, amongst other reactions.[40] As with the previous enzyme, this work will concentrate on the most prevalent reactions, e.g., aliphatic- and aromatic hydroxylation, aldehyde oxidation, double bond epoxidation and *N*- and S-oxidation.[1] The catalytic cycle for these reactions is briefly described in the following paragraph, but for a comprehensive overview of the catalytic cycle and the various CYP reaction types, the reader is referred to the work by Isin and Guengerich,[40] Coon,[41] Manikandan and Nagini[42] and Jung.[43]

The catalysis by CYPs requires the haem-iron center as a cofactor and the reduced nicotinamide adenine dinucleotide phosphate (NADPH) as an electron donor. The rate-limiting reaction step for CYP is presented in Figure 2. The cycle, however, begins with the haem in its resting state; a water molecule occupies the axial position, and the iron is in a low-spin ferric form. The first step involves the displacement of the axial water molecule and the association of the substrate molecule with the Fe^{III} (I).[44] This association causes a geometry change, and the iron is displaced below the plane of the porphyrin, inducing a change in the spin of Fe^{III} (low to high) and lowering the redox potential by around 100 mV. This change in redox potential facilitates a single electron transfer (SET) from a redox partner (NADPH) to produce a high-spin Fe^{II} species (II).[45,46] This species binds molecular oxygen, which oxidizes the iron back to the low-spin ferric form (III) and the iron returns to lie within the porphyrin plane. An additional SET yields the basic dioxo-dianion species (IV), which is doubly protonated, leading to the fission of the O−O bond and releasing a water molecule (V). The ferryl-oxo compound formed in this step is commonly known as "Compound I" and takes part in the rate-determining step. An oxygen atom is inserted into the R−H bond in step VI. Finally, the hydroxylated product is released, a water molecule returns to the ferric haem's axial position, and the starting complex is regenerated (VII).[1]

The importance of CYPs in drug metabolism, coupled with a wealth of experimental data, means that predicting the CYP metabolism of compounds has been a priority for the pharmaceutical industry. The natural choice was to create models of human CYP metabolism, allowing compounds to be screened virtually for potential metabolic liabilities. Successful models predicting regioselectivity and isoform specificity of CYPs for human isoforms have achieved accuracies of approximately 90%.[1,47] As discussed above, despite the success of current CYP models, tests are still conducted regularly using animal models, primarily for human safety. The aim is to produce all of the likely human metabolites of a test compound to identify any possible harmful effects in humans during later-stage trials. Test species are chosen to fulfill a list of criteria, including producing metabolites likely to be seen in humans, being able to survive in a laboratory, and being practical to handle and administer the test compound. Thus, in the current work, we expand our previously published models[1] to preclinical species such as rats, mice, and dogs.

Flavin-Containing Monooxygenases. The discovery of FMOs could be credited to Ziegler and Pettit, who in 1964 suggested that the oxidative *N*-dealkylation catalyzed in the mammalian liver homogenates is divided into partial reactions catalyzed by separate enzymes instead of a mixed-function oxygenase. According to the study, the two reactions were oxidation of the nitrogen atom and the subsequent dealkylation.[48] In 1966, the same research group was able to isolate the enzyme FMO, which catalyzed the oxidation of the nitrogen atom, proving their initial theory.[49] It is now known that FMOs are able to oxidize tertiary-, secondary-, and primary alkyl- and aryl amines, hydrazines and imidazoles.[15] S-oxidation by FMOs was proposed in 1974 by Poulsen et al.,[50] and today, the following sulfur-containing groups are known to be oxidized by FMOs: sulfides, thiols and disulfides, thiocarbamides and thioamides, mercaptopurines, and mercaptopyrimidine.[15] In addition, FMOs have been observed to oxidize a wide variety of atoms such as boron,[51] carbon (Bayer−Villiger oxidation),[52,53] phosphorus,[54] selenium[55] and iodine.[15,54] Furthermore, additional reaction types observed within humans include *N*-demethylation and desulfuration.[16] However, the prevalent FMO-mediated metabolites are *N*- and *S*-oxides; thus, this study concentrates on *N*- and S-oxidation by FMOs.

FMOs belong to the flavoprotein family of enzymes and require a single FAD to catalyze *N*- and S-oxidation. The

https://doi.org/10.1021/acs.jmedchem.2c01303
J. Med. Chem. XXXX, XXX, XXX−XXX

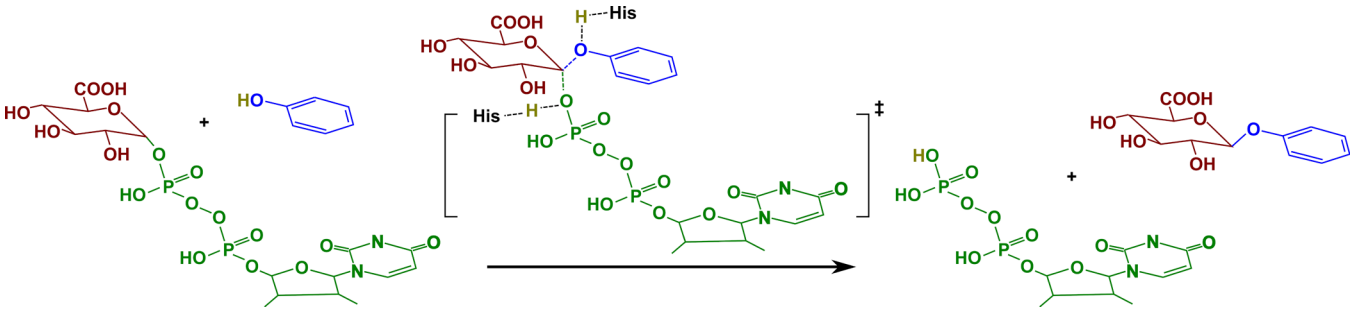


Figure 4. The transition state for glucuronidation by UGTs. The residues taking part in the reaction are based on the homology model of UGT isoform 1A1.[77]

catalytic cycle begins with FMO generating a stable peroxyflavin intermediate.[56] This is performed in two steps: first, the FAD undergoes a two-electron reduction utilizing the NADPH, and then it reacts rapidly with molecular oxygen to form the peroxyflavin. It is thought that FMOs in cells are predominately in a state where the peroxyflavin is ready to react with a substrate, and the system has been compared to a "cocked gun".[15] The oxidation works by transferring an oxygen atom from the peroxyflavin to the "soft-nucleophile" of the respective substrate, forming a hydroxyflavin and an oxidized substrate (Figure 3).[30] The final parts of the cycle of catalysis are the regeneration of FAD by releasing water and releasing nicotinamide adenine dinucleotide phosphate ($NADP^+$).

FMOs are an ancient gene family and can be found in all phyla examined, including the group chordate, to which humans belong.[57] In humans, there are five functionally active FMO isoforms, FMO1−5 and many nonfunctional pseudogenes (FMO6P−11P). FMOs are found in multiple tissues, but, as with AOs, they are mostly present in the liver, with FMO3 being the most highly expressed major contributor to the metabolism of xenobiotics. FMO1 is found in the fetal liver; however, this gene is switched off in the liver after birth, and its function is subsequently replaced by FMO3 as the child develops. FMO1 is still highly expressed in adult kidneys and is also found in the small intestine. FMO5 is mostly found in the liver but is also expressed in the stomach, pancreas, and small intestine. FMO2 and FMO4 are present in very low concentrations distributed across several organs. While more is known about FMO2 than FMO4, their contribution to metabolism is small, and in the case of FMO4, its contribution is negligible, and it can be disregarded.[16]

Historically, FMO metabolism, which contributes to the modification phase, has been underestimated, ignored, or attributed to CYPs due to the overlap of their substrate specificity. However, there are molecular entities that are predominantly or exclusively metabolized by FMOs.[58−62] Thus, disregarding FMO metabolism could lead to unexpected paths of metabolism or, worse, toxic metabolites—e.g., FMOs are known to produce sulfinic acids, and *S*-oxides and *S*,*S*-dioxides of thiocarbonyls.[15,63−67] In general, however, metabolites produced by FMOs are considered safer than CYP-mediated metabolites.[16] Predicting metabolism by FMOs could help researchers design drug candidates directed either away from or toward FMO-mediated metabolism to avoid toxic metabolites.

The number of studies regarding FMO metabolism is growing slowly compared to AOs, CYPs, or UGTs.[3] Computational studies focusing on the mechanism of *N*- and S-oxidation are very scarce, with only three published studies. There were two schools of thought as to how the substrate oxidation step proceeds. Ottolina et al. proposed an S_N2 reaction;[68] however, Bach proposed that the reaction proceeds via radical intermediates.[69] The latest results in our previously published work support the S_N2 reaction mechanism.[30] Only one model for predicting SoM for FMOs has been published by Fu and Lin, who used descriptors derived from quantum mechanics (e.g., Fukui reactivity indices) and circular fingerprints to train a support-vector machine classification model.[6]

Uridine 5′-Diphospho-glucuronosyltransferases. UGTs are considered the second most important enzymes for drug metabolism, after CYPs, and the most important enzymes of the conjugation phase. UGTs are estimated to participate in the metabolism of 15% of hepatically cleared drugs and approximately 40% of all conjugation reactions.[3,17,38,39] The UGTs have been actively studied since the 1960s, and it is one of the most actively studied enzyme families related to the metabolism of xenobiotics, with the number of studies dwarfed only by CYPs, reflecting their contribution to xenobiotic metabolism.[3] UGTs work by transferring a glucuronic acid (GA) moiety to a suitable functional group in the substrate, a reaction known as glucuronidation. Conjugation with a GA makes the substrate more polar; thus, in most cases, either deactivating the substrate or making it easier for the body to eliminate it. The most prevalent potential sites of metabolism are nitrogen atoms of amines, amides, and *N*-heterocycles (*N*-glucuronidation) and oxygen atoms of phenols, carboxylic acids, and alcohols (*O*-glucuronidation).[70] *C*- and *S*-glucuronides are known but are rare.[71,72] The current study concentrates only on *N*- and *O*-glucuronidation.

UGTs are a subclass of enzymes called glycosyltransferases, which are responsible for catalyzing the formation of glycosidic bonds to form glycosides. In general, the glycuronosyl reactions follow a mechanism where the sugar donor and the substrate are bound sequentially, followed by the sugar transfer, inverting the configuration at the anomeric center. The product is then released, followed by the release of the nucleotide moiety. In the case of UGTs, the sugar donor is uridine diphosphate GA (UDP-GA).[73,74] The generally accepted reaction for UGTs follows the S_N2 mechanism, where the nitrogen or the oxygen atom attacks the anomeric carbon of the GA, forcing the UDP to leave. Two residues of the enzyme act as the acid and base forming a "catalytic dyad" and stabilize the reaction as depicted in Figure 4.[30,75−78]

Most kingdoms in biology include species with UGTs.[79] A total of 31 UGT isoforms are found in humans—22 active isoforms and 9 pseudogenes. Based on the sequence similarity, the active isoforms are divided into four categories—UGT1,

https://doi.org/10.1021/acs.jmedchem.2c01303
J. Med. Chem. XXXX, XXX, XXX−XXX

UGT2, UGT3, and UGT4. In theory, the large number of different isoforms gives rise to broad substrate specificity, but in practice, the substrate specificity often overlaps between the isoforms. The isoforms can be found all over the body, ranging from the liver to the nasal cavity.[80] This work concentrates on the first two families, especially isoforms 1A1, 1A4, 1A9, and 2B7, which are primarily expressed in the liver and are responsible for the conjugation of the majority of xenobiotic UGT substrates.[79,81−83]

The first models, which explored the isoform-specific SoM prediction for UGTs, were published in 2006.[7] Sorich et al. developed naïve Bayes classifiers, using experimental data from the literature, for eight isoforms—1A1, 1A3, 1A4, 1A6, 1A8, 1A9, 1A10, and 2B7. Several other models[8−11] have emerged over the years, which have taken a different approach to predict site-specificity, discarding the isoform specificity and working with all known human UGT-catalyzed reactions. Such an approach allows the inclusion of additional data points since their origins are not restricted to isoform-specific studies. The number of data points within the referenced papers varied from around 1400 to 3300 unique SoM.

Modeling Drug Metabolism. There are many available modeling methods for predicting metabolism, ranging from empirical methods such as statistical modeling or machine learning to mechanistic approaches like molecular mechanics (MM), molecular dynamics (MD) or even quantum mechanics (QM).[84] For a model to be used for *in silico* screening, it must be fast; thus, empirical models, which are fast and relatively easy to set up (if one has enough data), should be preferred. The downsides of such models are that often, there are not enough data to train the model, and even where there are sufficient data, the models are nontransferable and qualitative. In cases where data is sparse, researchers may look toward models built using mechanistic methods. Such models can be built on smaller data sets, are more transferable due to the model's underlying physical principles and can be quantitative. The downside of such models is the execution time, which, even with modest methods, can be too long for practical use.

A different method of differentiating between computational techniques for predicting metabolism is to divide them into two distinct categories: ligand-based and structure-based models. In the former, structures and properties of known substrate or nonsubstrate compounds are modeled to develop structure−activity relationships. The second approach focuses on the structure of the metabolizing enzyme, its known reaction mechanisms, and its interactions with substrates. Structure-based methods include docking,[85] molecular dynamics simulations,[86,87] and QM/MM methods.[1,88]

In this work, we have chosen the ligand-based method since the available evidence suggests that structure-based methods, at present, have diminishing gains in accuracy and incur higher computational costs. Furthermore, we combine elements from the empirical modeling methods with elements from the mechanistic approaches to derive reactivity-accessibility models, which achieve a balance between computational cost and accuracy for modeling metabolism and have been used successfully by the authors of the current work[1] and others.[89,90] The reactivity-accessibility approach divides the metabolism of a compound into two parts—reactivity, which describes the pure reactivity of the potential SoM, and accessibility, which captures the accessibility of potential SoM to the catalytic center within the active site of the protein. The reactivity of a site is accounted for by calculating the activation energy of the rate-determining step of the corresponding reaction. This is a physical property, calculated using fundamental and empirical physical constants, and as such can be applied to any molecule. Furthermore, these models use the whole substrate to account for long-range electronic effects, which can play an important role in determining the reactivity of sites within a molecule. This gives the model a good domain of applicability compared with a purely statistical model, for which the domain of applicability is limited by the compound structures used to train the model. Previous studies have shown that activation energy is an important descriptor of whether a potential SoM will be metabolized experimentally.[1] The rationale is that rate of a reaction is related to the activation energy by the Arrhenius equation.[91] Since the reaction center is conserved across all isoforms of an enzyme class, the activation energy calculation is not dependent on the isoform, and the activation energy depends only on the substrate.

The accessibility of an SoM refers to how easily it may be approached by the enzyme's reactive center, which is affected by many factors. In this study, steric and orientation effects were considered—steric effects relate to features of the substrate itself that may hinder access to the SoM, while orientation effects capture effects of the binding site that may orient the substrate such that some sites are far from the reactive center. The steric aspect considers hindrance due to the bulk (or rigid structure) of the substrate itself. The orientation of the substrate in the binding pocket is important—some SoM may be distant from the reactive center in the bound conformation of the substrate—and is affected by functionalities of the substrate and the protein, e.g., hydrogen bonding, electrostatic interactions, and hydrophobicity. The effect of these factors on the rate of metabolism are accounted for with two-dimensional steric and orientation descriptors that provide information about key functional groups' locations relative to the potential SoM. The binding sites differ between the enzyme families and their isoforms, so accessibility is affected by both the isoform and the substrate. These steric and orientation descriptors require no knowledge of the three-dimensional structure of the binding pocket (an advantage over a docking study) and can be quickly calculated.[1] A statistical model of accessibility from the two-dimensional (2D) steric and orientation descriptors is used to correct the activation energies used to represent the reactivity.

For the reactivity-accessibility model, we assumed that the compound is bound to the active site since the experimental, site-specific data for metabolism include molecules, which are observed to be metabolized. Site-specific information for molecules known not to be metabolized was not considered because it is unclear why the molecule was not metabolized; a molecule may have a highly reactive site but would not be metabolized if it does not reach the binding site. Whether a compound is a substrate of a given isoform of an enzyme class can be addressed by a separate model.[92,93]

EXPERIMENTAL DATA

The data used herein were curated from sources that provide detailed information on the experimentally observed SoM. Since the models are meant to distinguish the experimentally observed SoM from all potential SoM, the molecules included in the data set, in the majority of the cases, have two or more potential SoM, out of which at least one is experimentally observed to be metabolized. To summarize, the

https://doi.org/10.1021/acs.jmedchem.2c01303
J. Med. Chem. XXXX, XXX, XXX−XXX

collected compounds were labeled according to which enzyme family and which isoform from that family is responsible for metabolizing the molecule (note that some molecules are metabolized by multiple isoforms or enzyme classes). Each potential SoM on a molecule was labeled as either observed to be metabolized or not by the corresponding isoform. The exception was the data for CYP metabolism by preclinical species, since in most cases, the published data did not include isoform-specific data. In this case, species-specific SoM data was curated by aggregating the influence of several CYP isoforms. Furthermore, compared to the other enzymes, the number of secondary and tertiary SoM was substantial for CYP substrates; thus, the sites were labeled as 1st, 2nd, 3rd, or "not observed" for primary, secondary, tertiary, or not metabolized SoM, respectively. In this curation, the emphasis was on high-quality data, retaining only data generated with appropriate experimental conditions. The data for AO, FMO, and UGT models were gathered only from *in vitro* experiments, where it was explicitly stated which isoform was studied (e.g., an isoform expressed in a cell line or isoform-specific microsomes). The experiments run with unphysiological substrate concentrations were rejected, where the lowest accepted concentration was 100 μM or less. If there were conflicting reports of the metabolism of a substrate (e.g., a primary site of metabolism in one paper was not recognized as a site of metabolism in another paper) then the substrate was rejected. Each metabolite included had to have an experimental confirmation (e.g., using mass-spectrometry or NMR studies); we did not include metabolites based only on expert opinions. If the site of metabolism was not explicitly confirmed (e.g., an aromatic ring was oxidated, but the researchers were not certain, which atom it was) then the substrate was rejected. The data for preclinical species followed the same rules with the exception of isoform specificity since these models were general. The four data sets are summarized in Table 1 and the following paragraphs describe the size and the content of the data sets for each isoform and enzyme family. The references from which the data were obtained are listed in the Supporting Information of this work.

Table 1. The Overview of Data for Building Reactivity-Accessibility Models

enzyme	isoform[a]	no. of substrates	no. of potential SoM	no. of SoM metabolized
AO	AO1	157	865	160
FMO	FMO1	56	172	56
	FMO3	67	209	69
UGT	UGT1A1	98	297	146
	UGT1A4	54	146	66
	UGT1A9	137	390	187
	UGT2B7	90	223	115
CYP	Mice	68	617	108
	Rats	163	1428	305
	Dogs	80	1091	154

[a]For CYPs the species instead of isoforms are mentioned.

For AO1, as in previous studies, all aromatic carbons are considered potential SoM.[5] The current work also included aldehyde SoM, although there are only eight molecules in this data set with this functionality that met the criteria for inclusion. To summarize, the data set for AOs consists of 157 molecules and 865 potential sites, of which 160 are observed experimentally to be metabolized: 155 primary and 5 secondary sites.

The FMO isoforms with sufficient data for building models are FMO1 and FMO3. The potential SoM include all nitrogen and sulfur atoms that could be oxidized according to the literature. Both the FMO1 and FMO3 data sets have a relatively small number of molecules (56 and 67 structures, respectively) and potential SoM (172 potential SoM out of which 56 are metabolized by FMO1 and 209 potential SoM out of which FMO3 metabolized 69), as can be seen in Table 1, compared to isoforms in other enzyme families. However, according to the literature, the smaller data sets should not hinder the model building process as FMO metabolism depends mainly on the reactivity of the sites.[15]

The data set for UGT isoform UGT1A1 contains 98 molecules with 297 potential SoM, and it features 146 potential SoM that are glucuronidated and 151 that are not. The majority of the potential SoM are phenols, followed by amines. The remaining SoM include carboxylic acids, alcohols, and a small number of other SoM types, which include nitrogen atoms. The data set for the UGT1A4 isoform is, overall, the smallest and contains only 54 molecules. However, it is the most balanced data set in terms of the SoM types, with amines being the most prevalent, followed by phenols and other SoM, including carboxylic acids and other sites which include nitrogen atoms. The structure of the UGT1A9 data set is similar to UGT1A1, mostly comprising phenolic SoM, followed by amines and other types. While the UGT1A9 data set is the largest amongst UGTs (137 molecules), it features a large number of flavonoids; thus, the variation within the neighborhood of the site types is similar to other data sets. The data set for UGT2B7 (90 molecules) is more balanced, with phenols still being the majority of the potential SoM, followed by amines, alcohols, carboxylic acids, and other sites featuring a nitrogen atom as the potential SoM.

For CYPs, three of the most common preclinical species and strains were selected: Sprague–Dawley (rat), beagle (dog), and various strains of mouse. Initially, the aim was to obtain site-specific rates for individual isoforms; however, it was found that information regarding isoforms is not commonly reported in the literature for non-human species and, as described above, all data for non-human species were aggregated by species and strain. Furthermore, such a wide variety of mouse strains were used in the literature that all of these strains were combined in this study to ensure the data set is sufficiently large for model building. The number of substrates in the data sets for mice, rats, and dogs is 68, 163, and 80. The data set for mice includes 617 potential SoM, out of which 108 are metabolized. The data set for rats is the biggest, with 1428 potential SoM, out of which 305 are metabolized. The data set for dogs features 1091 sites, out of which 154 are metabolized. Other species and strains that were considered but found to have comparatively fewer substrates with available data included Wistar rats, Cynomolgus monkeys, New Zealand White rabbits, and Göttingen minipigs.

In most cases, the literature searches yielded papers, which reported the detected metabolites as primary (1st), secondary (2nd), or tertiary (3rd) metabolites. However, in some cases, the papers contained the ideal data (rate of metabolism, V_{max}) for each potential site of metabolism in a molecule. Where this information was available, the experimentally observed rates were converted into a ranking within each molecule. The rates were ranked (i.e., 1st, 2nd, 3rd) within each molecule.

Aim of the Study. This study demonstrates the generalizability of the reactivity-accessibility approach by training isoform-specific SoM models for AO1, FMO1, and FMO3, and UGT1A1, UGT1A4, UGT1A9, and UGT2B7. Furthermore, we apply the same approach to train nonisoform-specific CYP models for preclinical species, such as mice, rats, and dogs. The in silico models are useful for predicting the modification and conjugation phases in humans. Modeling the metabolism of preclinical species could aid in ensuring the preclinical trials produce the likely human metabolites, using the model as an indicator for selecting the best preclinical species.

RESULTS AND DISCUSSION

The reactivity descriptor, E_a, is calculated using semi-empirical methods. In the following subsections, we provide a description of how to take the systematic errors for semi-empirical methods into account using correction factors for each enzyme family. We then describe how the corrected E_a values are combined with the steric and orientation descriptors and the results from the experimental studies to build models

https://doi.org/10.1021/acs.jmedchem.2c01303
J. Med. Chem. XXXX, XXX, XXX−XXX

for predicting the SoM. The model results are provided with data set splits, confusion matrices, and *y*-scrambled values.

GP Model for AO. We obtained the simplified reaction mechanism for the oxidation of azaheterocycles by AO from the work of Montefiori et al.[14] We describe additional work on expanding the simplified mechanism to aldehydes in the Supporting Information of the current study. We confirmed a correlation between E_a values calculated with PM6 and DFT for various SoM types to verify that PM6 is suitable for replacing DFT. To achieve that, the SoM were divided into seven environments for which correction factors were calculated, as described in detail in the Supporting Information. As can be seen from Figure 5, the initial squared correlation coefficient increases from 0.92 to 0.97 and most of the errors fall under 10 kJ per mol.

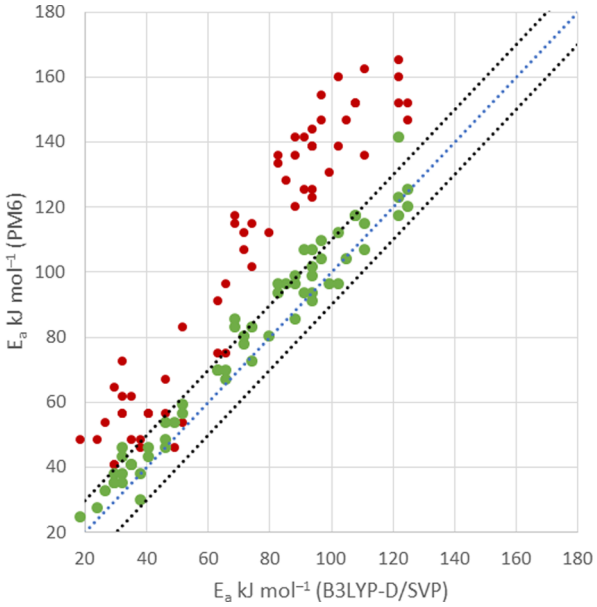


Figure 5. Correlation between DFT and semi-empirical E_a values. Red dots represent the E_a values, and the green points represent the corrected E_a values. The blue line is the identity line, and the black lines represent deviation of +10 and −10 kJ per mol from the identity line.

Applying the respective correction factors to the E_a values of different SoM environments, obtained using PM6, makes them directly comparable to each other because they are referencing the DFT energy scale. Out of the 159 cases, the corrected E_a alone was able to predict the experimentally observed primary SoM as the site with lowest E_a in 52% of cases. Since the AO substrates have, on average, over five potential SoM, the AUC provides a better indication of how well E_a alone describes the site-specificity of AOs. The average AUC for all molecules is 0.80, indicating that the E_a value is an important descriptor for predicting the SoM of AO metabolism, but we expect that supplementing this with the accessibility descriptors in order to take into account the steric and orientation effects will improve our ability to predict SoM.

The κ value for the test set for the Gaussian processes (GP) AO model is 0.83. E_a was amongst the most important descriptors; the most influential being the descriptor that recognizes the site as being ortho to a σ-bonded aromatic nitrogen atom (it is very common to azaheterocycles, which form the majority of the compounds in the data set). The balanced accuracy of prediction was 0.90 for the test sets. The confusion matrix for the test set is shown in Figure 6. The *y*-

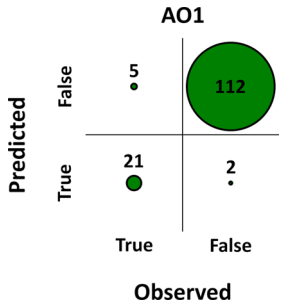


Figure 6. The confusion matrix of the test set of Gaussian processes model for AO1.

scrambled result had κ value of 0.05, which is considerably lower than the results from the test set, confirming that the models do not depend on spurious correlations between the observed experimental results and the measured descriptors.

GP Models for FMOs. We used the simplified reaction mechanism for calculating E_a for *N*- and S-oxidation by FMOs described in our previous work.[30] The initial tests, using AM1, demonstrated the feasibility of the mechanism, but unlike the correlation between the two methods for AOs, the correlation between semi-empirical methods and DFT for FMOs is only 0.26. The correlation did not improve after introducing separate correction factors for *N*- and S-oxidation, nor did it improve by dividing the SoM environments into further subenvironments (see Supporting Information). The low correlation can be explained by a hydrogen bond, which briefly forms between the cofactor and the leaving group during the transition state.[30] The hydrogen bond is observed in transition states optimized by DFT; however, it often does not form during the geometry optimization with the semi-empirical method AM1. The bond is often missing because AM1 is not as good at estimating the energetics of hydrogen bonding as DFT; thus, the correlation between the two methods is weak. For more information see Supporting Information for FMOs.

While the correlation between like-for-like sites was not sufficiently high, both AM1 and DFT correctly identified the experimentally observed site as that with the lowest calculated E_a when tested on a set of substrates in the data set. The corrected E_a alone was able to predict the experimentally observed primary SoM as the site with lowest E_a in 82% of cases for both FMO1 and FMO3. The AUC for both FMO1 and FMO3, for the whole data set, using AM1, was 0.91 and 0.92, respectively. Thus, the ranking of sites based on the E_a value calculated with AM1 is reliable for the reactivity-accessibility models. The reactivity descriptor alone could predict the experimentally observed primary sites in most cases.

The κ results for reactivity-accessibility GP models for predicting the SoM for the FMO1 and FMO3 test sets are 0.88 and 0.94, respectively. The confusion matrices can be seen in Figure 7. The balanced accuracies of the final models are 0.94 and 0.98 for FMO1 and FMO3, respectively. As with AOs, E_a and ΔE_a were amongst the most important descriptors in both models. The *y*-scrambled results were 0.00 and 0.03 for FMO1 and FMO3, respectively, demonstrating that the excellent performance of the models is unlikely due to chance correlation.

https://doi.org/10.1021/acs.jmedchem.2c01303
J. Med. Chem. XXXX, XXX, XXX−XXX

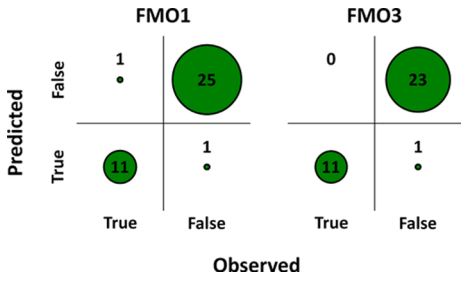


Figure 7. The confusion matrices of the test sets of GP models for FMOs.

GP Models for UGTs. As with FMOs, we used the simplified reaction mechanism for both *N*- and *O*-glucuronidation identified in our previous work.[30] The AM1 semi-empirical method used for predicting FMO metabolism was also used for UGTs; however, unlike FMOs, the correlation between AM1 and DFT was higher—0.58 before the corrections and 0.97 after applying the corrections (Figure 8). Interestingly, the E_a values for *O*-glucuronidation were much closer to the DFT values than those for *N*-glucuronidation. Thus, the correction factors for *O*-glucuronidation were very small compared to those for *N*-glucuronidation. The description of SoM environments and the derivation of the correction factors for *N*- and *O*-glucuronidation can be found in the Supporting Information.

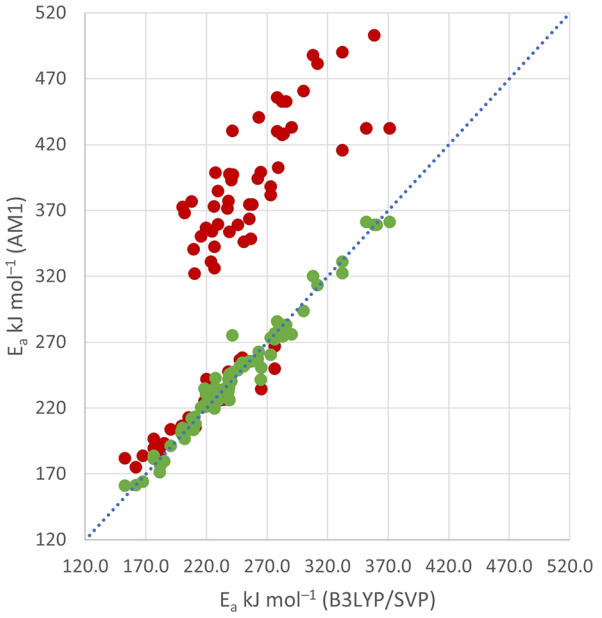


Figure 8. The correlation between DFT (B3LYP/SVP) and semi-empirical method (AM1) for *N*- and *O*-glucuronidation. Red points represent the uncorrected values, and green points represent the corrected values.

The AUC for the 1A1 isoform, using AM1, was 0.86. The GP model for 1A1 yielded a κ value of 0.81 with balanced accuracy of 0.90 (the confusion matrix for the 1A1 model can be seen in Figure 9). The *y*-scrambled κ result for 1A1 was 0.12, indicating that this result was unlikely to be due to random correlations with the data.

The AUC for the 1A4 isoform, using AM1, was 0.72, the lowest out of all sets. Since 1A4 is specialized for the metabolism of tertiary nitrogen atoms, it could be theorized that the accessibility descriptors play a bigger role compared to other UGT isoforms. The data set for building the GP model for 1A4 had the fewest data points amongst the chosen isoforms. The GP model had a κ value of 0.68 and a balanced accuracy of 0.84 (the confusion matrix for the 1A4 model can be seen in Figure 9). As before, the *y*-scrambled results, with a κ value of 0.02, proved that no random correlation exists in the data set.

The AUC for the 1A9 isoform, using AM1, was 0.78. The data set for building the GP model for 1A9 had the largest number of data points. This large data set yielded a result with a κ value of 0.63 and a balanced accuracy of 0.82 (the confusion matrix for the 1A9 model can be seen in Figure 9). The *y*-scrambled results had a κ value of −0.21, confirming that the result is unlikely to be due to chance correlations in the data set.

The AUC for the 2B7 isoform, using AM1, was 0.87. The Gaussian processes model yielded a κ value of 0.63 with a balanced accuracy of 0.82 with the *y*-scrambled results of 0.21 (the confusion matrix for the 2B7 model can be seen in Figure 9). It is surprising that the κ value of the GP model is relatively low, while the AUC is the highest amongst UGT data sets. This can partly be explained by exploring the data set of 2B7; the number of compounds in this test set is 18 while the amount of SoM, which get metabolized is 26. In a few cases, the model fails to recognize the secondary SoM, which in turn lowers the κ value of the model considerably.

WLS Models for CYPs. For the CYP models a 10-fold cross-validation weighted least squares (WLS) model was built. The cross-validation strategy was chosen to ensure that the model results are not dependent on a single training and validation split of the data. For each of the 10 models, the training and validation compounds were selected randomly, and a WLS model was trained.

Each trained model was applied to the test set. For each compound in the test set the model output a prediction for each potential SoM as a floating-point number between 1 and

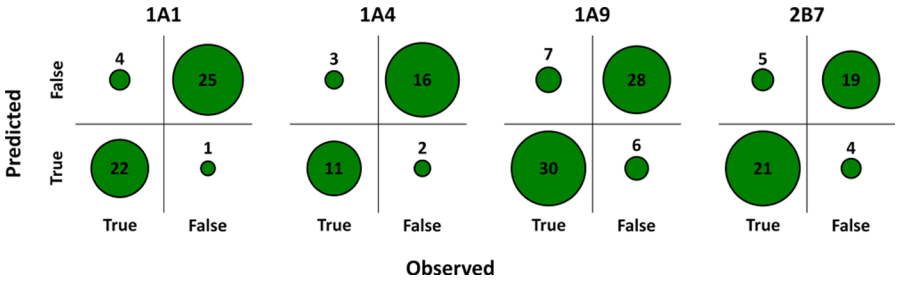


Figure 9. The confusion matrices of test sets of GP models for UGTs.

https://doi.org/10.1021/acs.jmedchem.2c01303
J. Med. Chem. XXXX, XXX, XXX−XXX

4 (where "1" indicates a primary site and "4" indicates not metabolized). For a given site, we used consensus modeling, where the predictions from the 10 models were averaged, and the resulting floating-point number was used as the final prediction.

The outputs of the model were ordered (lowest to highest) for the sites within a given compound and the AUC under the ROC curve calculated for each compound. The average of these AUCs for the compounds in the test set for each species is shown in Table 2 for the two types of activation energy calculations. See the Supporting Information for the detailed performances of individual models making up the 10-fold cross-validation.

Table 2. Average AUCs of Compounds on the Test Set for Three Species or Strains

species (strain)	AUC (standard deviation)
rat (Sprague–Dawley)	**0.89** (0.021)
mouse (any)	**0.92** (0.012)
beagle	**0.90** (0.016)

It is surprising that the accuracy of the preclinical general CYP models is comparable to the isoform-specific human AO, FMO, and UGT models. It is known that the pure reactivity for the potential SoM plays a critical role in CYP metabolism, but the highest accuracy is usually obtained by taking into account the isoform-specific steric and orientation effects.[1] While the experimental data for preclinical species did not specify individual isoforms, it is likely that the general CYP preclinical species models achieved such excellent results because the experimental data mostly consist of a single or small number of prevalent isoforms, e.g., the CYP3A family. Thus, the steric- and orientation component accounts for the aforementioned isoform(s).

CONCLUSIONS

This paper has described the prediction of the regioselectivity of metabolism by AOs, FMOs and UGTs for humans and CYPs for three preclinical species. The resulting models show excellent performance for the prediction of the primary SoM for isoforms of AOs, FMOs and UGTs for humans (Figure 10) and the prediction of primary, secondary, and tertiary SoM of enzyme families for mice, rats, and beagle dogs. While most of the models presented here cannot be directly compared to the already existing models due to their isoform-specific nature, the overall accuracy of the presented models is comparable with the best metabolism prediction models published. Furthermore, to the best of the authors' knowledge, the AO1 model is the only published model, which can predict both aldehyde- and aromatic (hetero) cycle oxidation, and the FMO1 and FMO3 models are the only isoform-specific FMO reactivity-accessibility models published to date.

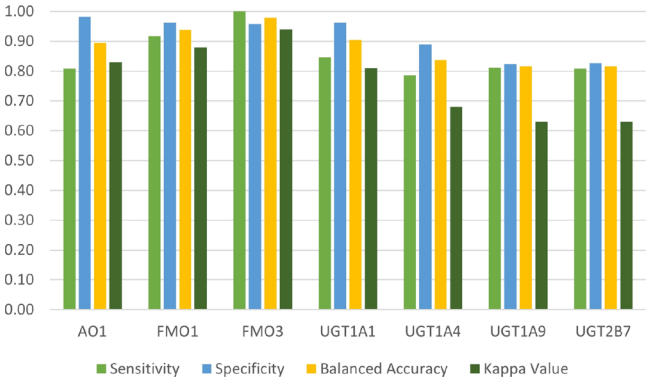


Figure 10. The sensitivity, specificity, balanced accuracy, and κ values for human isoforms of AO, FMO, and UGT.

The predictive models are based on a detailed understanding and simulations of the catalytic mechanisms of the respective enzyme families. The reactivity-accessibility approach used to build the 10 models applies semi-empirical methods to estimate the electronic activation energy of rate-limiting steps of the catalytic cycles. The simplified reaction mechanisms for the rate-limiting steps for the enzyme families have been validated previously using experimental data and DFT calculations. The activation energy was coupled with isoform-specific steric and orientation effects, which arise due to the interactions between the substrate and the binding pocket. The methods based on quantum mechanics offer generality and transferability since they are derived from fundamental physical principles. Furthermore, these models use the whole substrate molecule and consider long-range interactions, which play an important role in differentiating between sites within a molecule. This gives the model a good field of applicability compared with a purely statistical model, whose field of applicability would be limited by the chemistry used to train the model.

The seven models for human enzymes are isoform-specific and include the following isoforms: AO1 for AOs, FMO1, and FMO3 for FMOs and UGT1A1, UGT1A4, UGT1A9, and UGT2B7 for UGTs. The chosen isoforms represent the prevalent enzymes of their respective families in the human liver. The three models for preclinical species were for mice, rats, and dogs, but were not isoform-specific.

The isoform specificity of the models presented herein, sets them apart from previous studies and could be useful for researchers studying the metabolic fate of compounds through the modification and conjugation phases in humans. Furthermore, the models for preclinical species could help reduce, refine, and replace animal studies.

Future work in this field will include combining the substrate data for multiple enzyme families into a single model to predict which enzyme family(ies) and isoform(s) are most likely to be responsible for the metabolism of a compound. Isoform specificity models have already been published for CYPs,[92,93] and a similar model could also be useful for UGTs. Combining predictions of the enzyme(s) and isoform(s) responsible for the metabolism of a compound with the SoM predictions of the models described herein would enable the prediction of the metabolic fate of a compound based only on its chemical structure. The reactivity-accessibility method for modeling drug metabolism has proved to be generalizable, adding additional human enzymes from the conjugation phase. We believe a similar approach can be extended to additional enzyme families such as sulfo- and glutathione transferases.

EXPERIMENTAL SECTION

Reactivity-Accessibility Models. As described in the introduction, the reactivity-accessibility models consider the reactivity and accessibility of each potential SoM of a substrate molecule. Reactivity describes the inherent lability of a potential SoM, while accessibility describes how easily the reactive center can approach the potential SoM.[1] In this work, the reactivity is characterized using the E_a and the

https://doi.org/10.1021/acs.jmedchem.2c01303
J. Med. Chem. XXXX, XXX, XXX−XXX

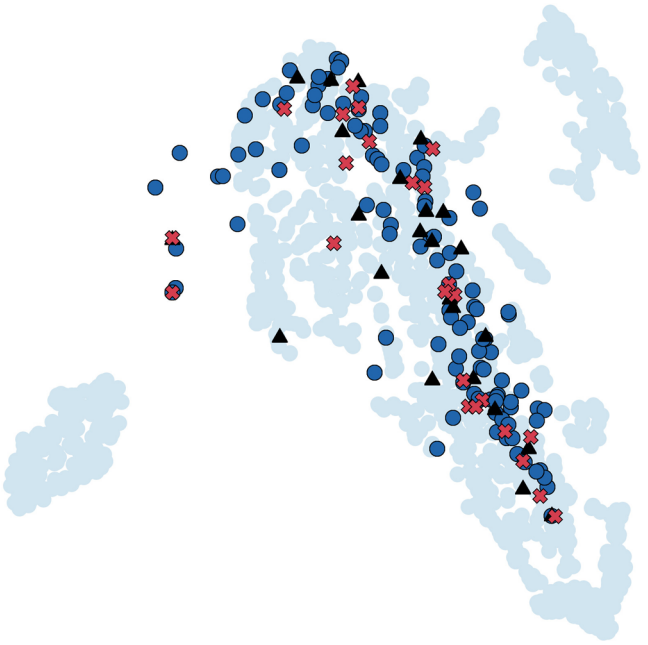


Figure 11. The chemical space plot representing the AO1 substrates. The light blue circles represent the 1300 launched drugs, the dark blue circles represent the compounds in the training set, the black triangles represent the compounds in the validation set, and the red crosses represent the compounds in the test set.

ΔE_a of a simplified transition state. The ΔE_a specifies the difference in E_a values between sites within a molecule. For example, the ΔE_a values for two potential sites in a molecule with the E_a values of 50 and 75 kJ mol^{-1} would be 0 and 25 kJ mol^{-1}, respectively. The simplified reaction mechanisms for AOs,[14] CYPs,[1] FMOs,[30] and UGTs,[30] with which the E_a values are calculated, have been previously published (with the exception of the oxidation of aldehydes, which can be found in the Supporting Information of the current work). However, the referenced work has used DFT to obtain the E_a values. In the current work, semi-empirical methods such as AM1[94] and PM6[95] are used to calculate E_a values. The semi-empirical methods are used because they are significantly faster than *ab initio* methods and therefore can be applied to an entire substrate on a routine basis.

The accessibility descriptors in this work are all based on the atom-pair descriptor concept, where distances from the potential SoM to specified functional groups are defined as counts of bonds. SMARTS patterns (SMILES arbitrary target specification, where SMILES stands for simplified molecular-input line-entry system) are used to define the groups that describe functionalities such as acidic and basic groups, hydrogen bond donors and acceptors, and lipophilic groups that may interact with key residues in the active site of a protein.[1] The reactivity and accessibility descriptors for each SoM are then associated with the data from the experiments (is a SoM observed to be metabolized or not), which enables us to build quantitative structure−activity relationship (QSAR) models for each aforementioned isoform or species.

Computational Methods. All potential substrate structures in this work were generated from SMILES using OEChem from OpenEye.[96,97] Transition state structures were based on previous work by the authors and others.[1,14,30] The calculations for obtaining the E_a values for the reactivity-accessibility models were performed using the semi-empirical methods AM1[94] and PM6[95] using the program package CP2K.[98] AM1 was chosen to calculate the E_a values because it had the best performance when testing it with our benchmark calculations (not published). It was, on average, the fastest and had the least number of failed calculations. Furthermore, it has been successfully implemented in our previously published reactivity-accessibility models.[1] Since the simplified mechanism for AO includes a molybdenum atom, the PM6 semi-empirical method is used for AO models, which has the necessary parameters for this element.

In many cases, the semi-empirical methods are subject to systematic errors due to the approximations they make to the Hamiltonian. Therefore, for the semi-empirical methods to be used confidently, corrections to account for these systematic errors are calculated by correlating the E_a values obtained with semi-empirical methods to the E_a values obtained with DFT. The potential SoM are divided into types based on the corrections they require (e.g., aliphatic and aromatic carbon atoms for CYP[1]) and the respective corrections are applied to the E_a values. The discovered SoM types can be recognized using SMARTS patterns and the application of corrections can be automated.

DFT calculations, were run using the B3LYP or B3LYP-D functionals[99−103] and the def2-SVP[104] basis set. An effective core potential was used for the molybdenum atom,[105] which was obtained from the Basis Set Exchange.[106] B3LYP was chosen because the presented reaction mechanisms feature organic molecules and geometry optimizations, including transition states, followed by frequency calculations by hybrid GGA functionals yield similar results to the more expensive hybrid meta-GGA functionals.[107] The B3LYP-D was used to study AO and B3LYP without the dispersion corrections was used to study FMO and UGT (see ref 30). The geometry optimizations were followed by frequency calculations to verify the local minima or the transition states. The DFT calculations were performed with the NWChem 6.8 package.[108]

Accessibility Descriptors. While the three-dimensional compound geometries are used for E_a calculations, the accessibility descriptors calculated are based only on the 2D compound structure.

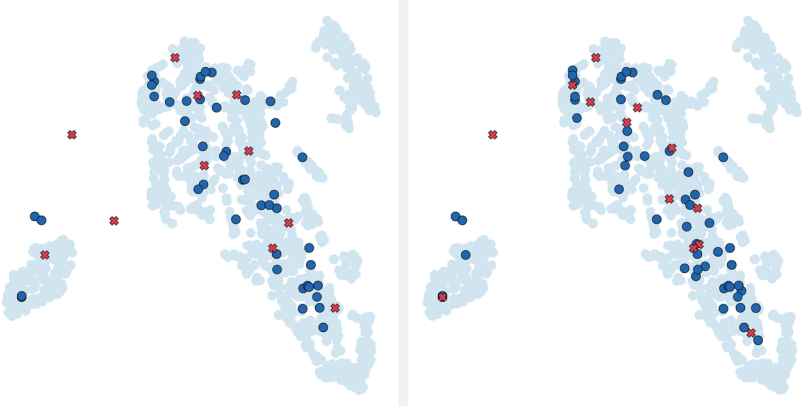


Figure 12. The chemical space plots representing the FMO1 (left) and FMO3 (right) substrates. The light blue circles represent the 1300 launched drugs, the dark blue circles represent the compounds in the training set, and the red crosses represent the compounds in the test set.

https://doi.org/10.1021/acs.jmedchem.2c01303
J. Med. Chem. XXXX, XXX, XXX−XXX

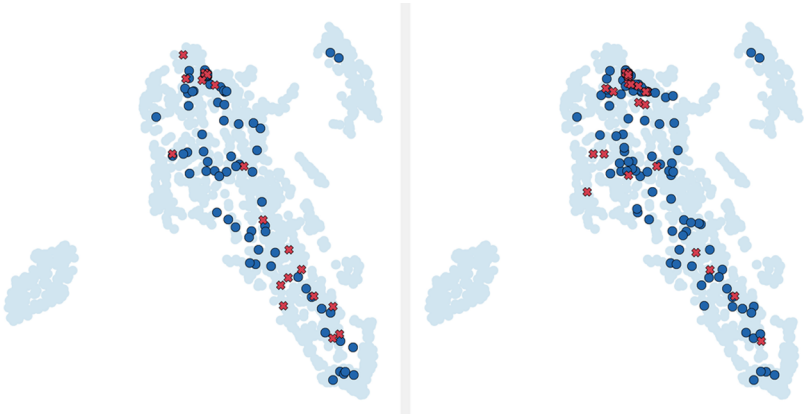


Figure 13. The chemical space plots representing the UGT1A1 (left) and UGT1A9 (right) substrates. The light blue circles represent the 1300 launched drugs, the dark blue circles represent the compounds in the training set, and the red crosses represent the compounds in the test set.

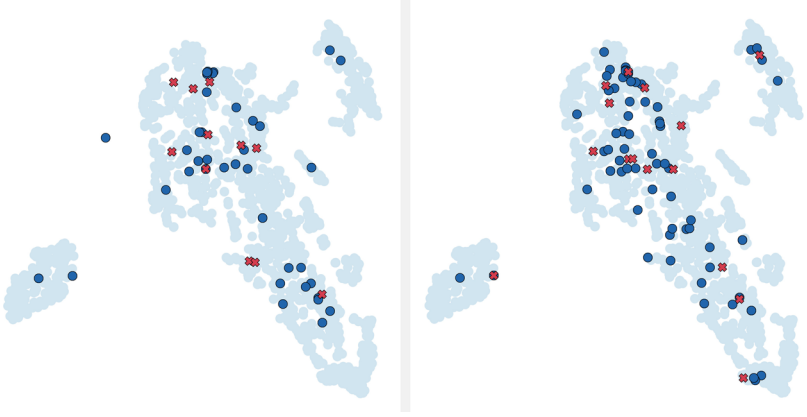


Figure 14. The chemical space plots representing the UGT1A4 (left) and UGT2B7 (right) substrates. The light blue circles represent the 1300 launched drugs, the dark blue circles represent the compounds in the training set, and the red crosses represent the compounds in the test set.

This decision was made due to the limited nature of three-dimensional descriptors—using a single conformation would not be appropriate since a particular substrate may adopt multiple conformations in the active site, which would require an extensive conformational sampling or molecular dynamics calculation *in situ* to average over all low-energy conformations. It should be noted that the reactivity model is not as sensitive to conformational variation; the energy differences between conformations cancel out because the reactant and product calculations use the same overall conformation of the compound. Using 2D atom-pair descriptors avoids the problem caused by conformational variability and has proven itself on multiple occasions.[1]

Machine Learning Methods. The GP method in StarDrop was used to train the majority of models described herein. GP is a powerful computational method for predictive QSAR modeling. Using a Bayesian probabilistic approach, the method is widely used in the field of machine learning but is not common in QSAR and ADMET (absorption, distribution, metabolism, excretion, and toxicity) modeling. This method overcomes many of the problems of existing QSAR modeling techniques, e.g., it does not require subjective a priori determination of parameters such as variable importance or network architectures and it is suitable for modeling nonlinear relationships. The method has built-in mechanisms to prevent overtraining and does not require cross-validation. In addition, the importance of each descriptor is reported; thus, the impact of E_a and ΔE_a can be directly measured. The details of the theory of Gaussian processes for QSAR modeling are described in a comprehensive study by Obrezanova et al.[109]

The CYP models were trained using the WLS technique[110] because, unlike other enzymes, CYP substrates frequently have multiple SoM with different relative rates (primary, secondary, tertiary), a regression model provides greater resolution for ranking the predicted sites. WLS is a linear regression that minimizes the residual sum of the squared deviations between model values and experimental data values. When fitting a line to the experimental data points, the weights allow each type of data point to be treated differently. The data point types that occur more frequently in the data (nonmetabolized sites and primary sites) are given lower weight and less common types (secondary and tertiary points) are given a higher weight. The weighting ensures the line is not fit to maximize its score (residual sum of squares) at the expense of the less common site types by fitting the line very well to only the major site types.

Data Splits. For small data sets, the data obtained for each isoform was split into training and test sets using the approximate ratio of 80:20, respectively. For larger data sets, the data were split into training, validation and test sets using the approximate ratio of 70:15:15. The split was made by compound; thus, all potential SoM of one substrate were either in the training, validation, or the test set. The compounds for the sets were chosen randomly, but the distribution of different sets was visually checked (without inspecting the individual structures) to ensure that the chemical space of the training set is roughly covered by the compounds in the validation and test sets (if compounds in either validation or test sets were found to be clustered in a specific region of chemical space, a new random split was performed). Since the models will not be based on molecules, but on the potential SoM within molecules, the leave-cluster-out split method was not considered. The training sets are used to build the model, the validation sets of larger data sets are used to compare models built in different ways, and the test sets are used to evaluate the model chosen in the validation step. It was ensured that

https://doi.org/10.1021/acs.jmedchem.2c01303
J. Med. Chem. XXXX, XXX, XXX−XXX

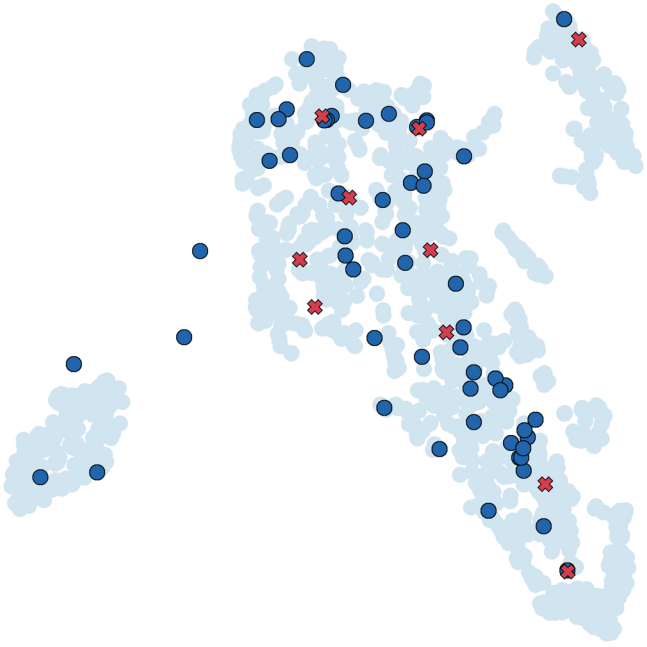


Figure 15. The chemical space plots representing the substrates metabolized by mice. The light blue circles represent the 1300 launched drugs, the dark blue circles represent the compounds in the 10-fold cross-validation set, and the red crosses represent the compounds in the test set.

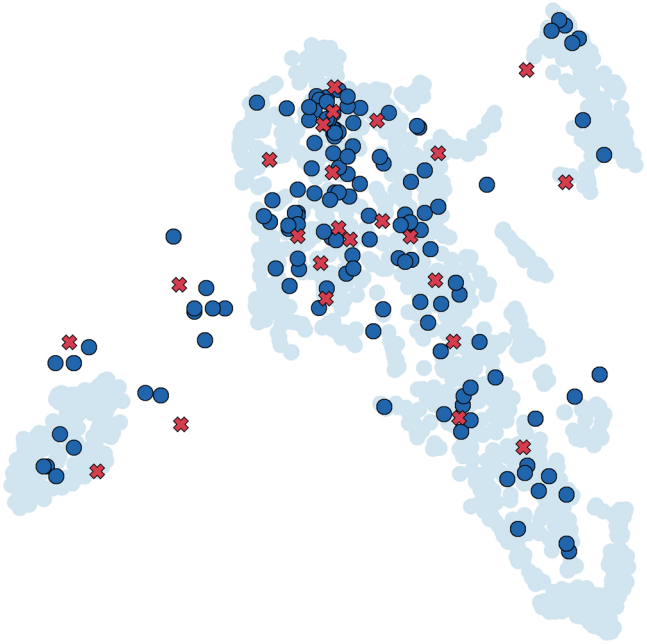


Figure 16. The chemical space plots representing the substrates metabolized by rats. The light blue circles represent the 1300 launched drugs, the dark blue circles represent the compounds in the 10-fold cross-validation set, and the red crosses represent the compounds in the test set.

the test sets would only contain molecules with two or more potential SoM. The models, where the validation sets were missing, are evaluated right after building the model and the step of comparing models built in different ways is skipped.

The splits are illustrated using the chemical space plots, where each compound is represented by a point and the similarity between two compounds by their proximity. The plots have been assembled using

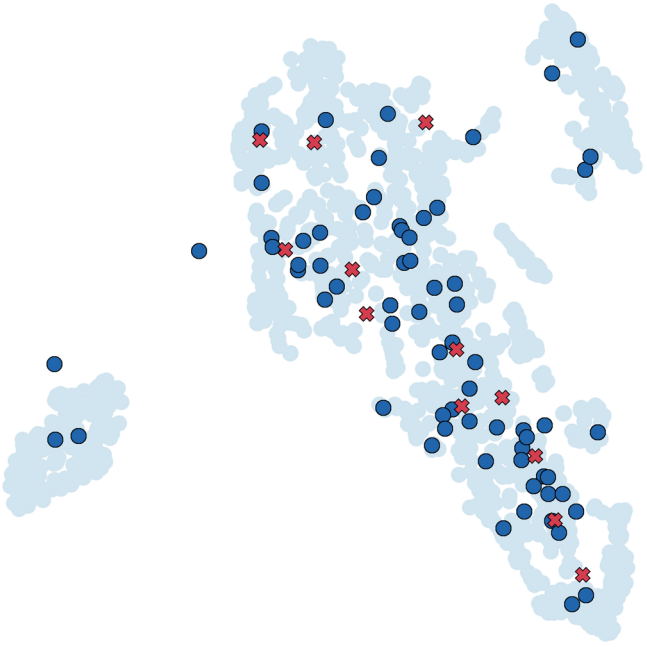


Figure 17. The chemical space plots representing the substrates metabolized by dogs. The light blue circles represent the 1300 launched drugs, the dark blue circles represent the compounds in the 10-fold cross-validation set, and the red crosses represent the compounds in the test set.

Table 3. Approximate Ranges for Evaluating κ Values

κ value	explanation
$\kappa < 0.5$	poor agreement
$0.5 \leq \kappa < 0.6$	moderate agreement
$0.6 \leq \kappa < 0.8$	good agreement
$0.8 \leq \kappa < 1.0$	very good agreement

the compound similarity fingerprint, constructed from the 2D path-based fingerprints, and the similarity is calculated using the Tanimoto similarity coefficient. The chemical spaces were created using a method called Visual Clustering in StarDrop, which uses an approach known as t-distributed Stochastic Neighbor Embedding—a nonlinear dimensionality reduction algorithm ideally suited to visualizing high-dimensional data in two dimensions.[111] The plots include data for approximately 1300 launched drugs, which gives a rough measure of the coverage of the given data sets and enables to compare different data sets to each other.

The chemical space of the substrates for AO1 can be seen in Figure 11. Since most of the substrates of AO1 are azaheterocycles, they tend to cover a narrow area (compared to other enzymes) on the given chemical space. There are exceptions, which are mostly aldehydes.

The following chemical space plots, in Figure 12, are for FMO1 and FMO3. Many substrates for both isoforms overlap; thus, the plots are very similar. Compared to AO1 chemical space, the data points for FMOs are sparser, but the location of the points varies more.

The data for UGT1A1 and UGT1A9 have been grouped together in Figure 13 because the enzymes are known for metabolizing phenolic compounds. While UGT1A1 is considered to be more varied regarding its substrates, then the data set of UGT1A9 features a number of very similar flavonoids, which can be seen on the plot of UGT1A9 (both training and test set data points gathered together).

Both the UGT1A4 and UGT2B7 (Figure 14) have fewer data points compared to the previous UGT isoforms. However, the isoforms are more geared toward *N*-glucuronidation and their substrates can be found from additional areas of the chemical space compared to the UGT1A1 and UGT1A9 isoforms.

https://doi.org/10.1021/acs.jmedchem.2c01303
J. Med. Chem. XXXX, XXX, XXX−XXX

The data sets of CYP substrates for mice, rats, and dogs are on Figures 15, 16, and 17, respectively. Since CYPs tend to metabolize a wide variety of compounds, then the datapoints are distributed more equally compared to the previous plots.

Model Statistics. The statistics that are used to report the inter-rater reliability of the Gaussian processes classification models is Cohen's kappa (kappa or κ). The κ value is a more robust measure than the percentage agreement since it is robust to biases in the representation of classes in the data set and takes into account the possibility of the agreement occurring by chance. For convenience, we also report balanced accuracy. Furthermore, confusion matrices for each model are provided. The rules of how κ values were evaluated are shown in Table 3.

The output for the CYP model differs from the other enzymes, and is an ordered list of potential SoM within a given compound—primary site being the first in the list, which is followed by the secondary SoM *etc.* Hence, the ROC-AUC is calculated for each compound in the set to evaluate the accuracy of the rank ordering, as was done for the human CYP models in our previous work.[1] The AUC is also used when evaluating the importance of the E_a value alone for each enzyme before building the reactivity-accessibility models. A greater AUC indicates a higher performance; the maximum possible AUC is 1 for a perfect classifier, and a value of 0.5 is equivalent to the performance of random selection.

Ideally, a validation set is used to fine-tune the model and the test set is used to make sure that the chosen model is predictive enough while not overtrained. However, as noted above, some data sets within this work are relatively limited in size and the validation set is missing. In such cases, the κ value of the test set might be satisfactory, but to reduce the risk of overtraining, additional tests such as *y*-scrambling were used. *Y*-scrambling is a simple test to explore the predictive power of a pure chance model. In *y*-scrambling, the values of the experimental data (the values to be predicted) were shuffled while the descriptor values were left intact. The scrambled data were then used to train a QSAR model. Cohen's κ value of an excellent model should be considerably higher than the κ value obtained from *y*-scrambling, which should be close to zero. Such tests are necessary because each data point has hundreds of descriptors that might correlate by chance.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jmedchem.2c01303.

Cartesian coordinates and additional data (PDF)

Data sets used for training and testing the models (CSV) (CSV) (CSV) (CSV) (CSV) (CSV) (CSV) (CSV) (CSV) (CSV)

■ AUTHOR INFORMATION

Corresponding Author

Mario Öeren − *Optibrium Limited, Cambridge CB25 9GL, U.K.;* orcid.org/0000-0003-4292-5557; Email: mario@optibrium.com

Authors

Peter J. Walton − *Optibrium Limited, Cambridge CB25 9GL, U.K.; School of Chemistry, University of Nottingham, Nottingham NG7 2RD, U.K.*

James Suri − *Optibrium Limited, Cambridge CB25 9GL, U.K.; School of Chemistry, University of St Andrews, St Andrews KY16 9ST, U.K.*

David J. Ponting − *Lhasa Limited, Leeds LS11 5PS, U.K.;* orcid.org/0000-0001-6840-2629

Peter A. Hunt − *Optibrium Limited, Cambridge CB25 9GL, U.K.;* orcid.org/0000-0003-0380-1893

Matthew D. Segall − *Optibrium Limited, Cambridge CB25 9GL, U.K.;* orcid.org/0000-0002-2105-6535

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jmedchem.2c01303

Notes

The authors declare the following competing financial interest(s): M.Ö., P.A.H., and M.D.S. work at Optibrium Ltd. P.J.W. worked at Optibrium Ltd. during the execution of the FMO project. J.S. worked at Optibrium Ltd. during the execution of the CYP project. D.J.P. works at Lhasa Ltd.

■ ABBREVIATION

AO, aldehyde oxidases; E_a, activation energy; FMO, flavin-containing monooxygenases; GA, glucuronic acid; GP, Gaussian processes; MM, molecular mechanics; MoCo, molybdenum-cofactor; QM, quantum mechanics; ROC, receiver operating characteristic; SMARTS, SMILES arbitrary target specification; SMILES, simplified molecular-input line-entry system; SoM, site of metabolism; UGT, uridine 5′-diphospho-glucuronosyltransferases; WLS, weighted least squares

■ REFERENCES

(1) Tyzack, J. D.; Hunt, P. A.; Segall, M. D. Predicting Regioselectivity and Lability of Cytochrome P450 Metabolism Using Quantum Mechanical Simulations. *J. Chem. Inf. Model.* **2016**, *56*, 2180−2193.

(2) Miners, J. O.; Smith, P. A.; Sorich, M. J.; McKinnon, R. A.; Mackenzie, P. I. Predicting Human Drug Glucuronidation Parameters: Application of In Vitro and In Silico Modeling Approaches. *Annu. Rev. Pharmacol. Toxicol.* **2004**, *44*, 1−25.

(3) 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*, No. e1323.

(4) Torres, R. A.; Korzekwa, K. R.; McMasters, D. R.; Fandozzi, C. M.; Jones, J. P. Use of Density Functional Calculations To Predict the Regioselectivity of Drugs and Molecules Metabolized by Aldehyde Oxidase. *J. Med. Chem.* **2007**, *50*, 4642−4647.

(5) Montefiori, M.; Lyngholm-Kjærby, C.; Long, A.; Olsen, L.; Jørgensen, F. S. Fast Methods for Prediction of Aldehyde Oxidase-Mediated Site-of-Metabolism. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 345−351.

(6) Fu, C.-w.; Lin, T.-H. Predicting the Metabolic Sites by Flavin-Containing Monooxygenase on Drug Molecules Using SVM Classification on Computed Quantum Mechanics and Circular Fingerprints Molecular Descriptors. *PLoS One* **2017**, *12*, No. e0169910.

(7) Sorich, M. J.; McKinnon, R. A.; Miners, J. O.; Smith, P. A. The Importance of Local Chemical Structure for Chemical Metabolism by Human Uridine 5′-Diphosphate−Glucuronosyltransferase. *J. Chem. Inf. Model.* **2006**, *46*, 2692−2697.

(8) Peng, J.; Lu, J.; Shen, Q.; Zheng, M.; Luo, X.; Zhu, W.; Jiang, H.; Chen, K. In Silico Site of Metabolism Prediction for Human UGT-catalyzed Reactions. *Bioinformatics* **2014**, *30*, 398−405.

(9) 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.

(10) Dang, N. L.; Hughes, T. B.; Krishnamurthy, V.; Swamidass, S. J. A Simple Model Predicts UGT-mediated Metabolism. *Bioinformatics* **2016**, *32*, 3183−3189.

(11) Cai, Y.; Yang, H.; Li, W.; Liu, G.; Lee, P. W.; Tang, Y. Computational Prediction of Site of Metabolism for UGT-Catalyzed Reactions. *J. Chem. Inf. Model.* **2019**, *59*, 1085−1095.

(12) Lepri, S.; Ceccarelli, M.; Milani, N.; Tortorella, S.; Cucco, A.; Valeri, A.; Goracci, L.; Brink, A.; Cruciani, G. Structure−Metabolism

Relationships in Human-AOX: Chemical Insights from A Large Database of Aza-aromatic and Amide Compounds. *Proc. Natl. Acad. Sci. U. S. A.* **2017**, *114*, E3178−E3187.
(13) Manevski, N.; King, L.; Pitt, W. R.; Lecomte, F.; Toselli, F. Metabolism by Aldehyde Oxidase: Drug Design and Complementary Approaches to Challenges in Drug Discovery. *J. Med. Chem.* **2019**, *62*, 10955−10994.
(14) Montefiori, M.; Jørgensen, F. S.; Olsen, L. Aldehyde Oxidase: Reaction Mechanism and Prediction of Site of Metabolism. *ACS Omega* **2017**, *2*, 4237−4244.
(15) Krueger, S. K.; Williams, D. E. Mammalian Flavin-containing Monooxygenases: Structure/Function, Genetic Polymorphisms and Role in Drug Metabolism. *Pharmacol. Ther.* **2005**, *106*, 357−387.
(16) Phillips, I. R.; Shephard, E. A. Drug Metabolism by Flavin-containing Monooxygenases of Human and Mouse. *Expert Opin. Drug Metab. Toxicol.* **2017**, *13*, 167−181.
(17) 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 (AUCI/AUC) Ratios. *Drug Metab. Dispos.* **2004**, *32*, 1201−1208.
(18) Walles, M.; Brown, A. P.; Zimmerlin, A.; End, P. New Perspectives on Drug-Induced Liver Injury Risk Assessment of Acyl Glucuronides. *Chem. Res. Toxicol.* **2020**, *33*, 1551−1560.
(19) Lemberg, R.; Wyndham, R. A.; Henry, N. P. On Liver Aldehydrase. *Aust. J. Exp. Biol. Med. Sci.* **1936**, *14*, 259−274.
(20) Gordon, A. H.; Green, D. E.; Subrahmanyan, V. Liver Aldehyde Oxidase. *Biochem. J.* **1940**, *34*, 764−774.
(21) Knox, W. E. The Quinine-Oxidizing Enzyme and Liver Aldehyde Oxidase. *J. Biol. Chem.* **1946**, *163*, 699−711.
(22) Knox, W. E.; Grossman, W. I. The Location of the Reactive Carbon in N^Methylnicotinamide. *J. Am. Chem. Soc.* **1948**, *70*, 2172.
(23) Hucker, H. B.; Gillette, J. R.; Brodie, B. B. Enzymatic Pathway for The Formation of Cotinine, a Major Metabolite of Nicotine in Rabbit Liver. *J. Pharmacol. Exp. Ther.* **1960**, *129*, 94−100.
(24) Pryde, D. C.; Dalvie, D.; Hu, Q.; Jones, P.; Obach, R. S.; Tran, T.-D. Aldehyde Oxidase: An Enzyme of Emerging Importance in Drug Discovery. *J. Med. Chem.* **2010**, *53*, 8441−8460.
(25) Garattini, E.; Terao, M. The Role of Aldehyde Oxidase in Drug Metabolism. *Expert Opin. Drug Metab. Toxicol.* **2012**, *8*, 487−503.
(26) Li, H.; Cui, H.; Kundu, T. K.; Alzawahra, W.; Zweier, J. L. Nitric Oxide Production from Nitrite Occurs Primarily in Tissues Not in the Blood. *J. Biol. Chem.* **2008**, *283*, 17855−17863.
(27) Paragas, E. M.; Humphreys, S. C.; Min, J.; Joswig-Jones, C. A.; Jones, J. P. The Two Faces of Aldehyde Oxidase: Oxidative and Reductive Transformations of 5-nitroquinoline. *Biochem. Pharmacol.* **2017**, *145*, 210−217.
(28) Sodhi, J. K.; Wong, S.; Kirkpatrick, D. S.; Liu, L.; Khojasteh, S. C.; Hop, C. E. C. A.; Barr, J. T.; Jones, J. P.; Halladay, J. S. A Novel Reaction Mediated by Human Aldehyde Oxidase: Amide Hydrolysis of GDC-0834. *Drug Metab. Dispos.* **2015**, *43*, 908−915.
(29) Garattini, E.; Fratelli, M.; Terao, M. Mammalian Aldehyde Oxidases: Genetics, Evolution and Biochemistry. *Cell. Mol. Life Sci.* **2008**, *65*, 1019−1048.
(30) Öeren, M.; Walton, P. J.; Hunt, P. A.; Ponting, D. J.; Segall, M. D. Predicting Reactivity to Drug Metabolism: Beyond P450s—Modelling FMOs and UGTs. *J. Comput.-Aided Mol. Des.* **2021**, *35*, 541−555.
(31) Skibo, E. B.; Gilchrist, J. H.; Lee, C. H. Electronic Probes of the Mechanism of Substrate Oxidation by Buttermilk Xanthine Oxidase: Role of the Active-site Nucleophile in Oxidation. *Biochemistry* **1987**, *26*, 3032−3037.
(32) Alfaro, J. F.; Jones, J. P. Studies on the Mechanism of Aldehyde Oxidase and Xanthine Oxidase. *J. Org. Chem.* **2008**, *73*, 9469−9472.
(33) Kaye, B.; Offerman, J. L.; Reid, J. L.; Elliot, H. L.; Hillis, W. S. A Species Difference in the Presystemic Metabolism of Carbazeran in Dog and Man. *Xenobiotica* **1984**, *14*, 935−945.
(34) Hutzler, J. M.; Cerny, M. A.; Yang, Y.-S.; Asher, C.; Wong, D.; Frederick, K.; Gilpin, K. Cynomolgus Monkey as a Surrogate for Human Aldehyde Oxidase Metabolism of the EGFR Inhibitor BIBX1382. *Drug Metab. Dispos.* **2014**, *42*, 1751−1760.
(35) Lolkema, M. P.; Bohets, H. H.; Arkenau, H.-T.; Lampo, A.; Barale, E.; de Jonge, M. J. A.; van Doorn, L.; Hellemans, P.; de Bono, J. S.; Eskens, F. A. L. M. The c-Met Tyrosine Kinase Inhibitor JNJ-38877605 Causes Renal Toxicity through Species-Specific Insoluble Metabolite Formation. *Clin. Cancer Res.* **2015**, *21*, 2297−2304.
(36) Jones, J. P.; Korzekwa, K. R. Predicting Intrinsic Clearance for Drugs and Drug Candidates Metabolized by Aldehyde Oxidase. *Mol. Pharmaceutics* **2013**, *10*, 1262−1268.
(37) Xu, Y.; Li, L.; Wang, Y.; Xing, J.; Zhou, L.; Zhong, D.; Luo, X.; Jiang, H.; Chen, K.; Zheng, M.; Deng, P.; Chen, X. Aldehyde Oxidase Mediated Metabolism in Drug-like Molecules: A Combined Computational and Experimental Study. *J. Med. Chem.* **2017**, *60*, 2973−2982.
(38) Guengerich, P. F. Cytochrome P450 and Chemical Toxicology. *Chem. Res. Toxicol.* **2008**, *21*, 70−83.
(39) Guengerich, F. P. Cytochrome P450s and Other Enzymes in Drug Metabolism and Toxicity. *AAPS J.* **2006**, *8*, E101−E111.
(40) Isin, E. M.; Guengerich, F. P. Complex Reactions Catalyzed by Cytochrome P450 Enzymes. *Biochim. Biophys. Acta* **2007**, *1770*, 314−329.
(41) Coon, M. J. CYTOCHROME P450: Nature's Most Versatile Biological Catalyst. *Annu. Rev. Pharmacol. Toxicol.* **2005**, *45*, 1−25.
(42) Manikandan, P.; Nagini, S. Cytochrome P450 Structure, Function and Clinical Significance: A Review. *Curr. Drug Targets* **2018**, *19*, 38−54.
(43) Jung, C. The Mystery of Cytochrome P450 Compound I: A Mini-review Dedicated to Klaus Ruckpaul. *Biochim. Biophys. Acta* **2011**, *1814*, 46−57.
(44) Munro, A. W.; Girvan, H. M.; Mason, A. E.; Dunford, A. J.; McLean, K. J. What makes a P450 tick? *Trends Biochem. Sci.* **2013**, *38*, 140−150.
(45) Sligar, S. G. Coupling of Spin, Substrate, and Redox Equilibriums in Cytochrome P450. *Biochemistry* **1976**, *15*, 5399−5406.
(46) Ruckpaul, K.; Rein, H., *Basis and Mechanisms of Regulation of Cytochrome P-450*; Taylor and Francis: London, 1989.
(47) 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.
(48) Ziegler, D. M.; Pettit, F. H. Formation of an Intermediate N-oxide in the Oxidative Demethylation of N,N-dimethylaniline Catalyzed by Liver Microsomes. *Biochem. Biophys. Res. Commun.* **1964**, *15*, 188−193.
(49) Ziegler, D. M.; Pettit, F. H. Microsomal Oxidases. I. The Isolation and Dialkylarylamine Oxygenase Activity of Pork Liver Microsomes. *Biochemistry* **1966**, *5*, 2932−2938.
(50) Poulsen, L. L.; Hyslop, R. M.; Ziegler, D. M. S-oxidation of Thioureylenes Catalyzed by a Microsomal Flavoprotein Mixed-function Oxidase. *Biochem. Pharmacol.* **1974**, *23*, 3431−3440.
(51) Jones, K. C.; Ballou, D. P. Reactions of the 4a-hydroperoxide of Liver Microsomal Flavin-containing Monooxygenase with Nucleophilic and Electrophilic Substrates. *J. Biol. Chem.* **1986**, *261*, 2553−2559.
(52) Chen, G. P.; Poulsen, L. L.; Ziegler, D. M. Oxidation of Aldehydes Catalyzed by Pig Liver Flavin-containing Monooxygenase. *Drug Metab. Dispos.* **1995**, *23*, 1390−1393.
(53) Fiorentini, F.; Geier, M.; Binda, C.; Winkler, M.; Faber, K.; Hall, M.; Mattevi, A. Biocatalytic Characterization of Human FMO5: Unearthing Baeyer−Villiger Reactions in Humans. *ACS Chem. Biol.* **2016**, *11*, 1039−1048.
(54) Ziegler, D. M.; Poulsen, L. L.; Duffel, M. W., *Kinetic Studies on Mechanism and Substrate Specificity of The Microsomal Flavin−Containing Monooxygenase*; Academic Press, 1980, 637−645.
(55) Ziegler, D. M.; Graf, P.; Poulsen, L. L.; Sies, H.; Stahl, W. NADPH-dependent Oxidation of Reduced Ebselen, 2-selenylbenzanilide, and of 2-(methylseleno)benzanilide Catalyzed by Pig Liver

https://doi.org/10.1021/acs.jmedchem.2c01303
J. Med. Chem. XXXX, XXX, XXX−XXX

Flavin-containing Monooxygenase. *Chem. Res. Toxicol.* **1992**, *5*, 163−166.
(56) Massey, V. Activation of Molecular Oxygen by Flavins and Flavoproteins. *J. Biol. Chem.* **1994**, *269*, 22459−22462.
(57) Hao, D. C.; Chen, S. L.; Mu, J.; Xiao, P. G. Molecular Phylogeny, Long-term Evolution, and Functional Divergence of Flavin-containing Monooxygenases. *Genetica* **2009**, *137*, 173−187.
(58) Lang, D. H.; Rettie, A. E. In Vitro Evaluation of Potential in Vivo Probes for Human Flavin-containing Monooxygenase (FMO): Metabolism of Benzydamine and Caffeine by FMO and P450 Isoforms. *Br. J. Clin. Pharmacol.* **2002**, *50*, 311−314.
(59) Mushiroda, T.; Douya, R.; Takahara, E.; Nagata, O. The Involvement of Flavin-Containing Monooxygenase but Not CYP3A4 in Metabolism of Itopride Hydrochloride, a Gastroprokinetic Agent: Comparison with Cisapride and Mosapride Citrate. *Drug Metab. Dispos.* **2000**, *28*, 1231−1237.
(60) Rawden, H. C.; Kokwaro, G. O.; Ward, S. A.; Edwards, G. Relative Contribution of Cytochromes P-450 and Flavin-containing Monoxygenases to the Metabolism of Albendazole by Human Liver Microsomes. *Br. J. Clin. Pharmacol.* **2000**, *49*, 313−322.
(61) Cashman, J. R.; Park, S. B.; Yang, Z. C.; Washington, C. B.; Gomez, D. Y.; Giacomini, K. M.; Brett, C. M. Chemical, Enzymatic, and Human Enantioselective S-oxygenation of Cimetidine. *Drug Metab. Dispos.* **1993**, *21*, 587−597.
(62) 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.
(63) Cashman, J. R.; Hanzlik, R. P. Microsomal Oxidation of Thiobenzamide. A Photometric Assay for the Flavin-containing Monooxygenase. *Biochem. Biophys. Res. Commun.* **1981**, *98*, 147−153.
(64) Dyroff, M. C.; Neal, R. A. Studies of the Mechanism of Metabolism of Thioacetamide S-oxide by Rat Liver Microsomes. *Mol. Pharmacol.* **1983**, *23*, 219−227.
(65) Chieli, E.; Malvaldi, G. Role of the Microsomal FAD-containing Monooxygenase in the Liver Toxicity of Thioacetamide S-oxide. *Toxicology* **1984**, *31*, 41−52.
(66) Ruse, M. J.; Waring, R. H. The Effect of Methimazole on Thioamide Bioactivation and Toxicity. *Toxicol. Lett.* **1991**, *58*, 37−41.
(67) Lee, J. W.; Shin, K. D.; Lee, M.; Kim, E. J.; Han, S.-S.; Han, M. Y.; Ha, H.; Jeong, T. C.; Koh, W. S. Role of Metabolism by Flavin-containing Monooxygenase in Thioacetamide-induced Immunosuppression. *Toxicol. Lett.* **2003**, *136*, 163−172.
(68) Ottolina, G.; de Gonzalo, G.; Carrea, G. Theoretical Studies of Oxygen Atom Transfer From Flavin to Electron-rich Substrates. *J. Mol. Struct.: THEOCHEM* **2005**, *757*, 175−181.
(69) Bach, R. Role of the Somersault Rearrangement in the Oxidation Step for Flavin Monooxygenases (FMO). A Comparison between FMO and Conventional Xenobiotic Oxidation with Hydroperoxides. *J. Phys. Chem. A* **2011**, *115*, 11087−11100.
(70) Hawes, E. M. N +-Glucuronidation, a Common Pathway in Human Metabolism of Drugs With a Tertiary Amine Group. *Drug Metab. Dispos.* **1996**, *26*, 830−837.
(71) Nishiyama, T.; Kobori, T.; Arai, K.; Ogura, K.; Ohnuma, T.; Ishii, K.; Hayashi, K.; Hiratsuka, A. Identification of Human UDP-glucuronosyltransferase Isoform(s) Responsible for the C-glucuronidation of Phenylbutazone. *Arch. Biochem. Biophys.* **2006**, *454*, 72−79.
(72) Buchheit, D.; Schmitt, E. I.; Bischoff, D.; Ebner, T.; Bureik, M. S-Glucuronidation of 7-mercapto-4-methylcoumarin by Human UDP Glycosyltransferases in Genetically Engineered Fission Yeast Cells. *Biol. Chem.* **2011**, *392*, 1089−1095.
(73) Lairson, L. L.; Henrissat, B.; Davies, G. J.; Withers, S. G. Glycosyltransferases: Structures, Functions, and Mechanisms. *Annu. Rev. Biochem.* **2008**, *77*, 521−555.
(74) Liang, D.-M.; Liu, J.-H.; Wu, H.; Wang, B.-B.; Zhu, H.-J.; Qiao, J.-J. Glycosyltransferases: Mechanisms and Applications in Natural Product Development. *Chem. Soc. Rev.* **2015**, *44*, 8350−8374.
(75) Radominska-Pandya, A.; Czernik, P. J.; Little, J. M.; Battaglia, E.; MacKenzie, P. I. Structural and Functional Studies Of UDP-glucuronosyltransferases. *Drug Metab. Rev.* **1999**, *31*, 817−899.
(76) Ouzzine, M.; Antonio, L.; Burchell, B.; Netter, P.; Fournel-Gigleux, S.; Magdalou, J. Importance of Histidine Residues for the Function of the Human Liver UDP-Glucuronosyltransferase UGT1A6: Evidence for the Catalytic Role of Histidine 370. *Mol. Pharmacol.* **2000**, *58*, 1609−1615.
(77) Locuson, C. W.; Tracy, T. S. Comparative Modelling of the Human UDP-glucuronosyltransferases: Insights into Structure and Mechanism. *Xenobiotica* **2007**, *37*, 155−168.
(78) Li, D.; Fournel-Gigleux, S.; Barré, L.; Mulliert, G.; Netter, P.; Magdalou, J.; Ouzzine, M. Identification of Aspartic Acid and Histidine Residues Mediating the Reaction Mechanism and the Substrate Specificity of the Human UDP-glucuronosyltransferases 1A. *J. Biol. Chem.* **2007**, *282*, 36514−36524.
(79) Mackenzie, P. I.; Owens, I. S.; Burchell, B.; Bock, K. W.; Bairoch, A.; Bélanger, A.; Fournel-Gigleux, S.; Green, M.; Hum, D. W.; Iyanagi, T.; Lancet, D.; Louisot, P.; Magdalou, J.; Chowdhury, J. R.; Ritter, J. K.; Schachter, H.; Tephly, T. R.; Tipton, K. F.; Nebert, D. W. Nomenclature Update for the Mammalian UDP Glycosyltransferase (UGT) Gene Superfamily. *Pharmacogenet. Genomics* **2005**, *15*, 677−685.
(80) Meech, R.; Hu, D. G.; McKinnon, R. A.; Mubarokah, S. N.; Haines, A. Z.; Nair, P. C.; Rowland, A.; Mackenzie, P. I. The UDP-Glycosyltransferase (UGT) Superfamily: New Members, New Functions, and Novel Paradigms. *Physiol. Rev.* **2019**, *99*, 1153−1222.
(81) Nair, P. C.; Meech, R.; Mackenzie, P. I.; McKinnon, R. A.; Miners, J. O. Insights into the UDP-sugar Selectivities of Human UDP-glycosyltransferases (UGT): a Molecular Modeling Perspective. *Drug Metab. Rev.* **2015**, *47*, 335−345.
(82) Bock, K. W. The UDP-glycosyltransferase (UGT) Superfamily Expressed in Humans, Insects and Plants: Animal-plant Arms-race and Co-evolution. *Biochem. Pharmacol.* **2016**, *99*, 11−17.
(83) Tukey, R. H.; Strassburg, C. P. Human UDP-Glucuronosyltransferases: Metabolism, Expression, and Disease. *Annu. Rev. Pharmacol. Toxicol.* **2000**, *40*, 581−616.
(84) Kazmi, S. R.; Jun, R.; Yu, M.-S.; Jung, C.; Na, D. In Silico Approaches and Tools for the Prediction of Drug Metabolism and Fate: A Review. *Comput. Biol. Med.* **2019**, *106*, 54−64.
(85) Feenstra, K. A.; De Graaf, C.; Vermeulen, N. P. E., *Drug-drug interactions*; Informa Healthcare, 2008.
(86) Nair, P. C.; McKinnon, R. A.; Miners, J. O. Cytochrome P450 Structure-function: Insights from Molecular Dynamics Simulations. *Drug Metab. Rev.* **2016**, *48*, 434−452.
(87) Panneerselvam, S.; Yesudhas, D.; Durai, P.; Anwar, M. A.; Gosu, V.; Choi, S. A Combined Molecular Docking/Dynamics Approach to Probe the Binding Mode of Cancer Drugs with Cytochrome P450 3A4. *Molecules* **2015**, *20*, 14915−14935.
(88) Dubey, K. D.; Wang, B.; Shaik, S. Molecular Dynamics and QM/MM Calculations Predict the Substrate-Induced Gating of Cytochrome P450 BM3 and the Regio- and Stereoselectivity of Fatty Acid Hydroxylation. *J. Am. Chem. Soc.* **2016**, *138*, 837−845.
(89) Rydberg, P.; Gloriam, D. E.; Olsen, L. The SMARTCyp Cytochrome P450 Metabolism Prediction Server. *Bioinformatics* **2010**, *26*, 2988−2989.
(90) 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.
(91) Laidler, K. J. The Development of the Arrhenius Equation. *J. Chem. Educ.* **1984**, *61*, 494.
(92) Hunt, P. A.; Segall, D. M.; Tyzack, J. D. WhichP450: a Multi-class Categorical Model to Predict the Major Metabolising CYP450 Isoform for a Compound. *J. Comput.-Aided Mol. Des.* **2018**, *32*, 537−546.
(93) Rostkowski, M.; Spjuth, O.; Rydberg, P. WhichCyp: Prediction of Cytochromes P450 Inhibition. *Bioinformatics* **2013**, *29*, 2051−2052.

https://doi.org/10.1021/acs.jmedchem.2c01303
J. Med. Chem. XXXX, XXX, XXX−XXX

(94) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and Use of Quantum Mechanical Molecular Models. 76. AM1: a New General Purpose Quantum Mechanical Molecular Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.
(95) Stewart, J. J. P. Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements. *J. Mol. Model.* **2007**, *13*, 1173−1213.
(96) Marcou, G.; Rognan, D. Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints. *J. Chem. Inf. Model.* **2007**, *47*, 195−207.
(97) Stahl, M.; Mauser, H. Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods. *J. Chem. Inf. Model.* **2005**, *45*, 542−548.
(98) Hutter, J.; Iannuzzi, M.; Schiffmann, F.; VandeVondele, J. CP2K: Atomistic Simulations of Condensed Matter Systems. *WIREs Comput. Mol. Sci.* **2014**, *4*, 15−25.
(99) Vosko, S. H.; Wilk, L.; Nusair, M. Accurate Spin-dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: a Critical Analysis. *Can. J. Phys.* **1980**, *58*, 80−1211.
(100) Lee, C.; Yang, W.; Parr, R. G. Development of the Colle-Salvetti Correlation-energy Formula Into a Functional of the Electron Density. *Phys. Rev. B* **1988**, *37*, 785−789.
(101) Becke, A. D. Density-functional Thermochemistry. III. The Eole of Exact Exchange. *J. Chem. Phys.* **1993**, *98*, 5648−5652.
(102) Stephens, P. J.; Devlin, F. J.; Chabalowski, C. F.; Frisch, M. J. Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields. *J. Phys. Chem.* **1994**, *98*, 11623−11627.
(103) Ehrlich, S.; Moellmann, J.; Reckien, W.; Bredow, T.; Grimme, S. System-Dependent Dispersion Coefficients for the DFT-D3 Treatment of Adsorption Processes on Ionic Surfaces. *ChemPhysChem* **2011**, *12*, 3414−3420.
(104) Weigend, F.; Ahlrichs, R. Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy. *Phys. Chem. Chem. Phys.* **2005**, *7*, 3297−3305.
(105) Andrae, D.; Häußermann, U.; Dolg, M.; Stoll, H.; Preuß, H. Energy-adjusted Ab Initio Pseudopotentials for the Second and Third Row Transition Elements. *Theor. Chim. Acta* **1990**, *77*, 123−141.
(106) Pritchard, B. P.; Altarawy, D.; Didier, B.; Gibson, T. D.; Windus, T. L. A New Basis Set Exchange: An Open, Up-to-date Resource for the Molecular Sciences Community. *J. Chem. Inf. Model.* **2019**, *59*, 4814−4820.
(107) Simón, L.; Goodman, J. M. How Reliable are DFT Transition Structures? Comparison of GGA, Hybrid-meta-GGA and Meta-GGA Functionals. *Org. Biomol. Chem.* **2011**, *9*, 689−700.
(108) Valiev, M.; Bylaska, E. J.; Govind, N.; Kowalski, K.; Straatsma, T. P.; Van Dam, H. J. J.; Wang, D.; Nieplocha, J.; Apra, E.; Windus, T. L.; de Jong, W. A. NWChem: A Comprehensive and Scalable Open-source Solution for Large Scale Molecular Simulations. *Comput. Phys. Commun.* **2010**, *181*, 1477−1489.
(109) Obrezanova, O.; Csányi, G.; Gola, J. M. R.; Segall, M. D. Gaussian Processes: A Method for Automatic QSAR Modeling of ADME Properties. *J. Chem. Inf. Model.* **2007**, *47*, 1847−1857.
(110) Ruppert, D.; Wand, M. P. Multivariate Locally Weighted Least Squares Regression. *Ann. Stat.* **1994**, *22*, 1346−1370.
(111) van der Maaten, L.; Hinton, G. Visualizing Data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579−2605.

https://doi.org/10.1021/acs.jmedchem.2c01303
J. Med. Chem. XXXX, XXX, XXX−XXX