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,* [†] *Matthew D. Segall* [†]

† Optibrium Limited, Cambridge Innovation Park, Denny End Road, Cambridge, CB25 9PB, UK

‡ School of Chemistry, University of Nottingham, University Park, Nottingham, NG7 2RD, UK

§ School of Chemistry, University of St Andrews, North Haugh, St Andrews, KY16 9ST, UK

^ Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS, UK

E-mail of the corresponding author: mario@optibrium.com

ORCID of the corresponding author: 0000-0003-4292-5557

Abstract

Unexpected metabolism in modification and conjugation phases can lead to the failure of many late-stage drug candidates or even withdrawals of approved drugs. Thus, it is critical to predict the sites of metabolism (SoM) for enzymes, which are known to interact with drug-like molecules, in the early stages of the research. The study presents methods for predicting the isoform-specific metabolism for human AOs, FMOs and UGTs and general CYP metabolism for pre-clinical 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 kappa values of up to 0.94 and AUC of up to 0.92.

Keywords

AO, Accessibility, CYP, DFT, FMO, Glucuronidation, Metabolism, ML, Oxidation, Reactivity, Semi-empirical, UGT

Declarations

Funding: The research was funded by Optibrium Ltd. and Lhasa Ltd.

Conflicts of interest: MÖ, PAH and MDS are employees of Optibrium Ltd. DJP is an employee of Lhasa Ltd. PJW was an employee of Optibrium Ltd. for the duration of the FMO project. JS was an employee of Optibrium Ltd. for the duration of the CYP project.

Ethics approval: The research does not involve human participation or personal data.

Availability of data and material: The data used in the project is available through the referenced publications.

Availability of data and material: The Supporting Information includes Cartesian coordinates, total DFT and semi-empirical energy values of the presented structures, and data sets used. In addition, the input files for the general DFT and semi-empirical calculations are provided.

Code availability: The software used in the project is freely available through the referenced publications.

Introduction

The characterisation 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, optimising 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, 8, 9, 10, 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 towards CYP metabolism but increases the likelihood of oxidation by AOs. The rapid clearance of molecules by AOs (not predicted by CYP-only modelling) has caused the discontinuation of multiple projects during clinical trials. [12, 13, 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 metabolised 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 pre-clinical 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* modelling of the metabolism of pre-clinical species could aid in ensuring the pre-clinical trials produce the likely human metabolites, using the model as an indicator for the best pre-clinical species. As well as the ethical benefits of this modelling 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 substrate space and reaction types – and the available data for building and validating models. Following this, we summarise the spectra of available modelling 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 catalysing the oxidation of aromatic heterocycles [21] and iminium ions. [22, 23, 24, 25] It is intriguing that AOs, which are considered to contribute to the modification phase, have also been observed to catalyse the reduction [26] of various molecules, e.g. nitro-compounds. However, with 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 catalyse 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] and [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 Alfaro et al. and Montefiori et al. have confirmed the proposed concerted reaction. [32, 14] The described transition state is depicted in Figure 1.

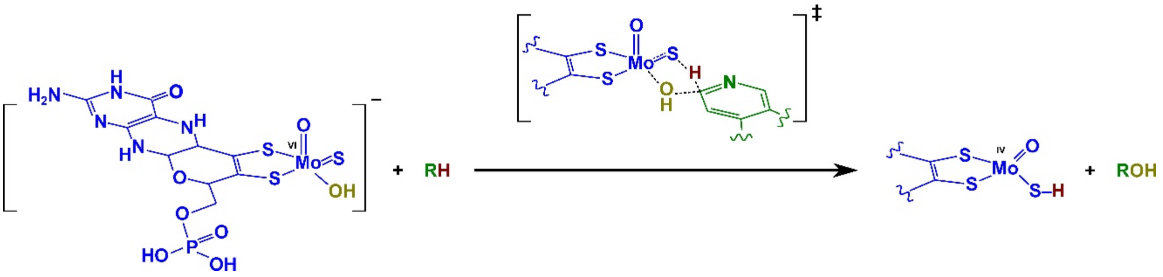

Fig. 1 The transition state for oxidation by AO.

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 et al. 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 dataset (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 kappa 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 recognising 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 to 90% of hepatically-cleared drugs in humans. [3, 38, 39] The catalytic action of CYPs is predominantly that of a monooxygenase (C-hydroxylation, heteroatom oxygenation, dealkylation) but also includes epoxide formation and aromatic dehalogenation, among 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 et al., [40] Coon, [41] Manikandan et al. [42] and Jung [43].

The catalysis by CYPs requires the haem-iron centre 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 oxidises 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]

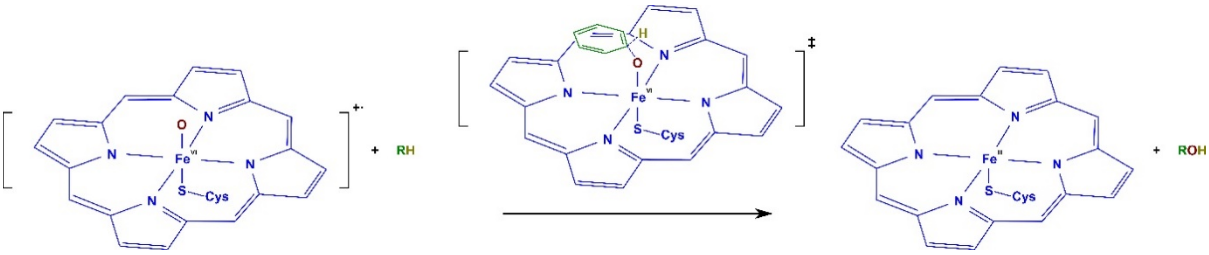

Fig. 2 The transition state for oxidation by CYP.

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 fulfil 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 pre-clinical species such as rats, mice, and dogs.

Flavin-containing Monooxygenases

The discovery of FMOs could be credited to Ziegler et al., who in 1964 suggested that the oxidative *N*-dealkylation catalysed in the mammalian liver homogenates is divided into partial reactions catalysed 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 catalysed the oxidation of the nitrogen atom, proving their initial theory. [49] It is now known that FMOs are able to oxidise 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 oxidised by FMOs: sulfides, thiols and disulfides, thiocarbamides and thioamides, mercaptopurines, and mercaptopyrimidine. [15] In addition, FMOs have been observed to oxidise a wide variety of atoms such as boron, [51] carbon (Bayer-Villiger oxidation), [52, 53] phosphorus, [54] selenium [55] and iodine [54]. [15] 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 catalyse *N*- and *S*-oxidation. The 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 oxidised 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$^+$).

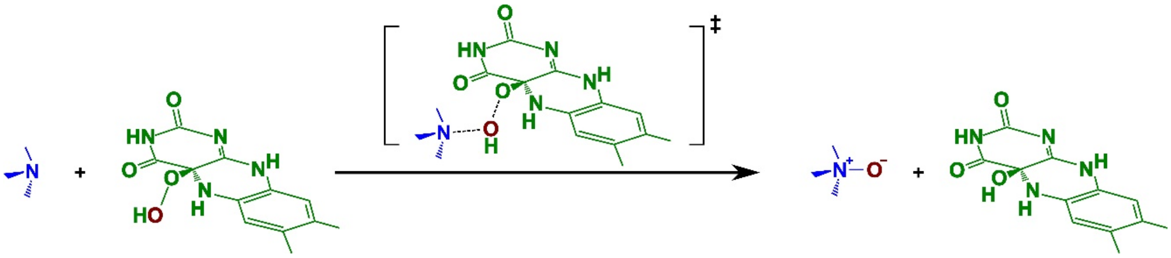

Fig. 3 The transition state for oxidation by FMO.

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 non-functional 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 foetal 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 metabolised by FMOs. [58, 59, 60, 61, 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 [63, 64, 65, 66, 67, 15]. 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 towards 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 et al. proposed that the reaction proceeds via radical intermediates. [69] The latest results in our previously published work supports the S_N2 reaction mechanism. [30] Only one model for predicting SoM for FMOs has been published by Fu et al., 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. [39, 3, 38, 17] 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 sub-class of enzymes called glycosyltransferases, which are responsible for catalysing 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 centre. 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 stabilise the reaction as depicted in Figure 4. [75, 76, 77, 78, 30]

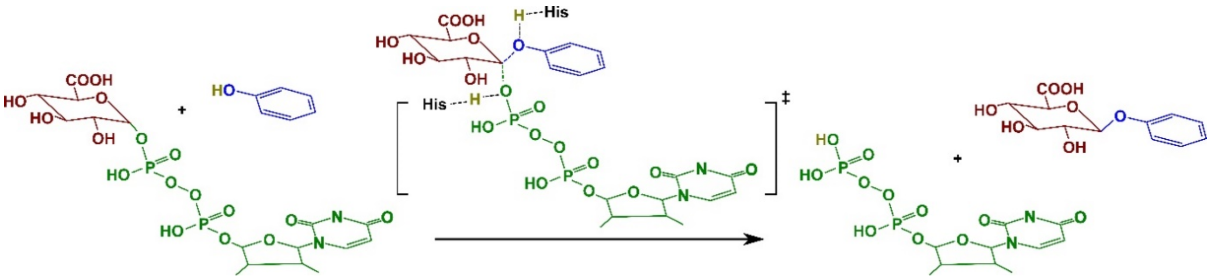

Fig. 4 The transition state for glucuronidation by UGTs. The residues taking part in the reaction are based on the homology model of UGT isoform 1A1. [77]

Most kingdoms in biology include species with UGTs. [79] There is a total of 31 UGT isoforms found in humans – 22 active isoforms and 9 pseudogenes. Based on the sequence similarity, the active isoforms are divided into four categories – UGT1, UGT2, UGT3 and UGT4. In theory, the large number of different isoforms give 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 liver and are responsible for the conjugation of the majority of xenobiotic UGT substrates. [79, 81, 82, 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, 9, 10, 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-catalysed 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.

Modelling Drug Metabolism

There are many available modelling methods for predicting metabolism, ranging from empirical methods such as statistical modelling 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 non-transferable and qualitative. In cases where data is sparse, researchers may look towards 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 manner 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 non-substrate compounds are modelled 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. [88, 1]

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 modelling methods with elements from the mechanistic approaches to derive reactivity-accessibility models, which achieve a balance between computational cost and accuracy for modelling 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 centre 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 the activation energy is an important descriptor of whether a potential SoM will be metabolised experimentally. [1]

The rationale being that rate of a reaction is related to the activation energy by the Arrhenius equation. [91] Since the reaction centre 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 a SoM refers to how easily it may be approached by the enzyme's reactive centre, 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 centre. 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 centre 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 includes molecules, which are observed to be metabolised. Site-specific information for molecules known not to be metabolised was not considered because it is unclear why the molecule was not metabolised; a molecule may have a highly reactive site but would not be metabolised 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 dataset, in the majority of the cases, have two or more potential SoM, out of which at least one is experimentally observed to be metabolised. To summarise, the collected compounds were labelled according to which enzyme family and which isoform from that family is responsible for metabolising the molecule (note that some molecules are metabolised by multiple isoforms or enzyme classes). Each potential SoM on a molecule was labelled as either observed to be metabolised or not by the corresponding isoform. The exception was the data for CYP metabolism by pre-clinical 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 labelled as 1st, 2nd, 3rd or "not observed" for primary, secondary, tertiary or not metabolised 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 was 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 recognised 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 pre-clinical species followed the same rules with the exception of isoform-specificity since these models were general. The four datasets are summarised in Table 1 and the following paragraphs describe the size and the content of the datasets for each isoform and enzyme family. The references from which the data were obtained are listed in the Supporting Information of this work.

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 summarise, the data set for AOs consists of 157 molecules and 865 potential sites, of which 160 are observed experimentally to be metabolised – 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 oxidised 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 metabolised by FMO1 and 209 potential SoM out of which FMO3 metabolised 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 dataset for the UGT1A4 isoform is, overall, the smallest and contains only 54 molecules. However, it is the most balanced dataset 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 neighbourhood 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 pre-clinical 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 dataset 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 metabolised. The data set for rats is the biggest, with 1428 potential SoM, out of which 305 are metabolised. The data set for dogs features 1091 sites, out of which 154 are metabolised. 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.

Table 1 The overview of data for building reactivity-accessibility models.

Enzyme	Isoform*	No. of Substrates	No. of Potential SoM	No. of SoM Metabolised
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

* For CYPs the species instead of isoforms are mentioned.

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 non-isoform specific CYP models for pre-clinical species, such as mice, rats, and dogs. The *in silico* models are useful for predicting the modification and conjugation phases in humans. Modelling the metabolism of pre-clinical species could aid in ensuring the pre-clinical trials produce the likely human metabolites, using the model as an indicator for selecting the best pre-clinical species.

Results and Discussion

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

The 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.

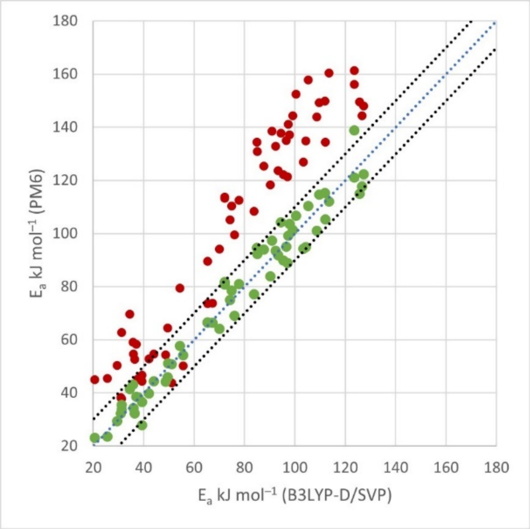

Fig. 5 Correlation between DFT and semi-empirical Ea values. Red dots represent the Ea values, and the green points represent the corrected Ea values. The blue line is the identity line, and the black lines represent deviation of +10 and −10 kJ per mol from the identity line.

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 kappa value for the test set for the Gaussian Processes (GP) AO model is 0.83. The E_a was amongst the most important descriptors; the most influential being the descriptor that recognises 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*-scrambled result had kappa 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.

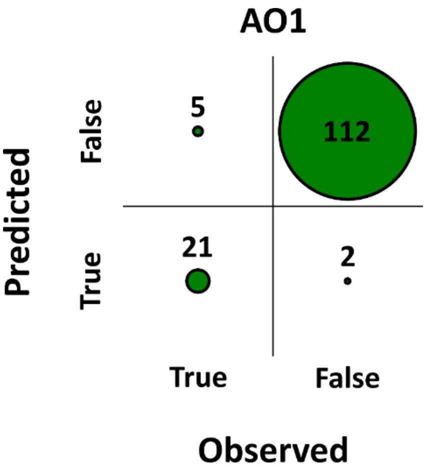

Fig. 6 The confusion matrix of the test set of Gaussian Processes model for AO1.

The GP Models for FMOs

We used the simplified reaction mechanism for calculating the 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 sub-environments (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 optimised by DFT; however, it often does not form during the geometry optimisation 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 kappa 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, the 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.

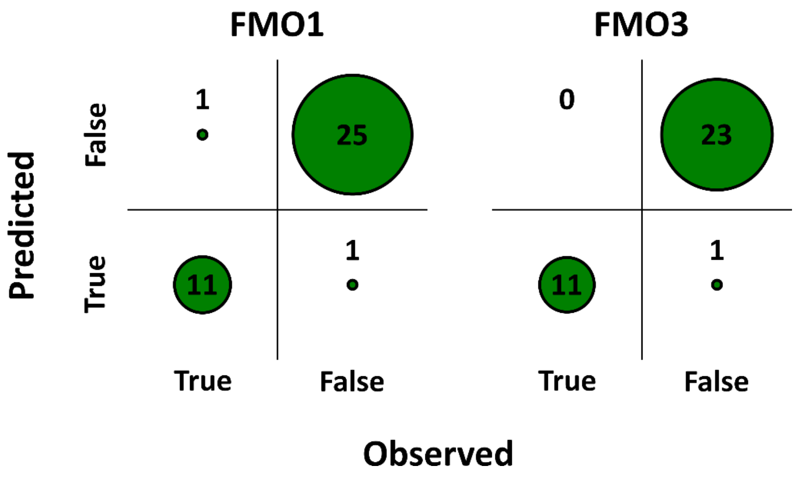

Fig. 7 The confusion matrices of the test sets of GP models for FMOs.

The 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.

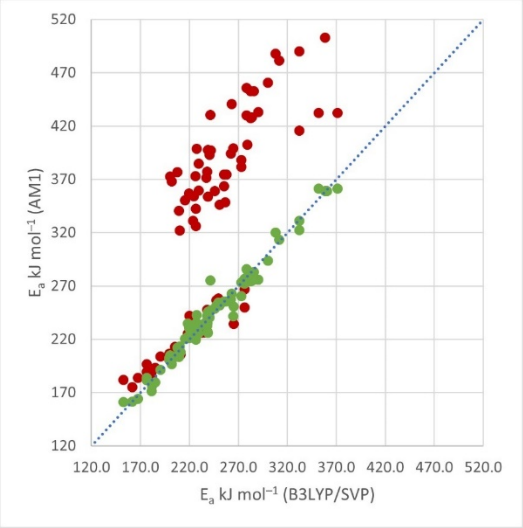

Fig. 8 The correlation between DFT (B3LYP/SVP) and semi-empirical method (AM1) for *N*- and *O*-glucuronidation. Red points represent the uncorrected values, and green points represent the corrected values.

The AUC for the 1A1 isoform, using AM1, was 0.86. The GP model for 1A1 yielded a kappa 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 kappa 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 specialised for the metabolism of tertiary nitrogen atoms it could be theorised 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 kappa 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 kappa 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 kappa 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 kappa 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 kappa 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 kappa value of the GP model is relatively low while the AUC is the highest among 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 metabolised is 26. In few cases the model fails to recognise the secondary SoM, which in turn lowers the kappa value of the model considerably.

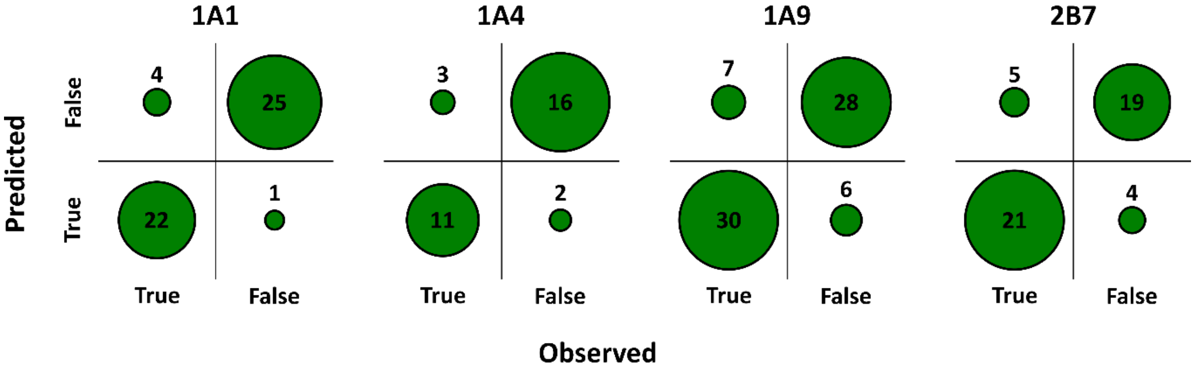

Fig. 9 The confusion matrices of test sets of GP models for UGTs.

The 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 4 (where a 1 indicates a primary site and a 4 indicates not metabolised). For a given site, we used consensus modelling, 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 pre-clinical 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 for 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 pre-clinical species did not specify individual isoforms, it is likely that the general CYP pre-clinical species models achieved such excellent results because the experimental data mostly consists 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 pre-clinical 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.

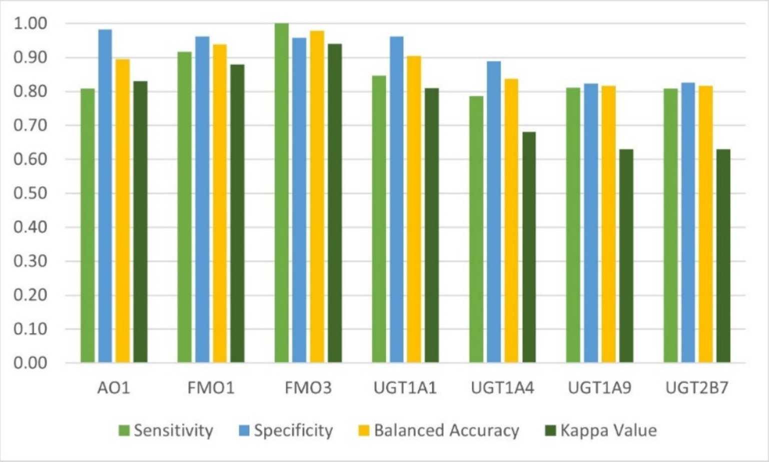

Fig. 10 The sensitivity, specificity, balanced accuracy and kappa values for human isoforms of AO, FMO, and UGT.

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 ten 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 pre-clinical 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 pre-clinical 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 or families and isoform(s) are most likely to be responsible for 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 metabolic fate of a compound based only on its chemical structure. The reactivity-accessibility method for modelling drug metabolism has proved to be generalisable, 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 centre can approach the potential SoM. [1] In this work, the reactivity is characterised using the Ea and the ΔE_a of a simplified transition state. The ΔE_a specifies the difference in E_a values between sites within a molecule. E.g. the ΔE_a values for two potential sites in a molecule with the E_a values of 50 kJ mol^{-1} and 75 kJ mol^{-1} would be 0 kJ mol^{-1} 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 which 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 metabolised 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 amount 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, in order 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 recognised using SMARTS patterns and the application of corrections can be automated.

DFT calculations, were run using the B3LYP or B3LYP-D functionals [99, 100, 101, 102, 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 optimisations, 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 reference [30]). The geometry optimisations 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. 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 modelling. 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) modelling. This method overcomes many of the problems of existing QSAR modelling 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 modelling non-linear relationships. The method has built-in mechanisms to prevent over-training 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 modelling 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 minimises 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 (non-metabolised 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 maximise 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 was 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 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 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 Neighbour Embedding – a nonlinear dimensionality reduction algorithm ideally suited to visualising 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.

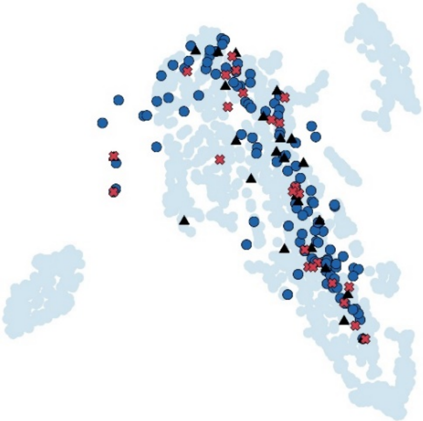

Fig. 11 The chemical space plot representing the AO1 substrates. The light blue circles represent the 1300 launched drugs, the dark blue circles represent the compounds in the training set, the black triangles represent the compounds in the validation set, and the red crosses represent the compounds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

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 more sparse, but the location of the points varies more.

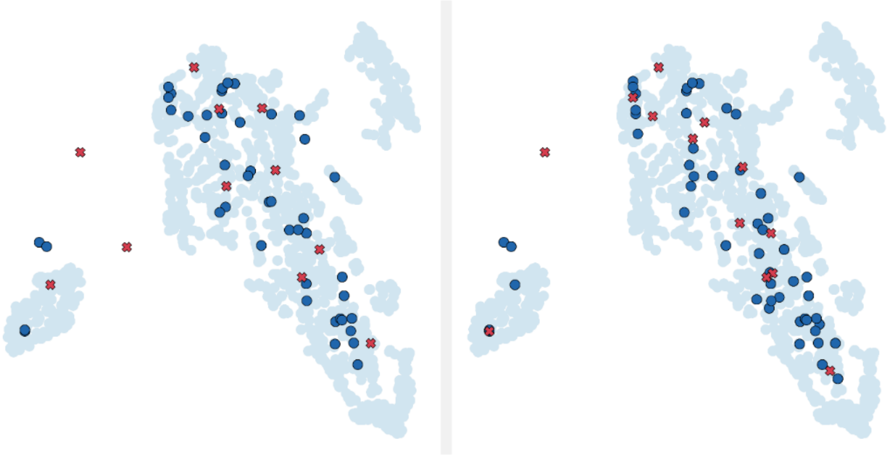

Fig. 12. The chemical space plots representing the FMO1 (left) and FMO3 (right) substrates. The light blue circles represent the 1300 launched drugs, the dark blue circles represent the compounds in the training set, and the red crosses represent the compounds in the test set.

The data for UGT1A1 and UGT1A9 have been grouped together in Figure 13 because the enzymes are known for metabolising 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).

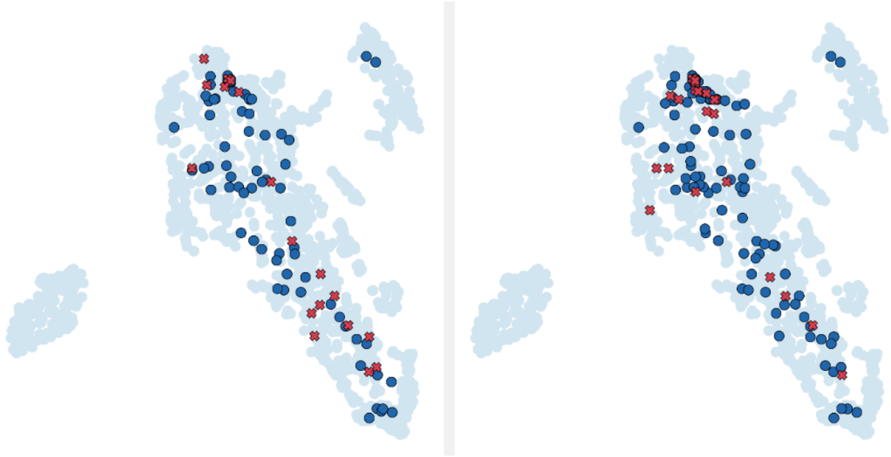

Fig. 13 The chemical space plots representing the UGT1A1 (left) and UGT1A9 (right) substrates. The light blue circles represent the 1300 launched drugs, the dark blue circles represent the compounds in the training set, and the red crosses represent the compounds in the test set.

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

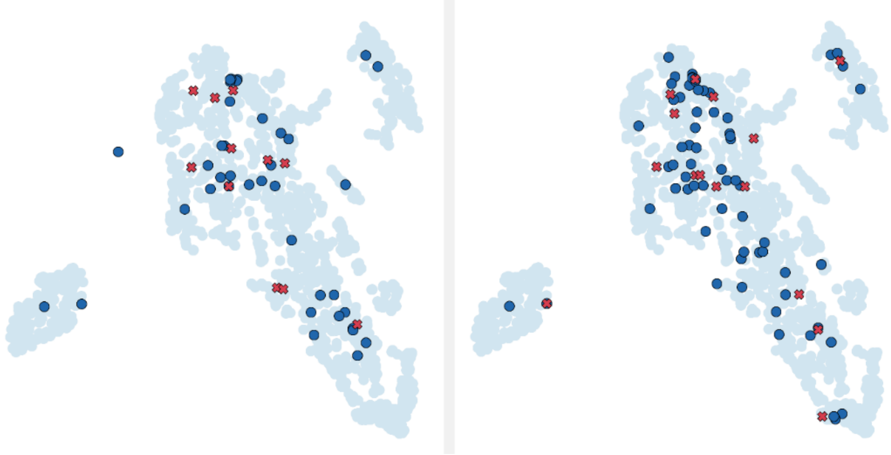

Fig. 14 The chemical space plots representing the UGT1A4 (left) and UGT2B7 (right) substrates. The light blue circles represent the 1300 launched drugs, the dark blue circles represent the compounds in the training set, and the red crosses represent the compounds in the test set.

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

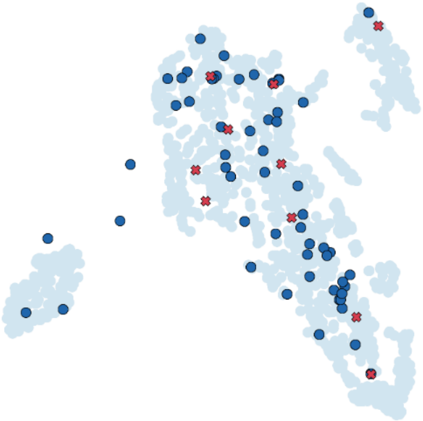

Fig. 15 The chemical space plots representing the substrates metabolised by mice. The light blue circles represent the 1300 launched drugs, the dark blue circles represent the compounds in the 10-fold cross-validation set, and the red crosses represent the compounds in the test set.

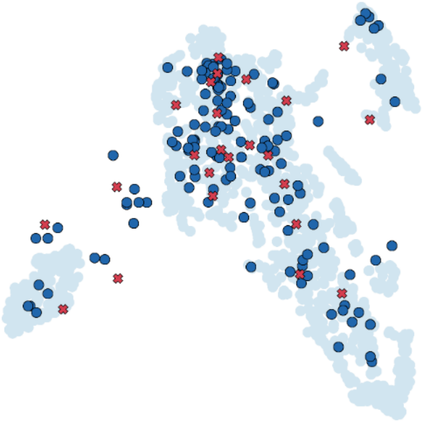

Fig. 16 The chemical space plots representing the substrates metabolised by rats. The light blue circles represent the 1300 launched drugs, the dark blue circles represent the compounds in the 10-fold cross-validation set, and the red crosses represent the compounds in the test set.

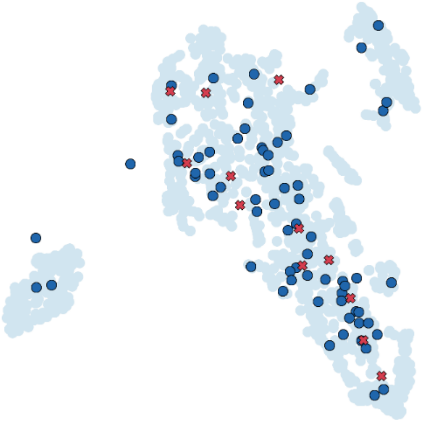

Fig. 17 The chemical space plots representing the substrates metabolised by dogs. The light blue circles represent the 1300 launched drugs, the dark blue circles represent the compounds in the 10-fold cross-validation set, and the red crosses represent the compounds in the test set.

Model Statistics

The statistics which are used to report the inter-rater reliability of the Gaussian Processes classification models is Cohen's kappa (kappa or κ). The kappa 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 kappa values were evaluated are shown in Table 3.

Table 3 Approximate ranges for evaluating kappa values.

$\kappa < 0.5$	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 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 Ea 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 kappa 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) are shuffled while the descriptor values were left intact. The scrambled data was then used to train a QSAR model. Cohen's kappa value of an excellent model should be considerably higher than the kappa value obtained from y- scrambling, which, should be close to zero. Such tests are necessary because each data point has hundreds of descriptors which might correlate by chance.

References

[1] J. D. Tyzack, P. A. Hunt and M. D. Segall, "Predicting Regioselectivity and Lability of Cytochrome P450 Metabolism Using Quantum Mechanical Simulations," Journal of Chemical Information and Modeling, vol. 56, no. 11, pp. 2180-2193, 2016.

[2] J. O. Miners, P. A. Smith, M. J. Sorich, R. A. McKinnon and P. I. Mackenzie, "Predicting Human Drug Glucuronidation Parameters: Application of In Vitro and In Silico Modeling Approaches," Annual Review of Pharmacology and Toxicology, vol. 44, pp. 1-25, 2003.

[3] V. A. Dixit, L. A. Lal and S. R. Agrawal, "Recent Advances in the Prediction of Non-CYP450-mediated Drug Metabolism," WIREs Computational Molecular Science, p. e1323, 2017.

[4] R. A. Torres, K. R. Korzekwa, D. R. McMasters, C. M. Fandozzi and J. P. Jones, "Use of Density Functional Calculations To Predict the Regioselectivity of Drugs and Molecules Metabolized by Aldehyde Oxidase," Journal of Medicinal Chemistry, vol. 50, no. 19, pp. 4642-4647, 2007.

[5] M. Montefiori, C. Lyngholm-Kjærby, A. Long, L. Olsen and F. S. Jørgensen, "Fast Methods for Prediction of Aldehyde Oxidase-Mediated Site-of-Metabolism," Computational and Structural Biotechnology Journal, vol. 17, pp. 345-351, 2019.

[6] C.-w. Fu and T.-H. Lin, "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, vol. 12, no. 1, p. e0169910, 2017.

[7] M. J. Sorich, R. A. McKinnon, J. O. Miners and P. A. Smith, “The Importance of Local Chemical Structure for Chemical Metabolism by Human Uridine 5′-Diphosphate–Glucuronosyltransferase,” Journal of Chemical Information and Modeling, vol. 46, no. 6, pp. 2692-2697, 2006.

[8] J. Peng, J. Lu, Q. Shen, M. Zheng, X. Luo, W. Zhu, H. Jiang and K. Chen, “In Silico Site of Metabolism Prediction for Human UGT-catalyzed Reactions,” Bioinformatics, vol. 30, no. 3, pp. 398-405, 2013.

[9] A. Rudik, A. Dmitriev, A. Lagunin, D. Filimonov and V. Poroikov, “SOMP: Web Server for In Silico Prediction of Sites of Metabolism for Drug-like Compounds,” Bioinformatics, vol. 31, no. 12, pp. 2046-2048, 2015.

[10] N. L. Dang, T. B. Hughes, V. Krishnamurthy and S. J. Swamidass, “A Simple Model Predicts UGT-mediated Metabolism,” Bioinformatics, vol. 32, no. 20, pp. 3183-3189, 2016.

[11] Y. Cai, H. Yang, W. Li, G. Liu, P. W. Lee and Y. Tang, “Computational Prediction of Site of Metabolism for UGT-Catalyzed Reactions,” Journal of Chemical Information and Modeling, vol. 59, no. 3, pp. 1085-1095, 2019.

[12] S. Lepri, M. Ceccarelli, N. Milani, S. Tortorella, A. Cucco, A. Valeri, L. Goracci, A. Brink and G. Cruciani, “Structure–Metabolism Relationships in Human-AOX: Chemical Insights from A Large Database of Aza-aromatic and Amide Compounds,” Proceedings of the National Academy of Sciences of the United States of America, vol. 114, no. 16, pp. E3178-E3187, 2017.

[13] N. Manevski, L. King, W. R. Pitt, F. Lecomte and F. Toselli, “Metabolism by Aldehyde Oxidase: Drug Design and Complementary Approaches to Challenges in Drug Discovery,” Journal of Medicinal Chemistry, vol. 62, no. 24, pp. 10955-10994, 2019.

[14] M. Montefiori, F. S. Jørgensen and L. Olsen, “Aldehyde Oxidase: Reaction Mechanism and Prediction of Site of Metabolism,” ACS Omega, vol. 2, no. 8, pp. 4237-4244, 2017.

[15] S. K. Krueger and D. E. Williams, “Mammalian Flavin-containing Monooxygenases: Structure/Function, Genetic Polymorphisms and Role in Drug Metabolism,” Pharmacology & Therapeutics, vol. 106, no. 3, pp. 357-387, 2005.

[16] I. R. Phillips and E. A. Shephard, “Drug Metabolism by Flavin-containing Monooxygenases of Human and Mouse,” Expert Opinion on Drug Metabolism & Toxicology, vol. 13, no. 2, pp. 167-181, 2016.

[17] J. A. Williams, R. Hyland, B. C. Jones, D. A. Smith, S. Hurst, T. C. Goosen, V. Peterkin, J. R. Koup and S. E. Ball, “Drug-drug Interactions For UDP-glucuronosyltransferase Substrates: A Pharmacokinetic Explanation For Typically Observed Low Exposure (AUCI/AUC) Ratios,” Drug Metabolism and Disposition, vol. 32, no. 11, pp. 1201-1208, 2004.

[18] M. Walles, A. P. Brown, A. Zimmerlin and P. End, “New Perspectives on Drug-Induced Liver Injury Risk Assessment of Acyl Glucuronides,” Chemical Research in Toxicology, vol. 33, no. 7, p. 1551–1560, 2020.

[19] R. Lemberg, R. A. Wyndham and N. P. Henry, “On Liver Aldehydrase,” Australian Journal of Experimental Biology & Medical Science, vol. 14, no. 4, pp. 259-274, 1936.

[20] A. H. Gordon, D. E. Green and V. Subrahmanyan, “Liver Aldehyde Oxidase,” Biochemical Journal, vol. 34, no. 5, pp. 764-774, 1940.

[21] W. E. Knox, “The Quinine-Oxidizing Enzyme and Liver Aldehyde Oxidase,” Journal of Biological Chemistry, vol. 163, no. 3, pp. 699-711, 1946.

[22] W. E. Knox and W. I. Grossman, “The Location of the Reactive Carbon in N^Methylnicotinamide,” Journal of American Chemical Society, vol. 70, no. 6, p. 2172, 1948.

[23] H. B. Hucker, J. R. Gillette and B. B. Brodie, “Enzymatic Pathway for The Formation of Cotinine, a Major Metabolite of Nicotine in Rabbit Liver,” The Journal of Pharmacology and Experimental Therapeutics, vol. 129, no. 1, pp. 94-100, 1960.

[24] D. C. Pryde, D. Dalvie, Q. Hu, P. Jones, R. S. Obach and T.-D. Tran, “Aldehyde Oxidase: An Enzyme of Emerging Importance in Drug Discovery,” Journal of Medicinal Chemistry, vol. 53, no. 24, pp. 8441-8460, 2010.

[25] E. Garattini and M. Terao, “The Role of Aldehyde Oxidase in Drug Metabolism,” Expert Opinion on Drug Metabolism & Toxicology, vol. 8, no. 4, pp. 487-503, 2012.

[26] H. Li, H. Cui, T. K. Kundu, W. Alzawahra and J. L. Zweier, “Nitric Oxide Production from Nitrite Occurs Primarily in Tissues Not in the Blood,” The Journal of Biological Chemistry, vol. 283, no. 26, pp. 17855-17863, 2008.

[27] E. M. Paragas, S. C. Humphreys, J. Min, C. A. Joswig-Jones and J. P. Jones, “The Two Faces of Aldehyde Oxidase: Oxidative and Reductive Transformations of 5-nitroquinoline,” Biochemical Pharmacology, vol. 145, pp. 210-217, 2017.

[28] J. K. Sodhi, S. Wong, D. S. Kirkpatrick, L. Liu, S. C. Khojasteh, C. E. C. A. Hop, J. T. Barr, J. P. Jones and J. S. Halladay, “A Novel Reaction Mediated by Human Aldehyde Oxidase: Amide Hydrolysis of GDC-0834,” Drug Metabolism and Disposition, vol. 49, no. 5, pp. 908-915, 2015.

[29] E. Garattini, M. Fratelli and M. Terao, “Mammalian Aldehyde Oxidases: Genetics, Evolution and Biochemistry,” Cellular and Molecular Life Sciences, vol. 65, pp. 1019-1048, 2008.

[30] M. Öeren, P. J. Walton, P. A. Hunt, D. J. Ponting and M. D. Segall, “Predicting Reactivity to Drug Metabolism: Beyond P450s—Modelling FMOs and UGTs,” Journal of Computer-Aided Molecular Design, vol. 35, pp. 541-555, 2020.

[31] E. B. Skibo, J. H. Gilchrist and C. H. Lee, “Electronic Probes of the Mechanism of Substrate Oxidation by Buttermilk Xanthine Oxidase: Role of the Active-site Nucleophile in Oxidation,” Biochemistry, vol. 26, no. 11, pp. 3032-3037, 1987.

[32] J. F. Alfaro and J. P. Jones, “Studies on the Mechanism of Aldehyde Oxidase and Xanthine Oxidase,” The Journal of Organic Chemistry, vol. 73, no. 23, pp. 9469-9472, 2008.

[33] B. Kaye, J. L. Offerman, J. L. Reid, H. L. Elliot and W. S. Hillis, “A Species Difference in the Presystemic Metabolism of Carbazeran in Dog and Man,” Xenobiotica, vol. 14, no. 12, pp. 935-945, 1984.

[34] J. M. Hutzler, M. A. Cerny, Y.-S. Yang, C. Asher, D. Wong, K. Frederick and K. Gilpin, “Cynomolgus Monkey as a Surrogate for Human Aldehyde Oxidase Metabolism of the EGFR Inhibitor BIBX1382,” Drug Metabolism and Disposition, vol. 42, no. 10, pp. 1751-1760, 2014.

[35] M. P. Lolkema, H. H. Bohets, H.-T. Arkenau, A. Lampo, E. Barale, M. J. A. de Jonge, L. van Doorn, P. Hellemans, J. S. de Bono and F. A. L. M. Eskens, “The c-Met Tyrosine Kinase Inhibitor JNJ-38877605 Causes Renal Toxicity through Species-Specific Insoluble Metabolite Formation,” Clinical Cancer Research, vol. 21, no. 10, pp. 2297-2304, 2015.

[36] J. P. Jones and K. R. Korzekwa, “Predicting Intrinsic Clearance for Drugs and Drug Candidates Metabolized by Aldehyde Oxidase,” Molecular Pharmaceutics, vol. 10, no. 4, pp. 1262-1268, 2013.

[37] Y. Xu, L. Li, Y. Wang, J. Xing, L. Zhou, D. Zhong, X. Luo, H. Jiang, K. Chen, M. Zheng, P. Deng and X. Chen, “Aldehyde Oxidase Mediated Metabolism in Drug-like Molecules: A Combined Computational and Experimental Study,” Journal of Medicinal Chemistry, vol. 60, no. 7, pp. 2973-2982, 2017.

[38] P. F. Guengerich, “Cytochrome P450 and Chemical Toxicology,” Chemical Research in Toxicology, vol. 21, no. 1, pp. 70-83, 2008.

[39] F. P. Guengerich, “Cytochrome P450s and Other Enzymes in Drug Metabolism and Toxicity,” An Official Journal of the American Association of Pharmaceutical Scientists, vol. 8, pp. E101-E111, 2006.

[40] E. M. Isin and F. P. Guengerich, “Complex Reactions Catalyzed by Cytochrome P450 Enzymes,” Biochimica et Biophysica Acta (BBA) - General Subjects, vol. 1770, no. 3, pp. 314-329, 2007.

[41] M. J. Coon, “CYTOCHROME P450: Nature's Most Versatile Biological Catalyst,” Annual Review of Pharmacology and Toxicology, vol. 45, pp. 1-25, 2005.

[42] P. Manikandan and S. Nagini, “Cytochrome P450 Structure, Function and Clinical Significance: A Review,” Current Drug Targets, vol. 19, no. 1, pp. 38-54, 2018.

[43] C. Jung, “The Mystery of Cytochrome P450 Compound I: A Mini-review Dedicated to Klaus Ruckpaul,” Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics, vol. 1814, no. 1, pp. 46-57, 2011.

[44] A. W. Munro, H. M. Girvan, A. E. Mason, A. J. Dunford and K. J. McLean, “What makes a P450 tick?,” Trends in Biochemical Sciences, vol. 38, no. 3, pp. 140-150, 2013.

[45] S. G. Sligar, “Coupling of Spin, Substrate, and Redox Equilibriums in Cytochrome P450,” Biochemistry, vol. 15, no. 24, pp. 5399-5406, 1976.

[46] K. Ruckpaul and H. Rein, Basis and Mechanisms of Regulation of Cytochrome P-450, London: Taylor and Francis, 1989.

[47] L. Olsen, M. Montefiori, K. P. Tran and F. S. Jørgensen, “SMARTCyp 3.0: Enhanced Cytochrome P450 Site-of-metabolism Prediction Server,” Bioinformatics, vol. 35, no. 17, pp. 3174-3175, 2019.

[48] D. M. Ziegler and F. H. Pettit, “Formation of an Intermediate N-oxide in the Oxidative Demethylation of N,N-dimethylaniline Catalyzed by Liver Microsomes,” Biochemical and Biophysical Research Communications, vol. 15, no. 2, pp. 188-193, 1964.

[49] D. M. Ziegler and F. H. Pettit, “Microsomal Oxidases. I. The Isolation and Dialkylarylamine Oxygenase Activity of Pork Liver Microsomes,” Biochemistry, vol. 5, no. 9, pp. 2932-2938, 1966.

[50] L. L. Poulsen, R. M. Hyslop and D. M. Ziegler, “S-oxidation of Thioureylenes Catalyzed by a Microsomal Flavoprotein Mixed-function Oxidase,” Biochemical Pharmacology, vol. 23, no. 24, pp. 3431-3440, 1974.

[51] K. C. Jones and D. P. Ballou, “Reactions of the 4a-hydroperoxide of Liver Microsomal Flavin-containing Monooxygenase with Nucleophilic and Electrophilic Substrates,” Journal of Biological Chemistry, vol. 261, no. 6, pp. 2553-2559, 1986.

[52] G. P. Chen, L. L. Poulsen and D. M. Ziegler, “Oxidation of Aldehydes Catalyzed by Pig Liver Flavin-containing Monooxygenase,” Drug Metabolism and Disposition, vol. 23, no. 12, pp. 1390-1393, 1995.

[53] F. Fiorentini, M. Geier, C. Binda, M. Winkler, K. Faber, M. Hall and A. Mattevi, “Biocatalytic Characterization of Human FMO5: Unearthing Baeyer–Villiger Reactions in Humans,” ACS Chemical Biology, vol. 11, no. 4, pp. 1039-1048, 2016.

[54] D. M. Ziegler, L. L. Poulsen and M. W. Duffel, Kinetic Studies on Mechanism and Substrate Specificity of The Microsomal Flavin–Containing Monooxygenase, Academic Press, 1980, pp. 637-645.

[55] D. M. Ziegler, P. Graf, L. L. Poulsen, H. Sies and W. Stahl, “NADPH-dependent Oxidation of Reduced Ebselen, 2-selenylbenzanilide, and of 2-(methylseleno)benzanilide Catalyzed by Pig Liver Flavin-containing Monooxygenase,” Chemical Research in Toxicology, vol. 5, no. 2, pp. 163-166, 1992.

[56] V. Massey, “Activation of Molecular Oxygen by Flavins and Flavoproteins,” The Journal of Biological Chemistry, vol. 269, no. 36, pp. 22459-22462, 1994.

[57] D. C. Hao, S. L. Chen, J. Mu and P. G. Xiao, “Molecular Phylogeny, Long-term Evolution, and Functional Divergence of Flavin-containing Monooxygenases,” Genetica, vol. 137, pp. 173-187, 2009.

[58] D. H. Lang and A. E. Rettie, “In Vitro Evaluation of Potential in Vivo Probes for Human Flavin-containing Monooxygenase (FMO): Metabolism of Benzydamine and Caffeine by FMO and P450 Isoforms,” British Journal of Clinical Pharmacology, vol. 50, no. 4, pp. 311-314, 2002.

[59] T. Mushiroda, R. Douya, E. Takahara and O. Nagata, “The Involvement of Flavin-Containing Monooxygenase but Not CYP3A4 in Metabolism of Itopride Hydrochloride, a Gastroprokinetic Agent: Comparison with Cisapride and Mosapride Citrate,” Drug Metabolism and Disposition, vol. 28, no. 10, pp. 1231-1237, 2000.

[60] H. C. Rawden, G. O. Kokwaro, S. A. Ward and G. Edwards, “Relative Contribution of Cytochromes P-450 and Flavin-containing Monoxygenases to the Metabolism of Albendazole by Human Liver Microsomes,” British Journal of Clinical Pharmacology, vol. 49, no. 4, pp. 313-322, 2001.

[61] J. R. Cashman, S. B. Park, Z. C. Yang, C. B. Washington, D. Y. Gomez, K. M. Giacomini and C. M. Brett, “Chemical, Enzymatic, and Human Enantioselective S-oxygenation of Cimetidine,” Drug Metabolism and Disposition, vol. 21, no. 4, pp. 587-597, 1993.

[62] S. Rendic and F. P. Guengerich, “Survey of Human Oxidoreductases and Cytochrome P450 Enzymes Involved in the Metabolism of Xenobiotic and Natural Chemicals,” Chemical Research in Toxicology, vol. 28, no. 1, pp. 38-42, 2014.

[63] J. R. Cashman and R. P. Hanzlik, “Microsomal Oxidation of Thiobenzamide. A Photometric Assay for the Flavin-containing Monooxygenase,” Biochemical and Biophysical Research Communications, vol. 98, no. 1, pp. 147-153, 1981.

[64] M. C. Dyroff and R. A. Neal, “Studies of the Mechanism of Metabolism of Thioacetamide S-oxide by Rat Liver Microsomes,” Molecular Pharmacology, vol. 23, no. 1, pp. 219-227, 1983.

[65] E. Chieli and G. Malvaldi, “Role of the Microsomal FAD-containing Monooxygenase in the Liver Toxicity of Thioacetamide S-oxide,” Toxicology, vol. 31, no. 1, pp. 41-52, 1984.

[66] M. J. Ruse and R. H. Waring, “The Effect of Methimazole on Thioamide Bioactivation and Toxicity,” Toxicology Letters, vol. 58, no. 1, pp. 37-41, 1991.

[67] J. W. Lee, K. D. Shin, M. Lee, E. J. Kim, S.-S. Han, M. Y. Han, H. Ha, T. C. Jeong and W. S. Koh, “Role of Metabolism by Flavin-containing Monooxygenase in Thioacetamide-induced Immunosuppression,” Toxicology Letters, vol. 136, no. 3, pp. 163-172, 2003.

[68] G. Ottolina, G. Gonzalo and G. Carrea, “Theoretical Studies of Oxygen Atom Transfer From Flavin to Electron-rich Substrates,” Journal of Molecular Structure: THEOCHEM, vol. 757, no. 1, pp. 175-181, 2005.

[69] R. Bach, “Role of the Somersault Rearrangement in the Oxidation Step for Flavin Monooxygenases (FMO). A Comparison between FMO and Conventional Xenobiotic Oxidation with Hydroperoxides,” The Journal of Physical Chemistry A, vol. 115, no. 40, pp. 11087-11100, 2011.

[70] E. M. Hawes, “N +-Glucuronidation, a Common Pathway in Human Metabolism of Drugs With a Tertiary Amine Group,” Drug Metabolism and Disposition, vol. 26, no. 9, pp. 830-837, 1996.

[71] “Identification of Human UDP-glucuronosyltransferase Isoform(s) Responsible for the C-glucuronidation of Phenylbutazone,” Archives of Biochemistry and Biophysics, vol. 454, no. 1, pp. 72-79, 2006.

[72] “S-Glucuronidation of 7-mercapto-4-methylcoumarin by Human UDP Glycosyltransferases in Genetically Engineered Fission Yeast Cells,” Biological Chemistry, vol. 329, no. 12, pp. 1089-1095, 2011.

[73] L. L. Lairson, B. Henrissat, G. J. Davies and S. G. Withers, “Glycosyltransferases: Structures, Functions, and Mechanisms,” Annual Review of Biochemistry, vol. 77, pp. 521-555, 2008.

[74] “Glycosyltransferases: Mechanisms and Applications in Natural Product Development,” Chemical Society Reviews, vol. 44, no. 22, pp. 8350-8374, 2015.

[75] A. Radominska-Pandya, P. J. Czernik, J. M. Little, E. Battaglia and P. I. MacKenzie, “Structural and Functional Studies Of UDP-glucuronosyltransferases,” Drug Metabolism Reviews, vol. 31, no. 4, pp. 817-899, 1999.

[76] M. Ouzzine, L. Antonio, B. Burchell, P. Netter, S. Fournel-Gigleux and J. Magdalou, “Importance of Histidine Residues for the Function of the Human Liver UDP-Glucuronosyltransferase UGT1A6: Evidence for the Catalytic Role of Histidine 370,” Molecular Pharmacology, vol. 58, no. 6, pp. 1609-1615, 2000.

[77] C. W. Locuson and T. S. Tracy, “Comparative Modelling of the Human UDP-glucuronosyltransferases: Insights into Structure and Mechanism,” Xenobiotica, vol. 37, no. 2, pp. 155-168, 2007.

[78] D. Li, S. Fournel-Gigleux, L. Barré, G. Mulliert, P. Netter, J. Magdalou and M. Ouzzine, “Identification of Aspartic Acid and Histidine Residues Mediating the Reaction Mechanism and the Substrate Specificity of the Human UDP-glucuronosyltransferases 1A,” Journal of Biological Chemistry, vol. 282, no. 50, pp. 36514-36524, 2007.

[79] P. I. Mackenzie, I. S. Owens, B. Burchell, K. W. Bock, A. Bairoch, A. Bélanger, S. Fournel-Gigleux, M. Green, D. W. Hum, T. Iyanagi, D. Lancet, P. Louisot, J. Magdalou, J. R. Chowdhury, J. K. Ritter, H. Schachter, T. R. Tephly, K. F. Tipton and D. W. Nebert, “Nomenclature Update for the Mammalian UDP Glycosyltransferase (UGT) Gene Superfamily,” Pharmacogenetics and Genomics, vol. 15, no. 10, pp. 677-685, 2005.

[80] R. Meech, D. G. Hu, R. A. McKinnon, S. N. Mubarokah, A. Z. Haines, P. C. Nair, A. Rowland and P. I. Mackenzie, “The UDP-Glycosyltransferase (UGT) Superfamily: New Members, New Functions, and Novel Paradigms,” Physiological Reviews, vol. 99, no. 2, pp. 1153-1222, 2019.

[81] P. C. Nair, R. Meech, P. I. Mackenzie, R. A. McKinnon and J. O. Miners, “Insights into the UDP-sugar Selectivities of Human UDP-glycosyltransferases (UGT): a Molecular Modeling Perspective,” Drug Metabolism Reviews, vol. 47, no. 3, pp. 335-345, 2015.

[82] K. W. Bock, “The UDP-glycosyltransferase (UGT) Superfamily Expressed in Humans, Insects and Plants: Animal-plant Arms-race and Co-evolution,” Biochemical Pharmacology, vol. 99, pp. 11-17, 2016.

[83] R. H. Tukey and C. P. Strassburg, “Human UDP-Glucuronosyltransferases: Metabolism, Expression, and Disease,” Annual Review of Pharmacology and Toxicology, vol. 40, pp. 581-616, 2000.

[84] S. R. Kazmi, R. Jun, M.-S. Yu, C. Jung and D. Na, “In Silico Approaches and Tools for the Prediction of Drug Metabolism and Fate: A Review,” Computers in Biology and Medicine, vol. 106, pp. 54-64, 2019.

[85] K. A. Feenstra, C. De Graaf and N. P. E. Vermeulen, Drug-drug interactions, Informa Healthcare, 2008.

[86] P. C. Nair, R. A. McKinnon and J. O. Miners, “Cytochrome P450 Structure-function: Insights from Molecular Dynamics Simulations,” Drug Metabolism Reviews, vol. 48, no. 3, pp. 434-452, 2016.

[87] S. Panneerselvam , D. Yesudhas, P. Durai , M. A. Anwar , V. Gosu and S. Choi, “A Combined Molecular Docking/Dynamics Approach to Probe the Binding Mode of Cancer Drugs with Cytochrome P450 3A4,” Molecules, vol. 20, no. 8, pp. 14915-14935, 2015.

[88] K. D. Dubey, B. Wang and S. Shaik, “Molecular Dynamics and QM/MM Calculations Predict the Substrate-Induced Gating of Cytochrome P450 BM3 and the Regio- and Stereoselectivity of Fatty Acid Hydroxylation,” Journal of the American Chemical Society, vol. 138, no. 3, pp. 837-845, 2016.

[89] P. Rydberg, D. E. Gloriam and L. Olsen, “The SMARTCyp Cytochrome P450 Metabolism Prediction Server,” Bioinformatics, vol. 26, no. 23, pp. 2988-2989, 2010.

[90] M. Šícho, C. de Bruyn Kops, C. Stork, D. Svozil and J. Kirchmair, “FAME 2: Simple and Effective Machine Learning Model of Cytochrome P450 Regioselectivity,” Journal of Chemical Information and Modeling, vol. 57, no. 8, pp. 1832-1846, 2017.

[91] K. J. Laidler , “The Development of the Arrhenius Equation,” Journal of Chemical Education, vol. 61, no. 6, p. 494, 1984.

[92] P. A. Hunt, D. M. Segall and J. D. Tyzack , “WhichP450: a Multi-class Categorical Model to Predict the Major Metabolising CYP450 Isoform for a Compound,” Journal of Computer-Aided Molecular Design, vol. 32, pp. 537-546, 2018.

[93] M. Rostkowski, O. Spjuth and P. Rydberg, “WhichCyp: Prediction of Cytochromes P450 Inhibition,” Bioinformatics, vol. 29, no. 16, pp. 2051-2052, 2013.

[94] M. J. S. Dewar, E. G. Zoebisch, E. F. Healy and J. J. P. Stewart, “Development and Use of Quantum Mechanical Molecular Models. 76. AM1: a New General Purpose Quantum Mechanical Molecular Model,” Journal of the American Chemical Society, vol. 107, no. 13, pp. 3902-3909, 1985.

[95] J. J. P. Stewart, “Optimization of parameters for semiempirical methods V: Modification of NDDO approximations and application to 70 elements,” Journal of Molecular Modeling, vol. 13, pp. 1173-1213, 2007.

[96] G. Marcou and D. Rognan, “Optimizing Fragment and Scaffold Docking by Use of Molecular Interaction Fingerprints,” Journal of Chemical Information and Modeling, vol. 47, no. 1, pp. 195-207, 2007.

[97] M. Stahl and H. Mauser, “Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods,” Database Clustering with a Combination of Fingerprint and Maximum Common Substructure Methods, vol. 45, no. 3, pp. 542-548, 2005.

[98] J. Hutter, M. Iannuzzi, F. Schiffmann and J. VandeVondele, “CP2K: Atomistic Simulations of Condensed Matter Systems,” WIREs Computational Molecular Science, vol. 4, no. 1, pp. 15-25, 2013.

[99] S. H. Vosko, L. Wilk and M. Nusair, “Accurate Spin-dependent Electron Liquid Correlation Energies for Local Spin Density Calculations: a Critical Analysis,” Canadian Journal of Physics, vol. 58, no. 8, pp. 80-159, 1980.

[100] C. Lee, W. Yang and R. G. Parr, “Development of the Colle-Salvetti Correlation-energy Formula Into a Functional of the Electron Density,” Physical Review B, vol. 37, no. 2, p. 785, 1988.

[101] "Density-functional Thermochemistry. III. The Eole of Exact Exchange," The Journal of Chemical Physics, vol. 98, no. 7, p. 5648, 1993.

[102] P. J. Stephens, F. J. Devlin, C. F. Chabalowski and M. J. Frisch, "Ab Initio Calculation of Vibrational Absorption and Circular Dichroism Spectra Using Density Functional Force Fields," The Journal of Physical Chemistry, vol. 98, no. 45, pp. 11623-11627, 1994.

[103] S. Ehrlich, J. Moellmann, W. Reckien, T. Bredow and S. Grimme, "System-Dependent Dispersion Coefficients for the DFT-D3 Treatment of Adsorption Processes on Ionic Surfaces," ChemPhysChem, vol. 12, no. 17, pp. 3414-3420, 2011.

[104] F. Weigend and R. Ahlrichs, "Balanced Basis Sets of Split Valence, Triple Zeta Valence and Quadruple Zeta Valence Quality for H to Rn: Design and Assessment of Accuracy," Physical Chemistry Chemical Physics, vol. 7, pp. 3297-3305, 2005.

[105] D. Andrae, U. Häußermann, M. Dolg, H. Stoll and H. Preuß, "Energy-adjusted Ab Initio Pseudopotentials for the Second and Third Row Transition Elements," Theoretica Chimica Acta, vol. 77, pp. 123-141, 1990.

[106] B. P. Pritchard, D. Altarawy, B. Didier, T. D. Gibson and T. L. Windus, "A New Basis Set Exchange: An Open, Up-to-date Resource for the Molecular Sciences Community," Journal of Chemical Information and Modeling, vol. 59, no. 11, pp. 4814-4820, 2019.

[107] L. Simón and J. M. Goodman, "How Reliable are DFT Transition Structures? Comparison of GGA, Hybrid-meta-GGA and Meta-GGA Functionals," Organic & Biomolecular Chemistry, vol. 9, pp. 689-700, 2011.

[108] M. Valiev, E. J. Bylaska, N. Govind, K. Kowalski, T. P. Straatsma, H. J. J. Van Dam, D. Wang, J. Nieplocha, E. Apra, T. L. Windus and W. A. de Jong, "NWChem: A Comprehensive and Scalable Open-source Solution for Large Scale Molecular Simulations," Computer Physics Communications, vol. 181, no. 9, pp. 1477-1489, 2010.

[109] O. Obrezanova, G. Csányi, J. M. R. Gola and M. D. Segall, "Gaussian Processes: A Method for Automatic QSAR Modeling of ADME Properties," Journal of Chemical Information and Modeling, vol. 47, no. 5, pp. 1847-1857, 2007.

[110] D. Ruppert and P. Wand, "Multivariate Locally Weighted Least Squares Regression," The Annals of Statistics, vol. 22, no. 3, pp. 1346-1370, 1994.

[111] L. van der Maaten and G. Hinton, "Visualizing Data using t-SNE," Journal of Machine Learning Research, vol. 9, no. 86, p. 2579–2605, 2008.