Predicting Regioselectivity of Cytosolic SULT Metabolism for Drugs

Mario Öeren,[†] Sylvia C. Kaempf, [†,‡] 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 St Andrews, North Haugh, St Andrews, KY16 9ST, UK

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

Abstract

Cytosolic sulfotransferases (SULTs) are a family of enzymes responsible for the sulfation of small endogenous and exogenous compounds. SULTs contribute to the conjugation phase of metabolism and share substrates with the Uridine 5'-diphospho-glucuronosyltransferase (UGT) family of enzymes. UGTs are considered to be the most important enzymes in the conjugation phase, and SULTs are an auxiliary enzyme system to them. Understanding how the regioselectivity of SULTs differs from that of UGTs is essential from the perspective of developing novel drug candidates. We present a general ligand-based SULT model trained and tested using high-quality experimental regioselectivity data. The current study suggests that, unlike other metabolic enzymes in the modification and conjugation phases, the SULT regioselectivity is not strongly influenced by the activation energy of the rate-limiting step of the catalysis. Instead, the prominent role is played by the substrate binding site of SULT. Thus, the model is trained only on steric and orientation descriptors, which mimic the binding pocket of SULT. The resulting classification model, which predicts whether a site is metabolized, achieved a Cohen's kappa of 0.65.

Keywords

Accessibility, DFT, Metabolism, ML, Reactivity, Semi-empirical, Sulfotransferase, SULT

Introduction

The general aim of the conjugation phase of metabolism is to increase the solubility of small molecules to prepare them for excretion. The solubility of a compound is increased by linking a polar moiety, such as glucuronic or sulfonic acid, from a cofactor to a potential site of metabolism (SoM). The conjugation phase is also called phase II metabolism, as it generally follows the modification phase (phase I), in which functional groups are introduced by enzymes such as Cytochrome P450s that are SoMs for the conjugation reactions [1, 2]. However, occasionally, the modification phase is omitted and a parent drug may be directly conjugated. [3]

Studying the modification and conjugation phases enables us to build models that predict the substrates of various metabolic pathways, which is valuable for developing drugs, agrochemicals, nutritional supplements, and cosmetics. The predictions from such models can help to optimize of the structures of new chemical entities and identify toxic metabolites early in a project, making the process more cost-effective. [4, 5] The prevalent enzyme families in the modification phase have been thoroughly studied, and several successful predictive models have been built for Aldehyde Oxidases [6, 7], Cytochrome P450s [5, 8-12], and Flavin-containing Monooxygenases (FMO) [7]. The most important enzyme family in the conjugation phase, for which several successful models have also been built, is Uridine 5'-diphospho-glucuronosyltransferase (UGT) [7, 13-16]. The present study concentrates on the cytosolic sulfotransferase (SULT) family of enzymes, which are considered to be the second most important enzyme family in the conjugation phase and share the substrate space with UGTs. [17] The general aim of this study is to expand the range of models for enzyme families in the conjugation phase by training a model which predicts the regioselectivity of SULT metabolism.

A second class of sulfotransferases, the membrane-associated sulfotransferases, which are known to sulfonate larger biomolecules, have not been considered in this study. [18, 19]

Sulfotransferases

In 1876, Baumann isolated phenyl sulfate from the urine of a patient who had been administered phenol. [20, 21] Since this discovery, sulfation has been proven to be an essential conjugation reaction for endo- and xenobiotics. Sulfation is catalyzed by SULTs, which are found in many organs, including the liver, where many xenobiotics are metabolized. The first human isoforms isolated were SULT2A1 by Otterness *et al.*, [22] and SULT1A1 by Wilborn *et al.*, [23] in 1992 and 1993, respectively. To date, there are at least 15 known human SULT isoforms: 1A1, 1A2, 1A3, 1A4, 1B1, 1C4, 1C2, 1C3a, 1C3d, 1E1, 4A1, 2A1, 2B1a, 2B1b, and 6B1. [24] Furthermore, four pseudogenes 1D1P, 1D2P, 2A1P, and 3A1P are also known. [25]

1A1 is generally recognized to have the greatest contribution to SULT mediated metabolism because it has a wide range of potential substrates; thus, it is the isoform of most interest. [26] Other isoforms prevalent in liver, kidneys, small intestine and lungs include 1A3, 1B1, 1E1, and 2A1. [27] Interestingly, 1A3, which shares 93% sequence identity with 1A1, [19, 18] is not as promiscuous as 1A1 and has a general preference towards bioamines such as dopamine. [19] 1B1 appears to have a major physiological role in thyroid hormone metabolism, [28] 1E1 has a high affinity for the sulfation of endogenous estrogens, and 2A1 is responsible for most bile acid sulfation. [26] The other SULT isoforms are not of as much interest as these isoforms do not play an important role in metabolizing xenobiotics within the aforementioned organs.

The Mechanism of Sulfation

Sulfation by SULTs transfers a sulfonic acid from a cofactor, 3'-phosphoadenosine-5'-phosphosulfate (PAPS), to the respective substrate, as depicted in Figure 1. The most commonly sulfated groups are phenols and alcohols. [19] Other functional groups, such as *N*-oxides, are known to be metabolized (*e.g.* minoxidil); [26] however, the experimental data for such groups is sparse, and they are not included in the following work. The substrate specificity of SULTs overlaps with UGT and where glucuronidation occurs, sulfation is usually a lesser metabolic pathway. Nonetheless, there are molecules which are primarily metabolized by SULTs (*e.g.* ethinyl Estradiol [29]). As with other metabolism enzymes, the amino acid sequence of the catalytic core is conserved in all SULT isoforms; thus, the reaction mechanism is consistant for all isoforms. [5, 18, 30]

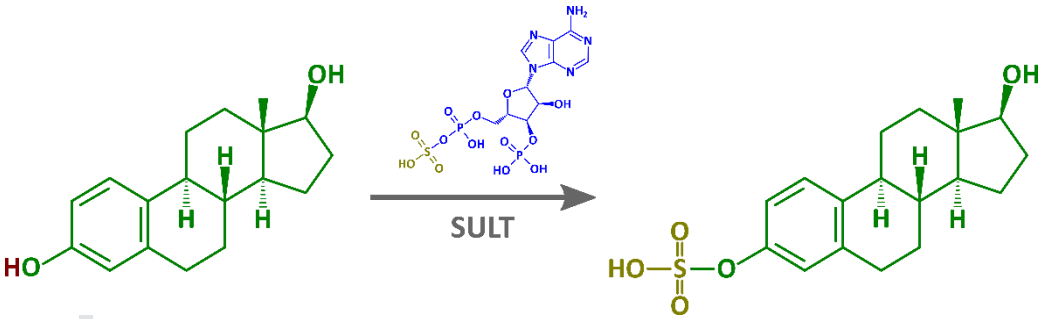

Figure 1. Sulfation of Estradiol, [31] a known SULT substrate. The cofactor PAPS is depicted on the reaction arrow.

The general catalytic cycle of sulfation follows a mechanism where the cofactor, PAPS, binds to the enzyme before the substrate. The sulfonic acid is transferred to the substrate once the ternary enzyme complex is formed. The products are released in an orderly fashion (metabolite first), and the catalytic cycle can start again. [18] The sulfation itself is thought to proceed by an in-line attack of the nucleophile at the sulfate group of PAPS; however, according to the DFT calculations by Bartolotti *et al.*, [32] it does not happen through the classical trigonal bipyramidal S_N2 transition state suggested by various experimental studies. [32-35] The study by Bartolotti *et al.* utilized the crystal structure of SULT1E1 (the SULT responsible for the sulfation of Estrogen) to model the transition state of the sulfation reaction and concluded that the lysine residues 48 and 106 act as possible charge stabilizers and also as possible sources of protons, while histidine 108 acts as a proton acceptor. The mechanism found is considered largely dissociative as it happens through two transition states. The first step is the the dissociation of the bond between the leaving group (3'-phosphoadenosine-5'-phosphate) and the sulfonic acid, during which a proton is carried over to the leaving group from either of the lysine residues. According to various studies, this step could, in theory, happen without the substrate and the SO_3 would be stabilized within the

protein. [32, 35, 36] The second transition state (Figure 2) involves the formation of the product, where the proton of the hydroxy group of the substrate is transferred to the histidine residue 108. The reaction ends with the protonation of the sulfate entity of the product. [32, 35]

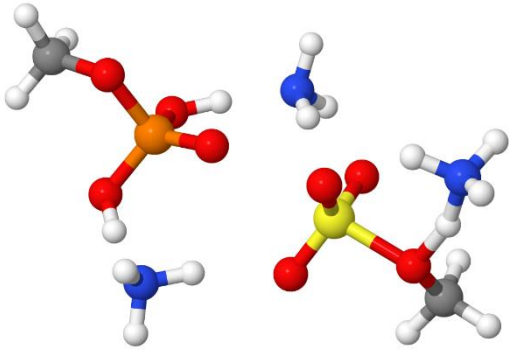

Figure 2. The formation of the product; reproduced from the work by Bartolotti *et al*. [32] The ammonia molecule accepting the proton is the aforementioned histidine 108 and remaining ammonia molecules are lysine residues 48 and 106.

Modelling Drug Metabolism

The modelling method we have used herein for the regioselectivity models is the reactivity-accessibility approach, which falls under the category of ligand-based models. This approach relies on experimental knowledge of molecules metabolized by the biological target of interest as opposed to structure-based methods, which use three-dimensional knowledge of the enzyme. The reactivity-accessibility approach divides the substrate-protein interaction into two conceptual parts; the first part, reactivity, describes the reactivity of a potential site of metabolism (SoM) to the reaction mechanism of the enzyme. In the current work we planned to use the activation energy value (E_a) of the rate-limiting step of the catalytic cycle to represent the reactivity. The E_a is related to the reaction rate, and previous studies of other enzymes have shown that experimentally measured reaction rates of metabolism correlate well with the calculated E_a values of the rate-limiting step. [5, 30] Since the E_a is often calculated using a simplified reaction mechanism, it does not consider the effects of the protein structure surrounding the reaction center; thus, additional descriptors are also required to describe the potential interactions between the substrate and the enzyme, and their impact on the accessibility of the SoM to the reaction center. In the current work, a set of two-dimensional structural descriptors are used to represent steric and orientation effects. [5]

The E_a is obtained by using fundamental and transferable quantum mechanical methods. While the structure of the cofactor is often simplified to decrease the timeframe of the calculations, the structure of the substrate should not be fragmented because the long-range electronic effects within the substrate can play an important role in estimating the E_a, particularly in conjugated systems. [5, 7, 30]

The accessibility descriptors are used to correct the E_a due to steric and orientation effects, which arise when the potential substrate enters the enzyme's active site. In the case of SULTs, the descriptors are used to describe the substrate accessibility to the sulfonate group of PAPS. The orientation of the substrate in the binding pocket of the protein depends on the interaction between the functional groups of the substrate and the residues of the protein and must be considered when predicting the observed SoMs of SULTs. Similarly, the steric descriptors capture the effects of steric hinderance within the substrate (e.g. a potential SoM is protected by bulky groups). These steric and orientation factors are described using two-dimensional, site-specific structural descriptors which capture where the potential SoM is located, in relation to the relevant functional groups of the substrate. [5, 7]

As mentioned above, the reaction mechanism for sulfation is conserved across the SULT isoforms, making the reactivity descriptor, the site-specific E_a, applicable to all SULT isoforms. The steric and orientation effects of individual isoforms would be captured by the two-dimensional structural descriptors. Thus, combining the reactivity and accessibility descriptors with isoform-specific experimental data allows us to train predictive Quantitative Structure-Activity Relationship (QSAR) models for every SULT isoform with enough publicly available experimental data. [5, 7, 12, 37]

The Aim of the Study

The initial aim of the study was to build a reactivity-accessibility model that predicts the SoM for substrates of SULT isoforms found in humans. First, a detailed understanding of the rate-limiting step of the SULT catalytic mechanism is necessary to assess the reactivity of each potential SoM. Such a study was conducted using density functional theory (DFT). It aimed to determine the simplified mechanism of the product formation step for SULT and to validate the mechanism identified by comparing the calculated E_a values with experimentally measured reaction rates and observed SoMs. Having determined the reaction mechanism using DFT simulations, semi-empirical quantum mechanical methods would typically be used to calculate E_a values within a reasonable time frame for use in a QSAR model; semi-empirical results used in this way are validated using DFT. [7, 30] However, as discussed below, our study, suggested that the SULT metabolism depends mostly on the accessibility of the SoM. Thus, the final predictive model was trained using only the accessibility descriptors using high-quality experimental regioselective SULT data.

Theory and Implementation

Computational Methods

Initial three-dimensional structures for all substrates in this work were generated with either Avogadro [38] or from SMILES by using OEChem by OpenEye. [39, 40] The transition state structures for DFT were generated manually. Both structure pre-optimization and potential energy surface scanning were done using the semi-empirical method AM1 [41] using the program package CP2K [42].

DFT calculations were run using the functional B3LYP-D [43-47] along with the def2-SVP [48] basis set. B3LYP was chosen because the presented reaction mechanisms calculated by hybrid GGA functionals yield similar results to the more expensive hybrid meta-GGA functionals. [7, 49] Geometry optimizations were performed on the initial structures, followed by frequency calculations to verify the local minima or the transition states. The DFT calculations were performed with the NWChem 6.8 program package [50]. The weak interactions between fragments within a transition state were calculated using Multiwfn. [51]

Machine Learning

As mentioned before, the descriptors used within the current work are E_a values describing the reactivity of a potential SoM and site-specific atom-pair descriptors for describing steric and orientation effects. The E_a values were calculated in a similar fashion to our previous work and are described below in detail for the SULT reaction mechanism. [5, 7, 30]. The semi-empirical calculations to obtain the E_a values are described in the results. The atom-pair fingerprints for potential SoMs were obtained using RDKit (https://www.rdkit.org) – the Python scripts used to calculate the fingerprints can be found from the Supporting Information.

The data obtained were split into a training and a test set in an approximate ratio of 80:20, respectively. The split was based on molecules; thus, all potential SoMs of each substrate were either in the training or the test set. The compounds for the training and test sets were chosen randomly, but a visual check was performed to confirm the chemical space (without inspecting the individual structures) of the training set was roughly covered by the compounds in the test set. The random forests (RF) method [52] from the Auto-Modeller module in StarDrop was used to train the model (default parameters were used).

The Cohen's kappa (kappa or κ) statistic is used to report the reliability of the RF classification models. The kappa value is a more robust measure than the percentage agreement since it takes into account the possibility of agreement occurring by chance, although percentages are also reported in the current work. Furthermore, the confusion matrix for the SULT model perdictions is provided. The rules of how kappa values were evaluated are shown in Table 1.

Table 1 Approximate ranges for evaluating kappa values.

κ value	Agreement value
$\kappa < 0.5$	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 most rigorous test of a model to avoid overtraining is to asses it's predictive accuracy on an external test set. However, the data set within this work is relatively moderate in size; thus, an external test set could not be used. In these cases, additional tests such as *y*-scrambling were applied. Such tests are necessary because each data point has hundreds of descriptors which might correlate for no reason. *Y*-scrambling is a simple test to explore the predictive power of pure chance in which the values of the experimental data (the values to be predicted) were shuffled while the descriptor values were left intact. The scrambled data were then used to train a QSAR model. The Cohen's kappa value for a model trained on the scrambled data should be close to zero, which corresponds to the degree of agreement that would be expected by chance. If a substantially higher Cohen's kappa is found, this would indicate that the model building and validation process is subject to overtraining. The Python scripts used to calculate the *y*-scrambling results can be found from the Supporting Information.

Experimental Data

The data used herein was curated from sources that provide detailed information on the experimentally observed SoMs for substrates of human SULTs. The dataset was initially curated by Lhasa Limited [53] to support Meteor Nexus, [54, 55] but has been expanded for this work. Since the model is intended to distinguish the experimentally observed SoMs from all potential SoMs, the molecules included in the dataset have two or more potential SoMs, out of which at least one is experimentally observed. Any experiments run with unphysiological substrate concentrations were rejected; the highest 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 an experimental confirmation (e.g. using mass-spectrometry or NMR studies); we did not include metabolites based only on expert opinions. Initially, the aim was to obtain site-specific data for individual isoforms; however, it was found that isoform-specific information was very limited within the publicly available sources. Therefore, the data were used to train a single, general, model of SULT metabolism covering multiple isoforms. Such a general model is inferior to the isoform-specific models, because of the variation introduced by combining different regioselectivities for multiple isoforms, but our previous research has demonstrated that a general model can have high predictive power. [7] The literature search yielded 75 compounds with 195 potential SoMs, out of which 84 were observed to be metabolized. The full dataset with relevant references for the compounds can be found in the supporting information.

Results and Discussion

Reaction Mechanism of SULTs

Since DFT calculations are computationally expensive compared to methods routinely applied to drug discovery and metabolism prediction, the system used to estimate the E_a (the reactivity) must be as small as possible but retain its chemical characteristics. In previous work by Bartolotti *et al.*, the adenosine phosphate group of PAPS was reduced to a methyl group and the interacting lysine and histidine residues were replaced with ammonia molecules, as shown in Figure 2. [32] However, as mentioned before, the reaction mechanism is dissociative and features two transition states: the dissociation of the bond between the cofactor and the SO_3 and the sulfation of the substrate. Between the two transition states is a minimum, where the SO_3 is trigonal planar – experiments have shown an "inactive" crystal structure which has been observed in SULT1E1 by replacing the leaving group (SO_3) with vanadate. [32 35] Furthermore, the dissociation can happen without the substrate; [36] thus, it can be presumed that the second step, sulfation of the substrate, is the rate-limiting step. The weak interactions between the lysine residues and the SO_3 in the intermediate structure can be considered a constant (irrespective of the substrate). Thus, the system to estimate the E_a can be further simplified by removing the leaving group

and the lysine residues. The ammonia in Figure 2, which represents the histidine 108 that accepts the proton from the substrate, has to be changed since it is susceptible to spurious hydrogen bonds with larger substrates; therefore, a simplified model for histidine (the imidazole ring alone, without the peptide backbone) will be used. [30] The resulting system includes the substrate, the SO_3, and the simplified histidine, as illustrated in Figure 3. We note that we explored alternatives to the reaction mechanism proposed by Bartolotti et al., [32] because other sources have proposed that the reaction mechanism could be S_N2. [18, 33] However, after exhaustive investigations (not included within this work) using the complete PAPS structure and a variety of simplified versions, we did not identify any structures which would converge into a valid transition state. Thus, we continued with the reaction system in Figure 3 (an example with Propofol).

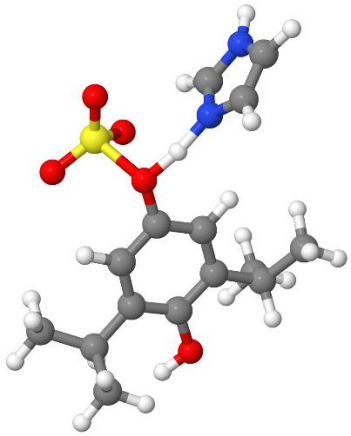

Figure 3. The simplified transition state for sulfation. The substrate shown is Propofol. [56]

In an ideal case, the preliminary study for testing the chosen simplification includes calculating the E_a for small and similar substrates for which the maximum reaction rate (V_{max}) has been measured within the same isoform. Such a scenario allows us to assess the correlation between the E_a and V_{max}, while minimizing the effects arising from variations in the the protein environment (steric and orientation effects). If V_{max} data are not available for the chosen compounds, the E_a value can also be compared to the order of observed SoMs – does the primary SoM have the lowest E_a *etc.*? [7, 30] However, for SULTs, the V_{max} values are often not reported, and the isoform responsible for metabolism is not known. Therefore, we will compare the order of the observed SoMs to the order of the calculated E_a values. In the following example, we chose five compounds (Figure 4) for a preliminary study using DFT to confirm that the SoMs with the lowest E_a values correspond to the experimentally observed SoMs, and that the system is not biased towards alcohols or phenols.

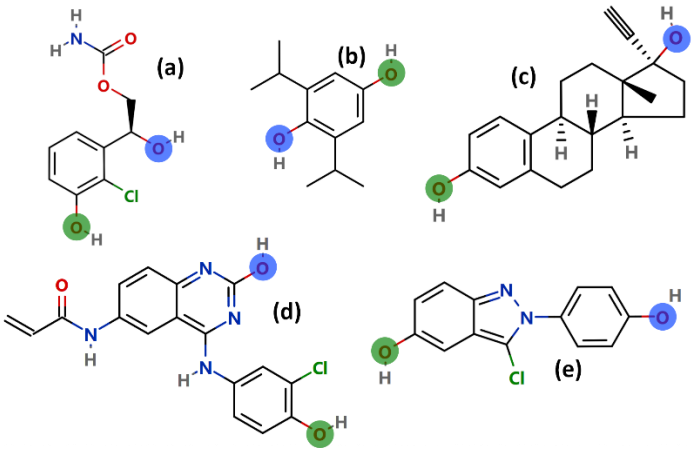

Figure 4. The SULT substrates used to study if E_a can be used to predict the experimentally observed SoMs. The molecules are (a) metabolite of Carisbamate, [57] (b) Propofol, [56] (c) Ethinyl Estradiol, [58] (d) Allitinib, [59] and (e) Indazole-Cl [60] The green circles represent experimentally observed sites of metabolism and the blue circles represent potential SoMs which are not observed in practice.

The results of the DFT study are shown in Table 2. These preliminary results do not show the desired clear relationship between the E_a and the experimentally determined SoMs metabolism (unlike the results in our previous studies of other drug-metabolizing enzymes [30]). In cases where the E_a agrees with the experimental

SoMs, the differences between the E_a values for potential SoM within a molecule were too small to distinguish between the sites clearly. While the spurious long-range interactions between the substrates and the SO_3 could account for some of this lack of correlation, taking weak interactions into account did not improve the relationship with the compounds in the preliminary study. Hence the study was expanded to all the available substrates.

Table 2. The DFT study results; green indicates when the calculated E_a agrees with the experimental results and blue is used when it does not.

Molecule	Experimental Result	E_a (kJ mol^{-1})	E_a without weak interactions (kJ mol^{-1})
Carisbamate	Metabolized	58.9	128.4
	Not Metabolized	36.1	106.0
Propofol	Metabolized	39.4	73.9
	Not Metabolized	28.2	76.9
Ethinyl	Metabolized	49.3	78.4
	Not Metabolized	66.8	124.3
Allitinib	Metabolized	51.0	125.7
	Not Metabolized	51.9	114.8
Indazole-Cl	Metabolized	58.6	66.7
	Not Metabolized	59.4	77.8

To enable all 75 compounds (199 potential SoMs) to be examined in a reasonable timeframe, we calculated the E_a using the semi-empirical AM1 method. However, to ensure the validity of this approach, we first confirmed that the DFT E_a values correlated with the ones calculated using AM1 for a set of alcohols and phenols. This was done by calculating the E_a for a set of compounds with both the DFT and the AM1 semi-empirical method and assessing the correlation between the two. The compounds do not have to be SULT substrates, but they should be small to avoid extraneous inter- and intramolecular interactions, and they should represent various potential environments for alcohols and phenols. We chose ethanol, trichloroethanol, phenylmethanol, 4-chlorocyclohexyl methanol, and 1,2-dihydro-4-pyridinylmethanol to represent alcohols and phenol, 4-chlorophenol, 4-chloro-1-naphthol, 2,6-dimethylphenol, and 1-naphthol to represent phenols.

The correlation results were excellent with an R^2 of 0.97. However, contrary to the E_a values for the potential SoMs for UGT, [7] the E_a values for alcohols are lower than the E_a values for phenols. Furthermore, the spread of the SULT E_a values for the simplified model is very small (20 – 55 kJ mol^{-1}) compared to the spread of the UGT E_a values for the simplified model (155 – 375 kJ mol^{-1}) [7].

While the correlation results were excellent, it can be seen in Figure 5 that the data points are not close to the identity line, and the slope of the trendline is less than one. In many cases, the semi-empirical methods are subject to systematic errors due to the approximations they make to the Hamiltonian. Therefore, for the semi-empirical methods to be used confidently, corrections to account for these systematic errors are calculated by correlating the E_a values obtained with semi-empirical methods to the E_a values obtained with DFT. The correction factor is found by using the best fit line equation.

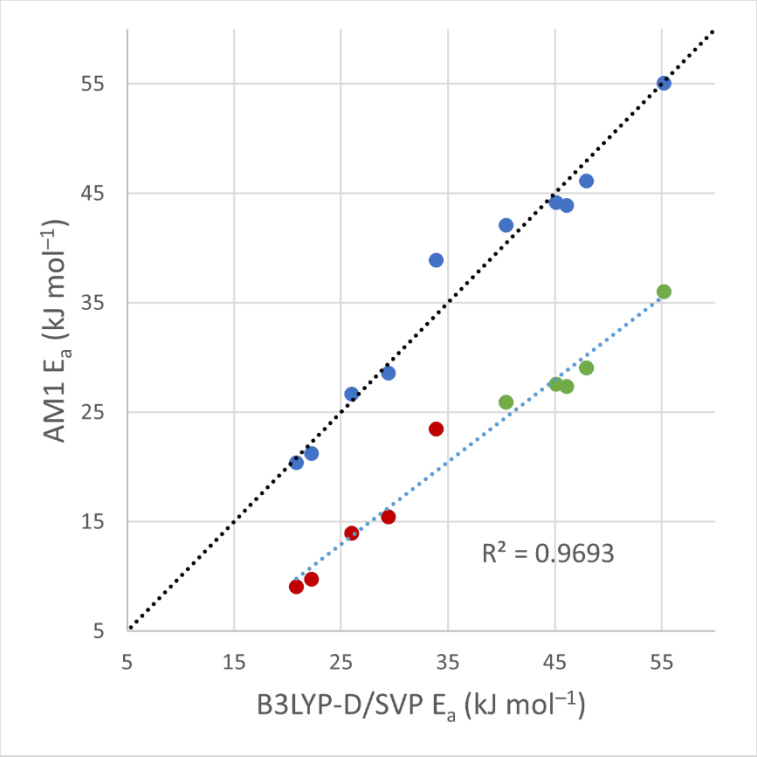

Figure 5. The correlation between DFT and semi-empirical methods. The red dots represent alcohols, the green dots represent phenols, the light blue dotted line represents the trendline and the black dotted line represents the identity line. The blue dots represent the data points after correcting them using the best fit line equation (1.2872x + 8.695).

Applying the respective correction factors to the E_a values of different SoM environments, obtained using AM1, makes them more accurate. However, out of the 75 cases, the lowest corrected E_a corresponded to the experimentally observed SoM in only 52% of the cases. The result is surprising because the corresponding value for UGT1A1, using a UGT-specific simplified reaction mechanism, was 82%. [7]

The results indicate that either the simplified transition state model does not adequately represent the energetics of the full system or, unlike other drug-metabolising enzymes we have studied, the SoM preference of SULTs does not depend on E_a. Since our calculated results agree with the previously published reaction mechanism [32], which in turn is consistent with the experiments [35, 36] we are inclined to accept that the SoM preference for SULTs does not depend on the E_a. Furthermore, the experimental studies using protein sequences in the substrate binding site of the enzyme have indicated that the enzymatic characteristics (regioselectivity) may correlate with the substrate-binding site confirming the results within this study. [24, 61] In conclusion, the low E_a values, small spread of the E_a values, weak correlation between the E_a and the experimentally determined SoMs metabolism, and previously published experimental results suggest that the SoM preference depends more on the structure of the substrate-binding site and hence the accessibility of the SoM.

Accessibility Model for SULTs

Since the previous results indicate that E_a values are not necessary to train the model, we explored the potential to build a model to predict SoMs using only the steric and orientation descriptors. [5] The data was split as explained in the Methodology section and is illustrated in Figure 6. [7, 62]

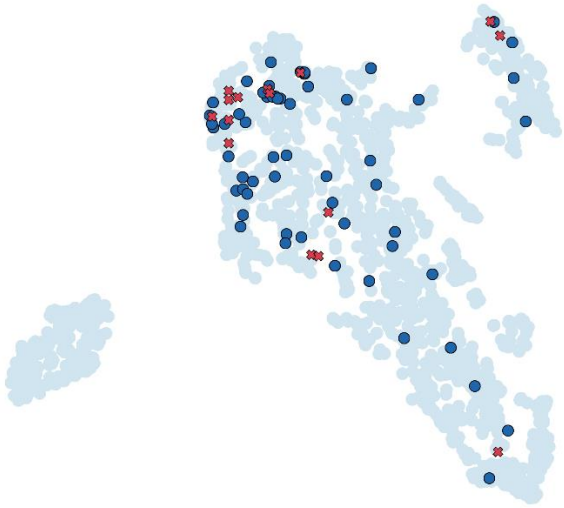

Figure 6. A chemical space plot representing the SULT substrates used for model training and test. The light blue circles represent 1300 launched drugs, the dark blue circles represent the compounds in the training set, and the red crosses represent the compounds in the test set. Each compound is represented by a point and the similarity between compounds by their proximity. These similarities are calculated using the Tanimoto similarity coefficient and a 2D path-based fingerprint. The chemical spaces were created in StarDrop, which using an approach known as t-distributed Stochastic Neighbour Embedding [62] – a nonlinear dimensionality reduction algorithm ideally suited to visualizing high-dimensional data in two dimensions.

The kappa value for the RF SULT model on the independent test set is 0.71, and the balanced accuracy is 0.85. The confusion matrix for the test set is shown in Figure 7. The result can be considered to be excellent, because the model was trained on data covering multiple isoforms and the accuracy is comparable to our previously published isoform- specific models of UGT metabolism. [7] While the experimental data for SULTs, in most cases, did not specify individual isoforms, it is likely that the model achieved such an excellent result because the experimental data mostly consists of results for a small number of prevalent isoforms. It is notable that including the E_a values in the training process resulted in a model with a lower kappa value on the independent test set (kappa of 0.65).

We built 1000 *y*-scrambled models and the results had an average kappa value of –0.01 and a median of 0.00., confirming that the SULT model does not depend on spurious correlations between the observed experimental results and the descriptors.

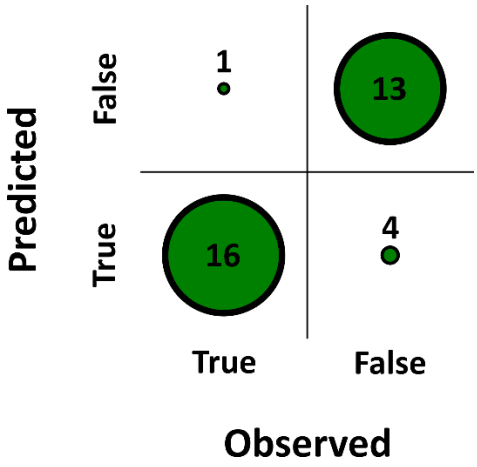

Figure 7. The confusion matrix for the test set of the SULT accessibility model.

Studying the dataset using a leave-cluster-out split method can give a useful insight into the transferability of the model between chemistries. The SULT substrates did not form distinct clusters, but we classified them into seven approximate clusters seen in Figure 8.

For each cluster, we used the SoMs of molecules in the cluster as a test set and the SoMs of the remaining compounds as a training set. The kappa values for clusters I, II, III, IV, V, VI, VII are 0.37, 0.42, 0.37, 0.43, 0.12, 0.55, and 0.38 respectively. It is expected that the results are lower and vary, especially because of the small number of compounds in each cluster, but the kappa value for the cluster V was notably lower than the overall model and other clusters. The poor result for cluster V is surprising, as Figure 8 shows, the molecules in cluster V are similar to other molecules in the dataset. However, it is possible that cluster V bundles compounds with uncharacteristic properties, which make it harder to predict the observed SoM. The results of the seven clusters can be explained; if the model is based only on structural descriptors it will not be as transferable as a model which include the E_a as a descriptor – the E_a is calculated with quantum mechanics, which is universally applicable (subject to the limitations of semi-empirical methods). Thus, the model achieves excellent results only when the training set includes a wide variety of structures.

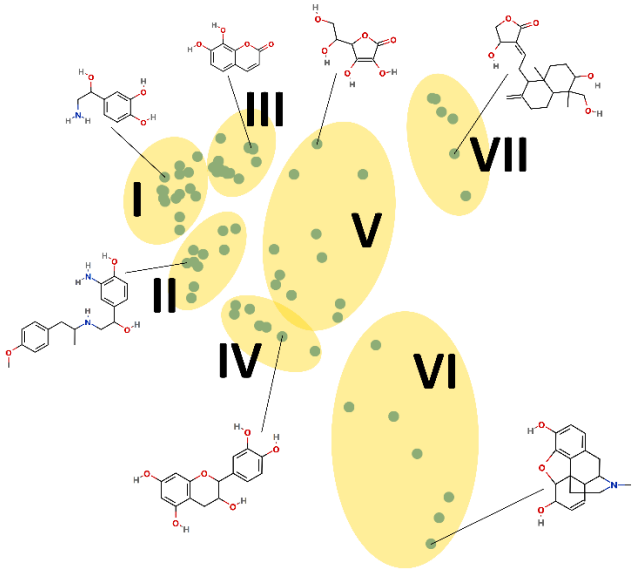

Figure 8. The chemical space of SULT substrates and the respective clusters with an randomly chosen example from each cluster. Each compound is represented by a point and the similarity between compounds by their proximity. These similarities are calculated using the Tanimoto similarity coefficient and a 2D path-based fingerprint. The chemical spaces were created in StarDrop, which using an approach known as t-distributed Stochastic Neighbour Embedding [62] – a nonlinear dimensionality reduction algorithm ideally suited to visualizing high-dimensional data in two dimensions.

Nonetheless, because the overall model was trained on a broad range of known SULT substrates and SoMs, we anticipate that it will be widely applicable to new compounds that are likely to be metabolised by SULTs. The model also provides a measure of its domain of applicability, so it can identify compounds and SoMs for which it cannot be confidently applied.

Conclusions

This paper described the prediction of the regioselectivity of metabolism by human SULTs. The resulting model shows excellent performance and good agreement between the experimentally observed and predicted data points. To our knowledge, previous models for SULT predict if a compound is a substrate or an inhibitor to the SULTs, and not the SoM. [63-66] Thus, the current work expands the capabilities of predictive models of SULT metabolism. The kappa value, balanced accuracy, sensitivity and specificity for the model are 0.71, 0.85, 0.94 and 0.76, respectively.

Unlike other predictive models we have published [5, 7], the SULT model does not use the results from simulations of the catalytic mechanism. On the contrary, the current study suggests that while the mechanism of sulfation is interesting, its energetics play a negligible role in the prediction of regioselectivity of SULTs. Leaving out the mechanism-specific descriptor, E_a, from the model has no impact on its performance. The insignificance of the E_a is probably because the SO_3 is a good leaving group, due to its stabilizing resonance. Furthermore, the intermediate structure, in which the SO_3 has left the PAPS is stabilized by SULT protein residues. Thus, the E_a values for breaking the bond with PAPS and forming a bond with the substrate are very low (around 29 and 8 kJ

mol^{-1} respectively for the system used by Bartolotti *et al*. [32]) and have a small spread. In conclusion, the calculations suggest that the SoM preference depends on the steric and orientation factors, which is consistent with the published experimental results [24, 35, 36, 61, 67]. Because the model is based solely on structural descriptors and not strongly influenced by the reaction energetics, we expect that it will not be as transferable as models trained using data from the simulations of the catalytic mechanism. However, because SULT models do not need extensive quantum mechanical calculations, the models are very fast.

Future work for SULTs will entail gathering isoform-specific data and trying to predict regioselectivity for specific isoforms, in particular 1A1 and 1E1. Since the results from a general model are in good agreement with the experimental data, we would anticipate that the models for individual isoforms would be further improved. Furthermore, training data for additional SULT substrates will further expand the domain of applicability of the models.

Ancillary Information

Supporting Information Availability:
(1) A PDF file with DOI strings for experimental data points, Cartesian coordinates, examples of input files, and additional data.
Corresponding Author Information:
Mario Öeren, mario@optibrium.com
Matthew D. Segall, matthew.segall@optibrium.com
Conflict of Interest Disclosure:
Mario Öeren, Peter A. Hunt, and Matthew D. Segall work at Optibrium Ltd.
Sylvia C. Kaempf worked at Optibrium Ltd. during the execution of the SULT project.
David J. Ponting works at Lhasa Ltd.

Data and Software Availability

The data sets (training and test set) used to train the SULT model can be found in the Supporting Information („1. SULT General Training Set.txt“ and „2. SULT General Test Set.txt“). The training and test sets used for the leave-cluster-out split method can be found in the Supporting Information as comma separated value files and are bundled together with the Python scripts („Python Scripts.zip”). The references for the experimentally observed sites of metabolism can be found in the Supporting Information („references.txt”). These files will include SMILES strings for the compounds, atom indices corresponding to a potential site of metabolism within the SMILES string, experimental results (is a site observed to get metabolized), DOI numbers for references, and the atom-pair fingerprints.

The 3D structures were generated with Avogadro, which is available for download at https://avogadro.cc/ or with software from OpenEye (https://www.eyesopen.com/). It must be mentioned that software from OpenEye was used to convert SMILES into 3D structures, but Avogadro can also be used for it. The DFT calculations were done using NWChem, which is available for download at https://nwchemgit.github.io/. The semi-empirical calculations were done using CP2K, which is available for download at https://www.cp2k.org/.
The atom-pair fingerprints for all potential sites of metabolism were generated using RDKit, which is available for download at https://www.rdkit.org/. The models were trained using StarDrop (https://optibrium.com/stardrop), but we have provided Python scripts („Python Scripts.zip” in the Supporting Information) to train the models using the same parameters using Python packages NumPy (https://numpy.org/), pandas (https://pandas.pydata.org/), and sklearn (https://scikit-learn.org/stable/).

References

(1) H. L. Liston, J. S. Markowitz and C. L. DeVane, “Drug glucuronidation in clinical psychopharmacology.,” Journal of Clinical Psychopharmacology, vol. 21, no. 5, p. 500–515, 2001.
(2) J. O. Miners and P. I. MacKenzie, “Drug glucuronidation in humans,” Pharmacology & Therapeutics, vol. 51, no. 3, pp. 347-369, 1991.

(3) C. Xu, C. Y.-T. Li and A.-N. T. Kong, “Induction of phase I, II and III drug metabolism/transport by xenobiotics.,” Archives of Pharmacal Research, vol. 28, pp. 249-268, 2005.
(4) 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.
(5) 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.
(6) 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.
(7) M. Öeren, P. J. Walton, J. Suri, D. J. Ponting, P. A. Hunt and M. D. Segall, “Predicting Regioselectivity of AO, CYP, FMO, and UGT Metabolism Using Quantum Mechanical Simulations and Machine Learning,” The Journal of Medicinal Chemistry, vol. 65, no. 20, p. 14066–14081, 2022.
(8) 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.
(9) G. Cruciani, E. Carosati, B. De Boeck, K. Ethirajulu, C. Mackie, T. Howe and R. Vianello, “MetaSite: Understanding Metabolism in Human Cytochromes from the Perspective of the Chemist,” Journal of Medicinal Chemistry, vol. 48, no. 22, pp. 6970-6979, 2005.
(10) M. Hennemann, A. Friedl, M. Lobell, J. Keldenich, A. Hillisch, T. Clark and A. H. Göller, “CypScore: Quantitative Prediction of Reactivity toward Cytochromes P450 Based on Semiempirical Molecular Orbital Theory,” ChemMedChem, vol. 4, no. 4, pp. 657-669, 2009.
(11) J. Zaretzki, M. Matlock and S. J. Swamidass, “XenoSite: Accurately Predicting CYP-Mediated Sites of Metabolism with Neural Networks,” Journal of Chemical Information and Modeling, vol. 53, no. 12, p. 3373–3383, 2013.
(12) 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.
(13) 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.
(14) 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.
(15) 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.
(16) 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.
(17) 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.
(18) E. Chapman, M. D. Best, S. R. Hanson and C.-H. Wong, “Sulfotransferases: Structure, Mechanism, Biological Activity, Inhibition, and Synthetic Utility,” Angewandte Chemie, vol. 43, no. 27, pp. 3526-3548, 2004.
(19) M. Negishi, L. G. Pederson, E. Petrotchenko, S. Shevtsov, A. Gorokhov, Y. Kakuta and L. C. Pederson, “Structure and Function of Sulfotransferases,” Archives of Biochemistry and Biophysics, vol. 390, no. 2, pp. 149-157., 2001.
(20) R. A. Al-Horani and U. R. Desai, “Chemical Sulfation of Small Molecules – Advances and Challenges,” Tetrahedron, vol. 66, no. 16, pp. 2907-2918, 2011.
(21) A. B. Roy, “Eugen Baumann and sulphate esters,” Trends in Biochemical Sciences, vol. 1, no. 10, pp. N322-N234, 1976.
(22) D. M. Otterness, E. D. Wieben , T. C. Wood , W. G. Watson, B. J. Madden, D. J. McCormick and R. M. Weinshilboum , “Human liver dehydroepiandrosterone sulfotransferase: molecular cloning and expression of cDNA.,” Molecular Pharmacology, vol. 41, no. 5, pp. 865-872, 1992.
(23) T. W. Wilborn, K. A. Comer, T. P. Dooley, I. M. Reardon, R. L. Heinrikson and C. N. Falany , “Sequence analysis and expression of the cDNA for the phenol-sulfating form of human liver phenol sulfotransferase.,” Molecular Pharmacology, vol. 43, no. 1, pp. 70-77, 1993.

(24) M. Suiko, K. Kurogi, T. Hashiguchi, Y. Sakakibara and M.-C. Liu, “Updated perspectives on the cytosolic sulfotransferases (SULTs) and SULT-mediated sulfation,” Bioscience, Biotechnology, and Biochemistry, vol. 81, no. 1, pp. 63-72, 2016.
(25) R. R. Freimuth, M. Wiepert, C. G. Chute, E. D. Wieben and R. M. Weinshilboum, “Human cytosolic sulfotransferase database mining: identification of seven novel genes and pseudogenes,” The Pharmacogenomics Journal, vol. 4, pp. 54-65, 2004.
(26) C. N. Falany , “Enzymology of human cytosolic sulfotransferases,” The FASEB Journal, vol. 11, no. 4, pp. 206-216, 1997.
(27) Z. Riches, E. L. Stanley, J. C. Bloomer and M. W. H. Coughtrie, “Quantitative Evaluation of the Expression and Activity of Five Major Sulfotransferases (SULTs) in Human Tissues: The SULT “Pie”,” Drug Metabolism and Disposition, vol. 37, no. 11, pp. 2255-2261, 2009.
(28) N. Gamage, A. Barnett, N. Hempel, R. G. Duggleby, K. F. Windmill , J. L. Martin and M. E. McManus, “Human Sulfotransferases and Their Role in Chemical Metabolism,” Toxicological Sciences, vol. 90, no. 1, pp. 5-22, 2006.
(29) M. L. Schrag, D. Cui, T. H. Rushmore, M. Shou, B. Ma and A. D. Rodrigues, “Sulfotransferase 1E1 is a Low Km Isoform Mediating the 3-O-sulfonation of Athinyl Estradiol,” Drug Metabolism and Disposition, vol. 32, no. 11, pp. 1299-1303, 2004.
(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) A. A. Adjei and R. M. Weinshilboum, “Catecholestrogen Sulfation: Possible Role in Carcinogenesis,” Biochemical and Biophysical Research Communications, vol. 292, no. 2, pp. 402-408, 2002.
(32) L. Bartolotti, Y. Kakuta, L. Pedersen, M. Negishi and L. Pedersen, “A quantum mechanical study of the transfer of biological sulfate,” Journal of Molecular Structure (Theochem), Vols. 461-462, pp. 105-111, 1999.
(33) L. C. Pedersen, E. Petrotchenko, S. Shevtsov and M. Negishi, “Crystal Structure of the Human Estrogen Sulfotransferase-PAPS Complex,” The Journal of Biological Chemistry, vol. 277, no. 20, p. 17928–17932, 2002.
(34) E. Chapman , M. D. Best, S. R. Hanson and C. H. Wong, Angewandte Chemie, vol. 43, pp. 3526-3548, 2004.
(35) Y. Kakuta, E. V. Petrotchenko, L. C. Pedersen and M. Negishi, “ The Sulfuryl Transfer Mechanism: Crystal Structure of a Vanadate Complex of Estrogen Sulfotransferase and Mutational Analysis,” The Journal of Biological Chemistry, vol. 273, no. 42, pp. 27325-27330, 1998.
(36) C. T. Bedford , A. J. Kirby, C. J. Logan and J. N. Drummond, “Structure-activity studies of sulfate transfer: the hydrolysis and aminolysis of 3’-phosphoadenosine 5’-phosphosulfate (PAPS),” Bioorganic & Medicinal Chemistry, vol. 3, no. 2, pp. 167-172, 1994.
(37) P. Rydberg, D. E. Gloriam and L. Olsen, “The SMARTCyp Cytochrome P450 Metabolism Prediction Server,” Bioinformatics, vol. 26, no. 23, pp. 2988-2989, 2010.
(38) M. D. Hanwell, D. E. Curtis, D. C. Lonie, T. Vandermeersch, E. Zurek and G. R. Hutchison , “Avogadro: an advanced semantic chemical editor, visualization, and analysis platform,” Journal of Cheminformatics, vol. 4, no. 17, pp. 1-17, 2012.
(39) 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.
(40) 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.
(41) 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.
(42) 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.
(43) 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.
(44) 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.

(45) "Density-functional Thermochemistry. III. The Eole of Exact Exchange," The Journal of Chemical Physics, vol. 98, no. 7, p. 5648, 1993.
(46) 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.
(47) 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.
(48) 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.
(49) 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.
(50) 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.
(51) T. Lu and F. Chen, "Multiwfn: A multifunctional wavefunction analyzer," The Journal of Computational Chemistry, vol. 33, no. 5, pp. 580-592, 2011.
(52) L. Breiman , "Random Forests," Machine Learning, vol. 45, p. 5–32, 2001.
(53) D. J. Ponting, E. Murray and A. Long, "Quantifying Confidence in the Reporting of Metabolic Biotransformations," Drug Discovery Today, vol. 22, no. 7, pp. 970-975, 2017.
(54) C. A. Marchant, K. A. Briggs and A. Long, "In Silico Tools for Sharing Data and Knowledge on Toxicity and Metabolism: Derek for Windows, Meteor, and Vitic," Toxicology Mechanisms and Methods, vol. 18, no. 2-3, pp. 177-187, 2008.
(55) D. J. Ponting, M. J. Burns, R. S. Foster, R. Hemingway, G. Kocks, D. S. MacMillan, A. L. Shannon-Little, R. E. Tennant, J. R. Tidmarsh and D. J. Yeo, "Use of Lhasa Limited Products for the In Silico Prediction of Drug Toxicity," in In Silico Methods for Predicting Drug Toxicity, New York City, Springer Publishing Company, 2022, pp. 435-478.
(56) R. J. Dinis-Oliveira, "Metabolic Profiles of Propofol and Fospropofol: Clinical and Forensic Interpretative Aspects," BioMed Research International, vol. 6852857, pp. 1-16, 2018.
(57) G. . S. J. Mannens, J. Hendrick, C. . G. M. Janssen, S. Chien, B. Van Hoof, T. Verhaeghe, M. Kao, M. F. Kelley, I. Goris, M. Bock, B. Verreet and W. Meuldermans, "The Absorption, Metabolism, and Excretion of the Novel Neuromodulator RWJ-333369 (1,2-Ethanediol, [1-2-Chlorophenyl]-, 2-carbamate, [S]-) in Humans," Drug Metabolism and Disposition, vol. 35, no. 4, pp. 554-565, 2007.
(58) M. L. Schrag, D. Cui, T. H. Rushmore, M. Shou, B. Ma and A. D. Rodrigues, "Sulfotransferase 1E1 is a Low Km Isoform Mediating the 3-O-sulfation Of Ethinyl Estradiol," Drug Metabolism and Disposition, vol. 32, no. 11, pp. 1299-1303, 2004.
(59) L. Lin, C. Xie, Z. Gao, X. Chen and D. Zhong, "Metabolism and Pharmacokinetics of Allitinib in Cancer Patients: The Roles of Cytochrome P450s and Epoxide Hydrolase in its Biotransformation," Drug Metabolism and Disposition, vol. 42, no. 5, pp. 872-884, 2014.
(60) C. Rakers, "Prediction of sulfotransferase specificity," in PhD Thesis, Lahr, Germany, 2016.
(61) A. Allali-Hassani, P. W. Pan, L. Dombrovski, R. Najmanovich, W. Tempel, A. Dong, P. Loppnau, F. Martin, J. Thonton, A. M. Edwards, A. Bochkarev, A. N. Plotnikov, M. Vedadi and C. H. Arrowsmith, "Structural and Chemical Profiling of the Human Cytosolic Sulfotransferases," PLOS Biology, vol. 5, no. 6, p. e165, 2007.
(62) 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.
(63) V. Y. Martiny, P. Carbonell, D. Lagorce, B. O. Villoutreix, G. Moroy and M. A. Miteva, "In Silico Mechanistic Profiling to Probe Small Molecule Binding to Sulfotransferases," PLOS ONE, vol. 8, p. e73587, 2013.
(64) I. Cook, T. Wang, C. N. Falany and T. S. Leyh, "High Accuracy in Silico Sulfotransferase Models," Enzymology, vol. 288, no. 48, pp. 34494-34501, 2013.
(65) C. Rakers, F. Schumacher, W. Meinl, H. Glatt, B. Kleuser and G. Wolber, "In Silico Prediction of Human Sulfotransferase 1E1 Activity Guided by Pharmacophores from Molecular Dynamics Simulations," Journal of Biological Chemistry, vol. 291, no. 1, pp. 58-71, 2016.

(66) S. Hwang, H. K. Shin, S. E. Shin, M. Seo, H.-N. Jeon, D.-E. Yim, D.-H. Kim and K. T. No, “PreMetabo: an in Silico Phase I and II Drug Metabolism Prediction,” Drug Metabolism and Pharmacokinetics, vol. 35, no. 4, pp. 361-367, 2020.

(67) I. Cook, M. Cacace, T. Wang, K. Darrah, A. Deiters and T. S. Leyh, “Small-molecule control of neurotransmitter sulfonation,” Journal of Biological Chemistry, vol. 296, pp. 1-10, 2021.

Table of Contents Graphic