Journal of
CHEMICAL INFORMATION
AND MODELING

Perspective

pubs.acs.org/jcim

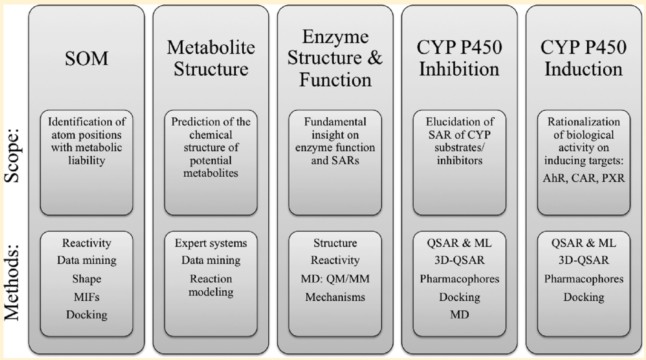
Computational Prediction of Metabolism: Sites, Products, SAR, P450 Enzyme Dynamics, and Mechanisms

Johannes Kirchmair,[†] Mark J. Williamson,[†] Jonathan D. Tyzack,[†] Lu Tan,[‡] Peter J. Bond,[†] Andreas Bender,[†] and Robert C. Glen[†,*]

[†]Unilever Centre for Molecular Science Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, CB2 1EW, Cambridge, United Kingdom

[‡]Department of Chemical Engineering and Biotechnology, University of Cambridge, Tennis Court Road, CB2 1QT, Cambridge, United Kingdom

ABSTRACT: Metabolism of xenobiotics remains a central challenge for the discovery and development of drugs, cosmetics, nutritional supplements, and agrochemicals. Metabolic transformations are frequently related to the incidence of toxic effects that may result from the emergence of reactive species, the systemic accumulation of metabolites, or by induction of metabolic pathways. Experimental investigation of the metabolism of small organic molecules is particularly resource demanding; hence, computational methods are of considerable interest to complement experimental approaches. This review provides a broad overview of structure- and ligand-based computational methods for the prediction of xenobiotic metabolism. Current computational approaches to address xenobiotic metabolism are discussed from three major perspectives: (i) prediction of sites of metabolism (SOMs), (ii) elucidation of potential metabolites and their chemical structures, and (iii) prediction of direct and indirect effects of xenobiotics on metabolizing enzymes, where the focus is on the cytochrome P450 (CYP) superfamily of enzymes, the cardinal xenobiotics metabolizing enzymes. For each of these domains, a variety of approaches and their applications are systematically reviewed, including expert systems, data mining approaches, quantitative structure−activity relationships (QSARs), and machine learning-based methods, pharmacophore-based algorithms, shape-focused techniques, molecular interaction fields (MIFs), reactivity-focused techniques, protein−ligand docking, molecular dynamics (MD) simulations, and combinations of methods. Predictive metabolism is a developing area, and there is still enormous potential for improvement. However, it is clear that the combination of rapidly increasing amounts of available ligand- and structure-related experimental data (in particular, quantitative data) with novel and diverse simulation and modeling approaches is accelerating the development of effective tools for prediction of in vivo metabolism, which is reflected by the diverse and comprehensive data sources and methods for metabolism prediction reviewed here. This review attempts to survey the range and scope of computational methods applied to metabolism prediction and also to compare and contrast their applicability and performance.



	SOM	Metabolite Structure	Enzyme Structure & Function	CYP P450 Inhibition	CYP P450 Induction
Scope:	Identification of atom positions with metabolic liability	Prediction of the chemical structure of potential metabolites	Fundamental insight on enzyme function and SARs	Elucidation of SAR of CYP substrates/ inhibitors	Rationalization of biological activity on inducing targets: AhR, CAR, PXR
Methods:	Reactivity Data mining Shape MIFs Docking	Expert systems Data mining Reaction modeling	Structure Reactivity MD: QM/MM Mechanisms	QSAR & ML 3D-QSAR Pharmacophores Docking MD	QSAR & ML 3D-QSAR Pharmacophores Docking

■ INTRODUCTION

In the discovery and development of new medicines, attrition rates are still very significant, despite the comprehensive measures taken by the chemical and pharmaceutical industry to lower the risk of failure. In pharmaceuticals, toxicity is a major contributor to the withdrawal of new drugs and often the underlying biological mechanism of toxicity is related to metabolism. Metabolic liability is not only a safety concern for drugs but is also highly relevant to a host of industries including nutritional supplements, cosmetics, or agrochemicals (basically any situation in which biology is exposed to chemistry).[1,2]

Metabolic liability can lead to a number of diverse issues, for example drug−drug interactions (DDIs),[3] including enzyme inhibition, induction, and mechanism-based inactivation,[4] resulting in substantial variations (one or more orders of magnitude) of drug concentrations present at target and antitarget sites.[5] These effects potentially lead to a loss of pharmacological efficacy due to enhanced clearance or toxic effects caused by accumulation. DDIs may also increase the rate of reactive, toxic intermediates formed.[6,7] The more the metabolism of a drug is specific to one enzyme, the more likely is the occurrence of DDIs.

DDIs caused by monoamine oxidase (MAO) inhibition often limit the coadministration of multiple drugs. This is problematic in the case of depression and infections, where coadministration of drugs is common.[8] Because of potentially lethal dietary and drug interactions, monoamine oxidase inhibitors have historically been reserved as a last line of treatment, used only when other classes of antidepressant drugs such as

Received: November 14, 2011
Published: February 17, 2012

ACS Publications © 2012 American Chemical Society

dx.doi.org/10.1021/ci200542m | J. Chem. Inf. Model. 2012, 52, 617−648

selective serotonin reuptake inhibitors and tricyclic antidepressants have failed. Tyramine metabolism can be compromised by dosing of MAO inhibitors, and in the case of dietary intake of large amounts of tyramine (e.g., aged cheese[9]), one theory is that tyramine displaces norepinephrine from the storage vesicles and may result in a cascade in which excess norepinephrine is released giving a hypertensive crisis. Many drugs are potentially lethal if ingested with MAO inhibitors. For example tryptamines, coadministered with an MAO inhibitor, can reach very high concentrations and result in serotonin syndrome.[10] The coadministration of drugs which are metabolized by MAOs requires great care as they may in combination saturate the capacity of MAO for metabolism, resulting in altered pharmacokinetics of the drugs and very high concentrations can be reached on multiple dosing. Another example is change of behavior, where transient behavioral sensitization to nicotine becomes long-lasting with addition of MAO inhibitors.[11]

Metabolic reactions may also be systematically exploited in drug design to optimize ADME and toxicity properties following a prodrug concept.[12] It may remain unclear whether the parent molecule is responsible for the entirety of the pharmacological effects observed or if one or several of its metabolites are contributing to the desired therapeutic effect. Another aspect to consider is that for a metabolism-activated prodrug, inhibition of the enzyme required for its activation may cause a loss of pharmacological efficacy or induce toxicity.

Identification of sites of metabolism (SOMs) on molecules and the structure of their metabolites can be decisive for the design of molecules with favorable metabolic properties. Medicinal chemistry driven ADME optimization programs can thus systematically address vulnerabilities in proposed drug molecules (Figure 1).

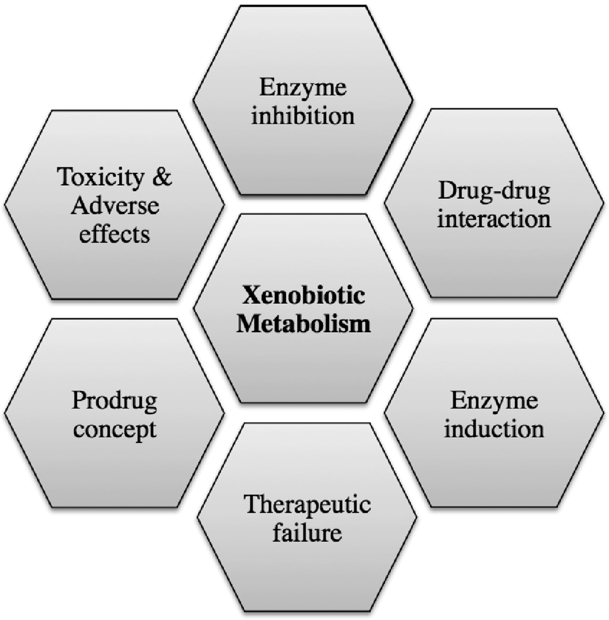


Figure 1. Xenobiotic metabolism and its broad spectrum of pharmacodynamic and pharmacokinetic effects. Potential issues of metabolic liability and biological activity of xenobiotics on metabolizing enzymes include DDIs (in particular, enzyme induction and inhibition), which in turn may cause therapeutic failure, toxicity, and adverse effects. Metabolic reactions can also be exploited for the rational design of prodrugs.

However, experimental techniques to detect and quantify DDIs, as well as to determine SOMs and structures of metabolites are still highly resource-demanding and challenging,[13] which is why in recent years substantial efforts have been applied to develop computational approaches to predict these metabolic outcomes.

In this review, we begin with a general perspective of the scope of ligand-based and structure-based methods for the prediction of xenobiotic metabolism. We then provide an overview of recent developments in computational methods for predicting (i) SOMs, (ii) structures of potential metabolites, and (iii) interactions of small molecules with metabolizing enzymes and metabolic pathways, focusing on CYPs. This section also includes a comprehensive overview on structure and MD (molecular dynamics) simulations of this enzyme superfamily.

We conclude with a perspective and outlook on the scope and limitations of in silico methods for the prediction of xenobiotic metabolism.

Overviews of computational methods discussed in this work are provided in Table 1 and Figure 2.

■ LIGAND-BASED AND STRUCTURE-BASED METHODS FOR PREDICTING XENOBIOTIC METABOLISM

Computational techniques for predicting xenobiotic metabolism can be classified into ligand-based and structure-based approaches. In the first approach, structures of known active and inactive compounds may be modeled to derive structure—activity relationships (SARs, where "activity" refers to a measure of metabolism) and other properties such as SOMs, etc. In the second approach, efforts are focused on the properties of the metabolizing enzyme itself, its interactions with the ligand, and the mechanism of the reaction.

Ligand-based methods need to deal with significant uncertainties about the binding site environment of a metabolizing enzyme, in comparison to structure-based approaches, because the properties of the proximate target environment a ligand binds to may remain largely speculative if no structural data are available. Minor modifications of the ligand may lead to clashes with the protein and, hence, to a loss of bioactivity. However, for ADME processes, small changes in ligand structure do not typically result in major changes in metabolism, presumably due to the flexibility of the metabolizing enzyme active sites.

There is increasing interest in methods considering the molecular shape of both ligands and enzyme active sites. While pharmacophoric interaction patterns have long been established as a major factor for bioactivity, in recent years, methods for the inclusion of molecular shape as a key component for molecular recognition have gained popularity. Besides metabolism prediction, shape-based methods are also applied in many areas of computational drug discovery such as protein active site comparisons, virtual screening, and lead optimization.

Of course, the relative utility of ligand-based and structure-based computational approaches is not as imbalanced as it may seem in the first instance. The substantially higher level of complexity introduced by the consideration of macromolecular structures raises demands in computational power, expert knowledge, and manual interaction, as well as reliance on the phenomenological descriptions of the protein, the ligand, and their environments.

Table 1. Overview of Methods for Predicting SOMs, Structures of Metabolites, and Interactions with Metabolizing Enzymes

Methods for predicting SOMs	Category	Description	Refs (Examples)
QMBO	Reactivity-based method	Derives likelihood of a metabolic reaction at a certain atom position from its hydrogen abstraction energy based on bond order, employing a DFT wave function. Considers accessibility of hydrogen atoms.	16
CypScore	Reactivity-based method	Uses AM1-based atom reactivity descriptors to estimate metabolic reactivity of a certain atom position. Six models to describe various generic CYP metabolic reactions.	17
Metaprint2D	Fingerprint-based data mining method	Derives likelihood of metabolic transformations for atoms with a defined atom environment from data mining of large biotransformation databases. Encodes atom environments using SYBYL[266] atom types in combination with circular fingerprints.	21−24
ADMET Predictor − Metabolite Module	Machine learning method	Derives the likelihood of metabolic reactions to happen at specific atom positions using ANN ensembles. Classification models allow identification of substrates for five CYP isoforms.	28
ROCS	Shaped-focused method	Uses shape-focused alignment of molecules to known CYP substrates in order to derive a potential geometric orientation to the catalytic heme iron. Atom positions in the proximity of the heme iron are considered potential SOMs.	37
Classic docking tools (AutoDock, FlexX, GLIDE, GOLD, etc.)	Protein−ligand docking-based method	Evaluate orientation of the ligand to the enzyme catalytic center in order to identify potential SOMs. Atom positions in the proximity of the heme iron are considered potential SOMs.	51, 52
MetaSite	Combined approach	Uses protein structural information, GRID-derived MIFs of protein and ligand, as well as molecular orbital calculations to estimate the likelihood of a metabolic reaction at a certain atom position.	40, 54, 55
Combined pharmacophore, homology modeling and quantum chemical approach	Combined approach	Combines a pharmacophore-based approach, homology modeling and molecular orbital calculations to pinpoint potential SOMs.	60
SMARTCyp	Combined approach	Utilizes a set of precalculated DFT activation energies in combination with topological accessibility descriptors for prognosis of potential SOMs (CYP3A4 and 2D6).	61, 62
StarDrop	Combined approach	Combines quantum chemical analysis and a ligand-based model of CYP substrates to highlight potential SOMs. Takes into account calculated logP values.	64
RS-Predictor	Combined approach	Utilizes a set of 148 topological and 392 quantum chemical atom-specific descriptors in combination with a SVM-like ranking and a multiple instance learning method to identify potential SOMs.	15
Machine learning-based multidescriptor approach	Combined approach	Takes into account quantum chemical, SASA and pharmacophoric descriptors using a random forest/ensemble decision tree approach to identify potential SOMs.	65
Machine learning-based multidescriptor approach	Combined approach	Employs electrostatic, inductive, energetic, topological, steric, and distance properties in combination with a SVM to predict potential SOMs of endogenous substrates.	67
Combined quantum chemical/docking/MD approach	Combined approach	Combination of quantum chemical methods with docking to account for reactivity, followed by MD simulations to predict potential SOMs.	69
MLite	Combined approach	Combines quantum chemistry-derived reactivity estimation with docking.	73
Ensemble-based/MD-supported docking	Protein−ligand docking-based method	Accounts for protein target flexibility with conformational ensembles of proteins generated using MD simulations and related techniques.	79−82
IDSite	Combined approach	Combination of an induced fit docking approach (GLIDE, PLOP) with a reactivity model (Jaguar).	84
Methods for predicting xenobiotic metabolites	Category	Description	References (Examples)
MetabolExpert	Expert system	Uses knowledge database of rules to predict the structures of likely metabolites. Predicts pathways in animals, plants, or through photodegradation.	88
META	Expert system	Uses a large dictionary of biotransformations to predict the structure of likely metabolites. Analyzes metabolite stability. Predicts pathways in mammals, through aerobic and anaerobic biodegradation.	89
Meteor	Expert system	Employs a collection of knowledge-based biotransformation rules defined using a dedicated structure representation language to derive the structure of likely metabolites. Considers calculated logP values for predictions.	90
University of Minnesota Pathway Prediction System (UM-PPS)	Expert system	Utilizes biotransformation rules to predict the structure of likely metabolites. Specific to microbial catabolic metabolism.	91
SyGMa	Expert system	Predicts structures of likely metabolites based on rules derived from the Accelrys Metabolite Database and assigns probability scores for each metabolite.	92
TIMES	Expert system	Employs a biotransformation library and a heuristic algorithm to generate metabolic maps.	93
JChem Metabolizer module	Expert system	Enumerates all possible metabolites of a given compound. Supports species-specific predictions of likely metabolites.	94
Metaprint2D-React	Fingerprint-based data mining approach	Predicts structures of likely metabolites based on the MetaPrint2D data mining approach.	21
Machine learning-based multidescriptor approach	Combined approach	See description of this software in Methods of Predicting SOMs.	67
MetaSite	Combined approach	See description of this software in Methods of Predicting SOMs.	40, 54, 55

Table 1. continued

Methods for predicting CYP binding affinity/inhibition by xenobiotics	Category	Description	References (Examples)
Linear interaction energy (LIE)	MD simulation	semi-empirical method for calculating free energy of binding for a ligand using ensemble averaged nonbonded interaction energies.	156−159
Free energy perturbation (FEP)/ thermodynamic integration (TI)	MD simulation	Simulations of unphysical states along a thermodynamic cycle connecting the bound and unbound ligand form for calculating the free energy of binding.	131, 161 −163
Decision tree, k-nearest neighbor, ANN, random forest, SVM, etc.	QSAR and machine learning method: Classification model	Classify compounds for enzyme inhibition. Allow conclusions to be drawn on isoform specificity.	202−206, 209
isoCyp	QSAR and machine learning method: Classification model	Classifies compounds for CYP3A4, 2D6, and 2C9 inhibition.	207, 208
PCA, PLS, multiple linear regression, etc.	QSAR and machine learning method: Regression model	Predict enzyme inhibition rates.	210, 212 −216
CoMFA, GRID/GOLPE	3D-QSAR method: Classification model	Classify compounds for enzyme inhibition. Allow conclusions to be drawn on isoform specificity.	222, 224
CoMFA, GRID/GOLPE	3D-QSAR method: Regression model	Predict enzyme inhibition rates. Allow derivation of 3D properties crucial for bioactivity.	224−229, 231, 232
Pharmacophore models	Pharmacophore-based method	Predict quantitative and qualitative enzyme inhibition. Allow conclusions to be drawn on isoform specificity.	213, 233, 236, 238 −240
Protein−ligand docking	Protein−ligand docking-based method	Predict binding mode and binding affinity. Allow conclusions to be drawn on isoform specificity.	159, 241, 242
Combined pharmacohore ensemble/ SVM approach	Combined approach	Uses an ensemble of pharmacophores to account for protein flexibility.	243
Proteochemometric analysis supported by GRIND and further physicochemical descriptors	Combined approach	Considers protein sequences of 14 CYPs as well as GRIND and further descriptors for substrates.	244
Combined machine learning, protein modeling, and docking approach	Combined approach	Uses simulated annealing to render the conformational space of the target protein and docking scores as attributes for subsequent ANN model generation.	246
VirtualToxLab	Combined approach	Uses flexible docking in combination with a multidimensional QSAR approach to predict ligand interaction with 16 antitargets, including CYP450 1A2, 2A13, 2C9, 2D6, and 3A4.	247

Methods for predicting CYP induction by xenobiotics	Category	Description	Examples (References)
PLS using VolSurf descriptors	QSAR approach	Uses atPLS and VolSurf descriptors for the development of a QSAR model for PXR/AhR interaction.	PXR: 251; AhR: 251
SVM, k-nearest neighbor, probabilistic neural network	Machine learning approach	Uses various machine learning methods to predict human PXR interaction.	PXR: 252
Pharmacophore modeling	Pharmacophore-based method	Uses pharmacophores to predict PXR/CAR interaction.	PXR: 253−256; CAR: 262
Protein−ligand docking	Protein−ligand docking-based method	Uses protein−ligand docking to predict PXR/AhR/CAR activation, rationalize SARs and gain insight on likely molecular interaction modes.	PXR: 256−258, 263; AhR: 259, 263; CAR: 263
GRID/GOLPE	3D-QSAR method	Uses MIF-derived 3D-QSAR models to predict CAR interaction.	CAR: 262
VirtualToxLab	Combined approach	Uses flexible docking in combination with a multidimensional QSAR approach to predict ligand interaction with AhR and 15 other targets (see above).	AhR: 247

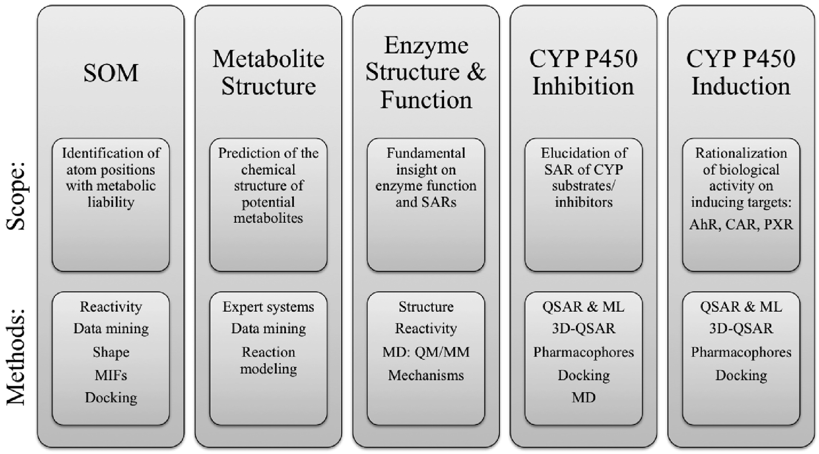

Scope:
Methods:
SOM
Identification of atom positions with metabolic liability
Reactivity
Data mining
Shape
MIFs
Docking
Metabolite Structure
Prediction of the chemical structure of potential metabolites
Expert systems
Data mining
Reaction modeling
Enzyme Structure & Function
Fundamental insight on enzyme function and SARs
Structure
Reactivity
MD: QM/MM
Mechanisms
CYP P450 Inhibition
Elucidation of SAR of CYP substrates/ inhibitors
QSAR & ML
3D-QSAR
Pharmacophores
Docking
MD
CYP P450 Induction
Rationalization of biological activity on inducing targets: AhR, CAR, PXR
QSAR & ML
3D-QSAR
Pharmacophores
Docking

Figure 2. Overview on topics and computational approaches covered in this review.

In general, structure-based methods have focused on deriving patterns from static structures of protein−ligand complexes without considering time-dependent conformational fluctuations, unless combined with even more resource-intensive MD simulations or similar methodologies (see the section on Molecular Dynamics Simulations on the Structure of CYPs). Receptor−ligand conformations observed in X-ray crystallographic experiments may include crystal packing effects and other experimental uncertainties.[14] Also, an observed ligand pose may not necessarily reflect the relevant binding mode for the metabolic transformation.

Structure-based methods that predict receptor−ligand binding geometry (e.g., for SOM prediction) should ideally take some of the changing properties upon complexation of both the ligand, the protein, and the ligand−protein complex into consideration. For example, desolvation and entropic components often play a dominant role in the energetics and geometry of ligand binding; however, both desolvation and entropic effects are difficult to model accurately (including diffusion, on−off rates and solvent reorganization). The approximations necessary to enable predictions to be made in a reasonable time-scale are significant. Additionally, ligands typically approach a binding/active site through a size-limiting access channel, and it is nontrivial for structure-based approaches to simulate the dynamic changes in geometry required for the approach of the ligand to the binding site. Ligands that are metabolized, however, may be assumed to include all these properties, which to some degree are intrinsically included in ligand-based models.

Despite the challenges, structure-based methods have significantly contributed to the rationalization of metabolic reactions and with rapidly advancing technologies for structure elucidation and computation of dynamics and reactivity structure-based methods are becoming ever more predictive and accurate. Certainly, the combination of both ligand-based and structure-based approaches is most promising, leading to synergistic effects between different algorithmic approaches that allow more comprehensive descriptions of metabolic reactions.

■ PREDICTING SITES OF METABOLISM

Identifying SOMs of a molecule may give decisive hints in the development of a medicinal chemistry strategy to optimize metabolic properties and, consequently, crucial parameters such as toxicity, bioavailability, and bioactivity. Chemical reactivity and orientation of the ligand bound to the catalytic site of the enzyme are key for the location of SOMs.

Today, a plethora of computational approaches are available, which attempt to identify the most likely SOMs. Algorithms include reactivity-based approaches, fingerprint-based data mining approaches, shape-focused techniques, molecular interaction fields (MIFs), protein−ligand docking, and combinations of methods (Figure 2). Most of the techniques available to date consider a single aspect of metabolic reactions, such as the reaction energy barrier, geometrical properties, or pharmacokinetic properties for predicting SOMs or potential metabolites.[13] In reality, however, a combination of factors (energetics, geometry, reaction mechanism, etc.) is decisive for a metabolic reaction to take place and hence combined approaches render a more realistic and more accurate view of biotransformations.

To quantify the SOM prediction performance of an algorithm, it is usually tested on a set of compounds for which at least one SOM has already been identified experimentally. Many testing protocols use the "Top-N" metric, which for a given compound is the occurrence of an experimentally confirmed SOM within the N top-ranked predicted SOMs. A major insufficiency of this metric is that it does not account for the fact that for some substrates the prediction of a SOM may be biased. This metric is for example strongly influenced by the size of a molecule, as larger molecules intrinsically have more putative SOMs. More recently, the "Lift" metric has been introduced in order to overcome this bias.[15] This measure includes the prediction accuracy of a computed model compared to a random model.

Reactivity-Based Approaches. Reactivity-based approaches utilize descriptors that are derived from the electronic structure of the molecule to predict its liability for metabolic transformations (see "Reactivity of CYPs" for more information on the reaction mechanism of CYPs). For computational efficiency, semi-empirical quantum mechanical methods can be used. As a preamble to the main calculation, a 2D representation of a ligand must be converted to a 3D representation and the correct hydrogen addition performed considering tautomeric states and pK_a values of ionizable groups, followed by geometry optimization.

QMBO[16] estimates hydrogen abstraction energies based on bond order. Starting with the premise that the hydrogen abstraction by Compound I (see "Reactivity of CYPs") from a substrate is the rate-limiting step, the method relates the reactivity of each hydrogen atom to the strength of its covalent bond. Using a wave function generated from density functional theory (DFT), bond orders for all C−H bonds in a substrate are calculated, and then normalized. Bond strength is correlated to deviations from average bond orders. Corrections are made for buried hydrogen atoms through scaling by a factor that is a function of the solvent accessible surface area (SASA) of the hydrogen atom.

Given that the highly electrophilic Compound I has received an electron from the substrate, the substrate will form a positively charged radical intermediate, and where this spin hole is localized, there will be a likely SOM. Using a DFT wave function, the QMSpin[16] tool calculates spin densities on all hydrogen atoms using Fermi contact values. The same SASA scaling approach used in QMBO is applied. Of these two methods, QMBO is reported to perform slightly better than QMSpin. Using the Top-3 metric for SOMs over 81 molecules, QMBO predicts 84% and QMSpin 78%.[16]

CypScore[17] makes use of a range of atomic reactivity descriptors generated from an AM1[18] wave function, with molecular properties derived from this using the software VAMP[19] and ParaSurf.[20] These descriptors are used for the generation of individual models for the most important reactions catalyzed by CYPs. Using the Top-3 metric, 87% of SOMs were identified from a test set of 39 molecules.

Most reactivity-based approaches put substantial weight on correction factors often related to topological properties describing accessibility. These are discussed in the section "Combined Approaches" below.

Fingerprint-Based Data Mining Approaches. Fingerprint-based data mining approaches describe the presence or absence of chemical features within a molecule and relate these to metabolic reaction sites and products. The chemical environment of a SOM is associated with a fingerprint descriptor enabling predictions to be made about novel compounds by searching for locations with the same or similar fingerprint within the target molecule. The relative likelihood of a fingerprint being associated with a SOM can be calculated from the number of occurrences of this transformation and the method can be trained on a data set of known metabolic transformations. This can suggest a ranking for different sites within a molecule according to their likelihood of being a SOM.

Metaprint2D[21−24] is a data-mining tool that identifies SOMs on the basis of circular fingerprints. The software package is Java-based, and an online service[25] allowing prediction of SOMs for uploaded or drawn structures is available. Metaprint2D mines large biotransformation databases such as the Accelrys Metabolite Database,[26] which contains more than 100,000 metabolic transformations. For each atom of each substrate−metabolite pair included in the database, the atom type at the SOM, as well as its proximate environment, is encoded in a fingerprint. In order to derive probabilities for a specific atom to be involved in a metabolic interaction, the occurrence of each of the encoded atom environments is calculated and compared to the number of biotransformations recorded in the database for this particular atom environment. The normalized occurrence ratio does not reflect the absolute probability of a metabolic reaction to occur, but can indicate the prevalence of a reported transformation in the literature. The Metaprint2D algorithm uses a memory cache to store the fingerprints, allowing results to be returned in under a second. A disadvantage of this data mining approach is that predictions cannot be made about novel atomic sites where the fingerprint does not exist in the data set. MetaPrint2D detects novel sites within a query molecule and acknowledges that they are outside the applicability domain when it presents the metabolic predictions in its graphical results output.

Tests of the method using compounds added to later versions of the Accelrys Metabolite Database (other than that used for model training) indicate that SOMs are correctly predicted in about 70−80% of cases among the three highest-scored atom positions.[21]

Metaprint2D-React[21] is an extension of MetaPrint2D that allows prediction of metabolic products on the basis of the structure of a substrate (see “Predicting Xenobiotic Metabolites”). The metabolites proposed by MetaPrint2D-React are ranked according to the relative occurrence of matching transformations in the training database associated with that fingerprint pattern. A freely accessible online version of the software is available.[27]

Machine-Learning Approaches. SOM prediction methods based on machine learning methods such as support vector machines (SVMs) and artificial neural networks (ANNs) have recently gained more attention for use in SOM prediction. Simulations Plus have introduced the Metabolite Prediction module[28] to their ADMET Predictor software. Like MetaPrint2D, this prediction tool is trained on the Accelrys Metabolite Database. This data set has been curated and some further sources for metabolic reactions were added. Models for the prediction of SOMs are based on artificial neural network ensembles (ANNEs) derived using atomic descriptors to generate an artificial neural network classification (ANNC) model. For each atom of a molecule, a score (0−1) for the likelihood of a metabolic reaction to happen is assigned. In addition to that, a substrate classification model can predict whether a compound is a substrate of five CYP isoforms.

Further examples include combining different methods, which are discussed in the section “Combined Approaches”.

Shape-Focused Approaches. The concept of shape-focused methods is based on the observation that compounds sharing a similar shape have a high probability of binding to the same receptor. The shape of a ligand is thought to resemble the partial complementary shape of the binding site. In this way, shape-focused methods attempt to identify bioactive molecules by molecular shape recognition. Shape similarity between the bioactive template molecule superimposed upon a candidate molecule is used for a quantitative calculation of the likelihood for metabolism, with the site identified from the proximity to the experimentally determined site on the template molecule.

Shape-based methods can be effective and have become increasingly popular in the area of similarity-based virtual screening[29−34] and bioactivity profiling.[35] Shape-focused methods are advanced versions of shape-based methods, where alignment and ranking are performed using a chemistry-aware algorithm. The inclusion of chemical information (resembling pharmacophoric constraints) to molecular shape can be an additional factor in improving the accuracy of these methods.[30]

The shape-focused screening method Rapid Overlay of Chemical Structures (ROCS)[36] was employed to predict the SOM of CYP2C9 substrates.[37] Flurbiprofen, a substrate of CYP2C9 that is hydroxylated by the enzyme at position 4′, was extracted in a protein-bound conformation from an X-ray structure for use as a structural template for a ROCS search. The assumption was that the SOM of any CYP2C9 substrate should be located in proximity to the SOM of flurbiprofen. ROCS was employed to superimpose 70 known CYP2C9 substrates to the flurbiprofen template and encouraging results were found: The SOMs of 60% of all investigated molecules were positioned within 3 Å of the 4′-hydroxylation site of flurbiprofen, with 39 (89%) out of the 44 top-ranked molecules correctly predicted. Additionally, alignments were further analyzed including the protein background to evaluate whether the positioning of the SOM also relates to the heme iron. The average distance between the known SOM and the heme iron was found to be 5.21 ± 0.95 Å, which is in agreement with experimentally observed substrate−heme distances. The approach is rapid, largely unbiased and does not require manual interaction, which allows large-scale profiling of drugs and other similar molecules to a defined substrate.

Molecular Interaction Fields. Molecular interaction fields (MIFs) are one of the most established and most versatile concepts in drug discovery. The idea of MIFs is to encode the variation of interaction energies between a target molecule and a chemical probe in 3D space. The probes are usually an atom, a point charge, or a molecular fragment and depending on the physicochemical properties of the probe, distinct characteristics about the target molecule can be derived from the calculated interaction energies.

In ligand-based design, MIFs may be used to derive 3D-QSAR models and pharmacophoric representations and to predict pharmacokinetic parameters such as cell permeability and metabolism. In structure-based research they are employed to analyze structural features of macromolecules such as protein−ligand and protein−protein interactions. MIFs have been reported to be useful to elucidate information for rational drug design on how to optimize protein−ligand interactions, areas of ligands vulnerable to metabolism, and ligand/isoform specificity. There are currently a number of implementations of MIFs in CYP-related metabolism of xenobiotics and examples covering many of these aspects have been published (as discussed in the following and in “Predicting Interaction of Xenobiotics with CYPs”).

GRID[38] has been developed as a tool to explore and characterize protein structures for areas favorable for interaction with small organic molecules. The software is well-known for its key role in the design of the anti-influenza drug zanamivir.[39] Common probes in GRID include DRY (representing hydrophobic interactions), N1 (representing hydrogen bond donors, derived from planar NH such as amides) and O (representing hydrogen bond acceptors derived from a sp2 carbonyl oxygen).

An example of a MIF-based software package for predicting SOMs is MetaSite.[40] Further approaches utilizing MIFs have been reported, in general combined with other computational techniques. For more detail on these approaches the reader is referred to "Combined Approaches".

Protein–Ligand Docking. Protein–ligand docking has become an integral part of today's rational lead identification and optimization process. By 2004, it is estimated that about 50 new entities had reached clinical trials and/or market approval guided by structure-based drug design.[41] With crystal structures available for the major human CYP isoforms, protein–ligand docking is showing promise for the analysis and prediction of CYP-induced metabolism.[7]

Docking consists of two parts. First, the ligand is docked into the receptor binding site. This is the actual docking process and involves sampling the conformational space of the ligand within the binding site. In the second step, the generated docking poses are evaluated and some measure of affinity or the fit of the ligand for the receptor is estimated. This step is also referred to as scoring.

By using protein structures directly as model systems, docking is largely unbiased by known SAR, offering opportunities to reveal entirely new binding modes. Current ligand docking methods are limited by many approximations including approximate descriptions of desolvation, entropic and enthalpic components, as well as dynamic structural changes of the protein upon ligand binding. Generally, the faster the method, the more significant are the approximations made. Packing forces in the protein template may also introduce bias to the model. Ligand-induced mobility of the binding site becomes apparent when comparing redocking with cross-docking results. In the first case, a docking algorithm is used to redock a ligand extracted from the binding pocket (no induced fit), while in the latter case, ligands are cross-docked to target structures derived from protein structures templated to an alternative ligand. Cross-docking in general shows significantly lower accuracy compared to redocking. The reasons for this include the conformational changes induced in both the protein and ligand.[42]

Docking may allow the derivation of SOMs by relating the proximity of ligand atoms of a docking pose to the catalytic center of the enzyme. Additionally, docking also allows investigation of binding properties of ligands to their target proteins and suggests SAR (see "Predicting Interaction of Xenobiotics with CYPs"). Direct consideration of the protein environment allows the elucidation of residues involved in ligand binding and can allow rationalization of different metabolic properties of enantiomers by analysis of the protein–ligand interactions. This information is valuable for the design and optimization of drug candidates.

Generally speaking, current docking algorithms offer adequate performance for elucidating the ligand-binding mode provided an appropriately complementary (to the ligand-binding pose) protein structure is applied that covers the relevant receptor conformation. Cross-docking still poses significant problems. Despite decades of research, some cardinal issues of docking—and in particular of scoring—have not yet been resolved.[7] Key aspects such as diffusion, entropy, and solvation effects remain largely neglected by scoring functions.[43]

A systematic investigation of the impact of water molecules present/absent in the active site of 19 CYP structures[44] has been reported for AutoDock,[45] FlexX,[46] and Genetic Optimization for Ligand Docking (GOLD).[47] Performance was compared for the process of docking into water-free binding sites, binding sites with crystallographic water present, and binding sites with water molecules placed using GRID. It was found that the consideration of crystallographic water molecules generally improves the docking accuracy of all three approaches and that the introduction of water molecules predicted with GRID leads to a further increase of accuracy with regard to root-mean-square deviations (RMSD) of the docked ligand from the experimentally determined position and SOM prediction accuracy.

In a follow-up study focusing on CYP2D6, considering several different scenarios for the inclusion of water, it was confirmed that the placement of water molecules almost always leads to an increase in docking accuracy.[48] Correctly placed water molecules improve docking-based SOM prediction (as well as virtual screening accuracy). The combination of GOLD with ChemScore[49] was found to perform particularly well.

Recently, Santos et al.[50] systematically investigated the impact of water molecules on protein–ligand docking of substrates of CYP2D6. Representative protein conformations used for docking were generated using MD simulations of an X-ray structure of (R)-3,4-methylenedioxy-N-ethylamphetamine (MDEA) bound to CYP2D6. Hydration sites were derived from the trajectory, and water molecules were placed at the computed positions for water in the protein frames used for docking. Eleven substrates similar to MDEA and 53 structurally distinct compounds were then docked into the representative protein conformations. These results indicated that water molecules placed into the binding site might have a beneficial impact on the docking performance of substrates closely related to the crystallographically determined ligand. However, this effect is strongly dependent on the protein and ligand conformation. Also, there seemed to be no benefit from the inclusion of MD-derived water molecules for substrates dissimilar to MDEA. Overall, in this case, water molecules were found rather weakly bound inside the CYP2D6 binding pocket, with an average residence time below 10 ps, which may contribute to the lack of beneficial effect on inclusion of specific water on docking and the observed substrate promiscuity of the enzyme.

Examples of the successful prediction of SOMs include a study by Vasanthanathan et al.,[51] who performed docking experiments on CYP1A2 employing GOLD with various docking scenarios. These scenarios differed in the treatment of the crystallographic water molecules mediating a hydrogen bond between the enzyme and the ligand and in the scoring function used (ChemScore or GOLDScore). Docking was deemed successful if the known SOM of a docking pose was computed to be within a 6 Å radius from the heme iron. While docking performed adequately to allow the elucidation of binding orientations and conformations of ligands as well as moderate enrichment rates during virtual screening, superior performance in classifying individual compounds into those that were or were not metabolized was observed for machine learning methods (reviewed later). In contrast to several other docking studies on CYPs, the authors did not detect a significant benefit from including water molecules in protein–ligand docking (see below).

Unwalla et al.[52] used Grid-based LIgand Docking with Energetics (GLIDE)[53] and a homology model for CYP2D6 to predict the SOMs of 16 substrates. Predictions were deemed correct if the known SOM is located less than 4.5 Å from the catalytic iron among the five top-ranked docking poses, which was the case for 85% of all investigated cases. Docking results obtained from this CYP2C5-derived homology model were significantly better correlated to the experiment than docking results obtained from a X-ray structure of apo CYP2D6.

Scoring ligand interactions with the heme prosthetic group is challenging, requiring specific parametrized scoring functions. Extensive hydrophobic surface areas that are present in CYPs contribute to the difficulty of docking ligands to this protein family. The lack of strong, directed hydrogen bonding interactions makes scoring functions highly dependent on the interpretation of weak van der Waals interactions between the ligand and the enzyme surface.

With today's computer hardware, computational resources are hardly an issue for most docking methods. A significant cost factor of docking is the degree of expert knowledge on the target under investigation and the interpretation of results from the docking algorithm itself, not only during setup but also during final data analysis. Relatively sparse coverage of structural information on CYPs (they are conformationally labile) limits the applicability of docking for this enzyme family. Techniques to extend the scope of docking include homology modeling and MD simulations, as discussed in the following section.

Combined Approaches. As pointed out above, chemical reactivity and accessibility are the decisive aspects for a metabolic reaction to take place. It would not be anticipated that a metabolic reaction would proceed at a highly reactive site of a molecule that, for steric reasons, cannot be oriented within the CYP to interact with the reaction center. In such a case, a method purely considering reactivity-related aspects is likely to fail. Consideration of both aspects is crucial, which is reflected in the variety of combined approaches that have been reported.[13]

MetaSite is an integrated software package for predicting CYP-mediated metabolism.[40] It evolved from earlier approaches combining GRID and ALMOND[54,55] and includes several different components to predict SOMs: MIF-based modules for the analysis of protein and ligand characteristics and quantum-chemical and knowledge-based components to account for metabolic reactivity. The software package encodes the characteristics of CYP enzymes in the form of a fingerprint that is independent of the initial conformation of side chains in order to account for protein adaption induced by ligand binding. Atoms of the small organic molecule under investigation are categorized by their GRID probe counterparts to reflect their hydrogen-bonding and charge capabilities as well as their hydrophobic characteristics. Subsequently, they are transformed into a distance-binned representation, resulting in a fingerprint for each GRID probe category. The descriptors of the CYP structure and of the ligand are then compared in order to quantify the exposure toward the catalytic heme moiety for any of the ligand atoms, which relates to the liability for metabolism. MetaSite includes a reactivity component that estimates the likelihood of a metabolic reaction on the basis of molecular orbital calculations, estimating the energetic barrier of hydrogen bond abstraction. The overall susceptibility at a certain atom position is then derived from a probability function accounting for both accessibility and reactivity. Only atoms obtaining high values for both measures are considered likely to be involved in metabolic reactions. The software also includes a knowledge-based reaction mechanism component for weighting preferences for types of metabolic reactions for the individual CYP isoforms.[56]

Results of a MetaSite analysis are reported as a histogram bar chart indicating the likelihood of metabolism for any of the ligand atoms. An interesting feature of MetaSite is the Partial Contribution plot, which indicates moieties of the compound that are expected to interact favorably with the CYP isoform under investigation. This can provide valuable hints on how to optimize the metabolic properties of a ligand.[57]

A structure-based approach such as this is limited to enzymes where a protein structure or homology model is available. A drawback of MetaSite is that it reports only the relative likelihood for a specific metabolic reaction to occur at a certain atom position. Results among different molecules cannot be directly compared. This is also the case for the Partial Contribution plot.

The authors of MetaSite claim a success rate of 85% for tagging a known SOM among the top-2 ranked atom positions. It was successfully employed to identify the main metabolites of the angiotensin receptor antagonists lorsartan, candesartan, valsartan, and tasosartan.[54] In a study by Zhou et al.,[58] MetaSite was found to correctly predict the SOM among the three highest-ranked atom positions for 78% of the investigated compounds. A more recent study employed MetaSite for the identification of SOMs to support experimental mass-spectrometric metabolite identification.[59] The SOM top-ranked by MetaSite were experimentally confirmed for 55% of the analyzed compounds. The hit rate improved to 84% when considering the three top-ranked atom positions.

De Groot et al.[60] reported one of the earliest approaches combining several methods to pinpoint SOM. They employed ligand-based pharmacophores, homology modeling, and molecular orbital calculations (AM1[18]) to identify the SOM of substrates of CYP2D6. On the basis of the information gained from the three domains, the SOM of six out of eight investigated ligands could be correctly identified.

Rydberg et al. have released an open-source, Java-based, SOM predictor called SMARTCyp.[61−63] This software package contains a database of precalculated DFT activation energies for various ligand fragments undergoing a CYP3A4-mediated transformation. SMARTCyp performs a lookup for a ligand using this database that contains SMiles ARbitrary Target Specification (SMARTS)-based fragments that are matched, in conjunction with an accessibility descriptor to provide a ranking of the SOM. Using the Top-2 metric, 76% of SOM are identified over a test set of 394 compounds. The tool is fast because of the precached QM results. Most recently, SMARTCyp was extended to predict reactivities for CYP2D6.[63]

StarDrop[64] predicts SOM on the basis of quantum chemical analyses of molecules. For each atom of a molecule, a metabolic site liability reflecting the potential metabolite formation is computed. The metabolic site liability of a ligand atom is denoted by values between 0 and 1. This is derived from an estimate of the hydrogen abstraction, which is related to the enzyme decoupling rate. While the same quantum chemical module is used for all CYP isoforms, specificity is assessed by alignment of the query molecule to a ligand-based model derived from known substrates of the respective isoform. Also logP, which appears to influence CYP-mediated metabolism (see "Expert Systems"), is taken into account. StarDrop was reported to give success rates comparable to SMARTCyp.[62]

More information on the performance of these tools is provided in ref 15.

RegioSelectivity (RS)-predictor[15] is a recently released, isoform-specific approach to predict SOM. It utilizes 148 topological and 392 quantum chemical atom-specific descriptors, a support vector machine (SVM)-based ranking, and a multiple instance learning method. Some of these descriptors are modified to include contributions from neighboring atoms. Using the Top-2 metric, 78% of SOMs were identified, outperforming SMARTCyp and StarDrop using either the Top-2 or Top-3 metrics. The authors also supply their 394 compound test set in their supporting materials. This is annotated with experimental SOMs in addition to StarDrop and SMARTCyp predicted SOMs. This is an excellent resource for comparison testing of future methods.

Hasegawa et al.[65] have developed a machine learning-based approach for the prediction of SOMs related to CYP3A4-based metobolism. It combines five quantum chemical descriptors related to reaction energies plus the activation energies calculated for the SMARTCyp approach to account for reactivity. Their approach also considers the solvent accessible surface area (SASA) to estimate the exposure of a ligand atom to the heme iron and pharmacophoric constraints to reflect the requirements of CYP3A4 for ligand binding. The latter were derived from known substrates using MOE.[66] Random forest/ensemble decision trees were employed to predict metabolic sites with these models, with k-nearest neighbor as a baseline method. The model considering only reactivity-related descriptors and SASA was outperformed by the models accounting for pharmacophoric requirements. The best-performing model included all three types of descriptors, with pharmacophoric constraints limited to a distance maximum of 10 Å, employing the random forest algorithm. Known SOMs were correctly predicted in 82% of all investigated cases among the two top-ranked atom positions. It should be kept in mind that only 10 molecules were contained in the test set, which is rather small when judging performance of the model in real-world situations, where new molecules will almost certainly be outside of the applicability domain.

A method to predict the metabolic reactivity of small endogenous metabolites has recently been presented that combines expert knowledge, computational chemistry, and machine learning.[67] On the basis of the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database,[68] 4843 reactions were characterized and classified into 80 different reaction classes. SMARTS patterns were used to define chemical substructures (reaction centers and surrounding moieties), and the physicochemical properties of these were encoded using electrostatic, inductive, energetic, topological, steric, and distance properties. A SVM binary classification model was trained for each reaction center SMARTS pattern in order to predict whether or not a certain atom position is a SOM. A score was obtained for candidate reaction centers of a molecule, which was then used to rank the likelihood of a metabolic reaction taking place. A 3-fold cross validation was performed, obtaining an average sensitivity of 0.74 and an average specificity of 0.87. Comparable results were found when applying the method to predict molecules that have been recently added to the KEGG database and which were outside of the training set. Interestingly, the four most important features identified were distance properties (encoding the geometrical position of a reaction center in a small organic molecule). The encouraging results obtained with this approach make it an interesting opportunity to extend its applicability domain to xenobiotics.

Several studies employed protein–ligand docking in combination with other approaches to predict metabolism. Lacking a component to account for reactivity, docking is frequently combined with quantum chemical methods in order to generate a more accurate description of the highly complex catalytic process. In this respect, the hydroxylation and O-dealkylation reactions of sirolimus (rapamycin) and everolimus (RAD-001) catalyzed by CYP3A4 were analyzed using a combined quantum chemical/docking/MD simulations approach.[69] The electronic properties of both CYP3A4 substrates were described employing B3LYP DFT. The substrates were also analyzed by docking on a CYP3A4 homology model using DOCK[70] and subsequent MD simulations using Assisted Model Building with Energy Refinement (AMBER).[71,72] The combined approach turned out to be useful to extend knowledge on electronic and orientation effects of substrates of CYP3A4.

MLite[73] is a combined model for CYP3A4-mediated metabolism, employing protein–ligand docking and the quantum chemical reactivity estimation method developed by Korzekwa et al.[74] and Jones et al.,[75] see above. Hydrogen atoms exposed to the heme iron were identified by docking and reactivity was considered on the basis of the results of quantum chemical calculations. In the case of O-dealkylation, a penalty for the activation energy was required to lower the number of incorrect predictions by the reactivity-based approach. The optimized model gave 76% correct predictions for the 25 compounds of the test set when considering the two highest-ranked atom positions for each molecule. This approach is related to a further method developed by this research consortium,[76] which also employed the semi-empirical (AM1[60]) method developed by Korzekwa et al. and Jones et al. for reactivity estimation, in this case for CYP1A2. AutoDock was used to orientate the ligands in the binding site to examine potential SOM and to estimate binding free energies. The overall prediction of SOM was derived from both measures. The SOMs of eight out of twelve substrates were correctly predicted as the primary SOM and for all of them as the secondary SOM.

Insufficient structural data on the conformational flexibility of CYPs is still limiting the applicability of docking approaches for this enzyme family. Homology modeling is one approach to address this problem, in order to derive structural models from closely related target structures that can serve as a template for model generation. Several CYPs, however, exhibit extraordinary levels of flexibility, which is a major problem for structure-based approaches in general. Docking algorithms are highly susceptible even to marginal conformational changes of the target structure. The approach benefits greatly from using collections of target structures crystallized in the presence of ligands of distinct shape and binding modes, so-called ensemble-based docking.[77] Apo structures of a target in general perform less well compared to holo structures due to the absence of a ligand inducing conformational shifts in the binding site.[52]

If the relevant conformational space of a target is not adequately covered by experimental structures, representative target conformations from MD simulations may boost docking performance, in particular in cases where heterogeneous binders induce conformational shifts in the binding region.[77] Thanks to its remarkable flexibility,[78] CYP3A4 is a prime target for ensemble-based docking approaches. A recent study employing an ensemble of CYP3A4 structures derived from several MD simulations to identify SOMs points out the importance of covering the conformational space of the catalytic binding site.[79]

Hritz et al.[80] found that the binding site of the apo structure of CYP2D6 is too tight to facilitate the binding of 80% of 65 investigated substrates of this isoform. By employing an ensemble-based approach, however, they were able to successfully identify the correct SOM for all of the ligands with at least one of the MD-derived target conformations. In this approach, the known substrates were docked to an ensemble of 2500 protein conformations of CYP2D6 derived by MD simulations from structures with five heterogeneous substrates bound in order to simulate the induced fit effect and thermal motion of the enzyme. To speed up future docking runs, a binary decision tree was derived that allows identification of the best suitable protein conformation for docking specific ligands. In this way, Hritz et al. were able to successfully predict the SOM of CYP2D6 substrates for 80% of all cases when docking to the single protein conformation selected by the decision tree model.

Instead of generating an ensemble of protein conformations for subsequent docking, MD simulations can also be employed for refinement of docking results. For example, Keizers et al.[81] found that automated docking of ligands to CYP2D6 structures is not accurate enough to reflect the ligand orientations identified in experiments. Using docked poses as a starting point for MD simulations, they found a significant improvement in ligand orientation and SOM prediction.

Moors et al.[82] derived an ensemble of 1000 protein conformations on the basis of a X-ray structure of CYP2D6 employing tCONCOORD[83] (which employs a set of distance constraints to explore the conformational space). Carbons or atom groups with one or more hydrogen atoms attached were considered potential SOMs. The closest distance between these hydrogen atoms and the heme iron was used to identify the likely SOMs. The optimum of that distance was determined to be 2.7 Å. Several different docking and scoring protocols were explored, obtaining AUC (area under the ROC curve) values of up to 0.93 for an independent test set.

IDSite[84] is a combined approach based on the Schrödinger software GLIDE, Protein Local Optimization Program (PLOP, implemented to the Schrödinger software suite as the Prime[85] package) and a reactivity model derived using Jaguar.[86] First, the ligand is placed into the CYP active site using docking (GLIDE). Then, two distinct refinements are carried out using PLOP. This includes both the refinement of protein side-chains and the conformation and orientation of the ligand itself. Poses are filtered considering their structures and energies and clustering by similarity of the ligand conformations is performed. In the last phase, the best poses are collected on the basis of a physical score, which considers the intrinsic chemical reactivity (based on a reactivity model derived from DFT calculations employing the B3LYP functional with the 6-31G* basis set) of potential SOM as well as the energy of the ligand poses. Its application on CYP2D6 has been reported recently[84] and models for other CYP isoforms are currently under development. For a test set of 56 CYP2D6 substrates, the correct SOM could be recovered for 83% of all cases. When using a training set of 36 compounds to fit four parameters (rescaling of the PLOP energy with two parameters, fitting radical intrinsic reactivities and score cutoff), 94% of the known SOMs of a test set of 20 compounds were successfully identified. Computational power appears to be limiting the application of this approach. The authors report a calculation time of approximately 450 h on a single 2.2 GHz AMD Opteron 6174 processor for a compound with three rotatable bonds.

■ PREDICTING XENOBIOTIC METABOLITES

The identification of metabolites of small organic molecules contributes to the understanding of ADME and pharmacological properties. Experimental identification of metabolites is highly demanding in infrastructural resources, expert knowledge, and time. If enough material is available, NMR allows structure elucidation of metabolites. Mass spectrometry (MS) is particularly valuable for the analysis of small amounts of material. Most common experiments are liquid chromatography/mass spectrometry (LC/MS) or liquid chromatography/tandem mass spectrometry (LC/MS/MS), which allows the identification of moieties of a molecule where the biotransformation has taken place (even if they do not always allow for the identification of SOMs). These experimental methods can decisively benefit from computational approaches supporting data analysis to pinpoint the actual SOM. For extensive reviews on these experimental approaches the reader is referred to refs 57, 59, 87.

While numerous computational methods aim to predict SOMs, only a very few approaches have been developed so far that allow prediction of the products of xenobiotic metabolism. Prediction methods are dominated by expert systems. Additionally, a fingerprint-based data mining approach is available and very recently an SVM-based technique has been reported, which, however, focuses exclusively on endogenous metabolites (see below).

Expert Systems. A number of expert systems have been developed that aim to predict the sites and products of metabolism using dictionaries of biotransformations. These approaches are based on knowledge rules, which are typically developed from a reservoir of *in vivo* and *in vitro* experimental results. They are inherently subjective in their nature as they are based on “expert” human knowledge rather than “robust and objective computational estimates”, although given the approximations involved in most computational methods, these can of course be poor substitutes for the real experimental conditions. So, expert systems have a strong history in this area and they can help bridge the gap between empirical knowledge and hard data. Potential metabolites are identified by searching a query molecule for the presence of a target fragment and converting this into a product fragment, as defined by a biotransformation dictionary.

The main drawback with this approach is the generation of a combinatorial explosion of predictions, as all possible combinations of transformations permitted by the dictionary are considered. Therefore, a key challenge is to prioritize the results and to stop when it is deemed that the molecule has been transformed into a sufficiently water-soluble state to enable excretion from the body. A further caveat is that most expert systems tend to contain combined data and rules for many different mammalian species and so are useful as a general indicator of biotransformations in the “average” mammal. However, metabolic pathways can be very different even in closely related mammalian species, so some expert systems have been developed to allow filtering of specific subsets of the data to a specific species.

Examples of expert systems include MetabolExpert,[88] META,[89] Meteor,[90] University of Minnesota Pathway Prediction System (UM-PPS),[91] Systematic Generation of potential Metabolites (SyGMa),[92] and TImes MEtabolism Simulator (TIMES),[93] as described below. They typically include functionality for the user to amend or add new rules to tailor the program to their specific needs and typically display the results as a tree of metabolites depicting multiple pathways and molecular structures. The main differentiating features apart from different data sources and knowledge rules include the methodologies used to prioritize predictions and terminate the chain of predictions.

MetabolExpert was one of the pioneering attempts at a metabolic site predictor, created in 1985. The knowledge database contains rules, which include substrate and metabolite listings but also contain lists of substructures, which inhibit or promote the reaction. The system lists all potential metabolites without applying any ranking system and predictions are terminated once known phase II metabolites are generated. Features of this system include the presentation of the results in a dendrogram showing chemical formulas along with hydrophobicity values and the generation of first- and second-order kinetics data. Additional enhancements include modules specific to animal and plant metabolism and soil degradation.

META uses a larger dictionary of biotransformations, each assigned with a priority enabling the metabolites to be arranged hierarchically to manage the combinatorial explosion problem. A genetic algorithm is used to prioritize the transformations dictionary. Additionally, this system uses quantum mechanical descriptors to test a proposed metabolite for stability and convert to a stable conformation if necessary. Further metabolic products are only generated after a stable conformation has been achieved. Subsets of the metabolic dictionary can be chosen to allow focus on specific areas such as mammalian metabolism, aerobic and anaerobic biodegradation.

Meteor extends the concept of biotransformation rules by using a structure representation language to describe atoms and bonds and is able to include descriptors such as charge and valency. This enables a more sophisticated description of the activating biophore rather than just functional groups. Overprediction is countered by classifying transformations as probable, plausible, equivocal, doubted or improbable and the use of relative reasoning rules to prioritize potentially competing reactions. Predictions are terminated once an external logP predictor deems that the molecule has become sufficiently water-soluble to be excreted. The biotransformation probabilities fall with decreasing lipophilicity as the ability of a molecule to cross membrane boundaries and undergo further biotransformations is reduced. Additional features of Meteor include a link to a metabolism database where examples of the biotransformation in the literature are reported with the ability to filter on sequence length, enzyme and species. The creators of Meteor note that despite the strategies described above Meteor still tends to generate a high volume of false positives. The results generated should be regarded as potential metabolites requiring further analysis, an observation that is generally true of all expert systems. They also make the observation that biological factors are interdependent in a nonlinear way rendering it a difficult problem to provide fail-safe rules to encompass all metabolic transformations.

UM-PPS uses a classification system similar to Meteor and biotransformations are classified as either very likely, likely, neutral, unlikely, or very unlikely. Overprediction remains a problem and further relative reasoning rules are used to rank and prioritize predictions. The system is specific to microbial catabolic metabolism and uses the University of Minnesota Biocatalysis/Biodegradation Database (UM-BBD)[91] as a source to generate the rules. The creators make similar observations to Meteor, that definitive prediction rules applied to a wide region of chemical space are hard to generate. They note that predictions work best where query molecules show similarity to those already in the database and where reaction conditions and concentrations are similar to those documented. The system tends to produce a wide range of predictions and so is more suited to general environmental metabolism rather than when applied to a specific microbial organism.

Some other systems use statistical analysis to evaluate a set of biotransformation rules when applied to a data set. SyGMa contains rules derived from inspection of the Accelrys Metabolite Database with each rule assigned a probability score relating to the number of correct predictions when applied to the database. Metabolites are assigned the probability score of the rule from which they were formed, allowing them to be ranked. This enables a more finely grained view of probabilities than the higher-level categorization performed in Meteor and UM-PPS. To allow for greater differentiation and specificity in predictions the rules are broken down into subsets and applied to specific chemical reaction series.

TIMES[93] utilizes a comprehensive library of biotransformations and employs a heuristic algorithm to generate plausible metabolic maps. Reaction rates from systematic and toxicity testing are used to generate transformation likelihoods, otherwise a combinatorial algorithm is used to translate known metabolic maps taken from reference systems into best-fit transformation probabilities.

Metabolizer[94] is a commercially available metabolism prediction module which enumerates all possible metabolites of a given compound and allows a prognosis to be made on metabolic pathways, major metabolites and metabolic stability. Besides human metabolism, it supports predictions on various species including rat, bacteria, and plants.

Fingerprint-Based Data Mining Approaches. MetaPrint2D-React[21] is an extension of MetaPrint2D (see "Predicting Sites of Metabolism") that allows prediction of the metabolites that are likely to be formed from a transformation. SMARTS patterns are used to classify the transformations on the basis of structural changes between substrate and metabolite and occurrence counts for each type of transformation are recorded. A separate occurrence ratio is calculated for each type of transformation and the structures of predicted metabolites are generated through reaction rules using SMIRKS patterns. Predictions can be selected from human, dog, rat or all species.

Combined Approaches. X-ray structures of CYPs indicate that the position and orientation of substrates and their metabolites are largely preserved. Consequently, metabolites would be expected to match the chemical and geometric constraints of the binding site, which would make them distinguishable from decoy molecules for docking algorithms. This is the hypothesis in recent work by Tarcsay et al.,[95] who used the expert system MetabolExpert to generate putative metabolites from known substrates. The docking program GLIDE[53] was subsequently employed to reduce the false positive rate. MetabolExpert was found to produce 74% of the known metabolites using the default setup and 82% following an enhanced rule set. Using the best setup for the combination of MetabolExpert with docking as a postprocessing filter, Tarcsay et al. report that their method is able to identify the correct metabolite among the three highest-ranked structures in 69% of all cases.

Also MetaSite allows the generation of likely metabolites[56] for selected CYP isoforms (see "Combined Approaches" in section "Predicting Sites of Metabolism").

■ THE CYP SUPERFAMILY OF METABOLIZING ENZYMES

The CYP family of monooxygenase enzymes facilitate phase I metabolism of endogenous and xenobiotic chemicals in many organisms and bacteria. CYPs are classified into families and subfamilies. Each family is indexed with a digit, a subfamily with a letter and an individual gene with a digit. In humans, many CYP isoforms exist, with CYP1A2, 2C9, 2D6, 2C19, 3A4, and 2E1 having been identified as playing the most significant roles in drug metabolism.

Structure of CYPs. CYPs are predominantly composed of α-helices, labeled A to L, commencing from the N terminus (Figure 3). Helices F and G form the roof of the embedded active site and between the I and L helices, lies the heme B prosthetic group, which is also termed iron protoporphyrin IX. This is present in all CYPs and provides the proximal ligand to the heme via the sulfur on the cysteine. Protoporphyrin IX and a small group of neighboring residues form a heme cradle region, which is highly conserved.

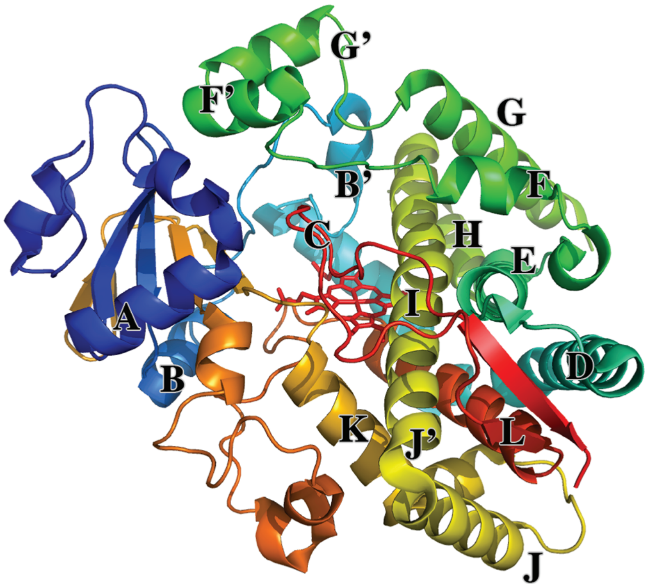


Figure 3. Structural overview of CYPs. Here, CYP3A4 (PDB 3NXU) is depicted, with helical structures labeled A–L according to the general scheme for CYPs.[96] The heme moiety is indicated in sticks mode (red color).

Helix I has a highly conserved kink above the active site, which may facilitate proton transfer. Otyepka et al.[96] have described mammalian (families 2 and 3) CYP isoforms on the basis of available crystal structures. An overview of available X-ray structures for CYPs is provided in ref 97.

Generally, the available crystal structures have had two sets of mutations carried out to facilitate their crystallization and purification. Helices F′, G′, F/G loop, and the N terminus putatively make contact with the membrane region; hence, a large hydrophobic transmembrane anchor at the N terminus is replaced with a shorter, more hydrophilic chain. Second, a four to five sequence polyhistidine-tag is inserted at the C-terminus. It is generally accepted that such changes do not modify the reactivity properties of the enzyme.[98]

The active site volume of CYP isoforms varies as a function of isoform. Cruciani et al.[99] investigated the structural characteristics of the binding sites of various isoforms using GRID. They point out the remarkable differences of the active site volumes computed using a hydrogen probe. While CYP1A2 shows a site volume of 630 Å^3, the binding site of the CYP3A4 model extends to 1500 Å^3. They also ranked CYP isoforms by the prevalence of hydrophilic regions in the binding pockets as CYP2D6 > 3A4 > 1A2 > 2C19 > 2C9 and highlighted the strong dependency of the hydrophilic patterns on protein flexibility. Otyepka et al.[96] partitioned the mammalian CYPs into three categories, starting with the largest: CYP2C5, 2C8, 2C9, 3A4 > 2B4, 2D6 > 2A6.

Various pathways from the interior active site, leading to the solvent exterior have been identified. For a comprehensive study making use of the CAVER[100] tool and a naming scheme of these pathways the reader is referred to ref 101.

Reactivity of CYPs. Upon arrival at the active site, a substrate will undergo chemical modification facilitated by the heme iron's unusual transient electronic state and the protein environment. However, there is a set order to this process. CYP functions within a catalytic cycle, with iron undergoing changes in its spin state, oxidation number and ligand coordination number as it proceeds around the ordered cycle. For more detail, the reader is referred to the review of Shaik et al.[102] The currently accepted mechanism is as follows. In its first state, the resting state, Fe(III) is in a low spin, doublet state and is hexacoordinated, with a water molecule occupying its distal axial position. Arrival of a substrate displaces the axial water leading to the formation of a pentacoordinate Fe(III) with a sextet state. The change in redox potential associated with this displacement allows it to accept an electron from a reductase, nicotinamide adenine dinucleotide phosphate (NADPH)-P450, thus forming the third state, Fe(II) that has a high spin state. Molecular oxygen is then bound, yielding state four, a singlet, oxy-ferrous complex. This singlet complex, readily accepts a second electron from the reductase, resulting in the formation of state five, a ferric-peroxo anion. The protein environment facilitates the shuttling of a proton to form state six, which is also referred to as Compound 0.

A second proton is accepted, followed by the heterolytic cleavage of the molecular oxygen and loss of a water molecule to form state seven, the electrophilic, high valence, iron-oxo compound. This is the active species, analogously termed Compound I, which then oxygenates the substrate. Next, the product leaves the active site and a water molecule returns to the distal axial position, completing the cycle.

Compound I is a ferryl-porphyrin-pi cation radical and has three unpaired electrons (a triradicialoid). Two unpaired electrons are localized on the ferryl group and the other is shared between the sulfur on the covalently bonded cysteine and the porphyrin. It is similar in all isoforms, and because of its electronic structure being a function of its protein environment, has earned itself the title of "Chemical Chameleon".[103,104] Recent electron paramagnetic resonance (EPR) studies have managed to provide experimental evidence for the long sought intermediate, Compound I.[105]

Molecular Dynamics Simulations on the Structure of CYPs. Classical MD simulations of various CYP isoforms have been used to gain insight into the proteins' flexibility, investigate channels into the active site, and generate ensembles of structures for higher level QM and QM/MM calculations. This first section focuses primarily on classical simulations used to specifically study properties arising from system dynamics. Simulations employed as a preamble to generating starting QM/MM structures will be covered in the subsequent section.

Classical Parameters Developed for CYPs. Most modern classic biological force fields have parameters for residues and base pairs found in proteins and DNA; hence, modeling a similar system can be easily accomplished. Entities that lie outside of this chemical space, such as small organic molecules or unusual residues, need special attention and additional preprocessing to develop missing parameters. Usually, there is a force field specific protocol for parameter development to ensure that the new parameters are consistent with the rest of the existing force field. Generally, parameter development focuses on bonded and electrostatic terms. A corollary to this is that a parametrization for a moiety within a specific force field is not transferable to a different force field.

The heme B prosthetic found in CYPs is such a group and various parameters for different force fields have been developed over time. A further complication to this arises due to iron's changing electronic state as it proceeds around the CYP catalytic cycle, since each state will have different classical parameters.[102] A range of parametrizations for the heme B prosthetic in CYPs have been carried out, some exclusively as a parametrization process, others for use within a specific study (Table 2).

In their study of the interaction of the P450cam isoform with valproic acid, Collins et al.[106] developed a set of heme B parameters for use with AMBER 3a. Schöneboom et al.[103] derived Chemistry at HARvard Molecular Mechanics (CHARMM)[107,108] parameters for Compound I on the basis of existing CHARMM heme parameters, by adding axial ligands and updating charge parameters. Bathelt et al.[109] carried out a similar process with the CHARMM22 parameter set. Oda et al.[110] developed AMBER parameters for the pentacoordinate Fe(III) in the sextet state, and Seifert et al.[111] reported AMBER parameters for CYP in the Compound I state. Favila et al.[112] presented improved heme B parameters to carry out MD simulations on CYP2C9 and CYP19 (also known as aromatase).

Although not CYP parameters, the classical force field parameters for the heme C prosthetic developed by Authenrith et al.[113] for both the CHARMM and AMBER force fields provided a foundation for derived CYP parameters by Skopalik et al.[114] (resting state) and Favia et al.[112] (ferrous low spin state).

Shahrokh et al.[115] have recently developed a set of AMBER compatible CYP parameters, which are readily accessible in their Supporting Information, for a range of heme states found in its catalytic cycle.

Classical Simulations of CYPs. The majority of simulations to date have been carried out with either the AMBER or CHARMM force field, although some have also utilized variants of the GROMOS or OPLS force field. All simulations have made use of explicit water solvent and followed a standard MD protocol. Recent simulations are on average 20 ns long, and in a few cases, are an order of magnitude greater than this. However, because simulation length is largely dictated by available computational power, older simulations are shorter. Generally, the simulations described here make use of a standard protocol for condensed phase biological MD as outlined below.

A crystal structure usually provides starting coordinates, but if there are missing residues or if a crystal structure is unavailable, homology modeling methods may be used. Missing hydrogens are then added. The correct protonation state of certain residues at physiological pH is assigned according to estimation of local pK_a before finally assigning force field terms to the system. Counterions are added to neutralize charge in the system and to provide a solvent salt concentration that reproduces physiological or experimental conditions as appropriate. The system is placed under periodic boundary conditions and Particle Mesh Ewald (PME) is used for accurate treatment of long-range electrostatic interactions. Next, energy minimization is carried out to relax any bad contacts between protein and solvent, followed by short thermalization (allowing the system to reach the desired temperature) and equilibration MD phases, prior to the production simulation. In some cases, extra biasing potentials or force field modifications may be introduced, for example, to calculate the thermodynamics of CYP binding site solvation and ligand binding, or to delineate substrate access pathways.

During normal production dynamics, the backbone RMSD relative to the original experimental or modeled structure may be monitored over time to check that the protein is stable under the influence of the simulation force field. Analysis approaches may include the calculation of the per-residue root-mean-squared fluctuation (RMSF), which provides a measure of backbone mobility and can be directly compared to crystallographically derived temperature factors. Large concerted motions may be identified using principal component analysis (PCA; the result of diagonalization of a covariance matrix that has been calculated from the dynamics), which filters out rapid thermal atomic fluctuations. Finally, the molecular details and energetic interactions between protein, ligand, and solvent may be analyzed, supplemented by visual analysis of the trajectory.

Early simulation studies focused on the local structure/dynamic relationships associated with alternative ligand binding modes in CYP isoforms. Wade and et al. created a series of camphor analogues designed to fill an empty region of the substrate binding pocket of P450cam and a mutant, and experimentally measured thermodynamic properties of binding were rationalized via substrate–protein/solvent interactions observed in MD simulations.[116] The same group applied multicopy MD simulations, in which several ligands are simulated in the presence of a single protein to locally enhance sampling, to explore multiple ligand binding modes of different enantiomers of camphor in P450cam, revealing differences in binding modes for similar but chemically different ligands.[117] The structural reorganization associated with differential substrate selectivity in CYP17 isoforms from various species has also been investigated via MD simulation approaches.[118] Moreover, simulations of a CYP2A4 model has helped to rationalize key interactions with active site residues shown via mutation to determine specificity for testosterone metabolism (in contrast with CYP2A5).[119] Strobel et al.[120] used simulations to provide a molecular explanation for the observation that inactivators are able to discriminate between the closely related CYP P450 isoforms 2B4 and 2B5 in terms of differential stabilities of inhibitor binding orientation.

Several simulation studies have sought to provide a molecular basis for the promiscuous nature of substrate specificity exhibited by CYPs in terms of their dynamics. Park et al.[121] used nanosecond-time scale MD simulations of CYP3A4 to propose that the high-amplitude flexibility of a loop near to an egress channel in the apo state (but damped when bound to a substrate or inhibitor) may be responsible for the enzyme's

Table 2. Overview of MD Simulation Setups for CYPs

Ref	FF used	Isoforms	Heme B parameters used	Simulation time/ns	Production ensemble	Comments
111	AMBER/FF99	CYP2C9	Developed within their study	5 (multiple)	NPT	
114	AMBER/FF99	CYP2A6, 2C9, 3A4	Own charge method with parameters taken from ref 113	5	NPT	FF99 used, missing residues in CYP3A4 (1TQN)
112	AMBER/FF99	CYP2C9, CYP19	refs 106, 113 Ferrous low spin state (S = 0) Fit of heme charges done at DFT (B3LYP/6-31G*) level	10	NPT	Cα RMSD plateaus in phases, culminating at ~4 Å DFT wave function used for RESP fitting HEME parameters not stated, only the RESP charges
128	CHARMM 27	CYP3A4	refs 103, 109	6	NVT	Cα RMSD plateaus at ~1.5 Å
130	AMBER/FF99	CYP1A2, 2A6, 2C9, 2D6, 3A4	Developed within the study	10	NPT	FF99 used, missing residues in 3A4 (1TQN) Heme parameters provided
140	GROMOS96	CYPC17	GROMOS96	20	NPT	Cα RMSD seems high at ~3 Å Berendsen thermostat
137	CHARMM 27 (protein) CHARMM 22 (heme)	CYP3A4	CHARMM22 Refit of heme charges done at DFT (B3LYP/MM/LACVP) level for sextet state	20	NVT	Cα RMSD plateaus at ~1.5 Å Thorough, with a rich set of supporting materials
139	GROMOS96	CYP3A4	–	5	NPT	Cα RMSD plateaus at ~2.5 Å
124	AMBER/FF03 with pi-pi stacking term modification	CYP2C9	Developed within their study, but used the method of ref 110	5	NPT	Cα RMSD plateaus at ~2.4 Å
169	CHARMM27	CYP51	–	20	NPT	Cα RMSD not reported
115	AMBER/FF99SB[267]	CYP3A4	Heme parameters for a wide range of states developed with this study.	25	NPT	Cα RMSD plateaus at ~1.7 Å
122	OPLS	CYP119	Heme parameters from ref 268	200	NVT	RMSD of F/G loop region indicates a large degree of flexibility. Berendsen thermostat
157	GROMOS 45A4	CYP1A2	Not stated.	1	NPT	
158	AMBER/FF99	CYP26A1 and CYP26B1	Not stated	0.5	NVT	
132	CHARM22	CYP101	Heme parameters from ref 269	300	NPT	Longest all-atom simulation to date

broad specificity. More recently, Lampe et al.[122] generated 200 ns trajectories of the thermophilic P450 CYP119 enzyme, providing evidence for a flexible conformational selection mechanism in substrate binding, and revealing from an analysis of correlated motions that the active site moves relatively independently from the rest of the protein, leading to "insulation" from external dynamics.[123] Seifert et al.[111] examined the interaction of CYP2C9 with its substrate, warfarin, using an AMBER FF99 model with their own heme B parameters. Multiple 5 ns trajectories were run and they observed a stable protein core and a mobile surface that gave rise to two channels, and concluded that this leads to a broad substrate profile with high regioselectivity. Sano et al.[124] studied three polymorphisms of CYP2C9 which involved the mutation of a single amino acid residue, using MD and docking methods, and related catalytic activity to deformations of this pocket.

Skopalík et al.[114] carried out simulations on three members of CYP3A4, 2C9, and 2A6. In addition to the standard MD simulations of each isoform, a MD run was carried out at 398 K. High temperature simulations enable a more extensive sampling of protein dynamics, and have previously been used, for example, to understand the origins of stability of the thermophilic and barophilic CYP119 enzyme via nanosecond-time scale simulations.[125] An analysis of the dynamics of the three isoforms revealed that CYP malleability is directly related to substrate specificity. CYP2A6 was the most rigid investigated CYP and consequently has a narrow range of substrates, while CYP3A4 was the most flexible and most promiscuous analyzed CYP. They also identified ten flexible regions and a rigid core as well as a malleable distal side and a less flexible proximal side. Such observations could not be concluded from previous X-ray structure analyses, as a comparison of calculated and experimental B-factors confirmed that protein dynamics are damped in the presence of crystal lattice contacts.

CYP substrate specificity and associated conformational flexibility has also been investigated via combined simulation and experimental approaches. A flexible loop around the active site has been observed in a CYP2C9 mutant, leading to a more open conformation consistent with spectroscopic data.[126] In another study, NMR-derived residual dipolar couplings were incorporated via torsional restraints within MD simulations of P450cam to identify a more open and accessible active site with a wider range of substrate orientations than those observed crystallographically.[127] Sigmoidal behavior in kinetic studies led Fishelovitch et al.[128] to examine the possibility of cooperative binding between two substrates. They modeled diazepam in CYP3A4, comparing simulations of this with the substrate-free state, and concluded that F304 on the proximal I helix plays a key role in the mechanism of co-operativity between substrate molecules.

Significant effort has been devoted to understanding the properties of internal CYP solvation. While the volumes of the active sites of different CYPs vary significantly, they have been shown to contain water that must be displaced upon substrate binding, and this must be considered when attempting to calculate the binding free energy of a particular substrate. Moreover, the reactivity of the active site is dependent upon water. Displacement of water favors the ferrous Fe(II) state and prevents electron uncoupling,[129] while water "wires" may relay proton transfers needed for heterolytic cleavage of dioxygen. Only limited information regarding water occupancy may be extracted from crystal structures, necessitating complementary simulation approaches. Several factors contribute to this issue, such as conformational flexibility of the protein, the difficulty of cryoprotection schemes, as well as the disordered nature of solvent molecules.

The simulations of Skopalík et al.[114] enabled observation of water exchange between the active site and bulk solvent, with the number of water molecules reaching a dynamic equilibrium after 1.5 ns. Analysis of access/egress channels (pw2e and pw2c) identified a solvent channel. In a more recent analysis, Hendrychová et al.[130] examined the effect of high pressures (300 MPa) upon CYP1A2, 2A6, 2C9, 2D6, and 3A4 both by experimental spectroscopy and molecular dynamics simulation. Using a similar protocol as in their previous paper, they found that with increased pressure, the flexibility of the CYPs decreased and the number of water molecules in the active site increased, with the exception of CYP3A4. They also concluded that of the five CYPs examined, CYP2A6 and 1A2 have the least malleable active sites.

Because of the relatively long time scales required to observe water exchange between internal cavities and bulk solvent, Helms and Wade[131] implemented a pioneering "alchemical" MD approach, in which a water molecule is gradually annihilated in bulk solvent and another is inserted into the protein cavity, enabling calculation of the relative free energy of hydration of the CYPcam active site via a simple thermodynamic cycle. This revealed that hydration by around six water molecules was thermodynamically most favorable, in agreement with experimental estimates, suggesting the active site water density is about half that of bulk solvent. Recent improvements in computational power enabled Miao and Baudry[132] to observe the unbiased water diffusion into and out of the active of P450cam over a 300 ns trajectory, confirming that on average, six water molecules are found in the camphor-free active site. However, between 0 and 12 water molecules could be present at any given time because solvent exchanged rapidly with bulk solvent via multiple channels identified with the CAVER[100] tool. Rydberg et al.[133] also showed that active-site waters could exchange with bulk during 4 ns simulations of six mammalian CYPs. The mean number of waters and exchange pathways varied significantly between isoforms, consistent with the variability in active site volume.[99]

MD simulations have also been utilized to study the effect of organic cosolvent molecules such as DMSO on the structure and dynamics of CYPs[134,135] for commercial exploitation, since many industrially important substrates are poorly soluble in water. Moreover, a series of 1 ns MD trajectories have been used to identify docking sites for CO binding to P450cam, rationalizing laser flash photolysis data.[136]

Because the active sites of the CYP, enzymes are substantially buried within the protein structure, specificity and kinetics are likely to be directly linked to the passage of water, substrates, and products into and out of the protein. For this reason, much focus has been placed on the identification of static or dynamic channels that connect the active site to the protein surface. Available crystal structures yield a wealth of information regarding access/egress channels,[101] but simulation approaches can complement this by providing insight into the dynamics and energetics of substrate passage. For example, Fishelovitch et al.[137] identified a gated water pathway and an additional substrate channel in CYP3A4 using their MolAxis tool.[138] A second MolAxis channel study on the same isoform[139] examined cooperative binding of ketoconazole and ligand-induced conformations changes. The authors also identified various bottleneck residues, and concluded that two channels could serve as egress routes for the substrate ketoconazole. Haider et al.[140] identified three channels that converged on the heme in a homology model of CYPC17, identified using the MOLE tool.[141] MD simulations demonstrated different substrate binding modes, and a correlation between low vibrational modes and active site gating was demonstrated. Accelerated MD approaches have also been developed to lower the energy barriers between different crystallographically identified open and closed conformational states of P450cam, providing insight into the likely mechanisms of ligand binding.[142]

The direct observation of ligand/product entry/exit is generally not possible using typical MD simulations due to limitations in accessible time scales. As a result, a variety of "biased" simulation approaches have been utilized to enhance sampling of ligand egress within CYPs.[143] The Wade group developed the random expulsion MD method, involving the application of a randomly oriented artificial force to an initially bound ligand, to identify channels in three bacterial CYPs.[144] Steered MD, in which a constant force is applied to a virtual spring attached to a molecule along a predefined vector, was subsequently used to drag ligands along each channel. A minimization-based approach identified a channel common to all three bacterial CYPs—with specific channel features in each isoform adapted for their respective substrate/product specificity—as the most likely route of ligand passage.[145] A similar approach was applied to a mammalian CYP2C5 enzyme and used to propose that hydrophilic substrates and products may pass through one channel, whereas lipophilic substrates may enter via a channel embedded within the membrane.[146] This was recently supported by the first reported MD simulations of a full-length mammalian CYP inserted within a lipid bilayer.[147] Coarse grained simulations were used to efficiently sample various orientations of the membrane-bound CYP, providing the starting point for subsequent all-atom simulations. Correlations were found in the permittivity of putative ingress and egress tunnels, and the presence of the membrane stabilized the open state of an internal aromatic gate in the previously identified channel connecting the active site to the bilayer phase.

The steered MD method and its variants have been applied to a number of other CYP systems including CYP2B1[148,149] and CYP3A4,[150−152] enabling the identification of possible channels and associated rupture forces during ligand passage, and to probe interactions with key residues identified by site-directed mutagenesis. As it becomes easier to generate longer time scales and steered MD simulation ensembles, it has even become possible to estimate the free energy pathways associated with ligand entry and egress.[153]

Molecular Dynamics Approaches for the Calculation of Affinity of Substrates and Metabolites. As well as providing qualitative information concerning protein−ligand interactions, a number of methods within the MD simulation framework may be utilized to directly estimate ligand binding energies. Although this tends to involve greater computational cost than, for example, docking, the results are normally more reliable due to the use of an accurate physical force field and extensive conformational sampling (if sufficient starting points are included). In more or less all the approaches, calculation of the ligand binding free energy (ΔG) relies on the observation that its value is a state function and is hence dependent only on its initial energy in solution and final energy following binding. Thus, a thermodynamic cycle may effectively be invoked, considering either the "transfer" of a ligand into solution and into the protein binding site, yielding the absolute binding energy, or the transfer of two different ligands to the binding site, yielding the relative binding energy.

For example, Åqvist et al.[154] devised the semi-empirical linear interaction energy (LIE) technique to estimate ligand-protein binding energies. In contrast with some more rigorous approaches (see below), only the physically relevant protein-bound and free solvated states of the ligand (i.e., the "corners" of the thermodynamic cycle) are considered with LIE, necessitating just two MD or Monte Carlo simulations.[155] From the equilibrium ensemble of nonbonded interaction energies between the ligand and its surroundings for these two states, ΔG may be calculated as an empirically weighted difference between the average bound and unbound energies. There are separate scaling coefficients for the nonpolar and electrostatic interactions and an extra independent constant may be included. While default values for these parameters are available, they are often optimized using a training set. Good correlation between experimental and calculated binding energies has been demonstrated in many cases when utilizing LIE, particularly for related series of ligands, and as early as 1996, it was shown that the approach could be successfully applied to the estimation of relative and absolute binding free energies for a range of substrates in P450cam.[156]

Vasanthanathan et al.[157] investigated ligand binding affinity of thirteen ligands with CYP1A2 using a LIE-based method. GOLD was initially used to place the ligand in the active site, followed by a standard MD protocol using the GROMOS 45A4 force field, totaling 1 ns of NPT production dynamics. From this trajectory ΔG was calculated using LIE. For comparison and validation, experimental IC_{50} values were converted to equilibrium constants (K_i), which in turn were used to estimate ΔG. They demonstrated a linear correlation between experimental and calculated values with a RMSD of 2.1 kJ/mol, and concluded that the van der Waals interactions dominated over the electrostatic interactions.

Karlsson et al.[158] constructed homology models of 26A1 and 26B1 and various ligands, including five retinoic acid metabolizing blocking agents (RAMBAs), which were docked into each respective site, followed by 0.5 ns of NVT MD simulation. The LIE method was used to calculate values of ΔG, which were compared to experimental IC_{50} values. The authors concluded that the homology models could distinguish between strong and weak inhibitors and identified various important residues in the active site.

Stjernschantz et al.[159] developed an iterative protocol employing multiple MD simulations accounting for an ensemble of generated docking poses. The approach was tested on a set of 12 known binders of CYP2C9, and the RMSD for predicted binding affinity was 2.9 kJ/mol. It was shown that this technique is able to lower the impact of the initial pose selection and it also takes into account the possibility of multiple binding modes.

Despite these reports, which often demonstrated impressive correlation between experimental and calculated ΔG, the inherently empirical nature of LIE may be undesirable in some cases. An alternative is to use "computational alchemy", which includes related methods such as free energy perturbation (FEP) and thermodynamic integration (TI). Such formally correct approaches tend to be computationally demanding but have the capacity to yield highly accurate relative or absolute ligand binding free energies, in principle limited only by the quality of the underlying force field and sufficient conformational sampling.[160]

The key feature of alchemical methods is that the configurations sampled between different states should have a significant degree of "overlap".[161] Thus, unlike in LIE where two ensembles must be generated, a multistep approach must be used, in which a path between the "corners" of the thermodynamic cycle (see above) is sampled. Practically, this is normally implemented by introducing a series of intermediate potential energy functions constructed as a linear combination of end states. Thus, one ligand is alchemically transformed to another (to yield the relative ΔG) or a ligand is alchemically "annihilated" (for the absolute ΔG).

Wade et al.[131] utilized an alchemical MD approach to calculate the relative ΔG of water molecule binding to the CYPcam active site and later extended this work to couple the exchange of six high-occupancy water molecules to the alchemical transformation of camphor inside the active site, yielding an accurate measure of its ligand binding ΔG.[162] Deng and Roux[161] pointed out that this strategy required prior knowledge of the solvation properties of the cavity in question. They therefore recently implemented a novel approach combining alchemical MD for ligand-binding ΔG calculation, with a grand canonical Monte Carlo algorithm, enabling the number of waters to vary freely during the simulation. This yielded extremely accurate estimations for the thermodynamics of camphor binding to P450cam,[163] and the method should be generalizable to any CYP site. It is clear that careful consideration of CYP solvation properties is necessary, even when combining MD with more empirical approaches to binding ΔG calculations.[157]

QM/MM Approaches with CYPs. Classical simulations can yield information about dynamics and conformation changes. However, this level of theory does not have the ability to represent bond breaking or making processes, whereas a richer quantum mechanical treatment of all the electrons within the system is still too computationally demanding.

A common approach is to partition the calculation into two regions that are treated at different levels of theory. Generally, a small quantum mechanical region is defined around the active site, where bond breaking is presumed to occur, and the rest of the system, usually the rest of the protein and bulk solvent, is treated classically. The two regions are not independent but are coupled. For example, the wave function in the QM region will be polarized by the influence of the point charges in the classical MM region, and conversely, the MM region will be influenced by a nonstatic point charge distribution emanating from the QM region. Many coupling methods exist and the QM/MM approach can be used to calculate single point potential energies or gradients can be evaluated for the system to be propagated dynamically (QM/MM MD).

The Thiel group has published several excellent papers characterizing the dynamics and catalytic mechanism of CYPs. Their work includes investigations on polarization and hydrogen bonding effects of the protein environment on Compound I.[103] They carried out 215 ps of MD using their protoporphyrin IX parameters to generate snapshots for subsequent QM/MM calculations and concluded that it transformed from a sulfur-centered radical to a porphyrin-centered radical cation. Further studies from this group include the QM/MM-based analysis of the C−H hydroxylation mechanism[164] and the theoretical investigation of the Compound I reactive intermediate.[165] This and additional work is discussed in their comprehensive review[102] on the structural properties, reactivity, and substrate selectivity of CYPs analyzed by QM/MM approaches, and therefore, we provide here a brief summary on this topic.

Starting configurations are usually generated from a classical MD simulation. A QM/MM hybrid approach is used whereby a truncated porphyrin (both propionate groups are removed) and a truncated cysteine (S) is set as the QM region. In the QM region, an UB3LYP method is used with the Los Alamos effective core potential and associated triple-ζ contracted basis, LACVP, on Fe and the 6-31G(d) basis on the remaining QM atoms.[166]

Bathelt et al.[109] examined the electronic properties of Compound I and concluded the lowest electronic state to be a near degenerate pair of doublet/quartet optimizations of A2u states, with two unpaired electrons localized on the Fe−O and the other delocalized on the porphyrin and thiolate ligand. In a later study,[167] they examined the mechanism of benzene hydroxylation facilitated by CYP2C9 and found that the energy barrier for the addition step is comparable to the experimental rate constants of the same family member, CYP2E1. They also compared the energetics of benzene approach toward the side or face (relative to the porphyrin plane) and concluded that these arene addition pathways are both accessible.

Fishelovitch et al.[168] explored the effect of a substrate on Compound I, showing that in a substrate-less Compound I state, the Fe−S bond elongated and localized the radical on the sulfur. Conversely, the presence of substrate shortened this bond.

Sen et al.[169] studied the dynamics of CYP51 with specific reference to the proton shuttling involved in the molecular oxygen activation. They conclude that the protein environment is important in tuning the electronic structure of intermediates and highlighted the range of proton shuttling routes within the CYP family. The study also demonstrates opportunities for experimental validation because much experimental data is available for the peroxo intermediate.

Lonsdale et al.[170] investigated the chemoselectivity of alkene oxidation with CYP101, finding that QM/MM reaction barrier heights were in qualitative agreement with experimental selectivities. Given the accessibility of both hydroxylation and epoxidation to the ferryl oxygen, it is the relative reactivity of Compound I that determines the relative ratio of products. Attention is also drawn to the complexities of such modeling and the need for a wide ensemble of starting geometries for QM/MM approaches. In a different study,[171] the same authors demonstrate the importance of a dispersion correction for B3LYP methods in the study of CYPs. Their work showed that inclusion of dispersion has a significant effect on the energies and geometries of transition states and encounter complexes. Furthermore, an improved agreement with experimental data was observed compared to their original CYP101 study.[170]

Li et al.[172] conducted a QM/MM study of CYP2A6 on (S)-(-)-nicotine. From calculating a minimum QM/MM energy path they determined a reaction mechanism: initial abstraction from the 5′-position to the oxygen on Compound I followed by an O-rebound step, the recombination of the nicotine moiety with the iron hydroxyl to produce the 5′-hydroxynicotine product. They conclude that nicotine exists in a less favorable protonation state when undergoing CYP catalysis.

The reaction mechanism of the conversion of tyramine to dopamine, as catalyzed by CYP2D6 was examined by Schyman et al.[173] The traditional Meisenheimer complex mechanism, i.e., direct aromatic hydroxylation, was ruled out on the basis of energetics. Instead, they deduced that the reaction path is entirely dictated by the local electric field of the protein environment. An initial phenolic H-abstraction of the protonated tyramine by Compound I is followed by a phenoxyl ring rebound, culminating in a keto−enol rearrangement to dopamine, outside of the protein. Interestingly, in this QM/MM study, the authors found that only a 3 Å radius sphere of the charges from the MM environment needed to be included in the QM/MM model for the H-abstraction intermediate to be stabilized to a value comparable to that of the full QM/MM model.

The mechanism of H-abstraction, followed by O-rebound is common between these two simulations. However, the tyramine is in a protonated state (protonated amino tail), whereas nicotine is in its unfavorable free base state.

A more recent study by Lonsdale et al.[174] investigated the electronic structure of the Compound I state between different CYP isoforms using a QM/MM approach, with the Compound I Fe−O bond enthalpy as the comparison metric. Interestingly, they find that there is little variation between the three isoforms they examined, and significantly more variation within a single isoform's ensemble of MD generated structures. They conclude that the hydrogen bonding, the polarizing environment and thermal fluctuations influence Compound I's electronic structure and stress the importance of using an ensemble of structures in QM/MM studies. They also state that the substrate's presence in the cavity alters the electronic structure of Compound I. It reduces the number of hydrogen bonds that ferryl oxygen has to the water solvent, increasing the spin density on ferryl oxygen, thus weakening the Fe−O bond.

Predicting Interaction of Xenobiotics with CYPs. Interaction of xenobiotics with metabolizing enzymes is likely to lead to substantial changes in biological effects, potentially causing a loss of activity or toxic effects. CYPs are key metabolizing enzymes, accounting for approximately 75% of total metabolism.[175] They play a pivotal role in drug metabolism, where seven of the 57 known human CYP isoforms facilitate more than 90% of the metabolism, as well as DDIs.[176] Polymorphism of certain CYP isoforms such as CYP1A2, 2D6, 2C9, or 2C19 add another layer of complexity to the problem.[177−179] Because of the wide range of metabolism rates observed for xenobiotics, determination of the kinetic profile and optimization of dosing (to remain within the therapeutic window) becomes an even more challenging and sometimes impossible task. Mutations of metabolizing enzymes and their impact on the kinetics of drug molecules are difficult to predict and would require a substantial computational effort for each compound and CYP. Consequently, prevention of interaction of compounds with polymorphic CYPs by rational design is a favorable strategy.

Predicting DDIs is a nontrivial and complex problem that has been traditionally addressed in elaborate clinical studies.[180] Even for the extrapolation of *in vitro* assay data to *in vivo* effects some major uncertainties and controversies exist.[181] Several *in vitro* CYP inhibitors, such as clotrimazole and other compounds sharing an imidazole scaffold, have been observed to induce these proteins *in vivo*.[182] One major challenge in predicting systemic effects derives from the crosstalk of receptors regulating metabolizing enzymes.[183] Though *in vitro* assays are becoming more readily available and more and more insight on the mechanism of inhibition and induction of metabolic enzymes has been gathered, a complete framework that would allow the accurate prediction of enzyme inhibition and induction is still missing.[5] Here, we provide an overview of computational methods aimed at (among other functions) the prediction of interactions between xenobiotics and CYPs. For more detail, the reader is referred to refs 7, 184, 185.

Predicting CYP Inhibition by Xenobiotics. CYP inhibition is in general considered to be more problematic than CYP induction. CYP inhibition may cause toxic effects by an increase of activity at targets and antitargets and is commonly evaluated by determining the inhibition constant, $K_{i,}$ using liver microsomes or cDNA-expressed microsomes.[179] CYP induction is a central focus for drug safety, potentially leading to a loss of therapeutic efficacy due to increased metabolism rates but also to an increased production of toxic metabolites (see below).

"QSAR and Machine Learning Methods". MD simulations are typically able to deal with previously unseen ligand-target pairs when attempting to predict metabolic sites on small molecules or their binding affinity (see "Molecular Dynamics Simulations on the Structure of CYPs"). However, they are at the same time computationally demanding and, consequently, they are not always applicable, particularly when time is at a premium. In real-world situations, very often an interactive, "immediate" prediction of the likelihood of being a substrate or an inhibitor of particular CYPs is desired, for instance, when medicinal chemists are interactively designing a series of structures. QSAR, machine learning, and other approaches are in many cases well suited to provide this kind of instant feedback to the user. At the same time, they are both based on prior knowledge. Hence, their main shortcoming is, generally speaking, their limited extrapolation ability to novel chemical series or molecules outside of their applicability domain.

QSARs attempt to derive quantitative relationships between the structure of a compound and its activity. Those relationships are based on the principle that the chemical structure of compounds—and similarities and trends of a chemical series—are related to the molecular properties exhibited.[186] Hence, in order to generate relationships between structure and function, a way to describe the molecule is needed (see ref 187 for a recent review), as well as a suitable statistical or machine learning method[188] to generate the particular SAR of interest. Activity in the particular context of this review means the inhibition of one or more metabolizing enzymes, in particular isoforms of the P450 enzymes by a small molecule, or its ability to act as a substrate of these enzymes.

What is remarkable about the area of QSAR in metabolism and also in toxicology is that guidelines for the validation of models have only relatively recently been published, as part of the Organisation for Economic Co-operation and Development (OECD) and European Union efforts to reduce the amount of animal testing performed in the context of exploring properties of novel chemical entities (NCEs). In order to trust QSAR models—which are essentially computational predictions of molecular properties—without resorting to experiment in every single case, requires validation and tests of robustness and is fundamental in order to gain trust in the models. In this context, it might be useful to consider a recently published summary of requirements pertaining to a practically useful QSAR model as follows (according to OECD guidelines): "The guidelines recommend that QSAR models should be associated with (i) a defined end point, (ii) an unambiguous algorithm, (iii) a defined domain of applicability, (iv) appropriate measures of goodness-of-fit, robustness, and predictivity, and (v) a mechanistic interpretation, if possible".[189,190] Certainly, these are essential criteria that should be applied to any trustworthy QSAR model upon which important practical decisions will be based.

The term "QSAR" does not immediately define which mathematical methods are used to derive those SARs. They can be statistical models or more complex machine learning models such as SVMs,[191] random forests,[192] and similar approaches. In the following, we will discuss approaches to predict CYP inhibitors and substrates using conventional (statistical) structure−activity modeling methods as well as machine learning methods applied to the area. Apart from reviewing work directly, the reader is also referred to previous reviews for further information, namely those focusing on QSAR modeling applied to CYP metabolism[178,184,185,193,194] as well as those focusing on the machine learning approaches to the field.[178,195−198] What is also remarkable is that in the field of metabolism or ADME/Tox prediction in general, "open" (free and often open-source) approaches are making progress, in tandem with public bioactivity databases,[199] etc. The reader is referred to a recent review of the field for an overview of many of the current available tools.[200]

Approaches to classify or quantify the interaction of chemical compounds with members of the CYP family can also be differentiated depending on how data is analyzed. Models can be generated for either inhibitors, substrates, or most generally, "ligands" or interaction partners of the enzyme. They can be either global or local, depending on the size and nature of the data set. While global models are built on large and diverse collections of compounds and therefore have a large domain of applicability, local models are specialized to a very specific chemical and biological space. Global models generally maintain their performance when predicting more diverse molecules (and properties), while local models may show superior prediction accuracy within the chemical space they are trained on.

Models can be either trained on molecules with measured activities for a whole enzyme family or for a specific isoform. Most challenging for the development of models is the fact that biological data such as IC_{50}, K_i, and K_m values depend very much on the experimental conditions (these are indeed different experimental measures which are often, through the sparsity of data, combined in single models), which leads to a significant variance in the data. Good models may thus be able to explain 65−85% of the variance in a data set. Models reported with higher accuracy are very likely to be overfitted.

Furthermore, different parameters can vary from model to model, such as the composition of the training set, the validation method, the choice of descriptors, and last but not least, the mathematical modeling method employed to generate SARs. Here again, models may be based on categorical or continuous variables. All of the above factors generally differ between models to be compared, and this needs to be kept in mind when relating the performance of models to each other. For more information on modeling toxicity data and its challenges, the reader is referred to a recent review.[201]

Classification Models. As an example of a relatively typical classification study, k-nearest neighbor, decision trees, random forests, ANNs, and SVMs using different kernels were employed to predict substrates, inhibitors, or "interactors" with the CYP1A2, 2D6, and 3A4 isoforms on the basis of a set of 335 structurally diverse compounds.[202] A total of 188 descriptors such as atom counts, charge properties, and connectivity indices were used for model building, employing 10-fold cross validation as a model validation method. Classification performances of 81.7−91.9% for CYP1A2, 89.2−92.9% for CYP2D6, and 87.4−89.9% for CYP3A4 were obtained. Interestingly, in this work various decision tree methods were found to outperform "numerical" methods such as SVMs—a result that seems not to hold when applying these algorithms to larger data sets.[203]

While single metabolizing enzyme predictions are easier to perform from the computational side, they do not reflect the situation that in reality many compounds are metabolized by more than one enzyme. To reflect this situation, in a more recent study[204] a total of 580 CYP substrates of seven different isoforms have been analyzed by applying multi- and single-label classification strategies, including SVMs, multi-label k-nearest neighbor classifiers, as well as ANN modeling methods. While single-label and multilabel classifiers at first sight seem to achieve similar performance, the authors state that "the multi-label approach more coherently reflects the real metabolism information", namely, that compounds are in many cases metabolized by different CYP isoforms, though at different rates and giving rise to different metabolites. In this particular case, out of the 580 compounds 488 were metabolized by a single CYP isoform, while the remaining 92 structures are metabolized by up to five enzyme variants.

The idea of having "true negatives" in a data set is also critical to address. However, compounds of unknown activity are often treated as "putative inactives". This was however approached explicitly in a recent study where for CYP3A4, 2D6, and 2C9 inhibitors and substrates were analyzed using SVMs.[205] For the prediction of inhibitors, only CYP3A4 and 2D6 were used, but substrate models have been generated for all isoforms. In particular, two consensus SVM methods, namely, "positive majority" and "positive probability" were employed in this work on a large data set comprising several hundred activity data points. In addition to the inhibitors and substrates, likely "non-inhibitors" and "non-substrates" were generated. The model obtained accuracies for classification of substrates and non-substrates, respectively, of 98.2% and 90.9% for CYP3A4, 96.6% and 94.4% for CYP2D6, and 85.7% and 98.8% for CYP2C9. Regarding the different machine learning methods employed, the consensus support vector machine methods were generally found to give better performance than those based on single SVM classification systems.

In many cases, data sets are not balanced, which poses additional challenges when developing machine learning models. In a recent study using CYP2D6 inhibitors,[206] a set of 185 training and 78 test data points were employed using "ensemble" descriptors (containing atom counts, constitutional descriptors, topological descriptors, etc.) in combination with a SVM. In this work, the influence of various oversampling and "threshold moving" techniques on classification performance was investigated and overall it was found that employing oversampling and threshold moving indeed had a positive impact on obtaining accurate classifiers.

isoCYP[207] is a software tool for the prediction of human CYP isoform specificity based on the work of Terfloth et al.,[208] who developed a number of classification models using multinomial logistic regression, decision trees and SVM. The models are trained on a set of 146 compounds known to be metabolized by human CYP3A4, 2D6, and 2C9 isoforms and the models were evaluated using an external validation data set comprising 233 compounds. The best model obtained a leave-one-out cross-validated predictivity of 83% (correct predictions) for the external validation set.

Another study employing SVM to identify and classify substrates of CYP1A2, 2C9, 2C19, 2D6, and 3A4 is based on a 17000 compounds data set from the National Institutes of Health Chemical Genomics Center (NCGC).[209] Classification models obtained area under the receiver operating characteristic (ROC) curves equal to or higher than 0.85 for any of the investigated CYP isoforms.

Quantitative Models. While classification models are sometimes preferred to numerical/regression models since they often have superior performance in validation experiments when only class labels are required, they are generally not able to make affinity predictions, which are at least in relative terms often needed when considering competing interactions in biological systems. Some of those quantitative models relating to metabolism prediction shall be discussed below.

Classical Quantitative QSAR Models. Lewis et al.[210] established quantitative models for ligands for a total of six P450 isoforms, namely, CYP1A2, 2B6, 2C9, 2C19, 2D6, and 3A4. In this work, the authors obtained correlation coefficients ranging from $R = 0.94$ to 0.99 between predicted and experimental P450 binding affinity, which in this case included a mixture of K_m, K_d, or K_i values. They found that including hydrogen bonding parameters in the QSAR analysis in combination with the number of pi-pi stacking interactions was crucial for model performance. Hence, given that interpretable variables were employed in the model in the first place, knowledge about interactions could also be derived—to some extent—from the resulting model.

For the isoforms CYP2D6, 1A2, 3A4, 2A6, 2C9, 2C8, 2C19, and CYP17, PLS regression as well as 18 machine learning methods as implemented in WEKA[211] were recently employed on a data set comprising a total of 797 ligand−CYP IC_{50} data points.[212] PLS regression as a baseline method was employed on a subset of six and 15 selected descriptors, respectively. From the resulting models, R^2 values from 0.69 to 0.94 were obtained for the six-descriptor subset, and from 0.78 to 0.99 for the set comprising 15 descriptors, respectively. However, a particular modeling technique, namely PLS "with mixed-integer linear programming-based hyperboxes", was able to improve upon those results using only six descriptors. The method involved homology modeling, docking, and finally ligand-based classification steps, which likely renders it time-intensive in practice. Also the descriptors selected to be the most significant are largely noninterpretable connectivity indices, which are not generally intuitively related to metabolic mechanisms.

In a similar fashion, the way data for CYP inhibition or substrate-likeness data are obtained experimentally can have a profound impact on the performance and applicability of CYP SAR models. In a recent study,[213] the data sets consisted of marketed drugs and drug-like compounds, all tested in four assays measuring the inhibition of the metabolism of four different substrates by the CYP3A4 enzyme, benzyloxycoumarin, testosterone, benzyloxyresorufin, and midazolam. It was found that by employing only a single one of the four data sets no reliable inhibition model could be generated. On the full data set, however, a multiple pharmacophore hypothesis could be developed which was able to model promiscuity, in particular the CYP3A4 isoform of the CYP family. Hence, care should be paid not only to the statistical model validation method employed, but also to the data used to generate models for P450 activity.

As a commercial software provider, ACD has a software suite on the market that is able to make regioselective predictions of metabolic transformations in its P450 Regioselectivity Module.[214] On the basis of more than 900 training set compounds, every atom is assigned a likelihood of undergoing one of five possible transformations, namely N-dealkylation, O-dealkylation, aliphatic hydroxylation, aromatic hydroxylation, S-oxidation. A reliability index for the prediction is provided, depending on the similarity to the training set atoms as well as the consistency of metabolic transformation information. Optionally, isoform-specific predictions can also be performed for CYP3A4, 2D6, 2C9, 1A2, and 2C19 using additional software modules. However, to the knowledge of the authors no standardized validation of the tool has been publicly disclosed.

In silico tools using different machine learning methods have been derived for various enzyme isoforms. For the CYP2D6 and 2C9 isoforms, multidimensional QSAR has been employed to quantify the binding affinity of 56 compounds and 85 compounds, respectively, to their metabolizing enzymes.[215] In this study, a cross-validated R^2 of 0.81 and 0.69 was obtained, and models were validated by different methods including Y scrambling and an additional external test set. This model is also available publicly at VirtualToxLab, which combines multidimensional QSAR with a flexible docking approach.[216]

VirtualToxLab supplies QSAR models for 16 proteins known or suspected to be responsible for adverse drug effects, including AhR, androgen receptor (AR), estrogen receptor (ER), etc., and reports a calculated toxic potential.[216] It was used to predict the toxic potential of more than 2000 xenobiotics and chemicals and to estimate the binding affinities of anthracene- and steroid-based compounds on various CYP isoforms and on AhR (see below).

3D-QSAR and Molecular Interaction Fields. 3D-QSAR analyses such as Comparative Molecular Field Analysis (CoMFA),[217] GRID/Generating Optimal Linear PLS Estimations (GOLPE)[218] and GRid alignment INDependent (GRIND) approaches have been employed to predict interaction partners of CYP enzymes. CoMFA as well as GRID descriptors depend on the alignment of molecules, which is problematic for many 3D-QSAR applications, as ligand alignment is nontrivial and time-consuming and induces a bias into the model. GRID alignment independent descriptors have been developed to overcome these shortcoming.[219,220] They are derived from MIFs by a data compression process that elucidates the most-relevant interaction regions of a molecule. GRIND descriptors are independent from the orientation of the molecule and hence only the internal coordinates of interaction patterns are encoded. Most importantly, GRIND descriptors can be back-converted into the primary MIF descriptors by an autocorrelation transform, which allows visualization and interpretation of data in the original 3D space. The ALMOND[220] software package allows the calculation, analysis, and interpretation of GRIND descriptors. It has recently been superseded by Pentacle.[221]

Crivori et al.[222] trained a classification model for the prediction of the metabolic stability of small organic molecules to CYP3A4 combining GRIND descriptors with VolSurf, a MIF-based program to calculate pharmacokinetic properties of small organic molecules.[223] The training set consisted of 1800 examples from the Pharmacia compound collection. The model derived using these data obtained a correct prediction rate of 75−85% of the test set compounds, with a precision of 86% correctly identified metabolically stable compounds. A further study reporting classification and regression models for CYP2C9[224] is reported in the next section.

CoMFA and GRID/GOLPE regression models were reported for CYP1A2 inhibition on the basis of a consistently assembled data set of 52 compounds comprising different scaffolds.[225] CoMFA models yielded a Q^2 of 0.69 (5-fold cross validation) and R^2 of 0.87, GRID/GOLPE yielded a Q^2 of 0.79 and R^2 of 0.90, respectively. In a similar approach, a CoMFA model for CYP2B6 inhibition[226] was prospectively validated, which led to the discovery of three potent pyridine-based novel CYP2B6 inhibitors. For a CYP2D6 isoform model based on the CoMFA method, a training set of 24 compounds with reported K_m values was developed which achieved a predictive R^2 of 0.62.

CoMFA/GOLPE models were also obtained for 50 steroid inhibitors of CYP19 (aromatase). The CoMFA fields are consistent with known, potent inhibitors of aromatase, not included in the model (R^2 of 0.885, cross-validated Q^2 of 0.673).[227] In a similar fashion, the CYP2A5 and 2A6 isoforms were analyzed using CoMFA and GRID/GOLPE.[228] From this work, it was suggested that the CYP2A5 binding site is likely larger than that of CYP2A6, as indicated by larger steric regions in the CoMFA coefficient arrays as well as a similar view of the corresponding GOLPE maps. From an analysis of electrostatic maps, it was deduced that CYP2A6 disfavors a negative charge near the lactone moiety of coumarin, which was in agreement with available data.

The GRID/GOLPE approach was also successfully employed to derive a 3D-QSAR model for CYP2C9 inhibitors.[229] In this study, the selection of ligand conformers was guided by docking a set of ligands to a CYP2C9 homology model employing GOLD. Following up on their prior effort, Afzelius et al.[224] used a more diverse data set to create regression (and classification models) for CYP2C9 inhibitors. The high diversity of the data, including ligands of MW 100−1000, rendered the elucidation of specific alignment rules unfeasible and lead to the application of a GRIND-based approach using ALMOND. In this study, 74% of the compounds of an external test set could be correctly classified. A subsequently derived quantitative model was able to predict the K_i values within half a log unit for the vast majority of compounds. This work was again extended by adapting the CORE method developed by Goodford[230] to handle ligand flexibility in order to account for ligand movement during the approach to the ligand binding site. Using this method the relevant conformational space is encoded in the form of a probability-of-interaction map. The resulting MIFs were transformed into GRIND descriptor values using ALMOND. The experimental K_i values of 11 out of 12 compounds representing the external test set were predicted within half a log unit by this confomer- and alignment-independent QSAR model.

A further 3D-QSAR study compared the GRID/GOLPE approach with COMparative BINding Energy (COMBINE) analysis, an approach that analyses a series of structures of protein−ligand complexes to derive 3D-QSAR models.[231] It was found that, provided an appropriate alignment is used for model generation, that the derived interaction patterns are largely consistent with each other.

A GRIND-based 3D-pharmacophoric model for CYP3A4 was reported by Cianchetta et al.[232] A data set of 331 compounds with inhibition data available for CYP3A4 was used to derive a 3D-QSAR model from MIFs. In parallel, GRIND descriptors were also calculated for four CYP3A4 crystal structures and subsequently compared with the pharmacophore model derived from the ligands. A clear correlation between the 3D-QSAR model and the MIFs generated from the experimental protein structures was found and three features crucial for binding affinity, two hydrophobic areas and one hydrogen-bond acceptor, were revealed.

"Classic Pharmacophore-Based Approaches". Interactions of ligands with biomolecules are driven by the complementarity of global and local physicochemical properties. Detection and analysis of patterns among ligands, the receptor, and ligand−receptor complexes form the fundamental basis for rational drug design. A pharmacophore defines a pattern of chemical and steric features essential for the interaction of a ligand with its receptor. The robustness of the approach infers from its simplicity. Only a very few feature types are required to characterize most of the commonly observed protein−ligand interaction types, such as hydrogen bond donors/acceptors, aromatic interactions, hydrophobic interactions, and ionic and metal interactions. These are usually depicted as colored spheres, disks, or graphs, and directed features (hydrogen bonding, pi-stacking and metal binding features) generally include a vector to indicate their orientations, if determinable.

Classic pharmacophore models are derived in a ligand-based or structure-based approach. For the ligand-based approach, the ligands are superimposed on the basis of matching chemically similar moieties, while for the structure-based approach analyses of the receptor or receptor−ligand complexes are used to derive pharmacophore models. In cases where the target itself or target structure are unknown or not amenable, ligand-based techniques are typically the only approaches available.

Development of a pharmacophore model is an iterative process, attempting to derive the optimum selection of ligands, chemical features, and geometric constraints during the model training phase. Rigid templates are particularly favorable for ligand-based modeling, as they allow the conformational degrees of freedom to be narrowed down. Once the model has been validated, large-scale virtual screening is rapid.

Pharmacophore models allow qualitative and quantitative predictions on the metabolism of small organic molecules. A pharmacophore model can only cover the chemical space considered for training of the model and, as a matter of fact, only a particular binding mode. This is in contrast to protein−ligand docking, where the protein−ligand interaction motifs are taken as a global model of the receptor. In general, the limitation of pharmacophore models to a specific binding mode is unlikely to affect the applicability of this approach. If required, multiple pharmacophore models can be constructed to account for the different interaction patterns. However, in the case of highly promiscuous binding sites such as in CYP3A4 this may be a limiting factor.[78]

Plasticity of the binding site can be reflected in classical pharmacophore models in part by adjustment of tolerances surrounding the pharmacophore feature points. For example, molecular dynamics simulations can be used to quantify and characterize target flexibility at an atomic level of detail. These observations can then be included in pharmacophore models for scoring of the fitness of molecules to the pharmacophoric constraints.

Reported pharmacophore models have been summarized in a number of comprehensive reviews,[176,233−237] and therefore, only a brief overview is presented here.

Qualitative and quantitative pharmacophore models have been developed for the classification/prediction of substrates and inhibitors of all CYP isoforms important to drug discovery, including CYP1A2, 2A6, 2B6, 2C9, 2D6, 3A4, 3A5, and 3A7. For example, CYP2D6 includes a characteristic positive charge about 5−7 Å distant from the oxidation site as well as an aromatic interaction.[233] Various pharmacophore models were proposed with respect to the underlying metabolic reaction.[233,236] Most of the CYP2C9 pharmacophore models include a hydrophobic/aromatic and a negatively charged interaction area.[233,238] However, nonanionic substrates of CYP2C9 are known,[239] which illustrates the possibility of bias introduced into a pharmacophore model by the training data. In contrast to CYP2C9, 3A4 appears to have no obvious specific pharmacophoric requirements. This isoform has often been reported as being a particularly challenging target for computational approaches, but at the same time, it is the most important CYP for xenobiotic metabolism.[176] Schuster et al.[240] have reported a collection of eleven structure-based and ligand-based pharmacophore models for the identification of substrates and inhibitors binding to various CYP isoforms.

A single pharmacophore model cannot characterize a promiscuous ligand binding site. Mao et al.[213] have shown for CYP3A4 that the generation of several local models trained on distinct data sets performs best.

Overall, pharmacophore-based methods have contributed significantly to the understanding of CYP ligand binding. They are powerful, accurate tools for local predictions, driving lead optimization and rapid prefiltering. Their moderate capabilities to reflect the global characteristics of promiscuous, highly plastic interaction sites may limit their applicability in CYP modeling.

"Protein−Ligand Docking". Protein−ligand docking can be used for predicting the binding mode of small organic molecules to their target as well as providing an estimate of the binding affinity, although most docking tools estimate the geometry of binding with a scoring function that is relevant to only this property and not to binding affinity. The ligand and protein conformations used as a starting configuration can be decisive for a docking algorithm.

Docking has also been successfully applied to predict and rationalize drug−drug interactions on CYP2D6.[241] Twenty established drugs were investigated for their ability to bind to a homology model of CYP2D6. Out of the13 drugs predicted to inhibit this CYP, 11 were experimentally confirmed. In the course of this study, an aromatic N-hydroxy metabolite of metoclopramide, a drug to treat nausea and vomiting, was identified and experimentally confirmed.

Using the same isoform, docking *via* AutoDock into CYP1A2 homology models has been performed, followed by COMBINE and GRID/GOLPE analyses of 12 heterocyclic amines. In this case, it was found that including structural information for the alignment of compounds did indeed improve model performance, in agreement with previous findings.[242]

"Combined Approaches for Interaction Prediction". Leong et al.[243] used a pharmacophore ensemble/SVM approach to increase the predictive power of a model for CYP2A6 interaction. Protein flexibility was simulated by employing three individual pharmacophore models.

An interesting method employing MIFs is the proteochemometric analysis of CYP sequences to relate these to the activity of inhibitors.[244] Using this approach, Kontijevskis et al. aimed to establish a general model for the prediction of isoform-specific inhibition rates and to gain knowledge about properties of the different CYP isoforms and their inhibitors. A data set of 375 structurally diverse inhibitors with reported IC_{50} values for 14 different CYP isoforms was collected and GRIND descriptors were calculated for all of these small organic molecules. The protein sequences of 14 CYP isoforms were aligned and analyzed for characteristic patterns. Six groups of aligned amino acid positions showed the same variance throughout the sequence alignment. These groups consist of two or more amino acid positions with binary sequence variability, encoded in binary descriptor values. Similar descriptors were also created for 20 additional amino acid positions that differed between two amino acids, but without covariation with any other sequence position. Finally, the remaining 388 amino acid positions were encoded by five z-scale properties,[245] which represent principal components for a set of measured physicochemical properties of amino acids. Principal component analysis (PCA) was employed to reduce the number of descriptors for both CYPs and ligands, cross-terms were generated to describe ligand-CYP interactions, and PLS was applied for generation of the models. Results of the model validation indicated a RMSD of prediction of a RMSD of 0.65−0.81 for the prediction of new ligand data on the investigated 14 CYP isoforms. Novel insights into SARs derived from the MIF-based analysis may help designing drugs with favorable metabolic profiles. The method could also be used for fast metabolic profiling in virtual screening.

Bazeley et al.[246] analyzed the binding affinity of substrates of CYP2D6 using a combination of machine learning, protein modeling and protein−ligand docking algorithms. Structural models of various CYPs were aligned and analyzed for structurally conserved regions, which were conformationally preserved during subsequent simulated annealing runs. This resulted in twenty distinct protein conformations reflecting the flexibility of CYP2D6. 82 small organic molecules with known affinities for CYP2D6 were docked to these protein conformations. The docking scores were used as attributes for ANN model generation. Also compound-specific descriptors were calculated and employed as attributes for model building. The best performing, optimized ANN model gave a prediction accuracy of 85%. Attribute selection for the ANN model identified docking scores of three of the 20 protein conformations as being dominant for the predictability of binding affinity. From the ligand perspective, the number of positive charges, ALogP and the number of aromatic rings were identified as the most important descriptors.

VirtualToxLab[247] uses a combination of multidimensional QSAR and flexible docking for the prediction of the interactions of small organic molecules with 16 antitargets, including CYP450 1A2, 2A13, 2C9, 2D6, and 3A4. Interactions of query molecules with any of the 16 target structures can be visualized.

Predicting CYP Induction by Xenobiotics. CYP induction is caused by either an amplified *de novo* synthesis of the protein or, less frequently, by a decelerated enzyme degradation pathway.[5] Depending on the CYP isoform involved in DDIs, enzyme induction can cause minor to substantial clinical effects.[177] CYP isoforms known to be induced in humans include CYP1A, 2A, 2B, 2C, 2E1, and 3A.[248]

Nuclear receptors are a primary control mechanism for gene transcription and, hence, a molecule that activates a nuclear receptor may function as an enzyme inducer. Among the nuclear receptors most important for the transcription of CYPs are the aryl hydrocarbon receptor (AhR), constitutive androstane receptor (CAR), and pregnane X receptor (PXR).[5,177]

These are also referred to as xenobiotic-activated receptors (XARs). They are known to be highly promiscuous sensors, binding a set of structurally diverse xenobiotics, in particular those that are hydrophobic.[182] For example, PXR binds rifampicin, dexamethasone, bosentan, and artimisinin antimalarials. Additionally, the peroxisome proliferator activated receptor (PPAR) and the glucocorticoid receptor (GR) have also been shown to mediate CYP enzyme expression,[7] and inducing xenobiotics have been reported to bind to the hepatocyte nuclear factor-4α (HNF4α) and the vitamin D receptor.[182] For more detail on xenobiotic inducers of CYPs, the interested reader is referred to excellent review on modeling CYP substrates, inhibitors, activators, and inducers by de Lisle et al.[249]

In silico methods have been reported for many of these receptors but their promiscuity, resulting from their flexibility, their extended binding pockets and their nonspecific, lipophilic interaction features make them challenging targets. A number of *in silico* methods in particular struggle with the problem of multiple binding modes for one target. In the case of pharmacophore models, individual models generally need to be created for each binding mode.

"Pregnane X Receptor (PXR)". PXR is one of the key xenobiotics sensing enzymes regulating the expression of CYP 3A4.[248] It is known for its promiscuity, being able to accommodate a large range of diverse chemical compounds within the binding site. Single ligands can be observed to bind in multiple modes, which adds to the complexity of understanding the binding of antagonists at this receptor. Xue et al.,[250] who aimed to develop antagonists by introducing sterically hindering moieties to known agonists, were unable to find any because of the enormous promiscuity of this protein.

QSAR studies including a PLS approach in combination with VolSurf descriptors have been used to derive models for ligands of human PXR and AhR.[251] Ung et al.[252] employed the machine learning methods SVM, k-nearest neighbor, and probabilistic neural network (PNN) for classifying activators and nonactivators. Prediction accuracy rates for a 10-fold cross validation test were found in the range of 61−87%, depending on the compound class (activators and nonactivators).

Also, a number of pharmacophore-based studies have been reported that aim to derive models identifying compounds binding to PXR (see refs 253−255). These models in general include three to five hydrophobic areas as well as one or two hydrogen bonding features. Yasuda et al.[256] used several pharmacophore models published in earlier studies to assess their potential in predicting PXR activators. They found two of these pharmacophore models obtained reasonable prediction accuracy, while the model based on the most diverse training set had inferior performance. This is likely to be caused by the presence of more than one binding mode, which cannot be adequately covered by a single model. Intriguingly, dicloxacillin was found to be the only PXR binder that could be successfully mapped with any of the three investigated pharmacophore models. The authors also employed the docking software FlexX to investigate the binding of known activators to PXR.

Gao et al.[257] introduced new hydrophilic moieties to PXR activators to lower their activity on this receptor. The optimization strategies were developed from results of a protein−ligand docking approach.

In a recently published study assessing the performance and reliability of PXR, using *in vitro* assays and *in silico* modeling approaches, the physicochemical properties of 37 marketed drugs and their interaction with PXR were examined.[258] The computational analyses pointed out that descriptors for molecular weight and shape of a ligand are decisive for its potential to bind to PXR. Compounds of molecular weight lower than 300 as well as compounds that have a molecular shape which does not correspond to the inverse shape of the protein binding site are unlikely to exhibit activity. Strong binders show favorable hydrophobic interactions and hydrogen bonding features within the binding pocket, as observed during docking studies using GLIDE.[53] Binding promiscuity is facilitated by the flexibility of the protein but also by the conformational space accessible to the binders.

"Aryl Hydrocarbon Receptor (AhR)". AhR is a cytosolic transcription factor that is a key regulator for CYP1A1 and CYP1A2 in humans.[248] Besides the induction of these CYP isoforms, AhR signaling also induces a number of other enzymes including UDP-glucuronosyltransferase 1A (UGT1A1) and glutathione S-transferase A2 (GSTA2).

In addition to the QSAR models for human AhR and PXR presented above, recent examples include the work of Bisson et al.,[259] who derived homology models of the AhR PAS domain to investigate intraspecies and interspecies differences in ligand binding using Internal Coordinate Modeling (ICM).[260] Among other observations, their approach led to the successful identification of two flavonoids, pinocembrin and 5-hydroxy-7-methoxyflavone, which are experimentally shown to be promoters of nuclear translocation and transcriptional activation of AhR and AhR-dependent induction of endogenous target genes. As an example of an implementation in a commercial software product, a model for the prediction of ligand activity on AhR is included in VirtualToxLab.

"Constitutive Androstane Receptor (CAR)". The constitutive androstane receptor (CAR) is responsible for the induction of CYP2B6, which is a key enzyme for the metabolism of a large number of drugs including human immunodeficiency virus (HIV) therapeutics, chemotherapeutics and opioids.[248]

Using a pharmacophore-based approach, CAR ligands were reported to show planar structures with a hydrogen bond acceptor and two to three hydrophobic areas.[261] Jyrkkärinne et al.[262] used pharmacophore-based virtual screening to identify novel agonists of CAR and employed these data to develop a GRID/GOLPE-based 3D-QSAR model in an attempt to explain activity of compounds on CAR. For a leave-20%-out validation method, the model obtained a Q^2 value of 0.68 and a standard error of prediction of 0.93. Further to that, docking studies of CAR agonists on the X-ray structure of human CAR revealed several key interactions for ligand binding.

More recently, homology models derived for CAR, AhR, and PXR (all in the rat) were used as a structural basis for docking a number of organic pollutants such as polybrominated dibenzofurans and polybrominated diphenyl ethers to these nuclear hormone receptors, with the aim of identifying potential binders to these proteins.[263] The results of this retrospective study were found to be in agreement with results of experiments and suggest that such structure-based modeling methods represent valuable tools for the evaluation of effects of small organic molecules on nuclear hormone receptors.

■ PERSPECTIVE AND OUTLOOK

Metabolic properties are decisive for the discovery, development, and market success of drugs, nutritional supplements, and agrochemicals. Interactions of xenobiotics with key metabolizing enzymes may lead to considerable changes in kinetics, which may cause toxic effects or failure of action. Genetic polymorphism adds another layer of complexity to the problem.

Numerous drug candidates and agrochemicals have failed in the past and novel entities are still failing because of unfavorable metabolic profiles, despite substantial efforts being made in industry to predict such problems early and to develop strategies to counter them.

A plethora of computational methods, both ligand-based and structure-based, have been developed, which can provide decisive insights on the metabolic fate of xenobiotics. They allow the prediction of likely sites of metabolism (SOMs), the chemical structure of potential metabolites, and inhibition and induction of key enzymes involved in xenobiotic metabolism. For all of these areas of research, a number of success stories have been published, as discussed in this review.

The complexity of predicting xenobiotic metabolism suggests that one particular algorithm will not show superior performance in all of the three domains discussed here. Rather, the combination of techniques is most promising to lead to valid conclusions.

SOM prediction is feasible, with state-of-the-art computational methods successfully identifying SOMs among the three highest-ranked atom positions in a range of 70–90% for most cases. Even though relative ranking of atom positions is feasible, current methods do not make accurate predictions for the absolute likelihood of a certain biotransformation. This limitation makes it difficult if not impossible to draw any quantitative conclusions on the metabolic liability of a certain molecule and, hence, to compare different molecules with respect to their metabolic stability. In particular, the combination of individual methods covering different aspects of the biotransformation process could be expected to lead to a potential boost in prediction accuracy. Systems biology approaches (not reviewed here) have a major contribution to make by incorporation of multiple models that take into account not only the likelihood of metabolism at a site in a molecule at a specified rate but also the molecules bioavailability and concentration at the tissue site of metabolism. A much more complicated scenario.

Knowledge-driven approaches such as expert systems and data mining techniques allow extrapolation of the structure to likely metabolites. Most challenging in this domain is addressing the combinatorial explosion problem, which arises from the virtually unlimited possibilities for processing generations of metabolites. Advanced reasoning rules and knowledge-based potentials considered for metabolite ranking, as well as consideration of physicochemical properties of the generated metabolites (such as logP) attempt to address this problem as well as possible.

Also, a respectable number of computational approaches have been developed and applied to the prediction of direct and indirect interactions of xenobiotics with key metabolic enzymes. CYPs represent the primary focus of research and inhibition of this enzyme family has been successfully predicted by a variety of computational approaches including (3D)-QSAR and machine learning methods, pharmacophore- and docking-based approaches, as well as combinations of these. The expression of CYPs is steered by nuclear receptors such as PXR, AhR, and CAR, to which similar methodologies have been applied but to a lower extent.

Methods developed in academia all too often lack maturation of the promising concepts they are based upon or an implementation of these algorithms into accessible—in the best case—user-friendly software packages. Tools freely available to the scientific community are still scarce. Further, the limited data related to metabolism (experimentally derived SOMs, structures of metabolites, enzyme inhibition, and induction data) are a bottleneck for method development. However, recent major releases of bioactivity data to the public, such as ChEMBL,[264] World Of Molecular BioAcTivity (WOMBAT),[265] and others,[199] give promising indications of a paradigm shift in the scientific community, fostering the open publication of experimental data.

Also with regard to structure-based approaches, there is still enormous potential left for further development. Even though several crystal structures of a variety of CYPs have become available, their numbers are not sufficient to adequately represent, in particular, flexibility of the enzyme catalytic site, though homology modeling and MD simulation techniques can help to overcome some of these limitations. For instance, for the development of agrochemicals it would be of utmost importance to have structures available of pest species. Expeditious technological progress in structure determination will improve this situation in the future and make enzyme-focused methods other than CYPs even more powerful and important. An improved data situation will also contribute to solving cardinal problems of structure-based approaches, such as target flexibility (in particular for CYPs), solvent, entropic effects, and as a consequence, affinity and rate of reaction prediction.

However, it is not expected that computational methods will substitute *in vitro* and *in vivo* methods in the foreseeable future. In fact, it is crucial for the generation of computational models to be based on further enhanced assay technology. Computational methods are still limited by the reliability of assay systems and the information content of the read out.

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: rcg28@cam.ac.uk. Phone: +44 (1223) 336 432.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Unilever for funding.

■ ABBREVIATIONS

ADME = absorption, distribution, metabolism, and excretion
AhR = aryl hydrocarbon receptor
AMBER = Assisted Model Building with Energy Refinement
ANN = artificial neural network
ANNC = artificial neural network classification
ANNE = artificial neural network ensemble
CAR = constitutive androstane receptor
CHARMM = Chemistry at HARvard Molecular Mechanics
COMBINE = COMparative BINding Energy
CoMFA = Comparative Molecular Field Analysis
CYP = cytochrome P450
DDI = drug–drug interaction
DFT = density functional theory
EPR = electron paramagnetic resonance
FEP = free energy perturbation
FF = force field
GLIDE = Grid-based LIgand Docking with Energetics

GRIND = GRid alignment INDependent
GOLD = Genetic Optimization for Ligand Docking
GOLPE = Generating Optimal Linear PLS Estimations
GSTA2 = glutathione S-transferase A2
HIV = human immunodeficiency virus
ICM = Internal Coordinate Modeling
KEGG = Kyoto Encyclopedia of Genes and Genomes
k-NN = k-nearest neighbor
LC = liquid chromatography
LIE = linear interaction energy
MAO = monoamine oxidase
MC = Monte Carlo
MD = molecular dynamics
MDEA = (R)-3,4-methylenedioxy-N-ethylamphetamine
MIF = molecular interaction field
ML = machine learning
MM = molecular mechanics
MS = mass spectrometry
NADPH = nicotinamide adenine dinucleotide phosphate
NCE = novel chemical entity
NCGC = National Institutes of Health Chemical Genomics Center
OECD = Organisation for Economic Co-operation and Development
PCA = principal component analysis
PLOP = Protein Local Optimization Program
PLS = partial least-squares
PME = Particle Mesh Ewald
PNN = probabilistic neural network
PXR = pregnane X receptor
QM = quantum mechanics
QM/MM = quantum mechanics/molecular mechanics
QSAR = quantitative structure−activity relationship
RAMBA = retinoic acid metabolizing blocking agent
RMSD = root-mean-square deviation
RMSF = root-mean-square fluctuation
ROC = receiver operating characteristic
ROCS = rapid overlay of chemical structures
RS-predictor = RegioSelectivity predictor
SAR = structure−activity relationship
SASA = solvent-accessible surface area
SMARTS = SMiles ARbitrary Target Specification
SOM = site of metabolism
SVM = support vector machine
SyGMa = Systematic Generation of potential Metabolites
TI = thermodynamic integration
TIMES = TImes MEtabolism Simulator
UGT1A1 = UDP-glucuronosyltransferase 1A
UM-BBD = University of Minnesota Biocatalysis/Biodegradation Database
UM-PPS = University of Minnesota Pathway Prediction System
WOMBAT = World Of Molecular BioAcTivity
XAR = xenobiotic-activated receptor

REFERENCES

(1) Munos, B. Lessons from 60 years of pharmaceutical innovation. *Nat. Rev. Drug Discovery* **2009**, *8*, 959−968.

(2) Swinney, D. C.; Anthony, J. How were new medicines discovered? *Nat. Rev. Drug Discovery* **2011**, *10*, 507−519.

(3) Mahmood, M.; Malone, D. C.; Skrepnek, G. H.; Abarca, J.; Armstrong, E. P.; Murphy, J. E.; Grizzle, A. J.; Ko, Y.; Woosley, R. L. Potential drug−drug interactions within veterans affairs medical centers. *Am. J. Health-Syst. Pharm.* **2007**, *64*, 1500−1505.

(4) Yeung, C. K.; Fujioka, Y.; Hachad, H.; Levy, R. H.; Isoherranen, N. Are circulating metabolites important in drug−drug interactions? Quantitative analysis of risk prediction and inhibitory potency. *Clin. Pharmacol. Ther.* **2011**, *89*, 105−113.

(5) Almond, L. M.; Yang, J.; Jamei, M.; Tucker, G. T.; Rostami-Hodjegan, A. Towards a quantitative framework for the prediction of DDIs arising from cytochrome P450 induction. *Curr. Drug Metab.* **2009**, *10*, 420−432.

(6) Hewitt, N. J.; Lecluyse, E. L.; Ferguson, S. S. Induction of hepatic cytochrome P450 enzymes: Methods, mechanisms, recommendations, and in vitro-in vivo correlations. *Xenobiotica* **2007**, *37*, 1196−1224.

(7) Sun, H.; Scott, D. O. Structure-based drug metabolism predictions for drug design. *Chem. Biol. Drug. Des.* **2010**, *75*, 3−17.

(8) Jones, S. L.; Athan, E.; O'Brien, D. Serotonin syndrome due to co-administration of linezolid and venlafaxine. *J. Antimicrob. Chemother.* **2004**, *54*, 289−290.

(9) Sullivan, E. A.; Shulman, K. I. Diet and monoamine oxidase inhibitors: A re-examination. *Can. J. Psychiatry* **1984**, *29*, 707−711.

(10) Gillman, P. K. Monoamine oxidase inhibitors, opioid analgesics and serotonin toxicity. *Brit. J. Anaesthesia* **2005**, *95*, 434−441.

(11) Lotfipour, S.; Arnold, M. M.; Hogenkamp, D. J.; Gee, K. W.; Belluzzi, J. D.; Leslie, F. M. The monoamine oxidase (MAO) inhibitor tranylcypromine enhances nicotine self-administration in rats through a mechanism independent of MAO inhibition. *Neuropharmacology* **2011**, *61*, 95−104.

(12) Stella, V. J.; Borchardt, R. T.; Hageman, M. J.; Oliyai, R.; Maag, H.; Tilley, J. W., Eds.; *Prodrugs: Challenges and Rewards*, Part 1; Springer: New York, 2007; Vol. 5.

(13) Tarcsay, Á.; Keseru, G. M. In silico site of metabolism prediction of cytochrome P450-mediated biotransformations. *Expert Opin. Drug Metab. Toxicol.* **2011**, *7*, 299−312.

(14) Kirchmair, J.; Markt, P.; Distinto, S.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Langer, T.; Wolber, G. The Protein Data Bank (PDB), its related services and software tools as key components for in silico guided drug discovery. *J. Med. Chem.* **2008**, *51*, 7021−7040.

(15) Zaretzki, J.; Bergeron, C.; Rydberg, P.; Huang, T.-W.; Bennett, K. P.; Breneman, C. M. RS-Predictor: A new tool for predicting sites of cytochrome P450-mediated metabolism applied to CYP 3A4. *J. Chem. Inf. Model.* **2011**, *51*, 1667−1689.

(16) Afzelius, L.; Arnby, C. H.; Broo, A.; Carlsson, L.; Isaksson, C.; Jurva, U.; Kjellander, B.; Kolmodin, K.; Nilsson, K.; Raubacher, F.; Weidolf, L. State-of-the-art tools for computational site of metabolism predictions: Comparative analysis, mechanistical insights, and future applications. *Drug Metab. Rev.* **2007**, *39*, 61−86.

(17) Hennemann, M.; Friedl, A.; Lobell, M.; Keldenich, J.; Hillisch, A.; Clark, T.; Göller, A. H. CypScore: Quantitative prediction of reactivity toward cytochromes P450 based on semi-empirical molecular orbital theory. *ChemMedChem* **2009**, *4*, 657−669.

(18) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. Development and use of quantum mechanical molecular models. 76. AM1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902−3909.

(19) *Materials Studio, VAMP Software Module*, version 6.0; Accelrys: San Diego, CA, 2011.

(20) *ParaSurf*, version 11; Cepos InSilico: Bedford, United Kingdom, 2011.

(21) Adams, S. E. Molecular Similarity and Xenobiotic Metabolism. Ph.D. Thesis, University of Cambridge, Cambridge, U.K., 2010.

(22) Boyer, S.; Arnby, C. H.; Carlsson, L.; Smith, J.; Stein, V.; Glen, R. C. Reaction site mapping of xenobiotic biotransformations. *J. Chem. Inf. Model.* **2007**, *47*, 583−590.

(23) Carlsson, L.; Spjuth, O.; Adams, S.; Glen, R. C.; Boyer, S. Use of historic metabolic biotransformation data as a means of anticipating metabolic sites using MetaPrint2D and Bioclipse. *BMC Bioinformatics* **2010**, *11*, 362.

(24) *MetaPrint2D*, version 1.0; Unilever Centre for Molecular Science Informatics, University of Cambridge: Cambridge, U.K., 2010.

(25) MetaPrint2D. http://www-metaprint2d.ch.cam.ac.uk/metaprint2d (accessed 01-12-2012).

(26) *Accelrys Metabolite Database*, version 2011.2; Accelrys: San Diego, CA, 2011.

(27) MetaPrint2D-react. http://www-metaprint2d.ch.cam.ac.uk/metaprint2d-react (accessed 01-12-2012).

(28) *ADMET Predictor, Metabolite Software Module*, version 5.5; Simulations Plus: Lancaster, CA, 2011.

(29) Carrieri, A.; Pérez-Nueno, V. I.; Fano, A.; Pistone, C.; Ritchie, D. W.; Teixidó, J. Biological profiling of anti-HIV agents and insight into CCR5 antagonist binding using in silico techniques. *ChemMedChem* **2009**, *4*, 1153−1163.

(30) Kirchmair, J.; Distinto, S.; Markt, P.; Schuster, D.; Spitzer, G. M.; Liedl, K. R.; Wolber, G. How to optimize shape-based virtual screening: Choosing the right query and including chemical information. *J. Chem. Inf. Model.* **2009**, *49*, 678−692.

(31) Kirchmair, J.; Ristic, S.; Eder, K.; Markt, P.; Wolber, G.; Laggner, C.; Langer, T. Fast and efficient in silico 3D screening: Toward maximum computational efficiency of pharmacophore-based and shape-based approaches. *J. Chem. Inf. Model.* **2007**, *47*, 2182−2196.

(32) Pérez-Nueno, V. I.; Pettersson, S.; Ritchie, D. W.; Borrell, J. I.; Teixidó, J. Discovery of novel HIV entry inhibitors for the CXCR4 receptor by prospective virtual screening. *J. Chem. Inf. Model.* **2009**, *49*, 810−823.

(33) Pérez-Nueno, V. I.; Ritchie, D. W.; Borrell, J. I.; Teixidó, J. Clustering and classifying diverse HIV entry inhibitors using a novel consensus shape-based virtual screening approach: Further evidence for multiple binding sites within the CCR5 extracellular pocket. *J. Chem. Inf. Model.* **2008**, *48*, 2146−2165.

(34) Venkatraman, V.; Pérez-Nueno, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods. *J. Chem. Inf. Model.* **2010**, *50*, 2079−2093.

(35) Pérez-Nueno, V. I.; Ritchie, D. W. Using consensus-shape clustering to identify promiscuous ligands and protein targets and to choose the right query for shape-based virtual screening. *J. Chem. Inf. Model.* **2011**, *51*, 1233−1248.

(36) Nicholls, A.; Grant, J. A. Molecular shape and electrostatics in the encoding of relevant chemical information. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 661−686.

(37) Sykes, M. J.; McKinnon, R. A.; Miners, J. O. Prediction of metabolism by cytochrome P450 2C9: Alignment and docking studies of a validated database of substrates. *J. Med. Chem.* **2008**, *51*, 780−791.

(38) Goodford, P. J. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *J. Med. Chem.* **1985**, *28*, 849−857.

(39) von Itzstein, M.; Wu, W. Y.; Kok, G. B.; Pegg, M. S.; Dyason, J. C.; Jin, B.; Van Phan, T.; Smythe, M. L.; White, H. F.; Oliver, S. W. Rational design of potent sialidase-based inhibitors of influenza virus replication. *Nature* **1993**, *363*, 418−423.

(40) Cruciani, G.; Carosati, E.; De Boeck, B.; Ethirajulu, K.; Mackie, C.; Howe, T.; Vianello, R. MetaSite: Understanding metabolism in human cytochromes from the perspective of the chemist. *J. Med. Chem.* **2005**, *48*, 6970−6979.

(41) Jorgensen, W. L. The many roles of computation in drug discovery. *Science (New York, N.Y.)* **2004**, *303*, 1813−1818.

(42) Mukherjee, S.; Balius, T. E.; Rizzo, R. C. Docking validation resources: Protein family and ligand flexibility experiments. *J. Chem. Inf. Model.* **2010**, *50*, 1986−2000.

(43) Kirchmair, J.; Spitzer, G.; Liedl, K. R. Consideration of Water and Solvation Effects in Virtual Screening. In *Virtual Screening: Principles, Challenges, and Practical Guidelines*; Sotriffer, C., Ed.; Wiley-VCH: Weinheim, Germany, 2011; pp 263−290.

(44) de Graaf, C.; Pospisil, P.; Pos, W.; Folkers, G.; Vermeulen, N. P. E. Binding mode prediction of cytochrome P450 and thymidine kinase protein-ligand complexes by consideration of water and rescoring in automated docking. *J. Med. Chem.* **2005**, *48*, 2308−2318.

(45) Goodsell, D. S.; Olson, A. J. Automated docking of substrates to proteins by simulated annealing. *Proteins* **1990**, *8*, 195−202.

(46) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(47) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(48) de Graaf, C.; Oostenbrink, C.; Keizers, P. H. J.; van Der Wijst, T.; Jongejan, A.; Vermeulen, N. P. E. Catalytic site prediction and virtual screening of cytochrome P450 2D6 substrates by consideration of water and rescoring in automated docking. *J. Med. Chem.* **2006**, *49*, 2417−2430.

(49) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.

(50) Santos, R.; Hritz, J.; Oostenbrink, C. Role of water in molecular docking simulations of cytochrome P450 2D6. *J. Chem. Inf. Model.* **2010**, *50*, 146−154.

(51) Vasanthanathan, P.; Hritz, J.; Taboureau, O.; Olsen, L.; Jørgensen, F. S.; Vermeulen, N. P. E; Oostenbrink, C. Virtual screening and prediction of site of metabolism for cytochrome P450 1A2 ligands. *J. Chem. Inf. Model.* **2009**, *49*, 43−52.

(52) Unwalla, R. J.; Cross, J. B.; Salaniwal, S.; Shilling, A. D.; Leung, L.; Kao, J.; Humblet, C. Using a homology model of cytochrome P450 2D6 to predict substrate site of metabolism. *J. Comput.-Aided Mol. Des.* **2010**, *24*, 237−256.

(53) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(54) Berellini, G.; Cruciani, G.; Mannhold, R. Pharmacophore, drug metabolism, and pharmacokinetics models on non-peptide AT1, AT2, and AT1/AT2 angiotensin II receptor antagonists. *J. Med. Chem.* **2005**, *48*, 4389−4399.

(55) Zamora, I.; Afzelius, L.; Cruciani, G. Predicting drug metabolism: A site of metabolism prediction tool applied to the cytochrome P450 2C9. *J. Med. Chem.* **2003**, *46*, 2313−2324.

(56) *MetaSite User Manual 3.1.2*; Molecular Discovery: Ponte San Giovanni, Perugia, Italy, 2008.

(57) Vaz, R. J.; Zamora, I.; Li, Y.; Reiling, S.; Shen, J.; Cruciani, G. The challenges of in silico contributions to drug metabolism in lead optimization. *Expert Opin. Drug Metab. Toxicol.* **2010**, *6*, 851−861.

(58) Zhou, D.; Afzelius, L.; Grimm, S. W.; Andersson, T. B.; Zauhar, R. J.; Zamora, I. Comparison of methods for the prediction of the metabolic sites for CYP3A4-mediated metabolic reactions. *Drug Metab. Dispos.* **2006**, *34*, 976−983.

(59) Trunzer, M.; Faller, B.; Zimmerlin, A. Metabolic soft spot identification and compound optimization in early discovery phases using MetaSite and LC-MS/MS validation. *J. Med. Chem.* **2009**, *52*, 329−335.

(60) de Groot, M. J.; Ackland, M. J.; Horne, V. A.; Alex, A. A.; Jones, B. C. Novel approach to predicting P450-mediated drug metabolism: Development of a combined protein and pharmacophore model for CYP2D6. *J. Med. Chem.* **1999**, *42*, 1515−1524.

(61) Rydberg, P.; Gloriam, D.; Olsen, L. The SMARTCyp cytochrome P450 metabolism prediction server. *Bioinformatics* **2010**, *26*, 2988−2989.

(62) Rydberg, P.; Gloriam, D. E.; Zaretzki, J.; Breneman, C.; Olsen, L. SMARTCyp: A 2D method for prediction of cytochrome P450-mediated drug metabolism. *ACS Med. Chem. Lett.* **2010**, *1*, 96−100.

(63) Rydberg, P.; Olsen, L. Ligand-based site of metabolism prediction for cytochrome P450 2D6. *ACS Med. Chem. Lett.* **2011**, *3*, 69−73.

(64) *StarDrop*, version 5.0; Optibrium: Cambridge, U.K., 2011.

(65) Hasegawa, K.; Koyama, M.; Funatsu, K. Quantitative prediction of regioselectivity toward cytochrome P450/3A4 using machine learning approaches. *Mol. Inf.* **2010**, *29*, 243−249.

(66) *Molecular Operating Environment (MOE)*, version 2011.10; Chemical Computing Group: Montreal, Quebec, Canada, 2011.

(67) Mu, F.; Unkefer, C. J.; Unkefer, P. J.; Hlavacek, W. S. Prediction of metabolic reactions based on atomic and molecular properties of small-molecule compounds. *Bioinformatics* **2011**, *27*, 1537−1545.

(68) Kanehisa, M.; Goto, S.; Sato, Y.; Furumichi, M.; Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **2012**, *40*, D109−D114.

(69) Kuhn, B.; Jacobsen, W.; Christians, U.; Benet, L. Z.; Kollman, P. A. Metabolism of sirolimus and its derivative everolimus by cytochrome P450 3A4: Insights from docking, molecular dynamics, and quantum chemical calculations. *J. Med. Chem.* **2001**, *44*, 2027−2034.

(70) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: Search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.

(71) Pearlman, D. A.; Case, D. A.; Caldwell, J. W.; Ross, W. S.; Cheatham, T. E. III; DeBolt, S.; Ferguson, D.; Seibel, G.; Kollman, P. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Comput. Phys. Commun.* **1995**, *91*, 1−41.

(72) Case, D. A.; Cheatham, T. E.; Darden, T.; Gohlke, H.; Luo, R.; Merz, K. M.; Onufriev, A.; Simmerling, C.; Wang, B.; Woods, R. J. The AMBER biomolecular simulation programs. *J. Comput. Chem.* **2005**, *26*, 1668−1688.

(73) Oh, W. S.; Kim, D. N.; Jung, J.; Cho, K.-H.; No, K. T. New combined model for the prediction of regioselectivity in cytochrome P450/3A4 mediated metabolism. *J. Chem. Inf. Model.* **2008**, *48*, 591−601.

(74) Korzekwa, K. R.; Jones, J. P.; Gillette, J. R. Theoretical studies on cytochrome P-450 mediated hydroxylation: A predictive model for hydrogen atom abstractions. *J. Am. Chem. Soc.* **1990**, *112*, 7042−7046.

(75) Jones, J. P. Computational models for cytochrome P450: A predictive electronic model for aromatic oxidation and hydrogen atom abstraction. *Drug Metab. Dispos.* **2002**, *30*, 7−12.

(76) Jung, J.; Kim, N. D.; Kim, S. Y.; Choi, I.; Cho, K.-H.; Oh, W. S.; Kim, D. N.; No, K. T. Regioselectivity prediction of CYP1A2-mediated phase I metabolism. *J. Chem. Inf. Model.* **2008**, *48*, 1074−1080.

(77) Amaro, R. E.; Li, W. W. Emerging methods for ensemble-based virtual screening. *Curr. Top. Med. Chem.* **2010**, *10*, 3−13.

(78) Ekroos, M.; Sjögren, T. Structural basis for ligand promiscuity in cytochrome P450 3A4. *Proc. Natl. Acad. Sci. U.S.A.* **2006**, *103*, 13682−13687.

(79) Teixeira, V. H.; Ribeiro, V.; Martel, P. J. Analysis of binding modes of ligands to multiple conformations of CYP3A4. *Biochim. Biophys. Acta* **2010**, *1804*, 2036−2045.

(80) Hritz, J.; de Ruiter, A.; Oostenbrink, C. Impact of plasticity and flexibility on docking results for cytochrome P450 2D6: A combined approach of molecular dynamics and ligand docking. *J. Med. Chem.* **2008**, *51*, 7469−7477.

(81) Keizers, P. H. J.; de Graaf, C.; de Kanter, F. J. J.; Oostenbrink, C.; Feenstra, K. A.; Commandeur, J. N. M.; Vermeulen, N. P. E. Metabolic regio- and stereoselectivity of cytochrome P450 2D6 towards 3,4-methylenedioxy-n-alkylamphetamines: In silico predictions and experimental validation. *J. Med. Chem.* **2005**, *48*, 6117−6127.

(82) Moors, S. L. C.; Vos, A. M.; Cummings, M. D.; Van Vlijmen, H.; Ceulemans, A. Structure-based site of metabolism prediction for cytochrome P450 2D6. *J. Med. Chem.* **2011**, *54*, 6098−6105.

(83) Seeliger, D.; Haas, J. r.; de Groot, B. L. Geometry-based sampling of conformational transitions in proteins. *Structure* **2007**, *15*, 1482−1492.

(84) Li, J.; Schneebeli, S. T.; Bylund, J.; Farid, R.; Friesner, R. A. IDSite: An accurate approach to predict P450-mediated drug metabolism. *J. Chem. Theory Comput.* **2011**, *7*, 3829−3845.

(85) *Prime*, version 3.0.111; Schrödinger: New York, 2011.

(86) *Jaguar*, version 3.0.111; Schrödinger: New York, 2011.

(87) Walker, G. S.; O'Connell, T. N. Comparison of LC-NMR and conventional NMR for structure elucidation in drug metabolism studies. *Expert Opin. Drug Metab. Toxicol.* **2008**, *4*, 1295−1305.

(88) Darvas, F. In *MetabolExpert: An Expert System for Predicting Metabolism of Substances*; Kaiser, K. L. E., Ed.; D. Reidel Publishing Co.: Dordrecht, Holland, 1987; pp 71−81.

(89) Klopman, G.; Dimayuga, M.; Talafous, J. META. 1. A program for the evaluation of metabolic transformation of chemicals. *J. Chem. Inf. Model.* **1994**, *34*, 1320−1325.

(90) Marchant, C. A.; Briggs, K. A.; Long, A. In silico tools for sharing data and knowledge on toxicity and metabolism: DEREK for Windows, METEOR, and VITIC. *Toxicol. Mech. Methods* **2008**, *18*, 177−187.

(91) Gao, J.; Ellis, L. B. M.; Wackett, L. P. The University of Minnesota Biocatalysis/Biodegradation Database: Improving public access. *Nucleic Acids Res.* **2010**, *38*, D488−491.

(92) Ridder, L.; Wagener, M. SyGMa: Combining expert knowledge and empirical scoring in the prediction of metabolites. *ChemMedChem* **2008**, *3*, 821−832.

(93) Mekenyan, O. G.; Dimitrov, S. D.; Pavlov, T. S.; Veith, G. D. A systematic approach to simulating metabolism in computational toxicology. I. The TIMES heuristic modelling framework. *Curr. Pharm. Des.* **2004**, *10*, 1273−1293.

(94) *JChem, Metabolizer Software Module*, version 5.7.1; ChemAxon: Budapest, Hungary, 2011.

(95) Tarcsay, A.; Kiss, R.; Keserű, G. M. Site of metabolism prediction on cytochrome P450 2C9: A knowledge-based docking approach. *J. Comput.-Aided Mol. Des.* **2010**, 399−408.

(96) Otyepka, M.; Skopalík, J.; Anzenbacherová, E.; Anzenbacher, P. What common structural features and variations of mammalian P450s are known to date? *Biochim. Biophys. Acta* **2007**, *1770*, 376−389.

(97) Lewis, D. F. V.; Ito, Y. Human CYPs involved in drug metabolism: Structures, substrates and binding affinities. *Expert Opin. Drug Metab. Toxicol.* **2010**, *6*, 661−674.

(98) Williams, P. A.; Cosme, J.; Ward, A.; Angova, H. C.; Vinkovic, D. M.; Jhoti, H. Crystal structure of human cytochrome P4502C9 with bound warfarin. *Nature* **2003**, *424*, 464−468.

(99) Cruciani, G.; Aristei, Y.; Vianello, R.; Baroni, M. GRID-derived molecular interaction fields for predicting the site of metabolism in human cytochromes. *Methods Princ. Med. Chem.* **2006**, *27*, 273−290.

(100) Petrek, M.; Otyepka, M.; Banas, P.; Kosinova, P.; Koca, J.; Damborsky, J. CAVER: A new tool to explore routes from protein clefts, pockets and cavities. *BMC Bioinf.* **2006**, *7*, 316.

(101) Cojocaru, V.; Winn, P. J.; Wade, R. C. The ins and outs of cytochrome P450s. *Biochim. Biophys. Acta, Gen. Subj.* **2007**, *1770*, 390−401.

(102) Shaik, S.; Cohen, S.; Wang, Y.; Chen, H.; Kumar, D.; Thiel, W. P450 enzymes: Their structure, reactivity, and selectivity-modeled by QM/MM calculations. *Chem. Rev.* **2010**, *110*, 949−1017.

(103) Schöneboom, J. C.; Lin, H.; Reuter, N.; Thiel, W.; Cohen, S.; Ogliaro, F.; Shaik, S. The elusive oxidant species of cytochrome P450 enzymes: Characterization by combined quantum mechanical/molecular mechanical (QM/MM) calculations. *J. Am. Chem. Soc.* **2002**, *124*, 8142−8151.

(104) Ogliaro, F.; Cohen, S.; de Visser, S. P.; Shaik, S. Medium polarization and hydrogen bonding effects on compound I of cytochrome P450: What kind of a radical is it really? *J. Am. Chem. Soc.* **2000**, *122*, 12892−12893.

(105) Rittle, J.; Green, M. T. Cytochrome P450 Compound I: Capture, characterization, and C-H bond activation kinetics. *Science* **2010**, *330*, 933−937.

(106) Collins, J. R.; Camper, D. L.; Loew, G. H. Valproic acid metabolism by cytochrome-P450 - a theoretical-study of stereo-electronic modulators of product distribution. *J. Am. Chem. Soc.* **1991**, *113*, 2736−2743.

(107) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.;

Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The biomolecular simulation program. *J. Comput. Chem.* **2009**, *30*, 1545−1614.

(108) Brooks, B. R.; Bruccoleri, R. E.; Olafson, B. D.; States, D. J.; Swaminathan, S.; Karplus, M. CHARMM: A program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* **1983**, *4*, 187−217.

(109) Bathelt, C. M.; Zurek, J.; Mulholland, A. J.; Harvey, J. N. Electronic structure of compound I in human isoforms of cytochrome P450 from QM/MM modeling. *J. Am. Chem. Soc.* **2005**, *127*, 12900−12908.

(110) Oda, A.; Yamaotsu, N.; Hirono, S. New AMBER force field parameters of heme iron for cytochrome P450s determined by quantum chemical calculations of simplified models. *J. Comput. Chem.* **2005**, *26*, 818−826.

(111) Seifert, A.; Tatzel, S.; Schmid, R. D.; Pleiss, J. Multiple molecular dynamics simulations of human P450 monooxygenase CYP2C9: The molecular basis of substrate binding and regioselectivity toward warfarin. *Proteins: Struct., Funct., Bioinf.* **2006**, *64*, 147−155.

(112) Favia, A. D.; Cavalli, A.; Masetti, M.; Carotti, A.; Recanatini, M. Three-dimensional model of the human aromatase enzyme and density functional parameterization of the iron-containing protoporphyrin IX for a molecular dynamics study of heme-cysteinato cytochromes. *Proteins: Struct., Funct., Bioinf.* **2006**, *62*, 1074−1087.

(113) Autenrieth, F.; Tajkhorshid, E.; Baudry, J.; Luthey-Schulten, Z. Classical force field parameters for the heme prosthetic group of cytochrome c. *J. Comput. Chem.* **2004**, *25*, 1613−1622.

(114) Skopalík, J.; Anzenbacher, P.; Otyepka, M. Flexibility of human cytochromes P450: Molecular dynamics reveals differences between CYPs 3A4, 2C9, and 2A6, which correlate with their substrate preferences. *J. Phys. Chem. B* **2008**, *112*, 8165−8173.

(115) Shahrokh, K.; Orendt, A.; Yost, G. S.; Cheatham, T. E. Quantum mechanically derived AMBER-compatible heme parameters for various states of the cytochrome P450 catalytic cycle. *J. Comput. Chem.* **2011**, 119−133.

(116) Helms, V.; Deprez, E.; Gill, E.; Barret, C.; Hui Bon Hoa, G.; Wade, R. C. Improved binding of cytochrome P450cam substrate analogues designed to fill extra space in the substrate binding pocket. *Biochemistry* **1996**, *35*, 1485−1499.

(117) Das, B.; Helms, V.; Lounnas, V.; Wade, R. C. Multicopy molecular dynamics simulations suggest how to reconcile crystallographic and product formation data for camphor enantiomers bound to cytochrome P-450cam. *J. Inorgan. Biochem.* **2000**, *81*, 121−131.

(118) Mathieu, A. P.; LeHoux, J.-G.; Auchus, R. J. Molecular dynamics of substrate complexes with hamster cytochrome P450c17 (CYP17): Mechanistic approach to understanding substrate binding and activities. *Biochim. Biophys. Acta, Gen. Subj.* **2003**, *1619*, 291−300.

(119) Gorokhov, A.; Negishi, M.; Johnson, E. F.; Pedersen, L. C.; Perera, L.; Darden, T. A.; Pedersen, L. G. Explicit water near the catalytic I helix Thr in the predicted solution structure of CYP2A4. *Biophys. J.* **2003**, *84*, 57−68.

(120) Strobel, S. M.; Szklarz, G. D.; He, Y. Q.; Foroozesh, M.; Alworth, W. L.; Roberts, E. S.; Hollenberg, P. F.; Halpert, J. R. Identification of selective mechanism-based inactivators of cytochromes P-450 2B4 and 2B5, and determination of the molecular basis for differential susceptibility. *J. Pharmacol. Exp. Ther.* **1999**, *290*, 445−451.

(121) Park, H.; Lee, S.; Suh, J. Structural and dynamical basis of broad substrate specificity, catalytic mechanism, and inhibition of cytochrome P450 3A4. *J. Am. Chem. Soc.* **2005**, *127*, 13634−13642.

(122) Lampe, J. N.; Brandman, R.; Sivaramakrishnan, S.; de Montellano, P. R. O. Two-dimensional NMR and all-atom molecular dynamics of cytochrome P450 CYP119 reveal hidden conformational substates. *J. Biol. Chem.* **2010**, *285*, 9594−9603.

(123) Brandman, R.; Lampe, J. N.; Brandman, Y.; de Montellano, P. R. O. Active-site residues move independently from the rest of the protein in a 200 ns molecular dynamics simulation of cytochrome P450 CYP119. *Arch. Biochem. Biophys.* **2011**, *509*, 127−132.

(124) Sano, E.; Li, W.; Yuki, H.; Liu, X.; Furihata, T.; Kobayashi, K.; Chiba, K.; Neya, S.; Hoshino, T. Mechanism of the decrease in catalytic activity of human cytochrome P450 2C9 polymorphic variants investigated by computational analysis. *J. Comput. Chem.* **2010**, *31*, 2746−2758.

(125) Chang, Y.-T.; Loew, G. Homology modeling, molecular dynamics simulations, and analysis of CYP119, a P450 enzyme from extreme acidothermophilic archaeon sulfolobus solfataricus. *Biochemistry* **2000**, *39*, 2484−2498.

(126) Roberts, A. G.; Cheesman, M. J.; Primak, A.; Bowman, M. K.; Atkins, W. M.; Rettie, A. E. Intramolecular heme ligation of the cytochrome P450 2C9 R108H mutant demonstrates pronounced conformational flexibility of the B-C loop region: Implications for substrate binding. *Biochemistry* **2010**, *49*, 8700−8708.

(127) Asciutto, E. K.; Dang, M.; Pochapsky, S. S.; Madura, J. D.; Pochapsky, T. C. Experimentally restrained molecular dynamics simulations for characterizing the open states of cytochrome P450cam. *Biochemistry* **2011**, *50*, 1664−1671.

(128) Fishelovitch, D.; Hazan, C.; Shaik, S.; Wolfson, H. J.; Nussinov, R. Structural dynamics of the cooperative binding of organic molecules in the human cytochrome P450 3A4. *J. Am. Chem. Soc.* **2007**, *129*, 1602−1611.

(129) Oprea, T. I.; Hummer, G.; Garcia, A. E. Identification of a functional water channel in cytochrome P450 enzymes. *Proc. Natl. Acad. Sci. U.S.A.* **1997**, *94*, 2133−2138.

(130) Hendrychová, T.; Anzenbacherova, E.; Hudecek, J.; Skopalík, J.; Lange, R.; Hildebrandt, P.; Otyepka, M.; Anzenbacher, P. Flexibility of human cytochrome P450 enzymes: Molecular dynamics and spectroscopy reveal important function-related variations. *Biochim. Biophys. Acta, Proteins Proteomics* **2011**, *1814*, 58−68.

(131) Helms, V.; Wade, R. C. Hydration energy landscape of the active site cavity in cytochrome P450cam. *Proteins: Struct., Funct., Bioinf.* **1998**, *32*, 381−396.

(132) Miao, Y.; Baudry, J. Active-site hydration and water diffusion in cytochrome P450cam: A highly dynamic process. *Biophys. J.* **2011**, *101*, 1493−1503.

(133) Rydberg, P.; Rod, T. H.; Olsen, L.; Ryde, U. Dynamics of water molecules in the active-site cavity of human cytochromes P450. *J. Phys. Chem. B* **2007**, *111*, 5445−5457.

(134) Roccatano, D.; Wong, T. S.; Schwaneberg, U.; Zacharias, M. Structural and dynamic properties of cytochrome P450 BM-3 in pure water and in a dimethylsulfoxide/water mixture. *Biopolymers* **2005**, *78*, 259−267.

(135) Roccatano, D.; Wong, T. S.; Schwaneberg, U.; Zacharias, M. Toward understanding the inactivation mechanism of monooxygenase P450 BM-3 by organic cosolvents: A molecular dynamics simulation study. *Biopolymers* **2006**, *83*, 467−476.

(136) Mouawad, L.; Tetreau, C.; Abdel-Azeim, S.; Perahia, D.; Lavalette, D. Co migration pathways in cytochrome P450cam studied by molecular dynamics simulations. *Protein Sci.* **2007**, *16*, 781−794.

(137) Fishelovitch, D.; Shaik, S.; Wolfson, H. J.; Nussinov, R. How does the reductase help to regulate the catalytic cycle of cytochrome P450 3A4 using the conserved water channel? *J. Phys. Chem. B* **2010**, *114*, 5964−5970.

(138) Yaffe, E.; Fishelovitch, D.; Wolfson, H. J.; Halperin, D.; Nussinov, R. MolAxis: Efficient and accurate identification of channels in macromolecules. *Proteins: Struct., Funct., Bioinf.* **2008**, *73*, 72−86.

(139) Krishnamoorthy, N.; Gajendrarao, P.; Thangapandian, S.; Lee, Y.; Lee, K. W. Probing possible egress channels for multiple ligands in human CYP3A4: A molecular modeling study. *J. Mol. Model.* **2010**, *16*, 607−614.

(140) Haider, S. M.; Patel, J. S.; Poojari, C. S.; Neidle, S. Molecular modeling on inhibitor complexes and active-site dynamics of cytochrome P450 C17, a target for prostate cancer therapy. *J. Mol. Biol.* **2010**, *400*, 1078−1098.

(141) Petrek, M.; Kosinova, P.; Koca, J.; Otyepka, M. MOLE: A Voronoi diagram-based explorer of molecular channels, pores, and tunnels. *Structure* **2007**, *15*, 1357−1363.

(142) Markwick, P. R. L.; Pierce, L. C. T.; Goodin, D. B.; McCammon, J. A. Adaptive accelerated molecular dynamics (Ad-AMD) revealing the molecular plasticity of P450cam. *J. Phys. Chem. Lett.* **2011**, *2*, 158−164.

(143) Wade, R. C.; Motiejunas, D.; Schleinkofer, K.; Sudarko; Winn, P. J.; Banerjee, A.; Kariakin, A.; Jung, C. Multiple molecular recognition mechanisms. Cytochrome P450: A case study. *Biochim. Biophys. Acta, Proteins Proteom.* **2005**, *1754*, 239−244.

(144) Luedemann, S. K.; Lounnas, V. r.; Wade, R. C. How do substrates enter and products exit the buried active site of cytochrome P450cam? 1. Random expulsion molecular dynamics investigation of ligand access channels and mechanisms. *J. Mol. Biol.* **2000**, *303*, 797−811.

(145) Luedemann, S. K.; Lounnas, V. r.; Wade, R. C. How do substrates enter and products exit the buried active site of cytochrome P450cam? 2. Steered molecular dynamics and adiabatic mapping of substrate pathways. *J. Mol. Biol.* **2000**, *303*, 813−830.

(146) Schleinkofer, K.; Sudarko; Winn, P. J.; Ludemann, S. K.; Wade, R. C. Do mammalian cytochrome P450s show multiple ligand access pathways and ligand channelling? *EMBO Rep.* **2005**, *6*, 584−589.

(147) Cojocaru, V.; Balali-Mood, K.; Sansom, M. S. P.; Wade, R. C. Structure and dynamics of the membrane-bound cytochrome P450 2C9. *PLoS Comput. Biol.* **2011**, *7*, e1002152.

(148) Li, W.; Liu, H.; Scott, E. E.; Graeter, F.; Halpert, J. R.; Luo, X.; Shen, J.; Jiang, H. Possible pathway(s) of testosterone egress from the active site of cytochrome P450 2B1: A steered molecular dynamics simulation. *Drug Metab. Dispos.* **2005**, *33*, 910−919.

(149) Scott, E. E.; Liu, H.; Qun, He, Y.; Li, W.; Halpert, J. R. Mutagenesis and molecular dynamics suggest structural and functional roles for residues in the n-terminal portion of the cytochrome P450 2B1 I helix. *Arch. Biochem. Biophys.* **2004**, *423*, 266−276.

(150) Fishelovitch, D.; Shaik, S.; Wolfson, H. J.; Nussinov, R. Theoretical characterization of substrate access/exit channels in the human cytochrome P450 3A4 enzyme: Involvement of phenylalanine residues in the gating mechanism. *J. Phys. Chem. B* **2009**, *113*, 13018−13025.

(151) Li, W.; Liu, H.; Luo, X.; Zhu, W.; Tang, Y.; Halpert, J. R.; Jiang, H. Possible pathway(s) of metyrapone egress from the active site of cytochrome P450 3A4: A molecular dynamics simulation. *Drug Metab. Dispos.* **2007**, *35*, 689−696.

(152) Yang, K.; Liu, X.; Wang, X.; Jiang, H. A steered molecular dynamics method with adaptive direction adjustments. *Biochem. Bioph. Res. Co.* **2009**, *379*, 494−498.

(153) Fukunishi, H.; Yagi, H.; Kamijo, K. Ä.; Shimada, J. Role of a mutated residue at the entrance of the substrate access channel in cytochrome P450 engineered for vitamin D3 hydroxylation activity. *Biochemistry* **2011**, *50*, 8302−8310.

(154) Åqvist, J.; Luzhkov, V. B.; Brandsdal, B. O. Ligand binding affinities from MD simulations. *Acc. Chem. Res.* **2002**, *35*, 358−365.

(155) Brandsdal, B. O.; Österberg, F.; Almlöf, M.; Feierberg, I.; Luzhkov, V. B.; Åvist, J. Free Energy Calculations and Ligand Binding. In *Advances in Protein Chemistry*; Valerie, D., Ed.; Academic Press: Burlington, MA, 2003; Vol. *66*, pp 123−158.

(156) Paulsen, M. D.; Ornstein, R. L. Binding free energy calculations for P450cam-substrate complexes. *Protein Eng.* **1996**, *9*, 567−571.

(157) Vasanthanathan, P.; Olsen, L.; Jørgensen, F. S.; Vermeulen, N. P. E.; Oostenbrink, C. Computational prediction of binding affinity for CYP1A2-ligand complexes using empirical free energy calculations. *Drug Metab. Dispos.* **2010**, *38*, 1347−1354.

(158) Karlsson, M.; Strid, Å.; Sirsjö, A.; Eriksson, L. A. Homology models and molecular modeling of human retinoic acid metabolizing enzymes cytochrome P450 26A1 (CYP26A1) and P450 26B1 (CYP26B1). *J. Chem. Theory Comput.* **2008**, *4*, 1021−1027.

(159) Stjernschantz, E.; Oostenbrink, C. Improved ligand-protein binding affinity predictions using multiple binding modes. *Biophys. J.* **2010**, *98*, 2682−2691.

(160) Durrant, J.; McCammon, J. A. Molecular dynamics simulations and drug discovery. *BMC Biology* **2011**, *9*, 71.

(161) Deng, Y.; Roux, B. Computations of standard binding free energies with molecular dynamics simulations. *J. Phys. Chem. B* **2009**, *113*, 2234−2246.

(162) Helms, V.; Wade, R. C. Computational alchemy to calculate absolute protein-ligand binding free energy. *J. Am. Chem. Soc.* **1998**, *120*, 2710−2713.

(163) Deng, Y. Computation of binding free energy with molecular dynamics and grand canonical Monte Carlo simulations. *J. Chem. Phys.* **2008**, *128*, 115103.

(164) Schöneboom, J. C.; Cohen, S.; Lin, H.; Shaik, S.; Thiel, W. Quantum mechanical/molecular mechanical investigation of the mechanism of C-H hydroxylation of camphor by cytochrome P450cam: Theory supports a two-state rebound mechanism. *J. Am. Chem. Soc.* **2004**, *126*, 4017−4034.

(165) Schöneboom, J. C.; Neese, F.; Thiel, W. Toward identification of the Compound I reactive intermediate in cytochrome P450 chemistry: A QM/MM study of its EPR and Mossbauer parameters. *J. Am. Chem. Soc.* **2005**, *127*, 5840−5853.

(166) Rassolov, V. A.; Ratner, M. A.; Pople, J. A.; Redfern, P. C.; Curtiss, L. A. 6-31G* basis set for third-row atoms. *J. Comput. Chem.* **2001**, *22*, 976−984.

(167) Bathelt, C. M.; Mulholland, A. J.; Harvey, J. N. QM/MM modeling of benzene hydroxylation in human cytochrome P450 2C9. *J. Phys. Chem. A* **2008**, *112*, 13149−13156.

(168) Fishelovitch, D.; Hazan, C.; Hirao, H.; Wolfson, H. J.; Nussinov, R.; Shaik, S. QM/MM study of the active species of the human cytochrome P450 3A4, and the influence thereof of the multiple substrate binding. *J. Phys. Chem. B* **2007**, *111*, 13822−13832.

(169) Sen, K.; Hackett, J. C. Molecular oxygen activation and proton transfer mechanisms in lanosterol 14 alpha-demethylase catalysis. *J. Phys. Chem. B* **2009**, *113*, 8170−8182.

(170) Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. Compound I reactivity defines alkene oxidation selectivity in cytochrome P450cam. *J. Phys. Chem. B* **2010**, *114*, 1156−1162.

(171) Lonsdale, R.; Harvey, J. N.; Mulholland, A. J. Inclusion of dispersion effects significantly improves accuracy of calculated reaction barriers for cytochrome P450 catalyzed reactions. *J. Phys. Chem. Lett.* **2010**, *1*, 3232−3237.

(172) Li, D.; Huang, X.; Han, K.; Zhan, C.-G. Catalytic mechanism of cytochrome P450 for 5 ′-hydroxylation of nicotine: Fundamental reaction pathways and stereoselectivity. *J. Am. Chem. Soc.* **2011**, *133*, 7416−7427.

(173) Schyman, P.; Lai, W.; Chen, H.; Wang, Y.; Shaik, S. The directive of the protein: How does cytochrome P450 select the mechanism of dopamine formation? *J. Am. Chem. Soc.* **2011**, *133*, 7977−7984.

(174) Lonsdale, R.; Olah, J.; Mulholland, A. J.; Harvey, J. N. Does Compound I vary significantly between isoforms of cytochrome P450? *J. Am. Chem. Soc.* **2011**, *133*, 15464−15474.

(175) Guengerich, F. P. Cytochrome P450 and chemical toxicology. *Chem. Res. Toxicol.* **2008**, *21*, 70−83.

(176) de Groot, M. J. Designing better drugs: Predicting cytochrome P450 metabolism. *Drug Discovery Today* **2006**, *11*, 601−606.

(177) Chu, V.; Einolf, H. J.; Evers, R.; Kumar, G.; Moore, D.; Ripp, S.; Silva, J.; Sinha, V.; Sinz, M.; Skerjanec, A. In vitro and in vivo induction of cytochrome P450: A survey of the current practices and recommendations: A pharmaceutical research and manufacturers of america perspective. *Drug Metab. Dispos.* **2009**, *37*, 1339−1354.

(178) Stjernschantz, E.; Vermeulen, N. P. E; Oostenbrink, C. Computational prediction of drug binding and rationalisation of selectivity towards cytochromes P450. *Expert Opin. Drug Metab. Toxicol.* **2008**, *4*, 513−527.

(179) Zhang, L.; Zhang, Y. D.; Zhao, P.; Huang, S.-M. Predicting drug-drug interactions: An FDA perspective. *AAPS J.* **2009**, *11*, 300−306.

(180) Hewitt, N. J.; de Kanter, R.; LeCluyse, E. Induction of drug metabolizing enzymes: A survey of in vitro methodologies and

interpretations used in the pharmaceutical industry - do they comply with FDA recommendations? *Chem.−Biol. Interact.* **2007**, *168*, 51−65.

(181) Wienkers, L. C.; Heath, T. G. Predicting in vivo drug interactions from in vitro drug discovery data. *Nat. Rev. Drug Discovery* **2005**, *4*, 825−833.

(182) Zhou, S.-F. Drugs behave as substrates, inhibitors and inducers of human cytochrome P450 3A4. *Curr. Drug Metab.* **2008**, *9*, 310−322.

(183) Pelkonen, O.; Turpeinen, M.; Hakkola, J.; Honkakoski, P.; Hukkanen, J.; Raunio, H. Inhibition and induction of human cytochrome P450 enzymes: Current status. *Arch. Toxicol.* **2008**, *82*, 667−715.

(184) Li, H.; Sun, J.; Fan, X.; Sui, X.; Zhang, L.; Wang, Y.; He, Z. Considerations and recent advances in QSAR models for cytochrome P450-mediated drug metabolism prediction. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 843−855.

(185) Roy, K.; Roy, P. P. QSAR of cytochrome inhibitors. *Expert Opin. Drug Metab. Toxicol.* **2009**, *5*, 1245−1266.

(186) Bender, A.; Glen, R. C. Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2*, 3204−3218.

(187) Bender, A.; Jenkins, J. L.; Scheiber, J.; Sukuru, S. C.; Glick, M.; Davies, J. W. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J. Chem. Inf. Model.* **2009**, *49*, 108−119.

(188) Mitchell, T. M. *Machine Learning*; McGraw-Hill: New York, 1997.

(189) Zvinavashe, E.; Murk, A. J.; Rietjens, I. M. Promises and pitfalls of quantitative structure-activity relationship approaches for predicting metabolism and toxicity. *Chem. Res. Toxicol.* **2008**, *21*, 2229−2236.

(190) OECD Quantitative Structure−Activity Relationships Project. http://www.oecd.org/document/23/0,3746,en_2649_34377_33957015_1_1_1_1,00.html (accessed 01-12-2012).

(191) Li, Q.; Bender, A.; Pei, J.; Lai, L. A large descriptor set and a probabilistic kernel-based classifier significantly improve drug-likeness classification. *J. Chem. Inf. Model.* **2007**, *47*, 1776−1786.

(192) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: A classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

(193) Chohan, K. K.; Paine, S. W.; Waters, N. J. Quantitative structure activity relationships in drug metabolism. *Curr. Top. Med. Chem.* **2006**, *6*, 1569−1578.

(194) Hansch, C.; Mekapati, S. B.; Kurup, A.; Verma, R. P. QSAR of cytochrome P450. *Drug Metab Rev* **2004**, *36*, 105−156.

(195) Burton, J.; Ijjaali, I.; Petitet, F.; Michel, A.; Vercauteren, D. P. Virtual screening for cytochromes P450: Successes of machine learning filters. *Comb. Chem. High Throughput Screening* **2009**, *12*, 369−382.

(196) Fox, T.; Kriegl, J. M. Machine learning techniques for in silico modeling of drug metabolism. *Curr. Top. Med. Chem.* **2006**, *6*, 1579−1591.

(197) Klon, A. E. Machine learning algorithms for the prediction of hERG and CYP450 binding in drug development. *Expert Opin. Drug Metab. Toxicol.* **2010**, *6*, 821−833.

(198) Kulkarni, S. A.; Zhu, J.; Blechinger, S. In silico techniques for the study and prediction of xenobiotic metabolism: A review. *Xenobiotica* **2005**, *35*, 955−973.

(199) Bender, A. Databases: Compound bioactivities go public. *Nat. Chem. Biol.* **2010**, *6*, 309−309.

(200) Gupta, R. R.; Gifford, E. M.; Liston, T.; Wallker, C. L.; Hohman, M.; Bunun, B. A.; Ekins, S. Using open source computational tools for predicting human metabolic stability and additional absorption, distribution, metabolism, excretion, and toxicity properties. *Drug Metab. Dispos.* **2010**, *38*, 2083−2890.

(201) Gleeson, M. P.; Modi, S.; Bender, A.; Marchese-Robinson, R. L.; Kirchmair, J.; Promkatkaew, M.; Hannongbua, S.; Glen, R. C. The challenges involved in modeling toxicity data in silico: A review. *Curr. Pharm. Des.* **2012**, article in press.

(202) Hammann, F.; Gutmann, H.; Baumann, U.; Helma, C.; Drewe, J. Classification of cytochrome P450 activities using machine learning methods. *Mol. Pharmaceutics* **2009**, *6*, 1920−1926.

(203) Vasanthanathan, P.; Taboureau, O.; Oostenbrink, C.; Vermeulen, N. P. E.; Olsen, L.; Jørgensen, F. S. Classification of cytochrome P450 1A2 inhibitors and noninhibitors by machine learning techniques. *Drug Metab. Dispos.* **2009**, *37*, 658−664.

(204) Michielan, L.; Terfloth, L.; Gasteiger, J.; Moro, S. Comparison of multilabel and single-label classification applied to the prediction of the isoform specificity of cytochrome P450 substrates. *J. Chem. Inf. Model.* **2009**, *49*, 2588−2605.

(205) Yap, C. W.; Chen, Y. Z. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* **2005**, *45*, 982−992.

(206) Eitrich, T.; Kless, A.; Druska, C.; Meyer, W.; Grotendorst, J. Classification of highly unbalanced CYP450 data of drugs using cost sensitive machine learning techniques. *J. Chem. Inf. Model.* **2007**, *47*, 92−103.

(207) *isoCYP*, version 1.0; Molecular Networks: Erlangen, Germany, 2007.

(208) Terfloth, L.; Bienfait, B.; Gasteiger, J. Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6, and 2C9 substrates. *J. Chem. Inf. Model.* **2007**, *47*, 1688−1701.

(209) Sun, H.; Veith, H.; Xia, M.; Austin, C. P.; Huang, R. Predictive models for cytochrome P450 isozymes based on quantitative high throughput screening data. *J. Chem. Inf. Model.* **2011**, *51*, 2474−2481.

(210) Lewis, D. F.; Modi, S.; Dickins, M. Quantitative structure−activity relationships (QSARs) within substrates of human cytochromes P450 involved in drug metabolism. *Drug Metabol. Drug Interact.* **2001**, *18*, 221−242.

(211) Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. The WEKA data mining software: An update. *SIGKDD Explorations* **2009**, *11*, 10−18.

(212) Dagliyan, O.; Kavakli, I. H.; Turkay, M. Classification of cytochrome P450 inhibitors with respect to binding free energy and pIC50 using common molecular descriptors. *J. Chem. Inf. Model.* **2009**, *49*, 2403−2411.

(213) Mao, B.; Gozalbes, R.; Barbosa, F.; Migeon, J.; Merrick, S.; Kamm, K.; Wong, E.; Costales, C.; Shi, W.; Wu, C.; Froloff, N. QSAR modeling of in vitro inhibition of cytochrome P450 3A4. *J. Chem. Inf. Model.* **2006**, *46*, 2125−2134.

(214) *ACD/ADME Suite, P450 Regioselectivity Module*, version 1.0; ACD/Labs: Toronto, ON, 2011.

(215) Rossato, G.; Ernst, B.; Smiesko, M.; Spreafico, M.; Vedani, A. Probing small-molecule binding to cytochrome P450 2D6 and 2C9: An in silico protocol for generating toxicity alerts. *ChemMedChem* **2010**, *5*, 2088−2101.

(216) Vedani, A.; Spreafico, M.; Peristera, O.; Dobler, M.; Smiesko, M. VirtualToxLab - in silico prediction of the endocrine-disrupting potential of drugs and chemicals. *Chimia* **2008**, *62*, 322−328.

(217) Cramer, R. D.; Patterson, D. E.; Bunce, J. D. Comparative molecular-field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(218) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating optimal linear PLS estimations (GOLPE): An advanced chemometric tool for handling 3D-QSAR problems. *Quant. Struct.−Act. Relat.* **1993**, *12*, 9−20.

(219) Fontaine, F.; Pastor, M.; Sanz, F. Incorporating molecular shape into the alignment-free grid-independent descriptors. *J. Med. Chem.* **2004**, *47*, 2805−2815.

(220) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. Grid-independent descriptors (GRIND): A novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.

(221) *Pentacle*, version 1.05; Molecular Discovery: Ponte San Giovanni, Perugia, Italy, 2010.

(222) Crivori, P.; Zamora, I.; Speed, B.; Orrenius, C.; Poggesi, I. Model based on GRID-derived descriptors for estimating CYP3A4 enzyme stability of potential drug candidates. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 155−166.

(223) Cruciani, G.; Pastor, M.; Guba, W. VolSurf: A new tool for the pharmacokinetic optimization of lead compounds. *Eur. J. Pharm. Sci.* **2000**, *11*, S29−S39.

(224) Afzelius, L.; Masimirembwa, C. M.; Karlén, A.; Andersson, T. B.; Zamora, I. Discriminant and quantitative PLS analysis of competitive CYP2C9 inhibitors versus non-inhibitors using alignment independent GRIND descriptors. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 443−458.

(225) Korhonen, L. E.; Rahnasto, M.; Mähönen, N. J.; Wittekindt, C.; Poso, A.; Juvonen, R. O.; Raunio, H. Predictive three-dimensional quantitative structure−activity relationship of cytochrome P450 1A2 inhibitors. *J. Med. Chem.* **2005**, *48*, 3808−3815.

(226) Korhonen, L. E.; Turpeinen, M.; Rahnasto, M.; Wittekindt, C.; Poso, A.; Pelkonen, O.; Raunio, H.; Juvonen, R. O. New potent and selective cytochrome P450 2B6 (CYP2B6) inhibitors based on three-dimensional quantitative structure-activity relationship (3D-QSAR) analysis. *Br. J. Pharmacol.* **2007**, *150*, 932−942.

(227) Oprea, T. I.; Garcia, A. E. Three-dimensional quantitative structure−activity relationships of steroid aromatase inhibitors. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 186−200.

(228) Poso, A.; Gynther, J.; Juvonen, R. A comparative molecular field analysis of cytochrome P450 2A5 and 2A6 inhibitors. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 195−202.

(229) Afzelius, L.; Zamora, I.; Ridderström, M.; Andersson, T. B.; Karlén, A.; Masimirembwa, C. M. Competitive CYP2C9 inhibitors: Enzyme inhibition studies, protein homology modeling, and three-dimensional quantitative structure-activity relationship analysis. *Mol. Pharmacol.* **2001**, *59*, 909−919.

(230) Goodford, P. In *Atom Movements during Drug−Receptor Interactions*; Alfred Benzon Symp.; Munksgaard International Publishers Ltd.: Munksgaard, Denmark, 1998, pp 215−230.

(231) Ortiz, A. R.; Pisabarro, M. T.; Gago, F.; Wade, R. C. Prediction of drug binding affinities by comparative binding energy analysis. *J. Med. Chem.* **1995**, *38*, 2681−2691.

(232) Cianchetta, G.; Li, Y.; Singleton, R.; Zhang, M.; Wildgoose, M.; Rampe, D.; Kang, J.; Vaz, R. J. Molecular interaction fields in ADME and safety. *Methods Princ. Med. Chem.* **2006**, *27*, 197−218.

(233) de Graaf, C.; Vermeulen, N. P. E; Feenstra, K. A. Cytochrome P450 in silico: An integrative modeling approach. *J. Med. Chem.* **2005**, *48*, 2725−2755.

(234) de Groot, M. J.; Ekins, S. Pharmacophore modeling of cytochromes P450. *Adv. Drug Delivery Rev.* **2002**, *54*, 367−383.

(235) Ekins, S.; Andreyev, S.; Ryabov, A.; Kirillov, E.; Rakhmatulin, E. A.; Bugrim, A.; Nikolskaya, T. Computational prediction of human drug metabolism. *Expert Opin. Drug Metab. Toxicol.* **2005**, *1*, 303−324.

(236) Ekins, S.; de Groot, M. J.; Jones, J. P. Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome P450 active sites. *Drug Metab. Dispos.* **2001**, *29*, 936−944.

(237) Ekins, S.; Mestres, J.; Testa, B. In silico pharmacology for drug discovery: Applications to targets and beyond. *Br. J. Pharmacol.* **2007**, *152*, 21−37.

(238) de Groot, M. J.; Alex, A. A.; Jones, B. C. Development of a combined protein and pharmacophore model for cytochrome P450 2C9. *J. Med. Chem.* **2002**, *45*, 1983−1993.

(239) Locuson, C. W.; Rock, D. A.; Jones, J. P. Quantitative binding models for CYP2C9 based on benzbromarone analogues. *Biochemistry* **2004**, *43*, 6948−6958.

(240) Schuster, D.; Laggner, C.; Steindl, T. M.; Langer, T. Development and validation of an in silico P450 profiler based on pharmacophore models. *Curr. Drug Discovery Technol.* **2006**, *3*, 1−48.

(241) Yu, J.; Paine, M. J. I.; Maréchal, J.-D.; Kemp, C. A.; Ward, C. J.; Brown, S.; Sutcliffe, M. J.; Roberts, G. C. K.; Rankin, E. M.; Wolf, C. R. In silico prediction of drug binding to CYP2D6: Identification of a new metabolite of metoclopramide. *Drug Metab. Dispos.* **2006**, *34*, 1386−1392.

(242) Lozano, J. J.; Pastor, M.; Cruciani, G.; Gaedt, K.; Centeno, N. B.; Gago, F.; Sanz, F. 3D-QSAR methods on the basis of ligand-receptor complexes. Application of combine and GRID/GOLPE methodologies to a series of CYP1A2 ligands. *J. Comput.-Aided Mol. Des.* **2000**, *14*, 341−353.

(243) Leong, M. K.; Chen, Y.-M.; Chen, H.-B.; Chen, P.-H. Development of a new predictive model for interactions with human cytochrome P450 2A6 using pharmacophore ensemble/support vector machine (phe/SVM) approach. *Pharm. Res.* **2009**, *26*, 987−1000.

(244) Kontijevskis, A.; Komorowski, J.; Wikberg, J. E. S. Generalized proteochemometric model of multiple cytochrome p450 enzymes and their inhibitors. *J. Chem. Inf. Model.* **2008**, *48*, 1840−1850.

(245) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **1998**, *41*, 2481−2491.

(246) Bazeley, P. S.; Prithivi, S.; Struble, C. A.; Povinelli, R. J.; Sem, D. S. Synergistic use of compound properties and docking scores in neural network modeling of CYP2D6 binding: Predicting affinity and conformational sampling. *J. Chem. Inf. Model.* **2006**, *46*, 2698−2708.

(247) Vedani, A.; Smiesko, M.; Spreafico, M.; Peristera, O.; Dobler, M. VirtualToxLab - in silico prediction of the toxic (endocrine-disrupting) potential of drugs, chemicals and natural products. Two years and 2,000 compounds of experience: A progress report. *ALTEX* **2009**, *26*, 167−176.

(248) Tompkins, L. M.; Wallace, A. D. Mechanisms of cytochrome P450 induction. *J. Biochem. Mol. Toxicol.* **2007**, *21*, 176−181.

(249) De Lisle, R. K.; Otten, J.; Rhodes, S. In silico modeling of p450 substrates, inhibitors, activators, and inducers. *Comb. Chem. High Throughput Screening* **2011**, *14*, 396−416.

(250) Xue, Y.; Chao, E.; Zuercher, W. J.; Willson, T. M.; Collins, J. L.; Redinbo, M. R. Crystal structure of the PXR-T1317 complex provides a scaffold to examine the potential for receptor antagonism. *Bioorg. Med. Chem.* **2007**, *15*, 2156−2166.

(251) Jacobs, M. N. In silico tools to aid risk assessment of endocrine disrupting chemicals. *Toxicology* **2004**, *205*, 43−53.

(252) Ung, C. Y.; Li, H.; Yap, C. W.; Chen, Y. Z. In silico prediction of pregnane X receptor activators by machine learning approaches. *Mol. Pharmacol.* **2007**, *71*, 158−168.

(253) Ekins, S. A pharmacophore for human pregnane X receptor ligands. *Drug Metab. Dispos.* **2002**, *30*, 96−99.

(254) Lemaire, G.; Benod, C.; Nahoum, V.; Pillon, A.; Boussioux, A.-M.; Guichou, J.-F.; Subra, G.; Pascussi, J.-M.; Bourguet, W.; Chavanieu, A.; Balaguer, P. Discovery of a highly active ligand of human pregnane X receptor: A case study from pharmacophore modeling and virtual screening to "in vivo" biological activity. *Mol. Pharmacol.* **2007**, *72*, 572−581.

(255) Schuster, D.; Langer, T. The identification of ligand features essential for PXR activation by pharmacophore modeling. *J. Chem. Inf. Model.* **2005**, *45*, 431−439.

(256) Yasuda, K.; Ranade, A.; Venkataramanan, R.; Strom, S.; Chupka, J.; Ekins, S.; Schuetz, E.; Bachmann, K. A comprehensive in vitro and in silico analysis of antibiotics that activate pregnane X receptor and induce CYP3A4 in liver and intestine. *Drug Metab. Dispos.* **2008**, *36*, 1689−1697.

(257) Gao, Y. D.; Olson, S. H.; Balkovec, J. M.; Zhu, Y.; Royo, I.; Yabut, J.; Evers, R.; Tan, E. Y.; Tang, W.; Hartley, D. P.; Mosley, R. T. Attenuating pregnane X receptor (PXR) activation: A molecular modelling approach. *Xenobiotica* **2007**, *37*, 124−138.

(258) Xiao, L.; Nickbarg, E.; Wang, W.; Thomas, A.; Ziebell, M.; Prosise, W. W.; Lesburg, C. A.; Taremi, S. S.; Gerlach, V. L.; Le, H. V.; Cheng, K.-C. Evaluation of in vitro PXR-based assays and in silico modeling approaches for understanding the binding of a structurally diverse set of drugs to PXR. *Biochem. Pharmacol.* **2011**, *81*, 669−679.

(259) Bisson, W. H.; Koch, D. C.; O'Donnell, E. F.; Khalil, S. M.; Kerkvliet, N. I.; Tanguay, R. L.; Abagyan, R.; Kolluri, S. K. Modeling of the aryl hydrocarbon receptor (AhR) ligand binding domain and its utility in virtual ligand screening to predict new AhR ligands. *J. Med. Chem.* **2009**, *52*, 5635−5641.

(260) Abagyan, R.; Totrov, M.; Kuznetsov, D. ICM - a new method for protein modeling and design: Applications to docking and

structure prediction from the distorted native conformation. *J. Comput. Chem.* **1994**, *15*, 488−506.

(261) Ekins, S.; Mirny, L.; Schuetz, E. G. A ligand-based approach to understanding selectivity of nuclear hormone receptors PXR, CAR, FXR, LXRalpha, and LXRbeta. *Pharm. Res.* **2002**, *19*, 1788−1800.

(262) Jyrkkärinne, J.; Windshügel, B.; Rönkkö, T.; Tervo, A. J.; Küblbeck, J.; Lahtela-Kakkonen, M.; Sippl, W.; Poso, A.; Honkakoski, P. Insights into ligand-elicited activation of human constitutive androstane receptor based on novel agonists and three-dimensional quantitative structure-activity relationship. *J. Med. Chem.* **2008**, *51*, 7181−7192.

(263) Wu, B.; Zhang, Y.; Kong, J.; Zhang, X.; Cheng, S. In silico predication of nuclear hormone receptors for organic pollutants by homology modeling and molecular docking. *Toxicol. Lett.* **2009**, *191*, 69−73.

(264) Overington, J. ChEMBL. An interview with John Overington, team leader, chemogenomics at the European Bioinformatics Institute Outstation of the European Molecular Biology Laboratory (EMBL-EBI). Interview by Wendy A. Warr. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 195−198.

(265) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Bora, A.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of molecular bioactivity. *Methods Princ. Med. Chem.* **2005**, *23*, 223−239.

(266) *SYBYL-X Suite*, version 1.3.2; Tripos: St. Louis, MO, 2011.

(267) Lindorff-Larsen, K.; Piana, S.; Palmo, K.; Maragakis, P.; Klepeis, J. L.; Dror, R. O.; Shaw, D. E. Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins: Struct., Funct., Bioinf.* **2010**, *78*, 1950−1958.

(268) Gogonea, V.; Shy, J. M.; Biswas, P. K. Electronic structure, ionization potential, and electron affinity of the enzyme cofactor (6R)-5,6,7,8-tetrahydrobiopterin in the gas phase, solution, and protein environments. *J. Phys. Chem. B* **2006**, *110*, 22861−22871.

(269) MacKerell, A. D.; Bashford, D.; Bellott; Dunbrack, R. L.; Evanseck, J. D.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiórkiewicz-Kuczera, J.; Yin, D.; Karplus, M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **1998**, *102*, 3586−3616.