OPEN ACCESS
IOP Publishing

Journal of Breath Research

J. Breath Res. **8** (2014) 046001 (14pp)

doi:10.1088/1752-7155/8/4/046001

Precursors for cytochrome P450 profiling breath tests from an *in silico* screening approach

Susanne von Grafenstein[1,5], Julian E Fuchs[1,5,6], Markus M Huber[1], Andrea Bassi[1,7], Alessandra Lacetera[1,8], Veronika Ruzsanyi[2,3], Jakob Troppmair[4], Anton Amann[2,3] and Klaus R Liedl[1]

[1] Department of Theoretical Chemistry and Center for Molecular Biosciences Innsbruck, University of Innsbruck, Innrain 80/82, A-6020 Innsbruck, Austria
[2] University Clinics for Anesthesia and General Intensive Care, Innsbruck Medical University Anichstraße 35, A-6020 Innsbruck, Austria
[3] Breath Research Institute, University of Innsbruck Rathausplatz 4, A-6850 Dornbirn, Austria
[4] Daniel Swarovski Research Laboratory, Department of Visceral- Transplant- and Thoracic Surgery, Innsbruck Medical University, Innsbruck, Innrain 66, A-6020 Innsbruck, Austria

E-mail: klaus.liedl@uibk.ac.at

Received 9 April 2014, revised 24 June 2014
Accepted for publication 1 July 2014
Published 18 September 2014

Abstract

The family of cytochrome P450 enzymes (CYPs) is a major player in the metabolism of drugs and xenobiotics. Genetic polymorphisms and transcriptional regulation give a complex patient-individual CYP activity profile for each human being. Therefore, personalized medicine demands easy and non-invasive measurement of the CYP phenotype. Breath tests detect volatile organic compounds (VOCs) in the patients' exhaled air after administration of a precursor molecule. CYP breath tests established for individual CYP isoforms are based on the detection of $^{13}CO_2$ or $^{14}CO_2$ originating from CYP-catalyzed oxidative degradation reactions of isotopically labeled precursors.

We present an *in silico* work-flow aiming at the identification of novel precursor molecules, likely to result in VOCs other than CO_2 upon oxidative degradation as we aim at label-free precursor molecules. The ligand-based work-flow comprises five parts: (1) CYP profiling was encoded as a decision tree based on 2D molecular descriptors derived from established models in the literature and validated against publicly available data extracted from the DrugBank. (2) Likely sites of metabolism were identified by reactivity and accessibility estimation for abstractable hydrogen radical. (3) Oxidative degradation reactions (O- and N-dealkylations) were found to be most promising in the release of VOCs. Thus, the CYP-catalyzed oxidative degradation reaction was encoded as SMIRKS (a programming language style to implement reactions based on the SMARTS description) to enumerate possible reaction products. (4) A quantitative structure property relation (QSPR) model aiming to predict the Henry constant *H* was derived from data for 488 organic compounds and identifies potentially VOCs amongst CYP reaction products. (5) A blacklist of naturally occurring breath components was implemented to identify marker molecules allowing straightforward detection within the exhaled air.

[5] These authors contributed equally to the study.
[6] Current address: Department of Chemistry, Unilever Centre for Molecular Science Informatics, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK
[7] Current address: Quality Assurance, Aminochemicals Ltd, A61 industrial estate, MRS3000 Marsa, Malta
[8] Current address: Faculty of Pharmacy, University San Pablo CEU, Campus de Montepríncipe, 28668 Madrid, Spain

Evident oxidative degradation reactions served as test case for the screening approach. Comparisons to metabolism data from literature support the results' plausibility. Thus, a large scale screening for potential novel breath test precursor using the presented five stage work-flow is promising.

S Online supplementary data available from stacks.iop.org/JBR/8/046001/mmedia

Keywords: personalized medicine, vocs, virtual screening, cytochrome p450, QSPR, oxidative degradation, breath test, CYP profiling, precursor identification

(Some figures may appear in colour only in the online journal)

1. Introduction

Exhaled breath analysis is based on volatile organic compounds (VOCs) which are produced through various biochemical pathways [1–3]. The family of cytochrome P450 enzymes (CYPs) metabolizes many endogenously produced biochemical substances as well as drugs and xenobiotics. This is of great importance in the activation and elimination of (pro-) drugs. Moreover, CYP activity is reflected in specific individual chemical profiles of exhaled breath [4, 5].

The vital role of cytochrome P450 enzymes (CYPs) in metabolism of xenobiotics makes the different isoforms to prominent players in pharmacokinetics and therefore also in the lead optimization process during drug development. Around three quarters of the clinical drug transformations are covered by oxidations catalyzed by CYP isoforms. A half dozen CYP isoforms (of the 57 human isoforms) contribute to 90% of these reactions (mainly CYP3A4, CYP2D6, CYP2C9, CYP2C19, CYP1A2, CYP2E1 [6, 7]).

High inter-individual variations in the activity profiles of these isoforms have clinical implications such as altered pharmacokinetics leading to increased adverse drug reactions or drug-drug interactions [8]. Failure of therapy occurs for drugs requiring activation by CYP-catalyzed reactions. A prominent example is the CYP2D6-catalyzed oxidation of the prodrug tamoxifen to the active metabolite endoxifen. CYP2D6 deficiency is related to ineffective therapy of breast cancer with tamoxifen [9–12]. To avoid such failures in medication and allow for an adaption of dosage, the consideration of the patient-individual CYP profile is an obvious and necessary step to improve therapies. Pharmacogenetic biomarkers can be used to assess the individual CYP profile originating from genome level variations. Several genomic mechanisms causing inter-individual differences in CYP profiles are known and well investigated as some members of the CYP family show high genetic variability [13]. Especially, the gene sequences of the CYP isoforms 2A6, 2D6, 2C9 and 2C19 are prone to gene polymorphism. Single nucleotide polymorphisms may result in single mutations at amino acid level and altered enzymatic properties of the corresponding enzymes [14, 15]. Additionally, the number of gene copies influences the CYP activity via so-called gene dose and gene concentration effects [13]. Gene deletion of CYP2D6 leads to a phenotype showing reduced enzyme activity (e.g. alleles CYP2D6*4, or CYP2D6*6) even in heterozygous combination with functional alleles [13, 15]. On the other hand, copy number variations showing gene multiplications lead to phenotypes with extensive enzyme turnover.

Pharmacogenetic biomarkers are suggested to be used during drug development and for patient stratification before selection of medication in clinical use. Prior to specific medication, for example application of tricyclic antidepressants in psychiatric indications, genotyping is considered for the identification of patients showing poor CYP2D6 or CYP2C19 activities [16, 17]. In addition to the high costs of genotyping, non-genetic molecular mechanisms causing patient-individual CYP profiles even for matching CYP genotypes limit benefit from pharmacogenetic CYP profiling [18].

CYP3A4 and the related CYP3A5 are known to be subject of transcriptional regulations. The 4- to 6-fold patient-individual differences in enzyme activities cannot be explained by genetic variations within the CYP3A4 gene but rather by altered expression levels [13]. These alterations in enzyme activity exceed variations caused by gene polymorphism as the frequency of CYP3A4 polymorphism is low [13]. However, genomic variations within the promoter region of CYP3A4 seem to explain these differences partially [13]. Additional factors modulating CYP expression can be environmental such as dietary or intake of drugs interacting with the transcriptional factors PXR, CAR, PPAR [19]. Besides such environmental factors diseases, sex, age as well as hormone status impact the individual CYP profile and modulate it over time [18, 20].

Phenotyping biomarkers reflect consequences of genetic variability and additionally allow to cover non-genetic variations. In consequence, the phenotyping complements genotyping or is even capable to overcome inherent limitations for highly regulated isoforms such as CYP3A4 [17, 21].

Established phenotypic stratification methods measure the metabolism of a given substrate by quantification of decreased substrate concentration or by increase of marked metabolites [21]. Detection via plasma serum or urinary samples suffers from low patient compliance. In contrast, breath analysis is a non-invasive approach, in which almost unlimited sample amount is available continuously. Thus, the detection of volatile biomarkers within the exhaled breath is an attractive alternative [22, 23] as we have already shown for other medical applications. e.g., cancer diagnosis [24–27].

Breath tests are established for the functional prediction of CYP3A4- or CYP2D6-mediated biotransformations. For example, in the erythromycin breath test specific for CYP3A4, the labeled precursor compound ^{14}C-erythromycin is metabolized by CYP3A4 resulting in $^{14}CO_2$ [28, 29]. Suitable precursor molecules for breath tests must release volatile compounds by the reaction of interest. Established

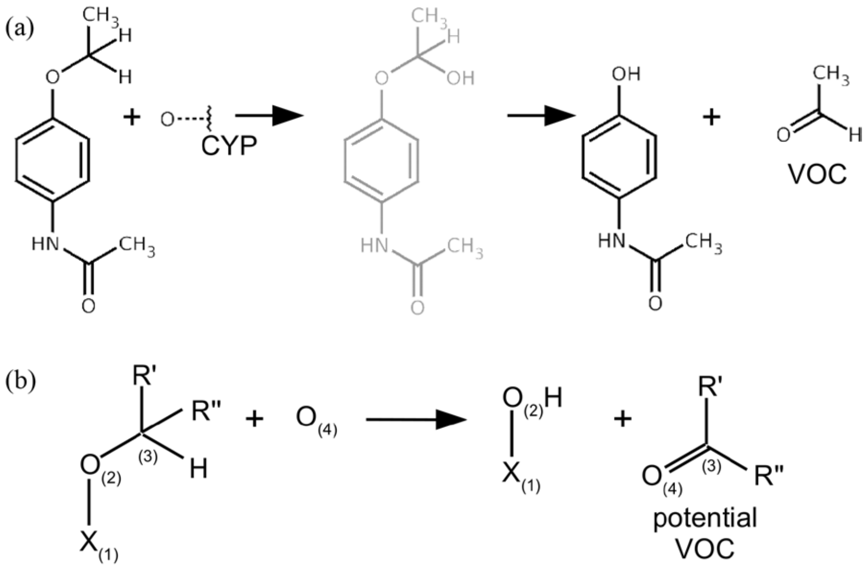

Figure 1. Oxidative degradations. (*a*) Exemplary O-dealkylation shown for the metabolic degradation of phenacetin by CYP1A2. The oxygen activated by the enzyme attacks next to the ether and forms an intermediate hemiacetal (grey) decomposing into the alcohol metabolite and acetaldehyde which is a volatile degradation product. (*b*) Implementation as SMIRKS reaction. The SMIRKS pattern is not a chemically complete reaction equation but visualizes the chemical requirements and restrictions for the identification of suitable substructures and enumeration of the products. In the processed molecules the atom with the index$_{(2)}$ must be an oxygen $O_{(2)}$ flanked by any atom $X_{(1)}$ except hydrogen or carbonyls and an alkyl rest$_{(3)}$ which has an abstractable hydrogen atom H. After addition of an oxygen atom $O_{(4)}$ the molecules is transformed into an alcohol and a ketone or aldehyde. $X_{(1)}$ as well as the further rests R′ and R″ are not affected by the encoded transformation. The corresponding SMIRKS formulation also considers amines and aromatic nitrogen atoms at index (2): [#1&!$(C=O):1][N,O,n:2][C&!$(C=O):3][H].[O:4] > > [*:1][N,O,n:2].[O:4]=[C:3]

breath tests rely on the release of CO_2 by CYP-catalyzed N- or O-demethylation and subsequent oxidation of the intermediate formaldehyde. Breath test read out quantifies the formation of CO_2 released by the specific enzymatic reaction in the patient's exhaled air. Due to the abundance of CO_2, the detection is only possible if the respective site of metabolism (SOM) in the precursor molecule is isotope-labeled and releases $^{13}CO_2$ thereby allowing quantification via infrared spectroscopy.

Reactions leading to a molecular breakdown of the parent molecule are especially promising to yield a volatile biomarker detectable in exhaled air. CYP-catalyzed reactions oxidize the parent compounds by the introduction of an oxygen atom from molecular oxygen. Hydroxylation of carbon atoms geminal to heteroatoms leads to labile intermediate hemiacetals or hemiaminals (figure 1(*a*)). In these cases the reaction results in an oxidative breakdown into a carbonyl compound and the free amine or alcohol metabolite. Such N- or O-dealkylations are also covered by the terms of heteroatom release or oxidative degradation [6].

Since by-products of oxidative degradations are apparently not limited to CO_2, development of breath tests with other volatile marker molecules should be feasible. Especially volatile ketones or aldehydes are suitable as marker molecules which do not occur as natural composites of human exhaled breath [1]. Those VOCs can be detected without being isotope-labeled using different mass spectrometric techniques for detection. Besides the advantage of label-free precursor molecules, alternative markers allow an extension of the approach towards a cocktail application for the parallel investigation of several CYP isoforms.

Here we present a virtual screening work-flow for the identification of suitable precursor molecules for a breath test based on the release of VOCs by CYP-catalyzed reactions.

2. Methods

2.1. Generation and preparation of data sets

The DrugBank 3.0 online database [30] was used as validation set for CYP metabolism. The DrugBank listed 6708 molecules (www.drugbank.ca database accession 24.10.2011). For some molecules metabolism data were annotated covering 878 reactions. Each annotation is composed of substrate, enzyme name, metabolite name, chemical reaction type and partially enzymatic parameters. Thereof, 739 listed metabolic reactions concerned CYP-catalyzed reactions as identified by enzyme name and were used as basis for further validation. These 739 metabolism annotations contain to some part multiple products of CYP metabolism as well as multiple CYP isoforms catalyzing the same reaction for the same compounds.

Compounds were extracted as SMILES (simplified molecular-input line entry system) and prepared with the MOE wash function to ensure relevant protonation states for strong acids and bases [31]. All 2D descriptors available in MOE were calculated for subsequent model generation (see [32] supporting information for a list of descriptors).

Besides the metabolism data set, two additional compound sets were generated based on literature: i) a quantitative training set for volatility and ii) a compilation of compounds naturally occurring in human breath. Both sets were also protonated in MOE. The training set of 488 organic chemicals for the volatility prediction was retrieved from the work of Raventos-Duran *et al* in SMILES notation [33]. This volatility set included experimental log$H$ values.

The other data set containing VOCs naturally occurring in human breath was based on a recent review [1]. This compilation provided CAS numbers (Chemical Abstract Services index) and names for 874 chemicals detected in breath. SMILES were automatically extracted via the Chemical Identifier Resolver hosted by the NCI cactus server[1] for the provided CAS numbers where possible. Two wrong assignments were identified and corrected manually (CAS 3855-78-5 and 7239-23-8). Likewise, SMILES for remaining 97 compounds for which automatic identification failed were assigned manually.

2.2. Decision tree for CYP isoform prediction

Ligand-based decision trees available from the literature were gathered and compared. Based on trees from Lewis *et al* [34]., Terfloth *et al* [35]., Yamashita *et al* [36]. and Zhang *et al* [37]. as well as substrate classifications of Ekins *et al* [38], we identified the following suitable molecular descriptors: molecular weight (MW), acid/base character (a_acid/a_base) and the logarithmic octanol-water partition coefficient calculated with an atomistic model (SlogP [39]). The overall structure of the tree was inspired by the literature and optimized to predict the data from the DrugBank metabolism set. Cutoff values at the branching points were determined by recursive partitioning.

Performance was evaluated for each isoform and the total set of molecules following equations (1)–(4) for sensitivity (*SN*), specificity (*SP*), accuracy (*TA*) and the Matthews correlation coefficient (*MCC*) ranging from −1 for an inverse correlation of prediction and experimental results over 0 for a random result to +1 for perfect agreement of the predictions. *TN* is the number of true negatives, *TP* the number of true positives, *FN* the number of false negatives, and *FP* the number of false positives. In case of multiple assignments in the DrugBank, each true CYP isoform assignment was counted as *TP*.

$$SN = \frac{TP}{TP + FN} * 100\% \tag{1}$$

$$SP = \frac{TN}{TN + FP} * 100\% \tag{2}$$

$$TA = \frac{TP + TN}{TP + TN + FP + FN} * 100\% \tag{3}$$

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP + FN)(TP + FP)(TN + FN)(TN + FP)}} \tag{4}$$

2.3. Prediction of site of metabolism

As the abstraction of a radical hydrogen atom (H·) is an essential step in the catalytic cycle of a CYP reaction, this energy of activation can be used to estimate the site of metabolism (SOM) within a CYP substrate [40]. Quantum mechanics allows to calculate the difference in the energetic state of the educt as well as the radical product of this step in the enzymatic reaction. A site in the molecule showing a relatively low energetic cost of hydrogen abstraction is more likely to be oxidized by a CYP [41, 42].

Molecules were prepared using MOE's [31] wash function and following energy minimization in MMFF94X [43] applying default settings. Positions of lowest energy difference between ground state and H·-deficient state were used to predict possible SOMs. We calculated energies for both states performing a geometry optimization using the semi-empirical method AM1 [44] with a STO-3G basis set [45] in Gaussian [46]. Positions where H·-abstraction resulted in an energy difference d$E < 30\,\mathrm{kcal\,mol^{-1}}$ were considered as potential SOMs.

As CYP-catalyzed oxidation reactions require accessibility of the SOM [47–49], we implemented a second selection criterion in our SOM prediction. Based on the energy-minimized structure from H·-abstraction calculations, we calculated solvent accessible surface areas (SASAs) for all hydrogens using MOE [31]. Accessible positions bearing at least one hydrogen with SASA exceeding $15\,\text{Å}^2$ were preserved, others were discarded.

2.4. Enumeration of CYP reaction products

Following the assumption that oxidative degradation is most likely to result in VOCs, we implemented a computational model for CYP-catalyzed degradation of xenobiotics (figure 1(*b*)). We generated a general SMIRKS pattern for O- and N-dealkylations to identify products of oxidative degradation resulting from oxidation at positions favored by a comparable low hydrogen abstraction energy. The SMIRKS format is a reaction encoding method developed by the chemoinformatics company Daylight [50] and was used here in the implementation in OEChem toolkit [51].

Using this enumeration step we are able to predict small molecules and hence potential VOCs from arbitrary compounds entered in SMILES format. O- or N-substituted carbons bearing an abstractable hydrogen yield a carbonyl function at the respective position and the O-, N-containing leaving group which is an alcohol or amine. Moreover, the pattern includes the further restriction that N/O neighboring atoms are not allowed to be carbonyl carbons. This was introduced to exclude respective esters or amides which are degraded by hydrolysis rather than CYP-catalyzed oxidation.

Chemically equivalent reactions leading to the same products were unified to a single result as they do not contain any additional structural information. If the degradation site is within a ring structure, only one product molecule is generated due to intramolecular cleavage; those reactions were neglected for further processing as the product is in general too large for VOC generation.

2.5. Estimation of volatility

Modeling of quantitative structure property relations (QSPR) was used to identify volatile compounds from CYP products [52]. Henry's law constant H was used to describe volatility in a physical metric following equation (5), where C represents

[1] http://cactus.nci.nih.gov/chemical/structure.

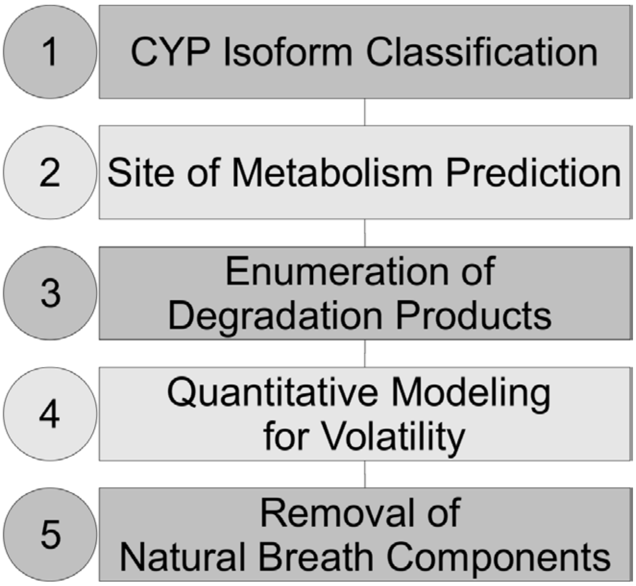
1
CYP Isoform Classification
2
Site of Metabolism Prediction
3
Enumeration of Degradation Products
4
Quantitative Modeling for Volatility
5
Removal of Natural Breath Components

Figure 2. The *in silico* screening work-flow for the identification of novel precursor molecules comprises five modules.

the solubility of a compound in water (mol/l) and P the partial pressure (atm).

$$H = \frac{C}{P} \tag{5}$$

To parameterize our QSPR model for log$H$ we extracted 488 experimental H-values for small organic compounds from a set of Raventos-Duran and co-workers [33]. To derive a linear QSPR equation by multi-linear regression, we selected molecular descriptors using forward selection until a threshold in the p-value of 0.05 for the next descriptor was exceeded [53]. At this stage we obtained a QSPR equation containing 20 descriptors. As such a model is expected to be over-fitted by the training set, we performed manual pruning steps. After several cycles of manual descriptor elimination we obtained a final model containing six descriptors resulting in a comparable fit for the test set, and nevertheless ensuring the model can be readily interpreted on a physical level.

2.6. Removal of natural breath components

Searching for breath test precursor molecules, we were especially interested in molecules releasing VOCs not present in the exhaled air. A blacklist of 874 molecules occurring in breath of healthy humans was generated based on the data in de Lacy-Costello *et al* [1]. The filtering script was implemented using the OEChem toolkit to canonicalize SMILES from the products of oxidative generation and SMILES from the blacklist [51]. Identical entries on the blacklist as well as predicted products from CYP degradation were discarded.

3. Results

The virtual screening work-flow we established comprises five modules (figure 2). Applying this screening strategy on larger databases, we yield hit lists of candidates for precursor molecules which subsequently need to be experimentally verified. Here, we present the validation of the individual modules as well as the application on the DrugBank metabolism set.

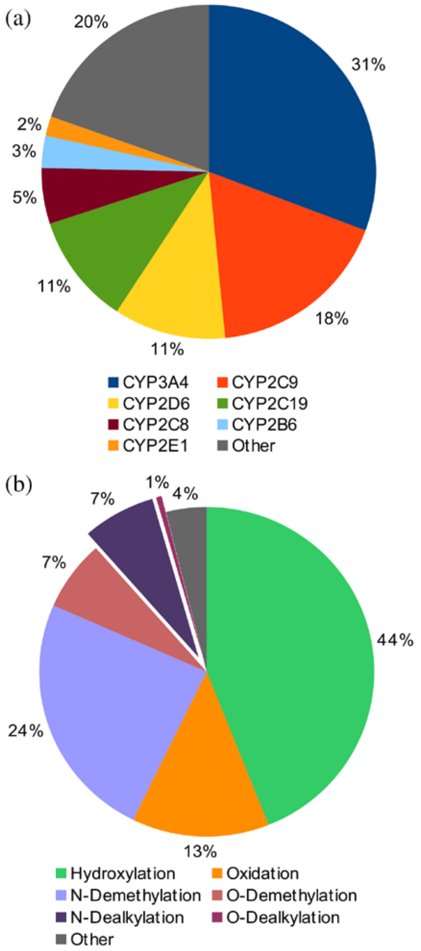

Figure 3. (*a*) CYP isoform distributions and (*b*) reaction distributions for the experimental metabolism data from DrugBank. 100% corresponds to 739 metabolic reactions in both diagrams.

Our analysis of the DrugBank metabolism set revealed that among extracted metabolism information for xenobiotics, 739 of 878 annotated reactions concerned CYP-related reactions. This percentage of 84% for CYP reactions nicely reproduces the distribution of phase I metabolic contributions for marketed drugs [6, 18]. Annotated reactions discarded for this study include metabolic contributions of e.g. aldehyde oxidase, cholinesterase or UDP-glucuronosyltransferase. In total, 739 CYP-catalyzed reactions were extracted for 374 individual substrates, being connected to 495 assignments to CYP isoforms after pooling of different formed metabolites catalyzed by the same isoforms.

The trends observed for the distribution of the CYP isoforms and the reactions confirm the validity of the DrugBank annotations as test set for the here presented approach. We focus on seven of the 16 isoforms according to metabolic contribution and availability of substrate characteristics in the literature (figure 3(*a*)). Most annotated reactions involve CYP3A4 followed by CYP2C9, CYP2D6, CYP2C19 reproducing widely accepted trends as reviewed by Guengerich [6]. CYP2C8 and

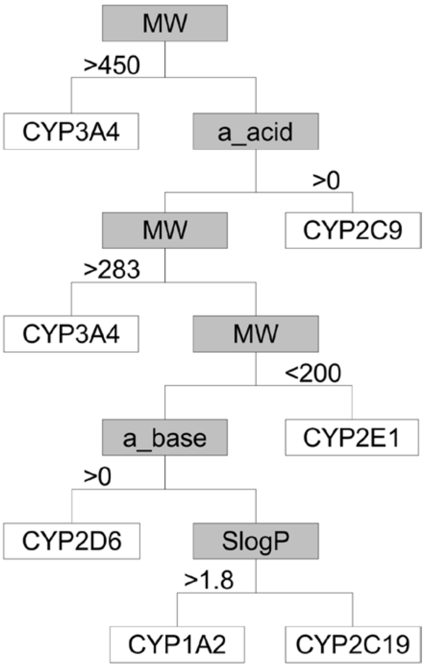

Figure 4. Decision tree for CYP isoform classification based on the molecular descriptors MW: molecular weight, a_acid: number of acidic atoms, a_base: number of basic atoms, and SlogP: calculated partition coefficient n-octanol/water.

CYP2B6 show a higher importance in the set of metabolism reaction extracted from the DrugBank. In contrast, the importance of CYP2E1 is underestimated in the DrugBank set [6].

Besides annotation of the involved metabolic enzymes, the DrugBank data set provides information on the formed metabolites (figure 3(*b*)). After manual curation and pooling of reaction types we observed a complex distribution of catalyzed reactions: 44% leading to a hydroxylation of the substrate, whereas 13% lead to a general oxidation of the xenobiotic. Less pronounced were reactions potentially leading to small volatile compounds: 24% N-demethylations and 7% O-demethylations were observed. These reactions would lead to CO_2 and are hence of less interest in the search for innovative breath test. Still, DrugBank also contains information on CYP-catalyzed dealkylation reactions producing fragments other than CO_2 after oxidative degradation: The DrugBank set includes 7% N-dealkylations as well as 1% O-dealkylations, providing in total 55 entries with dealkylations. These annotations cover 33 reactions and 32 molecules metabolized by the isoforms CYP3A4, CYP1A2, CYP2C8, CYP2C9, CYP2D6, CYP2C19, CYP2A6, CYP2B6, and CYP3A5 indicating that oxidative degradation is not specifically related to a particular CYP isoform (table S1). Those compounds could *per se* be suitable precursor molecules for a CYP profiling breath test. We will evaluate and discuss their potential as such within this study. However, the small number of only 32 unique molecules from the test set covering 374 molecules supports the need for a systematic screening strategy for further potential precursors.

We present an exemplary application of our virtual screening modules on the set of these 374 compounds (metabolism test set) and the subset of 32 compounds known to undergo oxidative degradation.

Table 1. Performance of the CYP isoform classification calculated on the training set of 216 molecules partially assigned to multiple isoforms. Statistical measures are sensitivity, specificity, total accuracy given as percent as well as Matthews correlation coefficient (MCC).

	P1A2	P2C9	P2C19	P2D6	P2E1	P3A4	Overall
Sensitivity (%)	11.32	34.12	14.55	33.33	70.00	72.30	43.14
Specificity (%)	98.16	98.47	97.52	96.56	97.09	79.41	96.17
Total accuracy (%)	76.28	73.15	76.39	80.09	95.83	74.54	79.48
MCC	0.20	0.45	0.23	0.42	0.59	0.48	0.50

3.1. Decision tree for CYP isoform classification

As first module of the virtual screening work-flow, we developed a decision tree which assigns a CYP isoform to each molecule of a desired data set based on 2D molecular descriptors. The final tree covers the six most abundant isoforms CYP3A4, CYP2C9, CYP2E1, CYP2D6, CYP1A2, and CYP2C19 according to Guengerich [6] (figure 4). Molecular weight (MW) was identified as suitable descriptor to form the branching point at several levels. Large compounds are likely to be metabolized by the very promiscuous CYP3A4 [34]. In the decision tree CYP3A4 is the endpoint for two different branches either for molecules heavier than 450 g mol^{-1} or for molecules not containing acidic atoms and heavier than 283 g mol^{-1}. The smaller binding pocket and the binding preferences for the isoform CYP2E1 are reflected in its branch collecting molecules with a molecular weight below 200 g mol^{-1} which were not classified as acidic before. The presence of an acid (a_acid) or basic (a_base) functionality directs the isoform prediction towards CYP2C9 or CYP2D6 respectively [34]. Logarithmic n-octanol/water partition coefficient logP(o/w) calculated by an atom-wise approach (SlogP [39]) was identified as a suitable descriptor to distinguish compounds metabolized by CYP1A2 or CYP2C19 as CYP1A2 shows a preference for aromatic compounds [34] which is reflected in a higher lipophilicity.

Performance of the decision tree was calculated individually for each isoform and as overall performance (table 1). For the evaluation we included only the 216 compounds of the DrugBank metabolism set metabolized by the seven CYP isoforms of the classification model. Multiple CYP isoforms are often responsible for the complex metabolism of one compound yielding in multiple assignments in the training set. The concept of our decision tree only allows to predict potential metabolism by one CYP isoform. In consequence, the number of false negative predictions is systematically high and therefore sensitivity is low for most isoforms. Evaluating only 106 compounds having single CYP assignments, the overall sensitivity rises from 43.14% to 78.3% (see table S2) without affecting specificity to a great extend (96.17 versus 95.05%). Thus, our decision tree allows to identify the primary CYP isoforms majorly contributing to drug metabolism. CYP isoform(s) and the prediction can be seen for exemplary compounds in table S2.

For the 32 molecules in the training set for oxidative degradation 84% are correctly predicted (table S1). The CYP

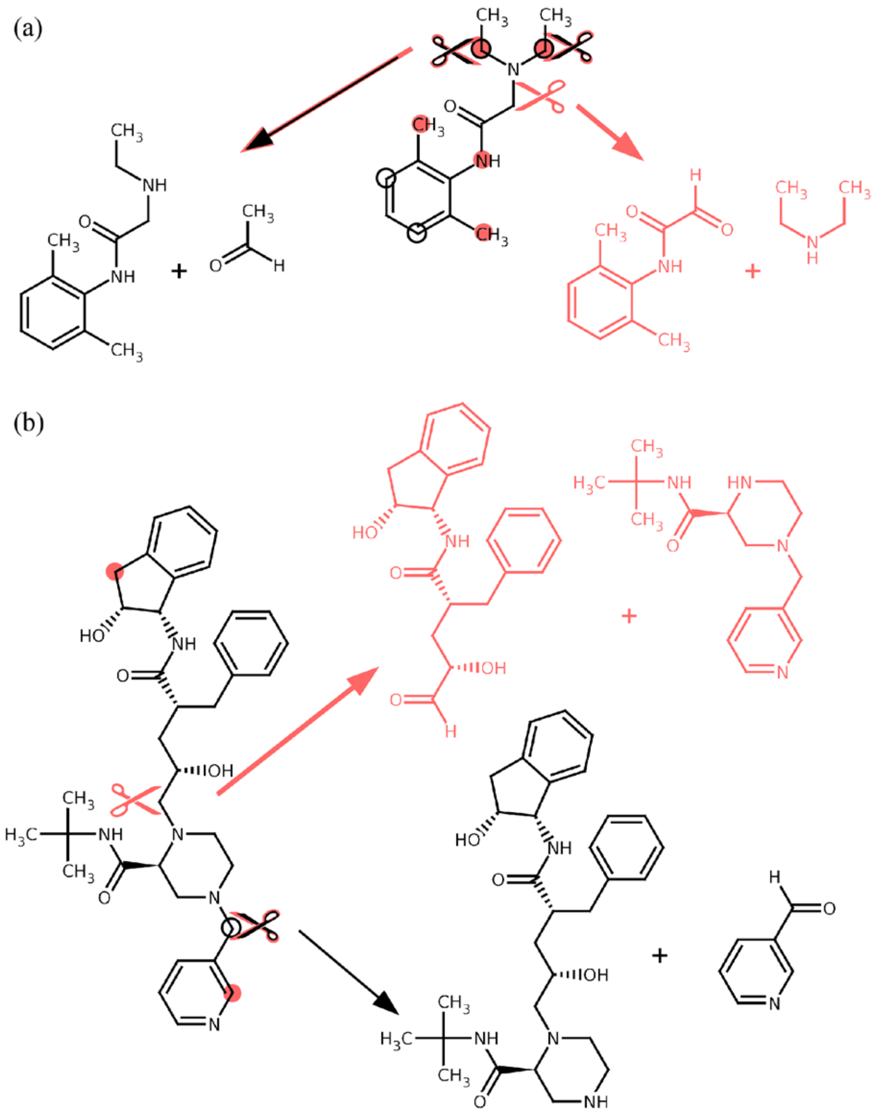

Figure 5. For lidocaine (*a*) and indinavir (*b*) we show exemplary the predicted sites of metabolism (red spheres) compared to experimental sites of metabolism (black open rings) as well as the proposed cleavage sites by the implemented reaction enumeration (red scissors). Experimentally verified reactions and products (black) were predicted by the product enumeration as well as additional degradation reactions (red).

isoforms responsible for the metabolism of amodiaquine and floxuridine are not covered in the decision tree. The wrong assignment for these molecules is an intrinsic limitation, as all molecules are classified to a covered CYP isoform without allowing a bypassing of the endpoints. Together with three wrong assignments for propranolol, ifosfamide and floxuridine (although their metabolizing isoforms would have been covered) we observe five errors in CYP isoform prediction.

3.2. SOM prediction

We estimated the reactivity of each potential site of radical hydrogen abstraction (H·) as this represents the major energetic barrier in the oxidation of aliphatic carbons catalyzed by CYPs. The difference in energy before and after H·-abstraction (d*E*) was calculated for each hydrogen atom. Positions yielding a d*E* smaller than 30 kcal mol^{-1} are considered as likely SOMs. This empirical cutoff value is of similar magnitude as values discussed in literature for aliphatic carbon centers [40]. Additionally, we ensured accessibility of predicted sites by discarding potential SOMs buried within the energy optimized conformer.

The method to predict the potential SOM was applied on the set of 32 molecules undergoing oxidative degradation (table S1 and figure 5). The presented approach for SOM prediction is not specific for sites of dealkylations but considers all reactions catalyzed by CYPs. Therefore, we additionally included all annotated CYP SOMs from the DrugBank for the evaluation of the SOM prediction.

Though inclusion of the accessibility criterion, our method is not very restrictive suggesting on average 4.13 SOMs per molecule whereas only 2.24 sites are assigned in the experimental set. Thereby, our approach predicts almost one sixth of all heavy atoms as potential SOMs (15.8%). Judging these numbers one should consider that the DrugBank metabolism set is not necessarily comprehensive. The more recent version DrugBank 4.0 [54] includes for example another two SOMs for the antiemetic drug aprepitant which were correctly assigned as SOMs by our algorithm. Furthermore, a more restrictive energy cutoff could be introduced to increase accuracy of prediction. Still, our *in silico* screening approach aims at identification of potential precursors. Thus, missing interesting candidate molecules by introduction of more stringent criteria would be counterproductive. Although being tolerant, the hydroxylation sites at aromatic substructures are frequently missed in the predictions (e.g. buprenorphine, buspirone, fluvastatin, lidocaine, propafenone, propanolol, and trazodone). In context of this study the latter is of minor importance as the focus lies on metabolism at sites next to amine or ether moieties as those sites result in degradations. Out of the 33 sites for oxidative degradation only two sites of N-dealkylations are discarded by the applied energy cutoff for SOMs (delaviridine and indinavir). Buried SOMs in letrozole, fluoxetine and verapamil are discarded via the accessibility criterion but are correctly identified with a low hydrogen abstraction energy. After removal letrozol is not assigned with a single predicted SOM.

3.3. Reaction enumeration for oxidative degradation

The implemented function for CYP product enumeration is limited to only oxidative degradations because we identified those reactions promising for the release of degradation fragments which might then be excreted via the lung. CYP-catalyzed oxidative degradation was encoded as SMIRKS reaction pattern. Our method identifies suitable substructures in molecules and generates the structures of the formed products. Resulting fragments are amines or alcohols on one side and ketones or aldehydes on the other side (figure 1).

All 33 reactions for the 32 drugs with experimental evidence for oxidative degradation are recovered with the product enumeration. In total 102 reactions were enumerated for 32 molecules in a first step. Out of these we discarded 29 intramolecular reactions and 15 reactions resulting in CO_2 and methanol or methylamine. The frequently occurring N- or O-demethylations resulting in degradation products of only one carbon are not of interest for the search of innovative VOCs and those reactions where not considered in the compilation of the test set (see above) and therefore neglected for enumeration. Thus, besides the correctly predicted reactions 25 additional reactions are proposed.

A restriction of false positive degradations can be achieved when integrating the results from the SOM prediction. For example for the tertiary amine in lidocaine all alkyl side chains are available for the product enumeration resulting in two different reactions where the smaller product is either acetaldehyde or diethylamine. Combining the information from the SOM prediction only the correct cleavage site is retained (figure 5(*a*)). Overall this restriction to predicted SOMs (including both reactivity and accessibility criteria) eliminates 22 reactions which cover also five of the correctly predicted cleavage sites. For example the correct cleavage of indinavir would be missed when including only degradation reactions at sites predicted with the proposed SOM filter. (figure 5(*b*))

3.4. Volatility prediction

A multi-linear regression model was developed to predict the training set log*H* values using descriptors based on the molecular composition and connectivity only.

$$\log H_{\text{pred}} = -0.777 + 1.59(\text{a\_don}) - 1.47(\log\text{P(o/w)}) + 1.47(\text{a\_acid}) + 0.30(\text{apol}) - 0.215(\text{bpol}) + 0.00764(\text{MW}) \quad (6)$$

The resulting equation (6) includes these six descriptors:

- the number of hydrogen bond donor atoms (a_don),
- logP(o/w) calculated according to a model from MOE [31],
- the number of acidic atoms (a_acid),
- the sum of atomic polarizabilities (apol) as well as
- the sum of the absolute value of differences between atomic polarizabilities (bpol) with polarizabilities calculated including implicit hydrogens [55],
- and the molecular weight (MW)

The contributions and directionalities of the individual coefficients follow expected trends. Higher mass and polar functions as hydrogen bond donors or acids increase logH_{pred}, thus reduce volatility. Also, the negative coefficient of logP(o/w) is expected as lipophilic compounds are expected to be more volatile. The terms for polarizability (apol and bpol) are highly correlated with each other (R^2 = 0.758). Here, the effect of apol—which is numerically higher and contributes with a greater coefficient to the equation—is reduced by the component including the descriptor bpol. However, in total, greater polarizability is a factor which reduces the volatility of a compound.

The use of different descriptors was kept minimal to avoid over-fitting. The success of this attempt is reflected in the small difference between R^2 = 0.898 for the complete set and Q^2 = 0.875 calculated in a k-fold cross-validation with a block size *k* of 10 (figure 6).

Acetaldehyde occurring seven times as product in the 33 reactions of oxidative degradation is listed in the training set with logH_{exp} = 1.1 mol/(l*atm) being higher than logH_{pred} = 0.279 mol/(l*atm). Also the second most frequent cleavage product acetone (five times in 33 reactions) is overestimated in volatility with logH_{pred} = 0.962 mol/(l*atm) compared to logH_{exp} = 1.5 mol/(l*atm). The prediction for the positively charged amino compounds are probably too low, such as N-methyl-4-phenylpropan-4-one amine (logH_{pred} = 3.76 mol/(l*atm)) resulting from the degradation of fluoxetine. Although a similar effect of basic and acidic functions in reducing volatility could be expected, this cannot be covered in the volatility predictions because no basic compounds were represented in the training set. In contrast, expectations are fulfilled for an oxidation product of 2-desoxyribose released upon the degradation of floxuridine with a logH_{pred} = 8.17 mol/(l*atm) suggesting no volatility.

3.5. Removal of natural breath components

The recent compilation of 874 VOCs which are present in the breath of healthy probands [1] allowed us to filter out degradation products which normally occur in the exhaled air. 16 of

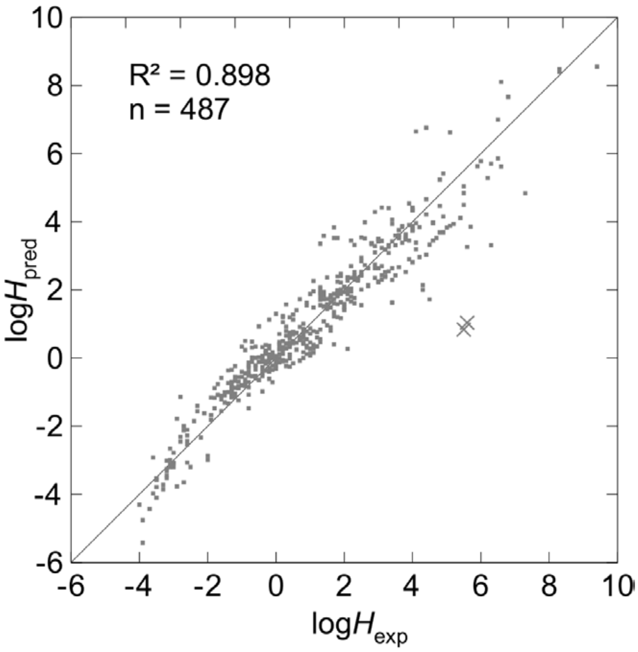

Figure 6. Experimentally determined and predicted logarithmic Henry constants (logH) for the training set. The two extreme values marked with X represent the two aldehydes ethandial and trichloroethanal mentioned in the discussion.

the 32 compounds in the test set are predicted to release small molecule reaction products which are not observed in the exhaled air. Moreover, some reactions result in compounds unlikely to be volatile.

A cutoff to exclude nonvolatile compounds based on the logH_{pred} could be defined using the breath components. The predictions for these compounds range from logH_{pred} = −7.34 mol/(l*atm) for octadecane to logH_{pred} = 9.22 mol/(l*atm) for glycol. The the log$H$ values forms a bell-like distribution with a mean logH_{pred} = −0.475 mol/(l*atm) and a standard deviation of 2.52 mol/(l*atm) (figure S1). We suggest to consider molecules with logH_{pred} < 4 mol/(l*atm) as volatile based on the data as this criteria cover ~95% of the volatile compounds in the breath. Based on this cutoff the set of compounds being potential precursors for breath tests was further reduced to eight compounds.

4. Discussion

4.1. Drugs as candidates for breath test precursors

We focused the initial investigation presented here on the DrugBank [54] a database covering drug molecules. Those molecules are expected to be studied in terms of pharmacokinetics and toxicological characteristics. Inconvenient aspects in selecting the precursors for a CYP breath test from a drug database are pharmacodynamic activities of the drug. For an application as an *in vivo* diagnostic tool the pharmacological activity should be minimal. However, for drug molecules activity as well as toxicity profile are expected to be well investigated and comprehensively understood. Pharmaceutical knowledge allows for example *a priori* exclusion of the class of anti-cancer drugs from *in vivo s*tudies due to expected and severe undesired side effects. In contrast, most drugs with antibiotic effect are considered to be well tolerable. Therefore, selection of candidate molecules for clinical studies is easier with drugs than with precursor molecules with unknown pharmacodynamic or toxicological profile. Preliminary studies with drug precursors could also be performed with a population taking the potential precursor for medical reasons. Moreover, details on the metabolism of drug compounds are also of interest from the pharmacological point of view and inclusion of exhaled air as compartment of excretion will complete the pharmacokinetic profile of a drug.

4.2. Suitable precursor candidates from the test set?

A major aim of this study was to proof the hypothesis that volatile metabolites – other than the VOCs naturally occurring in the exhaled air – can result from the CYP-catalyzed reactions. Such reactions are the basis for the development of a breath test which enables the surveillance of CYP activity with a label-free marker molecule released from a precursor compound.

We could show that the fraction of oxidative degradation within the test set of 374 drug molecules is rather low with < 10% of CYP-catalyzed reactions. While in large parts of these 32 cases the smaller cleavage product is a breath component still another 16 cases would generate molecules not present in the breath. Based on the cutoff derived from molecules detectable in breath, the QSPR model predicts the products of buprenorphine, cyclophosphamide, fluoxetine, haloperidol, ifosfamide, indinavir, terbinafine, and verapamil to be volatile. As discussed before, volatility predictions for the amino compounds as released from fluoxetine and udenafil have to be taken with caution. Moreover, two anticancer drugs, cyclophosphamide and ifosfamide must be excluded as potential precursor molecules due to their critical profile of side effects. Multiple CYP isoforms are involved in the metabolism of buprenorphine, haloperidol, and verapamil according to the metabolism assignments which limits their use as isoform-specific precursors.

Indinavir, an antiretroviral protease inhibitor used in HIV treatment, was identified as remaining precursor candidate. We predicted the VOC 3-pyridinecarboxaldehyde to be the byproduct from indinavir's degradation to the metabolite M6 [56]. CYP3A4 is most probably the one responsible enzyme in the degradation of indinavir as reviewed by Lin [57]. An aspect hampering the applicability of indinavir as precursor for a CYP3A4-specific breath test is the mechanism-based inhibition of CYP3A4 by indinavir [58]. Thus, in presence of indinavir the activity profiling for CYP3A4 is self-limited as indinavir actively decreases the CYP3A4 activity. Still, we expect that a proper calibration of a putative breath test could overcome this limitation.

4.3. The virtual screening work-flow for identification of breath test precursors

The lack of suitable precursor molecules in the test set of oxidative reactions highlights the need for a suitable screening approach. Here, we proposed an *in silico* strategy which could largely reproduce the experimental data of the experimental test set as outlined in the results section.

The modular concept allows a question-specific adaptation of the screening work-flow (figure 2). The sequence we present here follows the logical progression of drug metabolism (enzyme recognition, orientation in the active site, reaction, elimination of the products) and the post-processing of the product lists according to the required volatility and suitable marker molecules. For application to large scale molecular databases, modules can be rearranged to gain in computational performance. Modules 3) and 5) process the molecules as simple SMILES representations and can be applied on large databases. Only the resulting lists of interesting molecules can then be processed further. Also, the modules 1) and 4) represent fast chemoinformatic approaches which rely only on computationally cheap 2D-based descriptors. In contrast, the second module, SOM prediction, relies on quantum-mechanical calculations and is by far the module with highest cost in terms of computing time. Therefore, we propose to apply it only on compounds assigned to the CYP of interest and releasing promising VOCs in order to verify the reaction site in a post-processing application.

Moreover, the modular conception allows to adapt each step of the work-flow individually. We will consider exchange of here presented modules by alternative approaches when we expect an improvement in accuracy or performance. Potential for optimization will be discussed shortly for each of the five module. Additionally, we compare our results to comparable approaches from literature were applicable.

4.3.1. CYP isoform classification. CYP classification and metabolite enumerations are major questions in the computational assessment of metabolism prediction as covered in a recent review by Kirchmair *et al* [59]. Characteristics of the different binding pockets drive the molecular recognition and therefore allow an assignment of potential substrates based on their molecular descriptors [34]. We implemented a decision tree assigning a CYP isoform for each entered molecule using MOE's 2D molecular descriptors. Although more than 50 isoforms exist in humans, data is only available to derive characteristics for a selection of important, in our case seven, isoforms. Moreover, we do not allow to escape the tree without isoform assignment, which might be a way to include rare isoforms. Another systematical drawback of a decision tree model is the defiance in considering multiple isoforms often involved in the metabolism of a single molecule. Mishra *et al* propose a multi-label prediction relying on support vector machine models for five implemented CYP isoforms [60]. The training data of this publication was also extracted from the DrugBank. Although the study was only focused on five CYP isoforms, one can compare both approaches in terms of performance. The model performs very similar with accuracy values between 70% and 85% using more descriptors to describe each isoforms individually and an overall accuracy of 82.81% compared to 79.48% in our study. However, a decision tree has its attractiveness in terms of intuitive comprehension. Furthermore, our *in silico* work-flow does not focus on quantitative CYP isoform prediction, but focuses mainly on direct application in breath tests.

4.3.2. SOM prediction. Prediction of the site of metabolism (SOM) is implemented in our work-flow via a quantum mechanics-based estimation of hydrogen radical abstraction energy. The employed energy cutoff for susceptible sites is in agreement with other approaches relying on AM1 functionals [40, 61]. O- and N-dealkylations suitable for VOC generation are found to yield reaction intermediates of relative high stability [41]. Additionally, we discarded predicted SOMs expected to be too buried for CYP-catalyzed oxidation [40, 47, 49]. For the 32 analyzed substrates of oxidative degradation the SOMs for dealkylation could be covered in 27 cases. Still, the high number of reactive sites allows only a minor restriction of the dealkylation sites investigated in a subsequent step. As expected the presented approach is not valid for predicting hydroxylation sites at aromatic substructures [40, 62]. Although these sights are not relevant for prediction of N/O-dealkylation sites, we identified the SOM prediction as a module where alternative methods could have a beneficial effect for the screening performance. We consider to exchange the here presented approach for SOM prediction by the established program SMARTCyp [42, 47, 62, 63]. The open source tool represents a validated alternative. Methodologically, this tool relies on reactivity assignments by substructure rules derived from quantum mechanics-based precalculated fragments [63]. Accessibility information on putative SOMs is also included to exclude reactive but buried sites [47]. Advantages of this approach are related to the 2D-based implementation. In contrast to the approach we suggest, SMARTCyp is faster and not sensitive to conformational preparation.

In addition to these solely ligand-based strategies, other computationally more demanding approaches to prediction SOMs include the protein environment via docking [64–66]. On the other hand also data-driven approaches were found to yield accurate results at a comparably low computational cost. Random forest models are employed to estimate SOMs based on similarity to molecules with known CYP metabolism [67, 68]. We refer the reader to a recent computational study for a complete overview on strategies followed in SOM prediction and the available experimental training sets [69].

4.3.3. Enumeration of degradation purpose. In contrast to SOM prediction, direct comparison with previously published attempts to predict CYP metabolites is hampered: The process of product enumeration we present here has a very specific role as a module of the screening algorithm for VOC precursors only focusing on dealkylations. Attempts to predict the metabolites from CYP-catalyzed reactions rely on expert systems and/or data mining [59]. In comparison to SOM predictions, product enumeration is often not considered with the same attention. Approaches focusing on product enumeration are mainly follow-up applications of established SOM prediction tools, e.g, Metaprint2D-React [68] and MASS-Metasite focusing on prediction of metabolites and their fragment profiles in mass spectrometric analysis [70].

Here, the product enumeration is only considering the oxidative degradation, as this is likely to result in volatile side-products. Small metabolites of CYP-catalyzed degradation of xenobiotics are generally considered unimportant for further drug metabolism and are thus mostly neglected for CYP metabolite prediction. Still, we think that these small by-products of CYP-catalyzed degradation reactions might find an interesting application in the development of innovative CYP profiling breath tests.

We propose the elimination of aldehydes from the potential marker molecules as further potential restriction. Main aim of the metabolic transformations is the generation of detoxified and soluble metabolites ready for elimination and excretion from the body [6]. Aldehydes, however, are reactive products of CYP-catalyzed reactions and therefore often subject of further metabolic modifications. For example the aldehyde resulting as intermediate from N-dealkylation of haloperidol is further oxidized to the respective fluorobenzoylpropionic acid [71]. In contrast, ketones represent more stable products of oxidative reactions and therefore, they are better suitable marker molecules for a potential breath test. Inclusion of two explicit carbon atoms linked to $C_{(3)}$ in the SMIRKS pattern (figure 1(*b*)) adapts the implementation to identify reactions resulting in ketone products only. In course of this study aldehydes were kept included for a comparison with the set of 32 oxidative reactions showing predominantly aldehyde products.

4.3.4. Quantitative modeling for volatility prediction. In order to predict the volatility of the released molecules generated by oxidative degradation, we implemented a QSPR model predicting the Henry constant *H*. Recently, a combination of log*H* value and the logP(o/w) was proposed to predict the partition coefficient between blood and air P(b/a), which is part of the Fahri equation [5]. The latter allows to calculate the alveolar concentration of VOC in relation to the blood concentration and physiological parameters as ventilation and cardiac output. For a higher physiological importance in prediction of volatility, modeling of logP(b/a) could be beneficial. However, the limited access to consistent data sets of logP(b/a) values covering diverse chemical compounds is the main reason why we selected log*H* as target property.

Our multi-linear regression model was based on the data set of experimental *H* data compiled for the development of the GROMHE model [33]. The study especially considers that the experimental data provided is the effective Henry constant *H* and includes the solution state of carbonyls for the description where the attachment of a water molecule leads to geminal diols [33]. The effect is especially relevant for carbonyls with neighboring atoms having a pronounced inductive effect. Consequently for our model, volatility might be overestimated for individual compounds of the training set such as the two aldehydes ethandial and trichloroethanal. The large error for those compounds could origin in the wrong molecular description of those compounds in solution where they occur in equilibrium with the hydrate form. (figure 6)

In contrast to the 28 structural descriptors for the group contributions used in the GROMHE model, we actively limited the choice of descriptors to five descriptors in our model. This limitations clearly reduces the performance with a squared Pearson correlation coefficient between predicted and experimental log*H* of $R^2 = 0.898$ compared to $R^2 = 0.974$ for the GROMHE model. However, in analogy of QSAR modeling for enzyme inhibition data the model's error should be linked to the experimental error [72]. The experimental uncertainty is estimated with a factor of two for the *H* values by Raventos-Duran *et al* [33]. The risk of over-training is given for especially well performing models in terms of R^2 [72]. In our model the low difference between $R^2 = 0.898$ for the complete set and $Q^2 = 0.875$ for the cross-validation ensures that performance of the model is not largely driven by over-interpretation of experimental data.

Still, a drawback in the QSPR model can be seen in the fitting data set of experimental log*H* values which do not cover the chemical variability observed in the degradation products of drug molecules. For example the data set misses amine functions completely. Therefore, $\log H_{\text{pred}} = 2.58\,\text{mol/(l*atm)}$ for the degradation fragment of udenafil is underestimating the expected log*H* value, as the positively charged amine fragment is not expected to be volatile. This limitation is probably the most important one for the use of the QSPR model to identify novel breath test precursor molecules. A special attention on selecting hit compounds as well as a collection of a training set based on the special needs should be considered.

Based on a compilation of VOCs detected in the breath of healthy humans [1] we could benchmark the volatility model for the identification of a cutoff. According to the calculated $\log H_{\text{pred}}$ values for those compounds we defined compounds with a $\log H_{\text{pred}} < 4\,\text{mol/(l*atm)}$ as volatile.

4.3.5. Removal of natural breath components. Besides the valuable information for the interpretation of $\log H_{\text{pred}}$ values, the set of naturally occurring VOCs was used as filter. We explicitly remove reactions resulting in VOCs known to be breath components for the identification of novel breath test candidate molecules. Natural breath components are not suitable as markers without isotope labeling. Quantitative decomposition of the marker signal and the physiologically occurring amounts would be an analytical challenge due to interindividual variations of signal intensities. This requirement reduced the set of reactions we investigated here to eight potential precursor candidates (see discussion above).

Besides the screening and filtering approach we already discussed further requirements for the identification of a precursor for novel CYP breath tests. Criteria for post-processing of the hit lists is driven by the analytical detection of the released VOC or the pharmacological profile of the precursor. Especially, the latter eliminated several candidates from the set of 33 reactions we investigated in detail in this contribution.

However, the screening approach is designed to be applicable to any other databases of chemical compounds which results in larger hit lists of potential breath test precursors for a specific CYP isoform. Subsequent verification by experimental *in vitro* testing is planned.

5. Conclusions

We report on the development and a first application of an *in silico* work-flow to identify potential precursors for a CYP profiling breath test. The work-flow comprises of CYP isoform assignment, SOM prediction, metabolite prediction focusing on CYP-catalyzed dealkylation and volatility prediction for predicted metabolites.

Due to the strict requirements for a suitable precursor, the experimentally verified test set of 32 compounds did not contain any promising candidates to allow straightforward *in vivo* or *in vitro* verification. However, the set could be used to validate the results of the virtual screening approach. Potential for optimization is identified in the module for SOM prediction as well as in a restriction of the metabolite prediction on ketones as stable degradation products. Furthermore, an experimental screening is needed to verify the hit lists from the outlined virtual screening campaign.

In summary, the preliminary data presented here are a proof of concept that volatile compounds other than CO_2 can serve as potential markers for CYP specific reaction in exhaled air.

Acknowledgments

Klaus R Liedl and his team thank Openeye for providing a free academic license enabling the use of the OEChem Toolkit in this study. This work was in parts supported by a project of Klaus R Liedl funded by the Austrian science fund FWF: P23051 and a DOC fellowship awarded to Julian E Fuchs from the Austrian Academy of Science. Anton Amann and Veronika Ruzsanyi gratefully appreciate funding from the Oncotyrol-project 2.1.1. The Competence Centre Oncotyrol is funded within the scope of the COMET—Competence Centers for Excellent Technologies through BMVIT, BMWFJ, through the province of Salzburg and the Tiroler Zukunftsstiftung/Standortagentur Tirol. The COMET Program is conducted by the Austrian Research Promotion Agency (FFG).

We thank Roland G Huber, Hannes G Wallnoefer, Johannes Kirchmair, Bernd Wellenzohn and Barbara Simic, for fruitful discussion on the implementation of the virtual screening work-flow as well as the experimental potential of the hit compounds.

We declare no conflict of interest.

References

[1] de Lacy Costello B, Amann A, Al-Kateb H, Flynn C, Filipiak W, Khalid T, Osborne D and Ratcliffe N M 2014 A review of the volatiles from the healthy human body *J. Breath Res.* **8** 014001

[2] FilipiakW *et al* 2014 Comparative analyses of volatile organic compounds (vocs) from patients, tumors and transformed cell lines for the validation of lung cancer-derived breath markers *J. Breath Res.* **8** 027111

[3] Amann A, Miekisch W, Schubert J, Buszewski B, Ligor T, Jezierski T, Pleil J and Risby T 2014 Analysis of exhaled breath for disease detection *Annu. Rev. Anal. Chem.* **7** 455–82

[4] Haick H, Broza Y Y, Mochalski P, Ruzsanyi V and Amann A 2014 Assessment, origin, and implementation of breath volatile cancer markers *Chem. Soc. Rev.* **43** 1423–49

[5] Amann A, Mochalski P, Ruzsanyi V, Broza Y Y and Haick H 2014 Assessment of the exhalation kinetics of volatile cancer biomarkers based on their physicochemical properties *J. Breath Res.* **8** 016003

[6] Guengerich F P 2008 Cytochrome P450 and chemical toxicology *Chem. Res. Toxicol.* **21** 70–83

[7] Guengerich F P and Rendic S 2010 Update information on drug metabolism systems-2009, part I *Curr. Drug Metab.* **11** 1–3

[8] Lin J H and Lu A Y H 1998 Inhibition and induction of cytochrome p450 and the clinical implications *Clin. Pharmacokinet.* **35** 361–90

[9] Schroth W *et al* 2009 Association between CYP2D6 polymorphisms and outcomes among women with early stage breast cancer treated with tamoxifen *JAMA J. Am. Med. Assoc.* **302** 1429–36

[10] Borges S *et al* 2006 Quantitative effect of CYP2D6 genotype and inhibitors on tamoxifen metabolism: implication for optimization of breast cancer treatment *Clin. Pharm. Ther.* **80** 61–74

[11] Goetz M P *et al* 2007 The impact of cytochrome P450 2d6 metabolism in women receiving adjuvant tamoxifen *Breast Cancer Res. Treat.* **101** 113–21

[12] Opdam F L, Dezentje V O, den Hartigh J, Modak A S, Vree R, Batman E, Smorenburg C H, Nortier J W, Gelderblom H and Guchelaar H J 2013 The use of the C-13-dextromethorphan breath test for phenotyping CYP2D6 in breast cancer patients using tamoxifen: association with cyp2d6 genotype and serum endoxifen levels *Cancer Chemother. Pharmacol.* **71** 593–601

[13] Ingelman-Sundberg M, Sim S C, Gomez A and Rodriguez-Antona C 2007 Influence of cytochrome P450 polymorphisms on drug therapies: pharmacogenetic, pharmacoepigenetic and clinical aspects *Pharmacol. Ther.* **116** 496–526

[14] Zhou S-F 2009 Polymorphism of human cytochrome P450 2D6 and its clinical significance part I *Clin. Pharmacokinet.* **48** 689–723

[15] Zhou S-F 2009 Polymorphism of human cytochrome P450 2D6 and its clinical significance part II *Clin. Pharmacokinet.* **48** 761–804

[16] Sjoqvist F and Eliasson E 2007 The convergence of conventional therapeutic drug monitoring and pharmacogenetic testing in personalized medicine: focus on antidepressants *Clin. Pharmacol. Ther.* **81** 899–902

[17] Dahl M L 2002 Cytochrome P450 phenotyping/genotyping in patients receiving antipsychotics—useful aid to prescribing? *Clin. Pharmacokinet.* **41** 453–70

[18] Rendic S and Guengerich F P 2010 Update information on drug metabolism systems-2009, part ii. summary of information on the effects of diseases and environmental factors on human cytochrome P450 (CYP) enzymes and transporters *Curr. Drug Metab.* **11** 4–84

[19] Waxman D J 1999 P450 gene induction by structurally diverse xenochemicals: central role of nuclear receptors CAR, PXR and PPAR *Arch. Biochem. Biophys.* **369** 11–23

[20] Cotreau M M, von Moltke L L and Greenblatt D J 2005 The influence of age and sex on the clearance of cytochrome P450 3A substrates *Clin. Pharmacokinet.* **44** 33–60

[21] Mathijssen R H J and van Schaik R H N 2006 Genotyping and phenotyping cytochrome p450: perspectives for cancer treatment *Eur.J. Cancer* **42** 141–8

[22] Buszewski B, Kesy M, Ligor T and Amann A 2007 Human exhaled air analytics: biomarkers of diseases *Biomed. Chromatogr.* **21** 553–66

[23] Miekisch W, Schubert J K and Noeldge–Schomburg G F 2004 Diagnostic potential of breath analysis—focus on volatile organic compounds *Clin. Chim. Acta* **347** 25–39

[24] Bajtarevic A *et al* 2009 Noninvasive detection of lung cancer by analysis of exhaled breath *BMC Cancer* **9** 348

[25] Amann A, Corradi M, Mazzone P and Mutti A 2011 Lung cancer biomarkers in exhaled breath *Expert Rev. Mol. Diagn.* **11** 207–17

[26] Ligor M *et al* 2009 Determination of volatile organic compounds in exhaled breath of patients with lung cancer using solid phase microextraction and gas chromatography mass spectrometry *Clin. Chem. Lab. Med.* **47** 550–60

[27] Wehinger A, Schmid A, Mechtcheriakov S, Ledochowski M, Grabmer C, Gastl G A and Amann A 2007 Lung cancer detection by proton transfer reaction mass-spectrometric analysis of human breath gas *Int. J. Mass Spectrom.* **265** 49–59

[28] Watkins P B, Murray S A, Winkelman L G, Heuman D M, Wrighton S A and Guzelian P S 1989 Erythromycin breath test as an assay of glucocorticoid-induced liver cytochromes P-450. Studies in rats and patients *J. Clin. Invest.* **83** 688–97

[29] Watkins P B 1996 Erythromycin breath test and clinical transplantation *Ther. Drug Monit.* **18** 368–71

[30] Knox C *et al* 2011 Drugbank 3.0: a comprehensive resource for 'omics' research on drugs *Nucleic Acids Res.* **39** D1035–41

[31] Chemical Computing Group Inc 2011 *MOE Molecular Operating Environment, 2011.10 and 2010.10* (Montreal, Canada)

[32] Fuchs J E, Spitzer G M, Javed A, Biela A, Kreutz C, Wellenzohn B and Liedl K R 2011 Minor groove binders and drugs targeting proteins cover complementary regions in chemical shape space *J. Chem. Inform. Model.* **51** 2223–32

[33] Raventos-Duran T, Camredon M, Valorso R, Mouchel-Vallon C and Aumont B 2010 Structure-activity relationships to estimate the effective henry's law constants of organics of atmospheric interest *Atmos. Chem. Phys.* **10** 7643–54

[34] Lewis D F V 2000 On the recognition of mammalian microsomal cytochrome P450 substrates and their characteristics—towards the prediction of human p450 substrate specificity and metabolism *Biochem. Pharmacol.* **60** 293–306

[35] Terfloth L, Bienfait B and Gasteiger J 2007 Ligand-based models for the isoform specificity of cytochrome P450 3A4, 2D6 and 2A9 substrates *J.Chem. Inform. Model.* **47** 1688–1701

[36] Yamashita F, Feng C, Yoshida S, Itoh T and Hashida M 2011 Automated information extraction and structure-activity relationship analysis of cytochrome P450 substrates *J. Chem. Inform. Model.* **51** 378–85

[37] Zhang T, Dai H, Liu L A, Lewis D F V and Wei D 2012 Classification models for predicting cytochrome P450 enzyme-substrate selectivity *Mol. Inform.* **31** 53–62

[38] Ekins S, de Groot M J and Jones J P 2001 Pharmacophore and three-dimensional quantitative structure activity relationship methods for modeling cytochrome P450 active sites *Drug Metab. Dispos.* **29** 936–44

[39] Wildman S A and Crippen G M 1999 Prediction of physicochemical parameters by atomic contributions *J. Chem. Inform. Comput. Sci.* **39** 868–73

[40] Singh S B, Shen L Q, Walker M J and Sheridan R P 2003 A model for predicting likely sites of CYP3a4-mediated metabolism on drug-like molecules *J. Med. Chem.* **46** 1330–6

[41] Jones J P, Mysinger M and Korzekwa K R 2002 Computational models for cytochrome P450: a predictive electronic model for aromatic oxidation and hydrogen atom abstraction *Drug Metab. Dispos.* **30** 7–12

[42] Olsen L, Rydberg P, Rod T H and Ryde U 2006 Prediction of activation energies for hydrogen abstraction by cytochrome P450 *J. Med. Chem.* **49** 6489–99

[43] Halgren T A 1996 Merck molecular force field .1. Basis, form, scope, parameterization, and performance of MMFF94 *J. Comput. Chem.* **17** 490–19

[44] Dewar M J S, Zoebisch E G, Healy E F and Stewart J J P 1985 Development and use of quantum mechanical molecular models.76. AM1: a new general purpose quantum mechanical molecular model *J. Am. Chem. Soc.* **107** 3902–9

[45] Hehre W J, Stewart R F and Pople J A 1969 Self-consistent molecular-orbital methods.i. use of gaussian expansions of slater-type atomic orbitals *J. Chem. Phys.* **51** 2657–65

[46] Frisch M J *et al* 2009 *Gaussian 09, Revision D.01* (Wallingford, CT: Gaussian)

[47] Rydberg P, Rostkowski M, Gloriam D E and Olsen L 2013 The contribution of atom accessibility to site of metabolism models for cytochromes P450 *Mol. Pharm.* **10** 1216–23

[48] Hasegawa K, Koyama M and Funatsu K 2010 Quantitative prediction of regioselectivity toward cytochrome P450/3A4 using machine learning approaches *Mol. Inform.* **29** 243–9

[49] Tyzack J D, Mussa H Y, Williamson M J, Kirchmair J and Glen R C 2014 Cytochrome P450 site of metabolism prediction from 2D topological fingerprints using GPU accelerated probabilistic classifiers. *J. Cheminform.* **6** 29

[50] Daylight Chemical Information Systems Inc 2011 *Daylight Theory Manual* (Laguna Niguel, CA) (www.daylight.com/dayhtml/doc/theory/index.html)

[51] OpenEye Scientific Software Inc 2012 *Oechem TK—Python Release 1.9.0; Toolkit Release 2012* (Santa Fe, NM) (www.eyesopen.com)

[52] Katritzky A, R Lobanov V S and Karelson M 1995 QSPR—the correlation and quantitative prediction of chemical and physical-properties from structure *Chem. Soc. Rev.* **24** 279–87

[53] R Foundation for Statistical Computing 2008 *R: A Language and Environment for Statistical Computing* (Vienna, Austria: R Development Core Team) (www.R-project.org)

[54] Law V *et al* 2014 Drugbank 4.0: shedding new light on drug metabolism *Nucleic Acids Res.* **42** D1091–7

[55] Lide D R (Eds) 1994 *CRC Handbook of Chemistry and Physics* 75th edn (Boca Raton, FL: CRC Press)

[56] Balani S K *et al* 1994 Metabolism of 3-2-(benzoxazol-2-yl) ethyl -5-ethyl-6-methylpyridin-2(1H)-one (L-696,229), an HIV-1 reverse-transcriptase inhibitor, by rat-liver slices and in humans *Drug Metab. Dispos.* **22** 200–5

[57] Lin J H 1997 Human immunodeficiency virus protease inhibitors—from drug design to clinical studies *Adv. Drug Deliv. Rev.* **27** 215–33

[58] Ernest C S, Hall S D and Jones D R 2005 Mechanism-based inactivation of CYP3A by hiv protease inhibitors *J. Pharmacol. Exp. Ther.* **312** 583–91

[59] Kirchmair J, Williamson M J Tyzack J D Tan L Bond P J Bender A and Glen R C 2012 Computational prediction of metabolism: sites, products, SAR, P450 enzyme dynamics and mechanisms *J. Chem. Inf. Model.* **52** 617–48

[60] Mishra N K, Agarwal S and Raghava G P 2010 Prediction of cytochrome P450 isoform responsible for metabolizing a drug molecule *BMC Pharmacol.* **10** 8

[61] Korzekwa K R, Jones J P and Gillette J R 1990 Theoretical-studies on cytochrome-P-450 mediated hydroxylation: a predictive model for hydrogen-atom abstractions *J. Am. Chem. Soc.* **112** 7042–6

[62] Rydberg P, Ryde U and Olsen L 2008 Prediction of activation energies for aromatic oxidation by cytochrome P450 *J. Phys. Chem.* A **112** 13058–65

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

[64] Li J, Schneebeli S T, Bylund J, Farid R and Friesner R A 2011 Idsite: an accurate approach to predict P450-mediated drug metabolism *J. Chem. Theory Comput.* **7** 3829–45

[65] Tyzack J D, Williamson M J, Torella R and Glen R C 2013 Prediction of cytochrome P450 xenobiotic metabolism: tethered docking and reactivity derived from ligand molecular orbital analysis *J. Chem. Inform. Model.* **53** 1294–305

[66] Sun H and Scott D O 2010 Structure-based drug metabolism predictions for drug design *Chem. Biol. Drug Des.* **75** 3–17

[67] Kirchmair J, Williamson M J, Afzal A M, Tyzack J D, Choy A P K, Howlett A, Rydberg P and Glen R C 2013 FAst Metabolizer (FAME): a rapid and accurate predictor of sites of metabolism in multiple species by endogenous enzymes *J. Chem. Inform. Model.* **53** 2896–907

[68] Adams S, Carlsson L, Boyer S and Glen R 2010 *Metaprint2d Unilever Centre for Molecular Science Informatics* (University of Cambridge) (www-metaprint2d.ch.cam.ac.uk/metaprint2d-react/about.html)

[69] Campagna-Slater V, Pottel J, Therrien E, Cantin L-D and Moitessier N 2012 Development of a computational tool to rival experts in the prediction of sites of metabolism of xenobiotics by P450s *J. Chem. Inform. Model.* **52** 2471–83

[70] Cruciani G, Carosati E, De Boeck B, Ethirajulu K, Mackie C, Howe T and Vianello R 2005 Metasite: understanding metabolism in human cytochromes from the perspective of the chemist *J. Med. Chem.* **48** 6970–9

[71] Pan L P, Wijnant P, De Vriendt C, Rosseel M T and Belpaire F M 1997 Characterization of the cytochrome P450 isoenzymes involved in the *in vitro* N-dealkylation of haloperidol *Br. J. Clin. Pharmacol.* **44** 557–64

[72] Kramer C, Kalliokoski T Gedeck, P and Vulpetti A 2012 The experimental uncertainty of heterogeneous public k-i data *J Med. Chem.* **55** 5165–73