Evaluation of Three State-of-the-Art Metabolite Prediction Software Packages (Meteor, MetaSite, and StarDrop) through Independent and Synergistic Use


Laboratory of Medical Biochemistry and Clinical Analysis, Faculty of Pharmaceutical Sciences, Ghent University, Ghent, Belgium (H.T., K.B., J.F.V.B.); ADME-Tox Department, Janssen Research and Development, Janssen Pharmaceutica, Beerse, Belgium (R.J.M.-S.); ADME-Tox Department, Janssen Research and Development, San Diego, California (K.C.); and Department of Medicinal Chemistry (B.D.B.) and Drug Metabolism and Pharmacokinetics Department (G.M.), Janssen Research and Development, Beerse, Belgium

Received June 20, 2011; accepted August 10, 2011

ABSTRACT:
The aim of this study was to evaluate three different metabolite prediction software packages (Meteor, MetaSite, and StarDrop) with respect to their ability to predict loci of metabolism and suggest relative proportions of metabolites. A chemically diverse test set of 22 compounds, for which in vivo human mass balance studies and metabolic schemes were available, was used as basis for the evaluation. Each software package was provided with structures of the parent compounds, and predicted metabolites were compared with experimentally determined human metabolites. The evaluation consisted of two parts. First, different settings within each software package were investigated and the software was evaluated using those settings determined to give the best prediction. Second, the three different packages were combined using the optimized settings to see whether a synergistic effect concerning the overall metabolism prediction could be established. The performance of the software was scored for both sensitivity and precision, taking into account the capabilities/limitations of the particular software. Varying results were obtained for the individual packages. Meteor showed a general tendency toward overprediction, and this led to a relatively low precision (~35%) but high sensitivity (~70%). MetaSite and StarDrop both exhibited a sensitivity and precision of ~50%. By combining predictions obtained with the different packages, we found that increased precision can be obtained. We conclude that the state-of-the-art individual metabolite prediction software has many advantageous features but needs refinement to obtain acceptable prediction profiles. Synergistic use of different software packages could prove useful.

Introduction

In drug discovery, the metabolism of a drug compound is one of the key parameters to be investigated and optimized to obtain acceptable pharmacokinetic and/or safety profiles. In development, early prediction of relevant metabolites before introduction of the drug compound into man would substantially help preclinical and clinical development concerning the analysis of (phase I and/or II) metabolism data and decision making (Afzelius et al., 2007; Van Campen, 2009).

A variety of different approaches to predict drug metabolism have been described: rule-, ligand-protein-, and ligand-based methods (Cruciani et al., 2005; Tarcsey and Keseru, 2011). Other approaches include ab initio, classic pharmacophore, and three-dimensional quantitative structure-activity relationship pharmacophore-based methods (Testa, 2004; Cruciani et al., 2005; Czodrowski et al., 2009). Unfortunately, obstacles such as biological factors and substrate-product selectivity still hinder the widespread use of drug metabolism as a reliable tool (Testa, 2004).

Here, we describe the evaluation of three software tools: Meteor (Lhasa Ltd., Leeds, UK), MetaSite (Molecular Discovery Ltd., Middlesex, UK), and StarDrop (Optibrium Ltd., Cambridge, UK). We have chosen these tools because they are currently being used in both the discovery and development phases of the pharmaceutical industry and each tool is based on a different approach to predict metabolism. A comparison of the software packages is contained in Table 1.

Meteor is a rule-based (empirical) software tool (Langowski and Long, 2002). Its algorithm involves three successive steps. First, Meteor checks whether the query structure contains substructures that are labile toward any of the biotransformations contained in its knowledge base. Second, absolute reasoning rules evaluate the likelihood of a biotransformation taking place based on five levels: probable, plausible, equivocal, doubted, and improbable (Button et al., 2003). This classification depends on the logP of the query structure, whether the query structure is the unchanged drug or a metabolite, and in which
species the metabolism is to be predicted. Third, relative reasoning then is used to rank those biotransformations that can occur concomitantly on the same compound, based on a set of relative precedences (e.g., primary alcohols are oxidized in preference to secondary alcohols). Relative reasoning can be set at 0, 1, 2, 3 and can only be used when two or more biotransformations apply for the same compound: RRL0 means that no relative reasoning will be applied in the upcoming analysis; RRL1 means that only metabolites will be displayed for which there is no metabolite more likely (i.e., most likely); RRL2 means that most likely metabolites, and metabolites for which there is only one level of metabolites more likely, will be displayed. In essence, the reasoning engine uses further rules to avoid a combinatorial explosion of output resulting from unconstrained analyses of query structures (Ekins et al., 2007).

MetaSite is an automated docking model with reactivity correction (this correction considers the reactivity components of an atom related to the heme) and is designed to predict phase I cytochrome P450 (P450) metabolism. Based on GRID descriptors for the P450 enzymes and the potential substrate (Cruciani et al., 2005), metabolism is evaluated at all possible sites on the molecular structure, assigning every atom a likelihood of metabolism. Reactivity correction can be put in three different modes: 1) “off”; 2) “on for substrate”; and 3) “on for substrate and CYP” (Cruciani et al., 2005; Zhou et al., 2006; Vaz et al., 2010).

StarDrop uses a quantum mechanical approach for the prediction of the relative involvement of CYP3A4, 2D6, and 2C9 of the query compound. Its mechanism is based on calculation of the energy barrier to the electron removal, which is considered to be the rate-limiting step in product formation. All of the modeled P450 isoforms use the same model for the calculation of electronic lability, but each isoform has a different model for steric accessibility and orientation effects (different regioselectivity) (Earnshaw, 2010).

Metabolism prediction is scored using two performance parameters, sensitivity and precision, with the in vivo metabolic profile as a reference point. Sensitivity is a measure of how many in vivo metabolites are captured by the software; precision is a measure of how many predicted metabolites are actually observed in vivo (see Materials and Methods for a definition). Ideally, sensitivity as well as precision should be as high as possible for any given prediction. However, a prioritization of either sensitivity or precision will be necessary, depending on the scope of the intended research. Figure 1 correlates the location of a prediction result within each quadrant in the sensitivity/prediction plot with the value of the prediction result. In the Discussion, we will further comment on the usefulness of the prediction result in drug discovery and development.

### Materials and Methods

**The In Vivo Reference Set.** The goal of most in silico models is to predict what cannot directly be measured. In this project, the output of the metabolism prediction software packages was compared with the known metabolite profile obtained after administration in man. To understand the most important metabolic clearance pathways of a compound, the mass balance of the parent and its metabolites was determined in excreta after a single oral dose. A test set of 22 compounds (Fig. 2; Table 2) was chosen with divergent chemical structures, therapeutic doses, therapeutic indications, extent of metabolism, and number of metabolites and represented a range of metabolic lability from barely to extensively metabolized compounds.

---

**TABLE 1**

Overview of the three metabolite prediction software packages

<table>
<thead>
<tr>
<th>Mechanism</th>
<th>Displayed</th>
<th>Phase I</th>
<th>Phase II</th>
<th>Cytochrome P450 Isoforms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Meteor version 11</td>
<td>Rule-based</td>
<td>Metabolites</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>MetaSite version 3.0.1</td>
<td>Docking and reactivity correction</td>
<td>Metabolic hotspots</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>StarDrop version 3.4</td>
<td>Quantum mechanics</td>
<td>Metabolic hotspots</td>
<td>Yes</td>
<td>No</td>
</tr>
</tbody>
</table>

**Sensitivity/precision:**

*usefulness*

- **high sensitivity/low precision**: most in vivo metabolites are predicted, but also a lot of false predictions are made
- **results are sensitive, but not precise**
- **low sensitivity/low precision**: few in vivo metabolites are predicted and also a lot of false predictions are made
- **results are useless**
- **high sensitivity/high precision**: most in vivo metabolites are predicted and not many false predictions are made
- **results are sensitive & precise**
- **low sensitivity/high precision**: few in vivo metabolites are predicted, but not many false predictions are made
- **results are precise, but not sensitive**

Fig. 1. Sensitivity and precision classification and likely use of the output.
For every compound in the set of test compounds, the report of the human absorption, metabolism, and excretion trial was abstracted to provide a table of observed metabolites. Finally, the primary, main metabolic pathways were collated, highlighting metabolites that were expected to be formed directly from the parent compound (illustrated in the supplemental data).

The reference set contained 68 (phase I P450, phase I non-P450, and phase II) primary biotransformation pathways. A list of these biotransformations can

Fig. 2. Chemical structures of the 22 compounds of the test set.
be found per compound in Table 2. MetaSite and StarDrop are only able to predict phase I P450 metabolism. Consequently, only 50 (instead of 68 for Meteor) of the observed in vivo P450 biotransformations were taken as basis for the evaluation.

### Software Packages

The three different metabolite prediction software tools were Meteor (editor mode, version 11; Lhasa Ltd.), MetaSite (version 3.0.1; Molecular Discovery Ltd.), and StarDrop (version 3.4; Optibrium Ltd.). Chemical structures were drawn with CS ChemDraw Pro (version 4.5; Cam-

### Table 2

**Test set of 22 compounds that was submitted to the different software packages and for which in vivo mass balance studies were available as reference**

<table>
<thead>
<tr>
<th>Compound</th>
<th>Indication/Target</th>
<th>Single Oral Dose</th>
<th>Metabolic Clearance Normalized (TR-UD)/TR</th>
<th>Type of Biotransformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 RWJ-241947</td>
<td>Antihyperglycemic agent</td>
<td>80</td>
<td>85.1</td>
<td>O-dealkylation</td>
</tr>
<tr>
<td>2 Levocabastine</td>
<td>H1 antagonist</td>
<td>1</td>
<td>21.8</td>
<td>Thiiazolidinedione open</td>
</tr>
<tr>
<td>3 Risperidone</td>
<td>Antipsychotic agent</td>
<td>1</td>
<td>81.7</td>
<td>Acyl glucuronidation</td>
</tr>
<tr>
<td>4 Vorozole</td>
<td>Aromatase inhibitor</td>
<td>2.5</td>
<td>99.8</td>
<td>N-dealkylation</td>
</tr>
<tr>
<td>5 Cisapride</td>
<td>Gastroprokinetic agent</td>
<td>10</td>
<td>89.6</td>
<td>Aliphatic oxidation</td>
</tr>
<tr>
<td>6 Liarozole</td>
<td>Antitumor agent</td>
<td>150</td>
<td>85.6</td>
<td>N-glucuronidation</td>
</tr>
<tr>
<td>7 Ridogrel</td>
<td>Anticoagulant</td>
<td>5</td>
<td>98.9</td>
<td>N-dealkylation</td>
</tr>
<tr>
<td>8 Prucalopride</td>
<td>Chronic constipation</td>
<td>0.5</td>
<td>12.1</td>
<td>N-dealkylation</td>
</tr>
<tr>
<td>9 Sabeluzole</td>
<td>Alzheimer</td>
<td>10</td>
<td>82.1</td>
<td>O-dealkylation</td>
</tr>
<tr>
<td>10 Tipifarnib</td>
<td>Antitumor agent</td>
<td>50</td>
<td>93.6</td>
<td>N-dealkylation</td>
</tr>
<tr>
<td>11 Norcisapride</td>
<td>Gastroprokinetic agent</td>
<td>2.5</td>
<td>0</td>
<td>N-dealkylation</td>
</tr>
<tr>
<td>12 Carisbamate</td>
<td>Antiepileptic agent</td>
<td>500</td>
<td>98.3</td>
<td>Carbamyl hydrolysis*</td>
</tr>
<tr>
<td>13 JNJ-37822681</td>
<td>D2 antagonist</td>
<td>20</td>
<td>98.9</td>
<td>Glucuronidation</td>
</tr>
<tr>
<td>14 Canagliflozin</td>
<td>SGLT2 inhibitor</td>
<td>200</td>
<td>55.3</td>
<td>Epimerization</td>
</tr>
<tr>
<td>15 Doripenem</td>
<td>Broad spectrum carbapenem</td>
<td>500 (i.v.)</td>
<td>69.0</td>
<td>Benzylic oxidation</td>
</tr>
<tr>
<td>16 Darunavir</td>
<td>Anti-HIV-1 agent</td>
<td>400</td>
<td>89.6</td>
<td>Benzylic oxidation</td>
</tr>
<tr>
<td>17 Etravirine</td>
<td>Anti-HIV-1 agent</td>
<td>800</td>
<td>7.8</td>
<td>Glucuronidation</td>
</tr>
<tr>
<td>18 Rilpivirine</td>
<td>Anti-HIV-1 agent</td>
<td>150</td>
<td>61.2</td>
<td>Methyl oxidation</td>
</tr>
<tr>
<td>19 Galantamine</td>
<td>Acetylcholinesterase inhibitor</td>
<td>4</td>
<td>66.7</td>
<td>Methyl oxidation</td>
</tr>
<tr>
<td>20 R228060</td>
<td>Antidepressant</td>
<td>200</td>
<td>0.4</td>
<td>α,β-Unsaturated bond oxidation</td>
</tr>
<tr>
<td>21 Paliperidone</td>
<td>Antipsychotic agent</td>
<td>1</td>
<td>34.8</td>
<td>GSH conjugation*</td>
</tr>
<tr>
<td>22 JNJ-31001074</td>
<td>ADHD</td>
<td>10</td>
<td>87.8</td>
<td>O-demethylation</td>
</tr>
</tbody>
</table>

TR, total radioactivity in excreta; UD, unchanged drug; JNJ-37822681, N-[1-(3,4-difluorobenzyl)piperidin-4-yl]-6-(trifluoromethyl)pyridazin-3-amine; R228060, 2-aminoo 3-phenylpropyl carbamate; JNJ-31001074, 4-[1-[(4-cyclopentylpiperazin-1-yl)carbonyl]phenyl]morpholine; ADHD, attention deficit hyperactivity disorder.

* Carbamyl hydrolysis is considered as ester hydrolysis by esterases.

* GSH conjugation is considered phase II.
Testing of Individual Software Packages. Both Meteor and MetaSite allow the user to customize the prediction settings. Only StarDrop had no selectable settings. The outcome of each of the various settings investigated was evaluated, and these different outcomes were compared with the in vivo data. The goal of this software-per-software exercise was to find which settings provided the best prediction for a given software package (INTRA), 2) compare the three software packages to each other using their optimized settings (INTER), and 3) investigate whether there was any added prediction value in combining the outcomes of individual software packages (illustrated in the supplemental data).

Meteor. Each chemical structure was imported, and predictions were performed using each possible setting [absolute reasoning level (ARL) and relative reasoning level (RRL)]. Absolute reasoning was tested only at the “probable” and “plausible” levels. Relative reasoning was tested on levels 0, 1, and 2 for each absolute reasoning level. The outcomes for these six different setting combinations were compared with 68 in vivo biotransformations from the reference data. From now on, when quoting a setting in Meteor, we will use the format “ARL RRL” (e.g., “plausible 1”).

Meteor predicts biotransformations rather than metabolic hotspots. This is because per locus of metabolism, several metabolites can be formed. For example, following N-dealkylation, three metabolites are formed in silico: an amine, an alcohol, and further downstream oxidation of the alcohol to the acid. This complicates the comparison of the three packages, because where MetaSite or StarDrop would predict one site of metabolism correlating with N-dealkylation, Meteor would predict three metabolites.

MetaSite. Each chemical structure was imported into the software, and a prediction of hotspots was performed. MetaSite allows the user to toggle between various reaction settings post hoc. We have tested the settings “without reactivity correction,” “with reactivity correction for the substrate,” and “with reactivity correction for substrate and CYP.” Only “most probable,” and “intermediately probable” sites (correlating with relative percentage intervals [0.50; 1.00] and [0.25; 0.50], respectively) were considered. The metabolic hotspots obtained with these six different settings were compared with 50 in vivo P450 biotransformations from the reference data. From now on, when quoting a setting in MetaSite, we will use the format “considered SOM; reactivity correction setting” (e.g., “most probable and intermediately probable; on for substrate”).

StarDrop. Each chemical structure was imported into the software, and a prediction of metabolism was performed. StarDrop had no user-adaptable settings. Only “labile” and “moderately labile” positions were considered. The outcome was compared with 50 in vivo P450 biotransformations from the reference data.

Testing the Combination of Software Packages. The individual outcomes of the different software packages were also combined to examine whether this could increase the prediction reliability. Several combinational approaches were tested and reported using the following approaches (Table 3).

Intersection. All three software packages predict P450 reactions; only Meteor also predicts phase I non-P450 and phase II reactions. Consequently, when evaluating which metabolites were predicted in common by the various packages, only P450 reactions were taken into account. This “intersection” approach was used with the following combinations: Meteor-MetaSite; Meteor-StarDrop; and MetaSite-StarDrop.

Union. In the “union” approach, all the predictions that are made by each package are taken into account, including biotransformations that are predicted in common and biotransformations that are independently predicted by the software packages. We also investigated whether the different software packages were complementary to each other. By “complementary,” we mean the prediction of independently predicted biotransformations of a certain software package that were not seen in common to the two software packages. The following combinations were made: Meteor-MetaSite; Meteor-StarDrop; and MetaSite-StarDrop.

Intersection + n. “Intersection + 2” was the approach used to combine results from Meteor and MetaSite: P450 biotransformations that were common to these two packages, together with additional phase I non-P450 and phase II predictions from Meteor. “Intersection + 3” was the approach used to combine results from Meteor, MetaSite, and StarDrop: P450 biotransformations common to 1) the three packages and 2) Meteor and MetaSite, Meteor and StarDrop, and MetaSite and StarDrop, together with additional phase I non-P450 and phase II predictions from Meteor.

Evaluation. Each prediction of a metabolite or biotransformation route was compared with the experimentally observed in vivo biotransformations by scoring a predicted metabolite or biotransformation route as “TRUE,” “FALSE,” or “MISSED.”

In the case of binary classification of data (as in this case), performance testing parameters such as sensitivity, specificity, accuracy, and precision can be used. In the case of metabolism predictions compared with a reference set, accuracy and specificity cannot be used because the concept “true negative” (used to calculate accuracy and specificity) is not appropriate. A true negative would mean a metabolite NOT predicted in silico and NOT seen in vivo. This is an irrelevant situation in metabolism prediction, because one could then take any possible biotransformation, whether or not relevant to the query structure, into the comparison. For example, the human absorption, metabolism, and excretion report indicates that 1-{[6-(2-fluorobenzyl)-oxy-2-naphthyl]methyl}thioyl-2,4-thiazolidinedione (RWI-214947) is not acetylated in vivo (supplemental data). RWI-214947 acetylation is also not predicted in silico. Thus, acetylation would be a true negative. Nevertheless, true negatives do not make any sense in the evaluation of the performance of a metabolism prediction package.

Our prediction quality assessment was then as follows: true positive, a metabolite predicted and also seen in vivo was assigned the value TRUE; false positive, a metabolite predicted but NOT seen in vivo was assigned the value FALSE; false negative, a metabolite NOT predicted but seen in vivo was assigned the value MISSED.

Precision describes the proportion of correctly predicted metabolites in the population of all in silico metabolites of a test compound, and is defined as follows: precision = true positives/(true positives + false positives) × 100%.

Sensitivity describes the ability of a system to predict a TRUE metabolite in the population of all in vivo metabolites of a certain test compound and is defined as follows: sensitivity = true positives/(true positives + false negative) × 100%.

Results

Results for the Individual Software Packages. Meteor. Meteor was tested at six different settings: the ARL at probable with RRL 0,

<table>
<thead>
<tr>
<th>Combination Approach</th>
<th>Meteor</th>
<th>MetaSite</th>
<th>StarDrop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intersection</td>
<td>Plausible 2</td>
<td>Most probable; on for substrate</td>
<td>No</td>
</tr>
<tr>
<td>Intersection</td>
<td>Plausible 0</td>
<td>Most probable; on for substrate</td>
<td>No</td>
</tr>
<tr>
<td>Intersection</td>
<td>Plausible 0</td>
<td>Most probable; on for substrate</td>
<td>Yes</td>
</tr>
<tr>
<td>Intersection</td>
<td>No</td>
<td>Most probable; on for substrate</td>
<td>Yes</td>
</tr>
<tr>
<td>Intersection</td>
<td>Plausible 2</td>
<td>Most probable; on for substrate</td>
<td>No</td>
</tr>
<tr>
<td>Union</td>
<td>Plausible 0</td>
<td>Most probable; on for substrate</td>
<td>No</td>
</tr>
<tr>
<td>Union</td>
<td>Plausible 1</td>
<td>Most probable; on for substrate</td>
<td>No</td>
</tr>
<tr>
<td>Union</td>
<td>Plausible 1</td>
<td>Most probable; on for substrate</td>
<td>No</td>
</tr>
<tr>
<td>Union</td>
<td>No</td>
<td>Most probable; on for substrate</td>
<td>Yes</td>
</tr>
<tr>
<td>Intersection+2</td>
<td>Plausible 0 plus additional phase I non-P450 and phase II</td>
<td>Most probable; on for substrate</td>
<td>No</td>
</tr>
<tr>
<td>Intersection+3</td>
<td>Plausible 0 plus additional phase I non-P450 and phase II</td>
<td>Most probable; on for substrate</td>
<td>Yes</td>
</tr>
</tbody>
</table>
1, and 2; and the ARL at plausible with RRL 0, 1, and 2. For each of these settings, the absolute amounts of TRUE, FALSE, and MISSED biotransformations are shown in Fig. 3.

The influence of the parameter settings is best visualized through a precision versus sensitivity plot (Fig. 4). Using the probable level, the sensitivity is ~20% (20% of the experimentally observed metabolites are correctly predicted) with a precision of ~40% (40% of the predicted metabolites are actually observed in vivo). In the plausible level, the precision lowers a few percentage points, but there is a very substantial increase in sensitivity. Indeed, when moving from probable to plausible, many of the MISSED metabolites (not predicted, although experimentally observed) from the probable level do become TRUE metabolites (predicted and observed) in the plausible level.

Relative reasoning levels constrain the prediction output. The lower the relative reasoning level, the more restriction there is. The result of this constraint is exemplified in Fig. 4, where a decrease in RRL (RRL 1 compared with RRL 2) is associated with a strong decrease in sensitivity (especially at plausible absolute reasoning) and only a small increase in precision. That is to say, constraining the prediction output (lower RRL) results in slightly more precise, yet less sensitive, results. With the relative reasoning set to off (RRL = 0) we can isolate the effect of the absolute reasoning filter. In Fig. 4, the arrows clearly indicate that when moving from probable to plausible ARL, we gain

---

**The outcomes for the individual settings for Meteor**

![Diagram showing sensitivity and precision for different settings of Meteor](image-url)

**Fig. 3.** Absolute number of biotransformations in function of six different settings in Meteor. TRUE and FALSE predictions are made by the software; MISSED predictions are not seen in silico.

**Fig. 4.** Sensitivity/precision plot representing the outcomes for six different settings in Meteor. Most promising settings are encircled and will be used later on for making combinations of software packages. Note that dots 1 and 5 coincide.
Overview of the outcomes for the individual settings concerning CYP predictions

![Sensitivity/precision plot](image)

In Fig. 5, the sensitivity/precision plot representing the outcomes for six settings in MetaSite and one setting in StarDrop, all concerning P450 predictions. Most promising settings are circled per software package and will be used later on for making combinations of software packages.

Fig. 5. Sensitivity/precision plot representing the outcomes for six settings in MetaSite and one setting in StarDrop, all concerning P450 predictions. Most promising settings are circled per software package and will be used later on for making combinations of software packages.

**Discussion**

**Optimal Individual Software Settings.** Meteor in setting plausible 1 offered more precise, but less sensitive, predictions compared with the other five Meteor results (Fig. 4). In contrast, settings plausible 2 (default setting) and plausible 0 (Meteor’s most valuable setting) can be used to obtain more sensitive results. In our opinion the enhanced sensitivity is Meteor’s most valuable feature. However, it remains to be seen whether this enhanced sensitivity is really an advantage, especially in a discovery setting. Rule-based software such as Meteor is known to show a tendency for overprediction (high sensitivity/low precision) of metabolism (Cruciani et al., 2005), e.g., a piperidine moiety in a compound can undergo N-oxidation, but in another compound, the same piperidine moiety is oxidized on a carbon atom (lactam formation) of the heterocycle. Rule-based software that recognizes the piperidine moiety will generate these two metabolites (N-oxide and lactam). A high sensitivity/low precision is useless in lead optimization, because it is not able to guide the medicinal
Overview of best individual and combinational results concerning CYP predictions

1. Best individual settings for MetaSite
2. Best individual setting for StarDrop
3. Intersection (Meteor 'plausible 2' + MetaSite 'most probable_on for substrate')
4. Intersection (Meteor 'plausible 0' + MetaSite 'most probable_on for substrate')
5. Intersection (Meteor 'plausible 0' + StarDrop)
6. Intersection (MetaSite 'most probable_on for substrate' + StarDrop)
7. Union (MetaSite 'most probable_on for substrate' + StarDrop)

Overview of best individual and combinational results concerning phase I and phase II predictions

1. Meteor individual setting 'plausible 0'
2. Meteor individual setting 'plausible 2'
3. Union (Meteor 'plausible 2' + MetaSite 'most probable_on for substrate')
4. Union (Meteor 'plausible 0' + MetaSite 'most probable_on for substrate')
5. Union (Meteor 'plausible 1' + MetaSite 'most probable_on for substrate')
6. Union (Meteor 'plausible 1' + StarDrop)
7. Intersection+2 (Meteor 'plausible 0' + MetaSite 'most probable_on for substrate')
8. Intersection+3 (Meteor 'plausible 0' + MetaSite 'most probable_on for substrate' + StarDrop)

Combing Metabolite Prediction Software. The intersection combination of Meteor in the configuration plausible 0 with MetaSite in the configuration “most probable; on for substrate” enhances prediction reliability, because it provides a precision level practically unsurpassed by any other setting or combination (Fig. 6). Again, this is at the expense of sensitivity. By applying the intersection approach, an increase in prediction precision is achieved by a reduction of false-positive predictions made by the software packages and underscores the added value of combining software packages that are based on quite different algorithms.

The union combination of MetaSite and StarDrop (Fig. 6) was not useful as such but could provide us with some valuable insights: a
sensitivity of 65% is achieved by almost equal contributions of the two software tools, whereas in their intersection combination (Fig. 6), sensitivity represented only 37% (meaning few biotransformations in common). This would imply that MetaSite and StarDrop are, to a certain degree, complementary to each other in predicting in vivo metabolism. The union combinations from Fig. 7 did not yield more precise nor more sensitive results compared with Meteor’s best individual settings (plausible 0; plausible 2).

By making the intersection+n combinations (sum of the common P450 predictions by two or three programs and the phase I non-P450 and phase II predictions from Meteor), we wanted to investigate how the prediction would be affected if Meteor supplied information only on phase I non-P450 and phase II predictions, whereas the precision in P450 predictions is maintained by using the intersection approach. This procedure fits in neatly with the concept of taking the positive aspects of each individual software package and trying to circumvent, as much as possible, their limitations. Figure 7 demonstrates that precision is increased by nearly 20% for Meteor, 15% for StarDrop, and 5% for MetaSite compared with the individual packages when used in isolation.

In a recently published review article (Tarczay and Keseru, 2011), the performance of metabolite prediction software was also assessed. When we compared our investigation results with the studies reviewed in this publication, we found that our prediction quality assessment points toward a substantially lower prediction quality compared with what these authors report. This could be due to the fact that a great difference in scoring methodology exists between both our studies. The studies mentioned by Tarczay and Keseru (2011) report on scoring the accuracy of predictions (SOMs) of the different prediction tools based on the first, the first two, or the first three ranked predicted metabolites per software tool and per P450 isoform. In our article, we wanted to present a more generic approach, scoring the performance of the software outcome by letting it use its own constraints instead of the user choosing some cut-off value. We also wanted to evaluate the value of combining both oxidative and conjugative biotransformation predictions. We acknowledge the primary importance of P450-mediated metabolism, but reliable predictions of phase II metabolism are also of interest (Cubitt et al., 2011).

**Limitations and Pitfalls.** An important remark is that although we consider the in vivo metabolite profile in human as the golden standard, this approach has some limitations. The metabolite profile may change as the dose level (saturation of processes) or extent of absorption (formulation dependence) changes, or when transporters affect the exposure of the compound to the metabolizing enzymes. However, similar concerns also apply when comparing the software outputs against in vitro metabolite profiles in liver microsomes or hepatocytes.

In Meteor, we found that there is no correlation between the absolute reasoning level of predicted metabolites and their abundance in vivo, irrespective of the terminology probable, plausible, etc. Because Meteor does not consider the shape of the molecule in making predictions, Meteor overemphasizes simple oxidations. Indeed, allowing Meteor to evaluate metabolic transformations down to the "equivocal" level results in completely indiscriminate hydroxylation. Finally, N-glucuronidations at amine moieties are often found in vivo for the test set, but they are only predicted at the equivocal absolute reasoning level, which is much too low for that particular biotransformation in humans. N-acetylations are heavily overpredicted, whereas N-oxidations at morpholine and piperidine rings are underpredicted, again for this test set.

The reactivity correction provides the best results when set at on for substrate, thus without considering the reaction mechanism of the P450 enzyme. One would expect more precision of the prediction profile in MetaSite considering the P450 mechanism additionally, but, surprisingly, this is not the case. Sometimes MetaSite identifies an in silico SOM right next to what the in vivo results indicate as a SOM. The fine tuning of relative distances inside the active site of virtual P450 enzymes could possibly help to resolve this problem. In MetaSite, too, some important aliphatic/aromatic hydroxylations that were seen for the test set are not predicted.

For StarDrop, one cannot attach much value to the absolute factor in the metabolism prediction because it does not correlate with the in vivo abundance of metabolism, irrespective of the terminology labile positions, stable positions, etc. N-oxidations and some aliphatic hydroxylation, in our case, were not predicted by StarDrop.

**Conclusion.** In this study, we have evaluated and compared the metabolite prediction software packages Meteor, MetaSite, and StarDrop, in terms of precision and sensitivity. We conclude that the state-of-the-art individual metabolite prediction software has many advantageous features but still needs refinement to obtain an acceptably useful prediction profile. We found that intelligent combinations, i.e., combining packages based on different mechanistic principles, could prove useful and should be pursued to increase the prediction precision. Depending on the scope of the intended research, either sensitivity or precision should be prioritized. The answer to “what will happen” in drug metabolism should be sought in precise prediction tools; the answer to “what might happen” in drug metabolism should be sought in sensitive prediction tools.

**Authorship Contributions**

**Performed data analysis:** T’jollyn and Mannens.

**Wrote or contributed to the writing of the manuscript:** T’jollyn, Boussery, Mortishire-Smith, Coe, De Boeck, Van Boxelaer, and Mannens.

**References**


**Address correspondence to:** Geert Mannens, Senior Research Fellow, Drug Metabolism and Pharmacokinetics, Janssen Research and Development, Turnhoutsebaan 30, B-2340 Beerse, Belgium. E-mail: gmannens@its.jnj.com