DMD Fast Forward. Published on December 1, 2006 as doi:10.1124/dmd.106.013185

# CLASSIFICATION OF METABOLITES WITH KERNEL PARTIAL LEAST SQUARES (K-PLS)

Mark J. Embrechts and Sean Ekins[1]

Department of Decision Sciences and Engineering Systems, Rensselaer Polytechnic Institute, CII 5217, NY 12180, USA (M.J.E.) and GeneGo, Inc., 500 Renaissance Drive, Suite 106, St Joseph, MI 49085 and Department of Pharmaceutical Sciences, University of Maryland, 20 Penn Street, Baltimore, MD 21201, USA (S.E.)

Running title: **Metabolite Prediction**

Corresponding Author:

Sean Ekins M.Sc., Ph.D, D.Sc.

ACT LLC, 601 Runnymede Ave, Jenkintown, PA 19046

Phone 269-930-0974

Fax 215-481-0159

Email ekinssean@yahoo.com, sekins@arnoldllc.com

| | |
|---|---|
| Text pages: | 10 |
| Tables: | 1 |
| Figures: | 1 |
| References: | 16 |
| Words in Abstract: | 250 |
| Words in Introduction: | 490 |
| Words in Discussion: | 483 |

Abbreviations: K-PLS, Kernel-partial least squares; QSAR, quantitative structure activity relationship

DMD Fast Forward. Published on December 1, 2006 as DOI: 10.1124/dmd.106.013185
This article has not been copyedited and formatted. The final version may differ from this version.

## Abstract

Numerous experimental and computational approaches have been developed to predict human drug metabolism. As databases of human drug metabolism information are widely available these can be used to train computational algorithms and generate predictive approaches. In turn these may be used to assist in the identification of possible metabolites from a large number of molecules in drug discovery based on molecular structure alone. In the current study we have used a commercially available database (MetaDrug$^{TM}$) and extracted a fraction of the human drug metabolism data. This data was used along with augmented atom descriptors in a predictive machine learning model, Kernel Partial Least Squares (K-PLS). 317 molecules including parent drugs and their primary and secondary (sequential) metabolites were used to build these models corresponding to individual metabolism rules, representing the formation of discrete metabolites e.g. N-dealkylation. Each model was internally validated to assess the capability to classify other molecules that were left out. Using receiver operator curve statistics models for N-dealkylation, O-dealkylation, aromatic hydroxylation, aliphatic hydroxylation, O-glucuronidation and O-sulfation had area under the curve values from 0.75-0.84 and were able to predict between 61-79 % active molecules upon leave-one-out testing. This preliminary study indicates that K-PLS and possibly other similar machine learning methods (such as support vector machines) can be applied to predicting human drug metabolite formation in a classification manner. Improvements can be achieved using considerably larger datasets that contain more positive examples for the less frequently occurring metabolite rules as well as the external evaluation of novel molecules.

DMD Fast Forward. Published on December 1, 2006 as DOI: 10.1124/dmd.106.013185
This article has not been copyedited and formatted. The final version may differ from this version.

## Introduction

With the emphasis now on increasing the efficiency of drug discovery there is interest in using predictive computational approaches to complement in vitro and in vivo studies. In the area of metabolism prediction these techniques encompass pharmacophores (Ekins et al., 2001), quantitative structure activity relationships (QSAR) (Shen et al., 2003; Balakin et al., 2004), electronic models (Korzekwa et al., 2004), commercial drug metabolism databases (Borodina et al., 2004) as well as other methods that have been comprehensively reviewed elsewhere (de Graaf et al., 2005; Ekins et al., 2005a; de Groot, 2006). Some approaches have combined metabolite data and rules for suggesting metabolic pathways across multiple species (Erhardt, 2003). Such databases may also be useful for calculating the probability for a given metabolic reaction (Boyer and Zamora, 2002) to then indicate potential metabolites and the sites of metabolism using statistical or algorithmic approaches (Borodina et al., 2004). Although these types of comprehensive databases generally enable numerous search options to retrieve molecule structures and published information, the predictive capabilities seem limited at present (Wishart et al., 2006). A major limitation is that they are unlikely to have a complete dataset of reactions and molecular structures to extrapolate for a new molecule. In turn the user is reliant on the quality of the published in vitro or in vivo data which in many cases may predate modern analytical methods, such that older published metabolic pathways may be incomplete. In reality such database approaches provide knowledge of most published data and are perhaps limited to interpolation.

DMD Fast Forward. Published on December 1, 2006 as DOI: 10.1124/dmd.106.013185
This article has not been copyedited and formatted. The final version may differ from this version.

The combination of different approaches to drug metabolite prediction may balance the strengths and weaknesses of each approach and several commercial methods are now pursuing this direction. MetaDrug$^{TM}$ represents one such method combining a manually annotated database of human drug metabolism information including xenobiotic reactions, enzyme substrates and enzyme inhibitors with kinetic data (Ekins et al., 2005b; Ekins et al., 2006). This database has enabled the generation of rules for predicting likely metabolic reactions. The parent molecule and metabolites may then be scored through integrated QSAR models and rules for molecule reactivity before visualizing molecules as nodes on a network diagram (Ekins et al., 2005b; Ekins et al., 2006).

Such rule based metabolite predictions indicate that it is possible to generate many more metabolites than have been identified in the literature which may make the methods less useful (Ekins et al., 2006). We are therefore investigating approaches to limit the metabolites to those that are most likely. Recently a number of machine learning approaches including Support Vector Machines and Kernel-Partial Least Squares (K-PLS) (Rosipal and Trejo, 2001) have been implemented in a single software package (Analyze/Stripminer) and this was used with several benchmark datasets (Bennett and Embrechts, 2003) including protein binding and other physicochemical properties. The results with K-PLS indicated that it could be favorably applied to other datasets to enable QSAR model construction and aid drug discovery research. In the current proof of concept study we have used K-PLS to generate preliminary classification models to identify whether a metabolite is likely to be produced for a particular parent molecule.

## Materials and Methods

***Literature data.*** Three hundred and seventeen molecules were randomly extracted from the MetaDrug$^{TM}$ database (GeneGo Inc, St. Joseph, MI) (Ekins et al., 2006) and this represents a small fraction of the human drug metabolism content. These molecules were prepared as an sdf file containing data for the 65 metabolic pathways of interest (Ekins et al., 2005a) with binary data for the presence or absence of a metabolite.

***Descriptor calculation.*** ChemTree software (GoldenHelix, Bozeman, MT) running on a Pentium 4 processor was used to generate augmented atom molecular descriptors (Young et al., 2002) representing the presence or absence of a particular heavy atom with its immediately bonded neighbors. In total 61 descriptors were generated for the set of molecules.

***Data preprocessing*****.** Metabolic reactions with greater that 2 examples of the metabolite rule were then used for modeling, this narrowed down the dataset considerably. The matrix of molecular descriptors and biological activity data was then scaled (normalized) and variables with unchanging values were removed using feature selection with the StripMiner/Analyze software (software available from M.J.E. at http://www.rpi.edu/locker/82/001182/) (Embrechts et al., 2001). From the descriptors with more that 95 % correlation between each other (i.e., “cousin descriptors”), only the descriptors most correlated with the response was retained. In addition 4 sigma outliers were brought within 2.5 sigma.

***K-PLS Modeling method and testing***. The Analyze software uses the Kernel Partial Least Squares (K-PLS) method (Rosipal and Trejo, 2001) with two key parameters, the number of latent variables and the Parzen window or Gaussian kernel sigma. In this study the number of latent variables is held fixed at 5, and the Gaussian kernel sigmas are tuned using a second order Newton method where the performance criterion is the error minimization on the validation data using five-fold cross-validation. The sigmas were tuned just once, using the metabolite with the most positive instance cases. Sigma tuning on just one single metabolite is a conservative approach that prevents over-tuning. Furthermore the fact that the model still has a good predictive power on the other metabolites is another indication that over tuning did not occur in this case.

K-PLS uses kernels and can therefore be seen as a nonlinear extension of the PLS method. The commonly used radial basis function kernel or Gaussian kernel was applied, where the kernel is expressed as follows (Christianini and Shawe-Taylor, 2000):

$$K\left(\vec{x},\vec{x}_i\right)=e^{-\frac{\left\|\vec{x}-\vec{x}_i\right\|^2}{2\sigma^2}}$$

The K-PLS method can be reformulated to resemble support vector machines, but it can also be interpreted as a kernel with centering transformation of the descriptor data followed by a regular PLS method (Bennett and Embrechts, 2003).

For the predictive modeling on the other metabolites, the same sigmas were used. Sigma-tuning also allows for an identification procedure for pointing out the most relevant attributes by considering that the attributes with the larger sigma values are less relevant. After sigma tuning, the individual metabolites were predicted using K-PLS with a Gaussian kernel with multiple sigmas, using a leave-one-out procedure. Because the number of positive examples of a metabolite generally exceeded by the number of negative instances, the discrimination between positive and negative cases was made using a bias with a threshold of -0.5 for choosing the operating point on the receiver operator curve (ROC). The area under the curve (AUC) values were also calculated, with higher values approximating to better classifications. Because of the imbalance in the number of positive and negative examples, the balanced error rate was calculated taking the average of the number of correct that were positive and the number correct that were negative. In this case higher numbers are preferable.

## Results and Discussion

Tools for predicting potential metabolites of small molecule substrates in early drug discovery are important in guiding lead optimization to produce drug candidates with desirable metabolic and toxicological properties. We have recently developed and tested a computational tool that comprises a rule based method for metabolite prediction, integrated QSAR models and a database of human metabolic and signaling information (Ekins et al., 2006). In silico metabolite prediction typically generates many more potential metabolites than are actually observed. The emergence of machine learning tools combined with databases of human metabolism information represent methods for producing more reliable predictions of metabolites from an input structure alone. In the current study, for each of the over 300 molecules selected from the MetaDrug database with metabolism information, 2D molecular descriptors were calculated. Twenty three of the 65 reactions had sufficient binary data for modeling and a K-PLS model was produced for each using the Analyze/Stripminer software (Bennett and Embrechts, 2003). We evaluated the resulting classification models for predicting metabolic reactions after leave one out testing (Table 1). In general, we found that the reactions that are well populated with literature data (e.g. N-dealkylation, aromatic and aliphatic hydroxylation and O-glucuronidation) produced K-PLS models that perform well when assessed using the AUC value and the ROC plots (Table 1, Figure 1). As expected, those models that are sparsely populated with few positive instances of a metabolite being observed corresponding to a particular reaction (generally non-P450 related), are of poorer quality, indicative that no reliable classification can be made. Exceptions include N-hydroxylation and double bond peroxidation in which there are remarkably few positive examples but results are favorable for predictions, indicative that the examples provided generate useful rules based on path length descriptors.  This preliminary work with both phase I and II reactions indicates such an approach requires generally much larger databases than used here which will be available in later versions of MetaDrug$^{TM}$. Despite this, K-PLS models for N-dealkylation, O-dealkylation, aromatic hydroxylation, aliphatic hydroxylation, O-glucuronidation and O-sulfation reactions had AUC values between 0.75-0.84 and were able to predict between 61-79 % active molecules upon leave out testing while more importantly the balanced error predictions were between 70-82 %. Therefore this represents a useful method to classify the potential for an unknown molecule to undergo these particular metabolic reactions. However this approach requires further testing using considerably more data for the many sparsely populated metabolic reactions. In addition external validation of all models with a large test sets of molecules will be required alongside measures to ensure that a prediction is reliable such as those based on molecule similarity. This work represents the first occasion to our knowledge that K-PLS has been used for metabolite prediction, and the results obtained are promising with unbalanced datasets.  The integration of this K-PLS approach with rule-based and other QSAR methods could result in a more effective method for metabolite

DMD Fast Forward. Published on December 1, 2006 as DOI: 10.1124/dmd.106.013185
This article has not been copyedited and formatted. The final version may differ from this version.

prediction that would be useful in numerous drug discovery applications were reliable metabolite identification is important.

DMD Fast Forward. Published on December 1, 2006 as DOI: 10.1124/dmd.106.013185
This article has not been copyedited and formatted. The final version may differ from this version.

## References

Balakin KV, Ekins S, Bugrim A, Ivanenkov YA, Korolev D, Nikolsky Y, Skorenko SA, Ivashchenko AA, Savchuk NP and Nikolskaya T (2004) Kohonen maps for prediction of binding to human cytochrome P450 3A4. *Drug Metab Dispos* **32:**1183-1189.

Bennett KP and Embrechts MJ (2003) An optimization perspective on kernel partial least squares regression, in: *Advances in learning theory; methods, models and applications* (Suykens JAK, Horvath G, Basu S, Micchelli J and Vandewalle J eds), pp 227-250, IOS Press, Amsterdam.

Borodina Y, Rudik A, Filimonov D, Kharchevnikova N, Dmitriev A, Blinova V and Poroikov V (2004) A new statistical approach to predicting aromatic hydroxylation sites. Comparison with model-based approaches. *J Chem Inf Comput Sci* **44:**1998-2009.

Boyer S and Zamora I (2002) New methods in predictive metabolism. *J Comp-Aided Mol Des* **16:**403-413.

Christianini N and Shawe-Taylor J (2000) *Support vector machines and other kernel-based learning methods*. Cambridge University Press, Cambridge, MA.

de Graaf C, Vermeulen NP and Feenstra KA (2005) Cytochrome P450 in silico: an integrative modeling approach. *J Med Chem* **48:**2725-2755.

de Groot MJ (2006) Designing better drugs: predicting cytochrome P450 metabolism. *Drug Discov Today* **11:**601-606.

Ekins S, Andreyev S, Ryabov A, Kirillov E, Rakhmatulin EA, Sorokina S, Bugrim A and Nikolskaya T (2006) A Combined Approach to Drug Metabolism and Toxicity Assessment. *Drug Metab Dispos* **34:**495-503.

Ekins S, Andreyev S, Ryabov A, Kirilov E, Rakhmatulin EA, Bugrim A and Nikolskaya T (2005a) Computational Prediction of Human Drug Metabolism. *Exp Opin Drug Metab Toxicol* **1:**303-324.

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

Ekins S, Nikolsky Y and Nikolskaya T (2005b) Techniques: Application of Systems Biology to Absorption, Distribution, Metabolism, Excretion, and Toxicity. *Trends Pharmacol Sci* **26:**202-209.

Embrechts M, Arciniegas F, Ozdemir M and Momma M (2001) Scientific data mining with StripMiner, in: *Mountain workshop on soft computing in industrial applications*, Virginia Tech, Blacksburg, Virginia.

Erhardt PW (2003) A human drug metabolism database: potential roles in the quantitative predictions of drug metabolism and metabolism-related drug-drug interactions. *Current Drug Metabolism* **4:**411-422.

Korzekwa K, Ewing TJ, Kocher JP and Carlson TJ (2004) Models for cytochrome P450-mediated metabolism, in: *Pharmaceutical Profiling in Drug Discovery for Lead Selection* (Borchardt RT, Kerns.E.H., Lipinski CA, Thakker DR and Wang B eds), pp 69-80, AAPS Press.

Rosipal R and Trejo LJ (2001) Kernel Partial Least Squares regression in reproducing Kernel Hilbert Space. *J Machine Learning Research* **2:**97-123.

Shen M, Xiao Y, Golbraikh A, Gombar VK and Tropsha A (2003) Development and validation of k-nearest neighbour QSPR models of metabolic stability of drug candidates. *J Med Chem* **46:**3013-3020.

Wishart DS, Knox C, Guo AC, Shrivastava S, Hassanali M, Stothard P, Chang Z and Woolsey J (2006) DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res* **34:**D668-672.

Young SS, Gombar VK, Emptage MR, Cariello NF and Lambert C (2002) Mixture deconvolution and analysis of Ames mutagenicity data. *Chemo Intell Lab Sys* **60:**5-11.

DMD Fast Forward. Published on December 1, 2006 as DOI: 10.1124/dmd.106.013185
This article has not been copyedited and formatted. The final version may differ from this version.

## Footnotes Page

a) Unnumbered Footnote: The development of MetaDrug was supported by a National Institutes of Health Grant 1-R43-GM069124-01 and 2-R44-GM069124-02 “In silico Assessment of Drug Metabolism and Toxicity”.

b) Send reprint requests to: Sean Ekins, ACT LLC, 601 Runnymede Avenue, Jenkintown, PA 19046. Email ekinssean@yahoo.com

c) Numbered footnotes: [1] Current Address: ACT LLC, 601 Runnymede Ave, Jenkintown, PA 19046.

d) Competing Financial Interest: MetaDrug$^{TM}$ is a proprietary tool developed and licensed by GeneGo, Inc.

DMD Fast Forward. Published on December 1, 2006 as DOI: 10.1124/dmd.106.013185
This article has not been copyedited and formatted. The final version may differ from this version.

## Figure legend

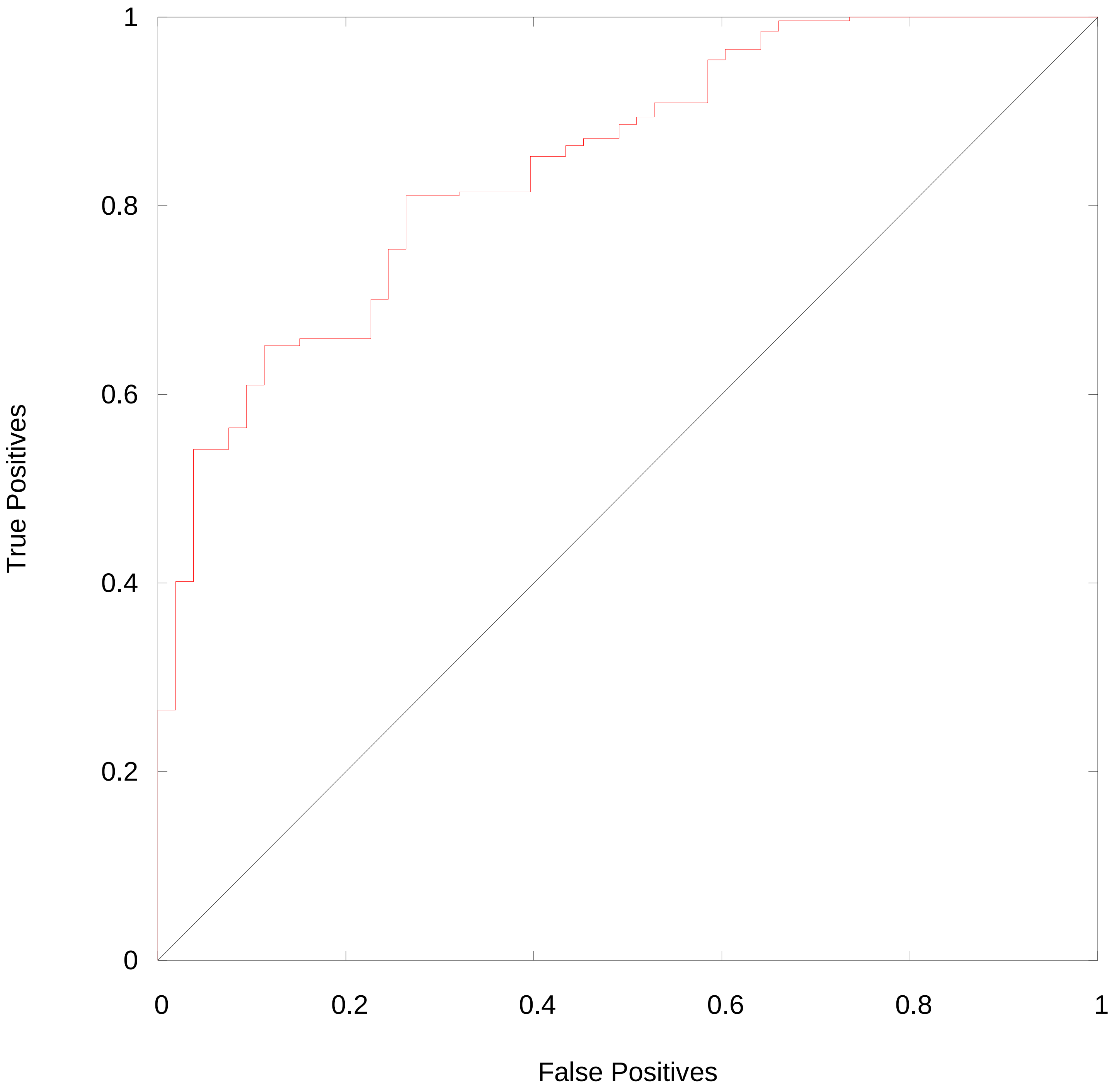
**Figure 1.** Representative Receiver Operator Curves to demonstrate the leave one out validation of K-PLS classification model for N-dealkylation (red line). Diagonal = random rate.

**Table 1**. Results of applying kernel partial least squares (K-PLS) models to human drug metabolism data for different reactions using 317 molecules. Percent correct represents the prediction for positive instances for a metabolite. Balanced error represents the average of the correct positive and correct negative predictions.

| | | | | **Actual** | |
|---|---|---|---|---|---|
| | | **Balanced** | **Number predicted** | **Number** | |
| **Metabolite type** | **AUC** | **error** | **active** | **active** | **Percent correct** |
| N-dealkylation | 0.843 | 74.49 | 36 | 53 | 68 |
| O-dealkylation | 0.815 | 74.82 | 17 | 28 | 61 |
| Aromatic hydroxylation | 0.756 | 65.9 | 70 | 95 | 74 |
| Aliphatic hydroxylation | 0.779 | 70.11 | 64 | 86 | 74 |
| Double bond peroxidation | 0.935 | 81.39 | 6 | 9 | 67 |
| Hydroxyl-carbonyl oxidation | 0.67 | 60.42 | 13 | 35 | 37 |
| Double bond formation (desaturation) | 0.776 | 73.8 | 9 | 17 | 53 |
| Aldehyde oxidation | 0.839 | 72.16 | 9 | 18 | 50 |
| Double bond epoxidation | 0.818 | 86.44 | 2 | 16 | 12 |

| | | | | | |
|---|---|---|---|---|---|
| N-oxide formation | 0.756 | 58.85 | 6 | 24 | 25 |
| N-hydroxylation | 0.928 | 82.21 | 4 | 6 | 67 |
| Carboxyl reduction | 0.878 | 49.52 | 0 | 3 | 0 |
| Ester hydrolysis | 0.849 | 49.52 | 0 | 3 | 0 |
| Epoxide hydrolysis | 0.841 | 69.85 | 5 | 12 | 42 |
| N-glucuronide transfer | 0.683 | 67.59 | 7 | 17 | 41 |
| O-glucuronide transfer | 0.773 | 70.71 | 89 | 112 | 79 |
| O-sulfate transfer | 0.848 | 82.17 | 15 | 21 | 71 |
| Glutathione S-transfer to benzyl | 0.8 | 49.52 | 0 | 3 | 0 |
| O-methyl transfer | 0.783 | 71.58 | 5 | 11 | 45 |
| N-acetyl transfer | 0.2675 | 49.36 | 0 | 3 | 0 |
| Sulfoxide oxidation | 0.6369 | 49.52 | 0 | 3 | 0 |
| Carbonyl reduction | 0.566 | 65.86 | 2 | 6 | 34 |
| Unsaturated bond hydration | 0.647 | 49.2 | 0 | 4 | 0 |
| Sulfide oxidation | 0.723 | 49.2 | 0 | 4 | 0 |

Figure 1


True Positives
False Positives