In Silico Metabolite Prediction Using Artificial Neural Network Ensembles

Marvin Waldman, Robert Fraczkiewicz, David Miller, Jinhua Zhang, and Robert D. Clark
Simulations Plus, Inc., Lancaster, CA 93534, USA (www.simulations-plus.com)

INTRODUCTION

Drug metabolism plays a crucial role in understanding bioavailability and drug-drug interactions, as well as in the design of prodrugs and in avoiding undesirable toxic metabolites. Cytochrome P450s (CYPs) are the major class of enzymes responsible for metabolism of most drugs. We have employed our state-of-the-art Artificial Neural Network Ensemble (ANNE) modeling methodology to develop *in silico* models for classifying drugs as substrates of nine CYP isoforms, 1A2, 2A6, 2B6, 2C8, 2C9, 2C19, 2D6, 2E1, and 3A4. Our models also predict CYP-specific likely sites of metabolic oxidation and the resulting metabolites.

The models make use of a new method that rapidly calculates atomic descriptors representing charge, reactivity, steric effects, and local atomic environment from 2D molecular representations. Direct calculation of the properties represented by these descriptors originally required extremely time-consuming density functional theory (DFT) molecular orbital calculations. Our method for approximate rapid calculation of charge and reactivity descriptors avoids the need for these time-consuming calculations, enabling over 130 property predictions per molecule to be made virtually instantaneously (dozens of compounds per second) on a modern laptop computer.

Our metabolism and metabolic site models have been trained on datasets larger than any previously reported. When combined with our approach for rapid computation of atomic descriptors (not chemical fragments), the resulting models are broadly applicable across a very wide range of chemistries while providing performance on par with or superior to previously reported literature models. We demonstrate this by showing performance results employing multiple metrics for both training and test sets. We also show examples of apparently incorrect predictions which were shown to be correct, in fact, by subsequent experiments, demonstrating the utility of the models as well as demonstrating the avoidance of overtraining by our protocol.

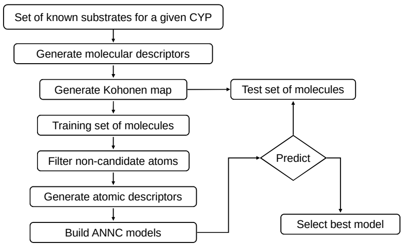
METHODS

Dataset Preparation:

- Obtain data from the following sources
  - Accelrys Metabolite Database
  - Literature datasets:
    - Sheridan et al., J Med Chem **50** 3173 (2007)
    - Rendic, Drug Metab Rev **34** 83 (2002)
  - Original literature sources, old and new
- Classify atoms of molecules as metabolized/not metabolized based on observed metabolites
- Generate atomic descriptors for each atom
- Build Artificial Neural Network Ensembles (ANNEs) to predict sites of metabolism

Flowchart of model building protocol:

- Note: The external test set was not used in any way during model training.



Set of known substrates for a given CYP

Generate molecular descriptors

Generate Kohonen map

Test set of molecules

Training set of molecules

Filter non-candidate atoms

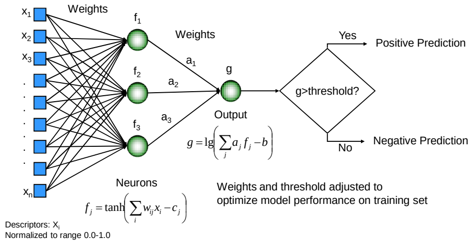
Predict

Generate atomic descriptors

Build ANNC models

Select best model

Classification Neural Network Architecture:



$$g = \lg\left(\sum_j a_j f_j - b\right)$$

$$f_j = \tanh\left(\sum_i w_{ij} x_i - c_j\right)$$

Weights and threshold adjusted to optimize model performance on training set

Descriptors: X_i
Normalized to range 0.0-1.0

PERFORMANCE

P450 Substrate Model	Train/Verify Set					Test Set				
	Total	Correct	Concord.	Sens.	Spec.	Total	Correct	Concord.	Sens.	Spec.
1A2	1074	861	80.0%	81.0%	80.0%	268	220	82.0%	84.0%	81.0%
2A6	532	413	78.0%	80.0%	77.0%	133	107	80.0%	77.0%	82.0%
2B6	566	447	79.0%	80.0%	79.0%	142	111	78.0%	77.0%	79.0%
2C8	558	428	77.0%	72.0%	79.0%	139	101	73.0%	74.0%	72.0%
2C9	1038	805	78.0%	78.0%	78.0%	260	185	71.0%	73.0%	70.0%
2C19	1011	773	77.0%	81.0%	75.0%	253	187	74.0%	73.0%	74.0%
2D6	1101	883	80.0%	80.0%	80.0%	275	232	84.0%	87.0%	82.0%
2E1	581	496	85.0%	85.0%	85.0%	145	123	85.0%	84.0%	85.0%
3A4	1243	1016	82.0%	82.0%	82.0%	311	251	81.0%	81.0%	81.0%

Performance of the CYP substrate classification models. The “Total” columns refer to the number of molecules in the training pool and test sets.

CYP Site Model	Train/Verify Set					Test Set				
	Excluded Atoms	Included Atoms	Concord.	Sens.	Spec.	Excluded Atoms	Included Atoms	Concord.	Sens.	Spec.
1A2	2448	3307	86.4%	86.9%	86.3%	299	525	86.5%	81.7%	87.4%
2A6	657	1086	86.8%	80.1%	88.0%	90	129	92.2%	96.8%	90.8%
2B6	1017	1725	90.4%	86.2%	91.0%	212	389	87.9%	82.9%	89.1%
2C8	1280	2008	88.0%	85.3%	88.4%	263	384	88.0%	80.8%	89.2%
2C9	2029	2985	89.9%	80.3%	91.2%	376	512	89.5%	76.1%	91.6%
2C19	1679	2720	91.0%	89.7%	91.2%	275	441	88.2%	81.3%	89.4%
2D6	2202	4118	91.0%	92.7%	90.7%	327	745	89.5%	83.7%	90.5%
2E1	755	1126	88.9%	87.0%	89.4%	161	229	87.8%	83.9%	89.0%
3A4	6635	9729	89.3%	82.0%	90.3%	1527	2519	88.0%	74.0%	89.7%

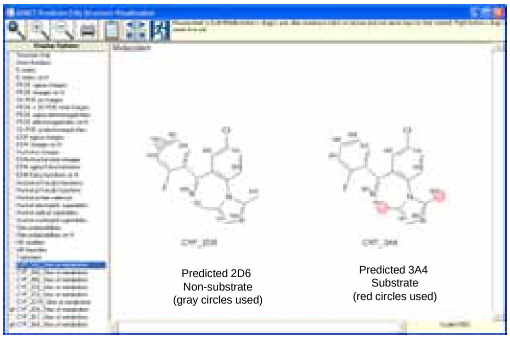
Atom-based performance of the CYP site models. The results are based on comparing predictions for the Included Atoms with the reported sites of metabolism for each CYP isoform. Excluded atoms were filtered out of the datasets based on simple chemical rules (e.g., being an oxygen) that preclude them from being CYP sites of metabolism.

CYP Site Model	Train/Verify Set				Test Set			
	No. Molecules	Top 1	Top 2	Top 3	No. Molecules	Top 1	Top 2	Top 3
1A2	281	78.3%	91.5%	96.1%	44	72.7%	93.2%	93.2%
2A6	99	65.7%	84.8%	90.9%	17	88.2%	100.0%	100.0%
2B6	138	76.8%	92.0%	97.8%	37	73.0%	86.5%	91.9%
2C8	139	74.8%	87.1%	92.8%	24	66.7%	83.3%	87.5%
2C9	224	73.7%	89.7%	95.5%	43	81.4%	88.4%	93.0%
2C19	210	83.3%	96.7%	99.0%	38	73.7%	87%	89.5%
2D6	290	78.3%	92.8%	97.6%	58	87.9%	96.6%	98.3%
2E1	115	77.4%	93.9%	97.4%	26	80.8%	92.3%	100.0%
3A4	604	74.8%	87.6%	94.5%	138	77.5%	87.7%	92%

Molecule-based performance of the CYP site models. Shown are the percent of molecules with a reported site of metabolism correctly identified among the Top 1, 2, or 3 scoring atoms in the training/test sets.

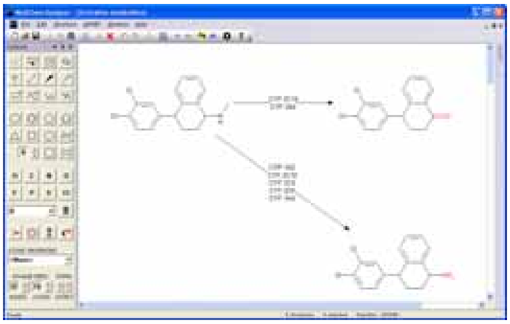
IMPLEMENTATION

ADMET Predictor™



Candidate metabolic sites are shown with propensity scores ranging from 0-1000. High scoring atoms are shown with hashed red circles and are predicted sites of metabolism. Site predictions (in gray) are also shown for molecules predicted to be non-substrates in case they may be known experimentally to actually be substrates.

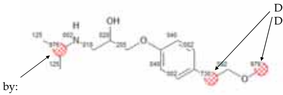
MedChem Designer™



Predicted sites are converted to metabolite predictions in MedChem Designer using SMIRKS-based rules, which may be customized by the user.

EXAMPLES

CYP2D6 Site Predictions for Metoprolol

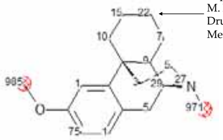


Reported by:
D S Mautz et al.,
Drug Metab Dispos **23** 513 (1995)

Reported by:
G.P. Hayhurst et al.,
Biochem. J. **355** 373 (2001)

Our preliminary model’s training set for metoprolol did not include the N-dealkylation site of the isopropyl group reported by Hayhurst et al. Nevertheless, this site was predicted by the model and subsequently confirmed upon discovery of the Hayhurst article. Shown above are the current model predictions.

CYP2D6 Site Predictions for Dextromethorphan

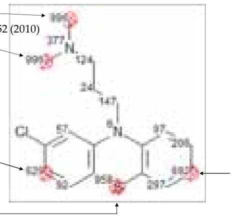


Reported by:
M. Matsunaga et al.,
Drug Metab Dispos **37** 699 (2009)
Metabolized by CYP2D6.49

The O-methyl and N-methyl sites have been reported by multiple researchers. The preliminary model’s training set also included the site marked with the arrow as reported by Matsunaga et al., but the model assigned this a very low score. Referring to the original article revealed this site was oxidized by a mutant form of CYP2D6, not the native form. Shown above are the current model predictions.

CYP1A2 Site Predictions for Chlorpromazine

Reported by:
J. Wojcikowski et al.,
Biochem. Pharmacol. **80** 1252 (2010)



Weak False positive?

Reported by:
K. Yoshii et al.,
Life Sciences **67** 175 (2000)

The initial model was built using only the site reported by Yoshii et al. in 2000. Nevertheless, this model predicted additional sites on the sulfur, N-methyl groups, and ring carbon. Subsequently, a publication appeared in 2010 confirming the sulfur and N-methyl groups as sites. Thus, “apparent” false positives became true positives. The additional predicted ring carbon site is as yet unreported, but may someday be confirmed(!)

DISCUSSION

We have built CYP site models of metabolism for 9 CYP isoforms in which the model predictions in the form of propensity scores may be displayed on all candidate sites using our ADMET Predictor and MedChem Designer programs. Additionally, MedChem Designer supports the display of the actual metabolites resulting from any of the CYPs using SMIRKS-based rules. The models are among the most accurate reported in the literature as well as covering more CYPs than any so far reported in the literature.

Using our ANNE technology and rapid and accurate atomic charge and reactivity descriptors, model predictions are extremely rapid (typically dozens of molecules per second in addition to ~100 other properties), and the models have proven robust and resistant to overtraining, as demonstrated in the above examples. In each of the above cases, the initial model made predictions that disagreed with the initial training set data, but was subsequently found to be correct based either on subsequent literature, discovery of literature sources not initially identified, or closer examination of the reported literature.

In the case of metoprolol, we found that the predicted N-dealkylation site (flagged as a non-site in the initial training set) was confirmed in a previously unidentified reference not part of our initial data sources. In the case of dextromethorphan, our data sources identified the ring carbon site as being mediated by 2D6. Upon closer examination of the original source literature, we found this was a 2D6 mutant, and not native 2D6. Thus, an “apparent” false negative became a true negative. Finally, in the case of chlorpromazine, several “apparent” false positives became true positives when the 2010 publication by Wojcikowski et al. appeared.