Bioinformatics, 31(12), 2015, 1966–1973
doi: 10.1093/bioinformatics/btv100
Advance Access Publication Date: 19 February 2015
Original Paper

OXFORD

Structural bioinformatics

Extending P450 site-of-metabolism models with region-resolution data

Jed M. Zaretzki, Michael R. Browning, Tyler B. Hughes and S. Joshua Swamidass*

Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, MO 63130, USA

*To whom correspondence should be addressed.

Associate Editor: Janet Kelso

Received on March 3, 2014; revised on February 10, 2015; accepted on February 11, 2015

Abstract

Motivation: Cytochrome P450s are a family of enzymes responsible for the metabolism of approximately 90% of FDA-approved drugs. Medicinal chemists often want to know which atoms of a molecule—its metabolized sites—are oxidized by Cytochrome P450s in order to modify their metabolism. Consequently, there are several methods that use literature-derived, atom-resolution data to train models that can predict a molecule's sites of metabolism. There is, however, much more data available at a lower resolution, where the exact site of metabolism is not known, but the region of the molecule that is oxidized is known. Until now, no site-of-metabolism models made use of region-resolution data.

Results: Here, we describe XenoSite-Region, the first reported method for training site-of-metabolism models with region-resolution data. Our approach uses the Expectation Maximization algorithm to train a site-of-metabolism model. Region-resolution metabolism data was simulated from a large site-of-metabolism dataset, containing 2000 molecules with 3400 metabolized and 30 000 un-metabolized sites and covering nine Cytochrome P450 isozymes. When training on the same molecules (but with only region-level information), we find that this approach yields models almost as accurate as models trained with atom-resolution data. Moreover, we find that atom-resolution trained models are more accurate when also trained with region-resolution data from additional molecules. Our approach, therefore, opens up a way to extend the applicable domain of site-of-metabolism models into larger regions of chemical space. This meets a critical need in drug development by tapping into underutilized data commonly available in most large drug companies.

Availability and implementation: The algorithm, data and a web server are available at http://swami.wustl.edu/xregion.

Contact: swamidass@wustl.edu

1 Introduction

Cytochrome P450 (abbreviated as CYP or P450) enzymes are a family of proteins responsible for the metabolism of ~90% of FDA-approved drugs (Guengerich, 2006; Nebert and Russell, 2002) The P450-mediated metabolism of a drug affects its clinical efficacy and safety. Knowing which atoms of a molecule—its sites of metabolism—are oxidized by P450s enables medicinal chemists to rationally design new molecules with improved metabolic profiles.

Determining which exact atoms of an individual molecule are metabolized is often difficult and expensive. The mass spectrometry experiments typically used in drug discovery most often can only identify *regions* of a molecule that are metabolized, rather than identify the specific metabolized atoms that medicinal chemists need to know. Analyzing the mass spectra of a P450-substrate reaction to determine metabolic products is a complicated problem, the

complete description of which lies outside the scope of this article (Castro-Perez, 2007; Scheubert *et al.*, 2013; Siegel *et al.*, 2013; Xiao *et al.*, 2012;). The key point is that although there are many ways of interpreting mass spectrometry data, none can reliably identify atom-resolution sites of metabolism for all molecules (Gerlich and Neumann, 2013; Heinonen *et al.*, 2012; Kerber *et al.*, 2001; Rasche *et al.*, 2012; Stein, 1995; Wolf *et al.*, 2010).

For these two reasons—the value of knowing sites of metabolism and the difficulty of determining them at atom-resolution—substantial effort has been invested in building computational models that can predict a molecule's sites of metabolism (Huang *et al.*, 2013; Kirchmair *et al.*, 2012; Rudik *et al.*, 2014; Zaretzki *et al.*, 2013). There are several methods, but the best performing methods all use machine learning to train models from literature-curated databases of hundreds of molecules with known atom-resolution sites of metabolism (Zaretzki *et al.*, 2012, 2013).

The reliance of these models on literature-derived, atom-resolution data has two important consequences. First, the molecules used to train site of metabolism models are, for the most part, limited to only those available in the literature. Literature-derived data are based, primarily, on molecules that are structurally different from proprietary-lead molecules. Consequently, models trained on literature data are often not as accurate on molecules in drug development, the exact molecules for which accurate predictions are most important (Dapkunas *et al.*, 2009). Second, these models do not make use of the large number of molecules with known, region-resolution sites of metabolism, which can include data from thousands of molecules more similar to the molecules currently in development.

To overcome these shortfalls, we introduce XenoSite-Region, the first reported method that uses region-resolution data to train atom-resolution site of metabolism prediction models. Region-resolution data are much easier to experimentally obtain than atom-resolution data, and it is already collected in drug discovery on a regular basis. For example, Pfizer collects ADME data, including region-resolution sites of metabolism, on 2000 molecules every week, more than triple the total amount of atom-resolution data in literature-derived datasets (Hop *et al.*, 2008). By including this region-resolution data, site-of-metabolism models could extend into areas of chemical space that are poorly represented in the public domain.

Our strategy is to modify an existing site-of-metabolism prediction method, XenoSite (Zaretzki *et al.*, 2013), to train from region-resolution data. XenoSite uses a neural network to predict which atoms in the molecule are metabolized, and is a suitable starting point for our approach for two reasons. First, XenoSite is accurate, picking out a correct site of metabolism in the top two predictions for every molecule with an accuracy of 87% on literature-derived data. This accuracy surpasses those of other site-of-metabolism models like RS-Predictor (Zaretzki *et al.*, 2011), SMARTCyp (Rydberg *et al.*, 2010), StarDrop (Optibrium Ltd., 2009) and Schrodinger (Schrödinger LLC, 2011)—with performances of 84.3, 82.1, 75.8 and 68.2%, respectively—on the same dataset (Zaretzki *et al.*, 2012).

Second, Xenosite can both train from probabilistically labeled training data and also outputs a score that is a well-scaled probability. In addition to using binary 1/0 labels on each atom, Xenosite can also train from data where the probability is used as a training target. Similarly, XenoSite's outputs range from 0 to 1 and correspond closely to the probability that an atom is a site of metabolism. Both these capabilities are necessary for the approach we will use to adapt XenoSite to use region-level data.

We hypothesize that a statistical method called Expectation Maximization (EM) (Dempster *et al.*, 1977) can train XenoSite models with region-resolution data. XenoSite, like all reported machine learning methods in the literature, requires a training set with every atom labeled as a site or not-a-site-of-metabolism. For region-resolution data, however, these labels are hidden from us. EM is designed to work in just these situations, enabling statistical models to be trained with key variables hidden from view. The EM algorithm is iterative, using the output of the model to successively improve an estimate of the true labels for the training data, which is, in turn, used to improve the output of the model.

2 Materials and methods

In this section, we start by describing the atom- and region-resolution data used in this study. Next, we specify the model used to predict sites of metabolism, describing how it is trained from atom-resolution data. Finally, we explain how this model is extended with the EM algorithm to train from region-resolution data.

2.1 Atom-resolution data

For this study, we use the largest publicly available repository of P450 substrates, which contains 680 molecules distributed across nine isozymes (Table 1) (Zaretzki *et al.*, 2012). Each CYP substrate is a molecule with multiple sites, at least one of which is a site of metabolism (a positive) and the rest are non-metabolized sites. Sites of metabolism are metabolized CYP enzymes, and the rest are not (Table 1). These data have two important nuances that must be understood (Fig. 1). First, some sites of metabolism actually correspond to more than one atom in cases where two atoms are involved in the CYP reaction. Second, atoms in molecules with local or global symmetries are topologically equivalent to one another. These groupings are automatically calculated for all molecules, and are available in the supporting information.

2.1.1 Multi-atom sites

Even when the exact site of metabolism is known, this site sometimes maps to more than one atom in a molecule. Using a strategy described in prior work, we group atoms representing the same metabolic reaction together into the same multi-atom site (Korolev *et al.*, 2003; Zaretzki *et al.*, 2011). In this study, any halogen or oxygen bound to a single atom is grouped with the atom to which it is bound, forming a multi-atom site.

2.1.2 Topologically-equivalent atoms

It is also necessary to track atoms that are topologically equivalent. These atoms are modeled separately during training, so each one

Table 1. This study predicts the sites of metabolism of substrates of nine P450 isozyme, which are composed of sites of metabolism (SOMs) and un-metabolized sites (UMSs)

Isozyme	Number of substrates	Number of SOMs	Number of UMSs
1A2	271	526	3812
2A6	105	160	1127
2B6	151	232	1962
2C19	218	338	3300
2C8	142	248	2193
2C9	226	391	3406
2D6	270	436	4144
2E1	145	240	1526
3A4	475	890	8677

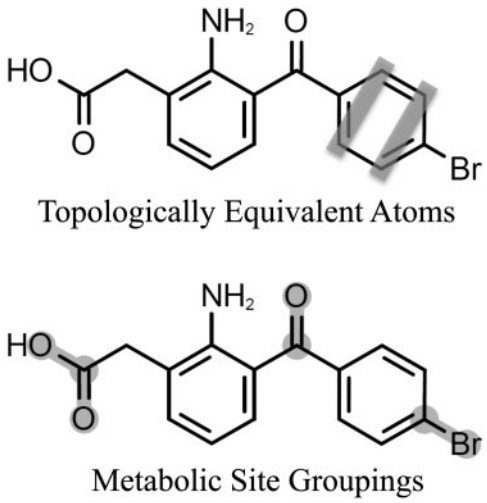

Fig. 1. Topological equivalence and multi-atom sites. This sample molecule (bromfenac) illustrates both multi-atom sites and topologically equivalent atoms, two types of atom groupings that are important during training and assessment of site-of-metabolism models. (Top) Topologically equivalent atoms (grouped by gray boxes) arise from either global or local symmetries in the molecule. They are treated as multiple sites during training, but are grouped together when evaluating predictions. (Bottom) Multi-atom sites (grouped by gray shading) are formed by grouping halogens and singly-bound oxygens with the heavy atom to which they are bound

yields its own distinct prediction in the model. During assessment, however, only the topologically equivalent site with the highest prediction is kept, with the others being discarded. This prevents topologically equivalent sites from being double counted.

2.2 Region-resolution data

Pharmaceutical companies generate large amounts of region-resolution metabolism data (Hop *et al.*, 2008), which is entirely unused for modeling because there does not exist methods of using it. In public databases, there are hundreds of molecules with atom-resolution data, but in these private datasets there are thousands of molecules with region-resolution data. Unfortunately, these large databases of region-resolution data are not publicly available. Instead, we used a public database of atom-resolution metabolism data to simulate region-resolution data.

Region-resolution data are simulated from the atom-resolution site of metabolism data for use in this study by using a graph partition algorithm (Karypis and Kumar, 1999) to splis each molecule in the training sets into regions of approximately three, five, seven or nine atoms (Fig. 2). In practice, regions of size five are most similar to those observed in real region-resolution data from mass spectrometry experiments, and the range of sizes enables us to assess the sensitivity of our approach to region size. Each of these regions is labeled with either the exact number of observed sites of metabolism (for the 'exact' region data) or a binary label indicating if there is one or more observed sites of metabolism (for the 'inexact' region data). These labels give rise to region constraints that are used during the training.

2.3 Site-of-metabolism model

XenoSite constructs models using a neural network with five hidden nodes. These models take as input a vector of numerical descriptors for each heavy atom in a molecule, and output a score for each atom.

2.3.1 Atom descriptors

In this study, we use the five classes of descriptors to numerically characterize each atom. First, we use topological descriptors that encode features about the local graph structure of the molecule in the

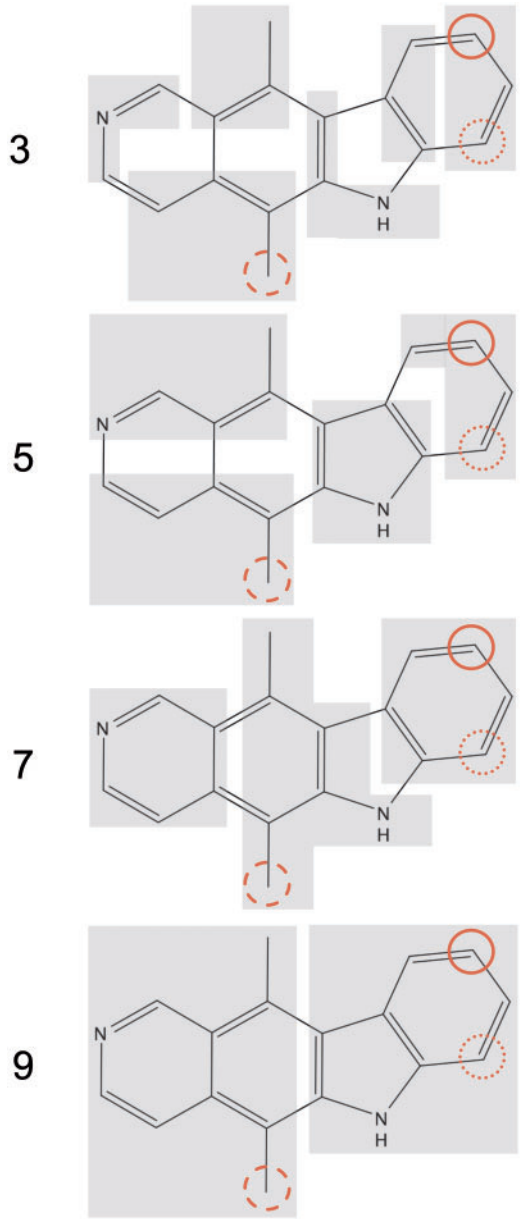

Fig. 2. Regions for an example molecule. Regions of approximately three, five, seven and nine atoms are marked on an example molecule, ellipticine, with grey polygons. For reference, the primary, secondary and tertiary atomic sites of P450 metabolism are marked, respectively, with solid, dashed and dotted circles. Qualitatively, regions of size five appear to be most similar to those obtained from mass spectrometry data

atom's neighborhood. Second, we use quantum descriptors that quantify the electronic structure near the atom as computed by MOPAC. Third, we include the output of SMARTCyp as an input descriptor, which quantifies each atom's reactivity with a heme iron in a quantum chemical simulation. Fourth, we use atom fingerprints to include information about other atoms in the training set that are in a similar environment. Fifth and finally, we include descriptors of the molecule as a whole. The exact specification of these descriptors is detailed in Zaretzki *et al.* (2013).

2.3.2 Training target

To train the model, we need a target for each atom. For atom-resolution data, this target is just a one or zero indicating if an atom

is a member of an observed site of metabolism or not. We can also use a probability (a real number ranging from 0 to 1) in place of a binary target.

2.3.3 Training algorithm

The neural nework's weights are calibrated with gradient descent on the cross-entropy error (Baldi and Brunak, 2001), For each training run, three random restarts were performed, and the weights with the best training set accuracy were kept. Models created with this protocol produce output scores between 0 and 1 that can be interpreted as probabilities. Probabilistically interpretable output is a key feature of this approach, as it enables us to apply the EM algorithm.

2.3.4 Multi-atom sites

The model directly assigns a prediction for every atom in the molecule. Some sites, however, include multiple atoms. We define a site prediction y_s as the probability that at least one of the associated atoms are metabolized,

$$y_s = 1 - \prod_{a \in s}[1 - y_a], \tag{1}$$

where the product is over the predictions of all the atoms y_a in the site.

2.3.5 Region prediction

For each region, we can compute the predicted number of sites of metabolism within a region as

$$y_r = \sum_{s \in r} y_s, \tag{2}$$

where the summation is over the sites within the region.

2.4 The EM algorithm

We hypothesize that it is possible to use EM to learn P450 models from region-level data. EM is a statistical method that allows the parameters of a model to be fit, even when the outputs of the model cannot directly be observed in the training data (i.e. when they are hidden) (Dempster *et al.*, 1977). Biologists may be familiar with EM because it is a technique frequently used to train Hidden Markov Models for sequence alignment and to fit multimodal distributions (Lawrence and Reilly, 1990; Redner and Walker, 1984).

In our case, the hidden variables are the elements of a vector of numbers, associated with every site in the training data. Each variable should be a 1 for sites of metabolism and 0 otherwise, but we do not know what the correct value is, they are hidden. First, this vector is initialized to reasonable starting values. Next, during the Maximization step (M-step) we train a neural network to create a mapping between each atom's descriptors and the current estimate of the hidden variables. Then, during the Expectation step (E-step), we re-estimate our hidden variables using the trained model and the region constraints that specify how many sites of metabolism we expect in each region. Finally, we alternate between the E- and M-steps until convergence.

2.4.1 Initialization and progression of the algorithm

The EM algorithm alternates between E- and M-steps until convergence. To initialize the algorithm, we pick a random output vector $Y = (y_i)$ by sampling from a uniform distribution ranging from 0 to 1. Applying the E-step to this vector gives an initial guess for our hidden variables, $K = (k_i)$, which is consistent with the training data (Equation 3). Here, by consistent, we mean that the sum of k_is associated with each region is within the range of known sites of metabolism for the region. Each element of K is our current estimate of the exact atom-resolution sites of metabolism. Next, the M-step fits a neural network using K as targets (Equation 4),

$$K_1 \leftarrow \text{E-step} \quad (Y_{\text{init}}) \tag{3}$$

$$W_1 \text{ and } Y_1 \leftarrow \text{M-step} \quad (K_1, D), \tag{4}$$

where W_1 is the tuned weights of the neural network and Y_1 is the output of the model using these weights with the data. This output vector is used in the M-step to compute the next guess for the hidden variables K. The next iteration repeats the E- and M-steps,

$$K_2 \leftarrow \text{E-step} \quad (Y_1) \tag{5}$$

$$W_2 \text{ and } Y_2 \leftarrow \text{M-step} \quad (K_2, D). \tag{6}$$

Subsequent iterations repeat these steps until convergence. Convergence is most easily measured by waiting for the K vector to stabilize, which usually takes no more than 10 iterations.

As the algorithm progresses, both the K and Y vectors should converge to the true, atom-level sites of metabolism. They should label the metabolized atoms with high probabilities, and the non-metabolized atoms with low probabilities. This behavior is a key feature we hope to observe in our empirical studies, which would indicate atom-level sites of metabolism can be recovered from region-level training data.

2.4.2 Expectation step

The E-step computes the expected values of the hidden variables K from the outputs Y conditioned on the region constraints associated with the training data. The expectation of the hidden variables is the probability-weighted average of all binary realizations of binomial distributions parameterized by Y that assign the right number of metabolized sites to each region.

Conceptually, the expectation is computed by, first, enumerating all binary realizations of Y, each denoted as a vector of boldface variables $\mathbf{Y} = (\mathbf{y}_i)$, from which binary site and integer region $\mathbf{y}$ variables are computed with Equations (1) and (2), respectively. Second, vectors that do not have the right number of sites in each region are rejected. Third, the remaining vectors are scored by their probability according to Y, and, finally, a probability-weighed average of the binary vectors is computed. This average vector is the expectation, and is assigned to K. While conceptually clear, computing K in this way is very slow because there are exponentially many realizations of Y that must be enumerated.

Fortunately, the expectation is computable in polynomial time. Here, we treat the sites of metabolism labels as binary random variables following binomial distributions parameterized by Y. For each region, the expectation of these variables is assigned to elements of K. For sites in an individual region r known to have between m_r and n_r sites of metabolism, this update can be derived from Baye's Rule,

$$k_s = E[\mathbf{y}_s \mid m_r \le \mathbf{y}_r \le n_r] \tag{7}$$

$$= P(\mathbf{y}_s \text{ is } 1)\frac{P(m_r - 1 \le \mathbf{y}_r - \mathbf{y}_s \le n_r - 1)}{P(m_r \le \mathbf{y}_r \le n_r)}, \tag{8}$$

where $\mathbf{y}_r$ is the integer sum of the binary labels associated with the sites in region r which contains site s, $\mathbf{y}_s$ is the binary label associated with site s, $P(\mathbf{y}_s \text{ is } 1)$ is probability according to Y that the site s is metabolized (or y_s, using Equation 1 as required), the numerator is the probability that the number of sites of metabolism in the region,

ignoring site s, ranges from $m_r - 1$ to $n_r - 1$, and the denominator is the probability that the number of metabolized sites in the region will range from m_r to n_r. For the exact data, m_r and n_r are equal to the known number of sites of metabolism in the region. For the inexact data, m_r is one and n_r is infinity for regions with at least one known site of metabolism, otherwise both m_r and n_r are 0.

The summation distributions necessary to compute the probabilities in the formula are constructed by convolving Y parameterized binomial distributions associated with the region's sites. This requires only a few lines of code to implement, and is best understood by studying the Python implementation included in the supporting information.

Multi-atom sites. For multi-atom sites, the expectation of individual atoms is conditioned on the expectation of its associated site

$$k_a = E[\mathbf{y}_a \mid E[\mathbf{y}_s] = k_s] \tag{9}$$

$$= k_s \frac{P(\mathbf{y}_a \text{ is } 1)}{P(1 \leq \mathbf{y}_s)}, \tag{10}$$

where k_s is the expectation of the site from Equation (7), $P(\mathbf{y}_a$ is $1)$ is y_a (the output of the model on atom a) and the denominator is the probability that at least one atom in the site is metabolized. Once again, the summation distribution necessary to compute the denominator is constructed by convolving binomial distributions corresponding to the summation terms.

Topologically equivalent atoms. The expectation step estimates the atom-resolution sites of metabolism, assigning a k_a to all atoms in the molecule, for use as a target during training. Topologically equivalent atoms may be assigned different targets, which is not consistent. Consequently, as a final step, topologically equivalent atoms are assigned the maximum k_a of all the k_a's associated with them.

2.4.3 Maximization step

The M-step finds a maximum *a posteriori* estimate of the neural network's parameters using gradient descent on the negative log likelihood of the data according the model,

$$\underset{W}{\text{minimize}} \quad H(K, Y(W, D)) + W^T W. \tag{11}$$

The network's parameters (an array W) are adjusted so the outputs of the network (an array Y), using the descriptor data (a matrix D) as an input, matches our current guess for the sites of metabolism (an array K). The final term is a weight decay term that keeps the network's parameters from becoming too large. H is the sum of the cross-entropy of all elements of K and Y,

$$H(K, Y) = \sum_a -k_a \log y_a + (k_a - 1)\log(y_a - 1), \tag{12}$$

where the summation is over all atoms in the training data and is equivalent to the negative log likelihood of the data conditional on the model. This minimization computes a maximum likelihood estimate of the network's parameters given the data. This is a standard neural network training with gradient descent to match the network's output to our guesses for the atom-level sites of metabolism K.

3 Results and discussion

In the following sections, we study the behavior and performance of site-of-metabolism models trained on region-resolution data.

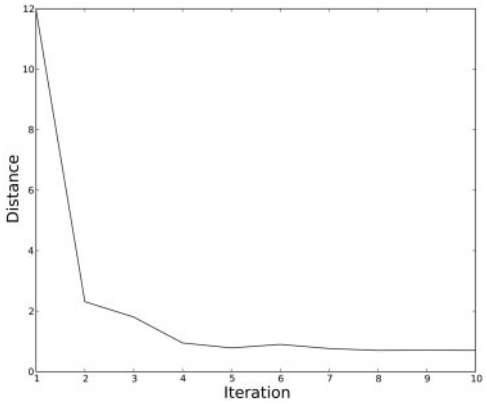

Fig. 3. Algorithm convergence. The EM algorithm converges in five to six iterations on our site of metabolism datasets. The figure plots the Euclidian distance of the model's output vector from each iteration with the output vector of the prior iteration. The most change happens in the first few iterations, after which the change quickly stabilizes at a low value. The distance never reaches 0 because of small variations from run to run and floating point errors. These convergence behaviors are consistent across all isozymes and region sizes

3.1 Convergence

As a first test, we verified that the model would converge within a reasonable number of iterations. Convergence was tracked by plotting how much either the Y or K vector changed during each iteration. Running the algorithm with 5-atom sized regions, we observed that the algorithm usually converged in five to six iterations (Fig. 3). This is an encouraging result, which was consistent across all isozymes and region sizes (data not shown). We therefore decided to iterate the EM algorithm 10 times in all further experiments—more than enough to reach convergence—rather than construct more complicated convergence criteria.

3.2 Identifying sites of metabolism in regions

As a second test, we verified that the K and Y vectors converge to the atom-resolution sites of metabolism during training. Put another way, we expected that region-trained models should identify atom-resolution sites of metabolism in the training data.

In many specific cases, region-trained models identify sites of metabolism in metabolized regions (Fig. 4). There are also cases where the correct site of metabolism is not identified. These examples are only qualitative evidence, but we were able to quantify this behavior more systematically. Using the area under the ROC curve (AUC), we extracted all the sites within metabolized regions, and measured how well the output of the trained model separates sites of metabolism from non-metabolized sites. All AUCs were divided by the AUC performance of an atom-trained model, which serves as a good upper limit on the best performance we could expect from any of the region-trained models (Fig. 4).

The region-trained models are never shown which sites within a metabolized region are the true sites of metabolism, so this is a valid assessment. Unfortunately, defining regions in different ways changes the atoms over which the performance is measured. The raw AUCs between different region sizes, therefore, are not directly comparable. However, normalizing performance by dividing by the atom-trained performance makes the AUCs comparable. The atom-trained models, in this sense, are trained with the exact sites of metabolism and are only included here to make the region AUCs comparable.

In this assessment, we see that, across all isozymes, region-trained models can identify sites of metabolism in metabolized

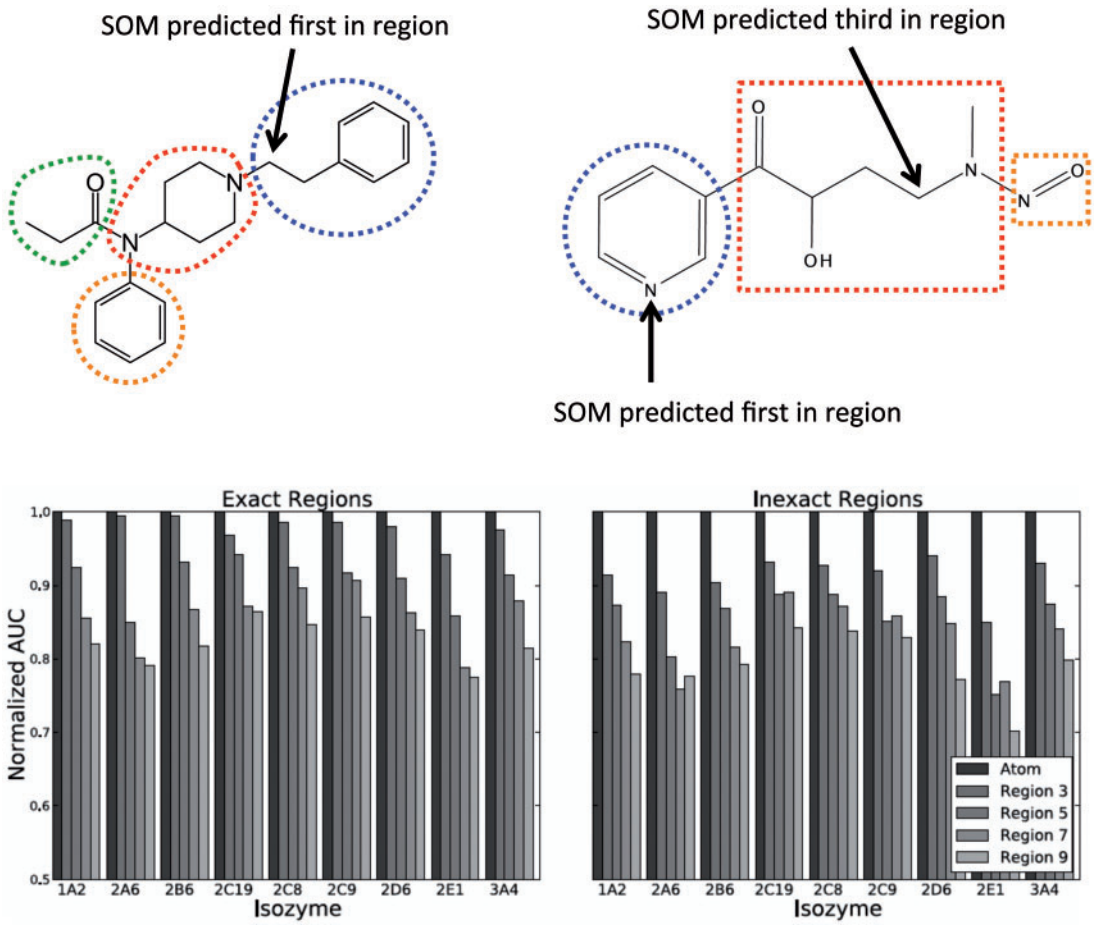

Fig. 4. Identifying atomic sites of metabolism in region-resolution training data. Sample molecules from the 7-atom region dataset (with regions circled) demonstrate how the correct site of metabolism is often identified in training molecules, even though this information is not given to the algorithm (top left and right molecules). Sometimes, however, the correct site of metabolism is not identified (top right molecule). Using the AUC normalized by the performance of atom-trained models, we quantified this behavior by extracting all the sites within metabolized regions, and measuring how well the output of the trained model could separate sites of metabolism from non-metabolized sites. This is true for models trained with both exact and inexact region data

regions with almost the same accuracy as atom-trained models. As would be expected, their ability to identify sites of metabolism is not as strong as atom-trained models and decreases as region size increases. This is a critical result, because it shows that the EM algorithm is working as we would hope, and we expect that the region-trained models will work nearly as well as atom-trained models.

We see nearly the same performance in both the exact and inexact datasets. The inexact data only mark regions as containing more than one site of metabolism or not, without specifying exactly how many sites are metabolized. Inexact regions, for this reason, give the algorithm less information. It is not always possible to exactly specify how many sites of metabolism are in a region from mass spectrometry data, so the performance on the inexact data is important to track. The fact that the inexact-trained performance is nearly the same as the exact-trained performance is encouraging.

3.3 Cross-validated accuracy

For the next test, we assessed the accuracy of region-trained models using leave-one-out cross-validation. In this assessment, we hold out one molecule from the training set to test on for each fold. In turn, every molecule is held out from training and then has its sites of metabolism predicted by the trained model. Predictions are assessed using the Top-2 metric, where the percentage of molecules where a correct SOM is predicted within the top two sites of the molecule. They are also assessed using the average AUC, where the AUC for each molecule is computed and averaged across the whole dataset. The performance of region-trained models is compared with atom-trained models configured with identical parameters. There are no other region-trained methods with which to compare because this is the first published study on this problem. We expect that region-trained models will perform nearly well as the atom-trained models in this assessment.

This is exactly what we see. Region-trained models—using either exact or inexact regions—perform quite well, having results that are comparable to those of atom-trained models (Fig. 5). Using the Top-2 metric, we see exact region-level models using regions of size 3, 5, 7 and 9 to have respective accuracies 1, 10, 17 and 25%, respectively, lower than those of atom-level models, averaged across all isozymes. With inexact regions, we observe average performance drops of 5, 9, 13 and 19%, respectively. Using the AUC metric, we see similar patterns, with an average performance loss across all isozymes of 2, 9, 14 and 18% for exact regions of size 3, 5, 7 and 9, respectively. With inexact regions the performance drop is comparably small; we see an average performance drop of 9, 15, 17 and 21% for the same region sizes.

As we would expect, the atom-trained models tend to perform the best, having the highest accuracy for all but two of the datasets. Likewise, the accuracy of the region-trained models falls off as the

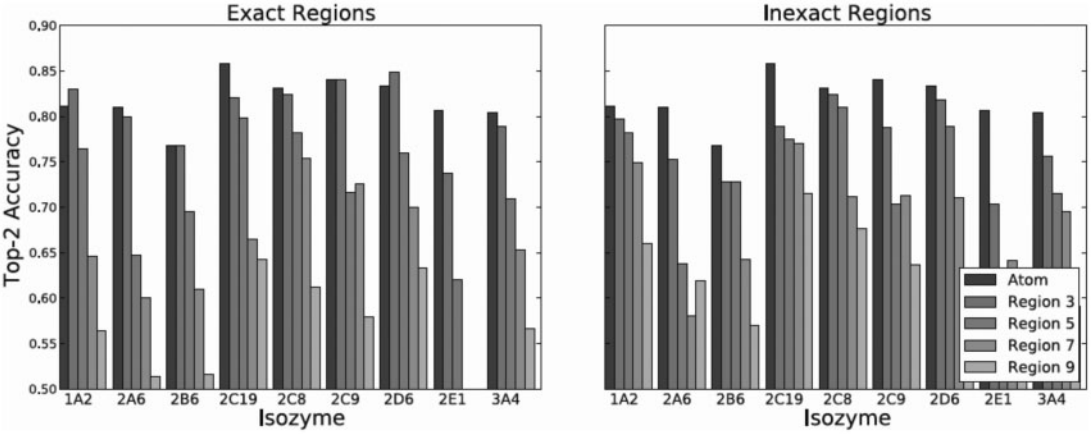

Fig. 5. Top-2 performance using leave-one-molecule-out cross-validation. We quantified the performance of atom- and region-trained models using the Top-2 metric: the percentage of molecules for which a site of metabolism is ranked within the top two sites of the molecule. Region-trained models perform nearly as well as atom-trained models, though their accuracy does decrease as the size of regions increases

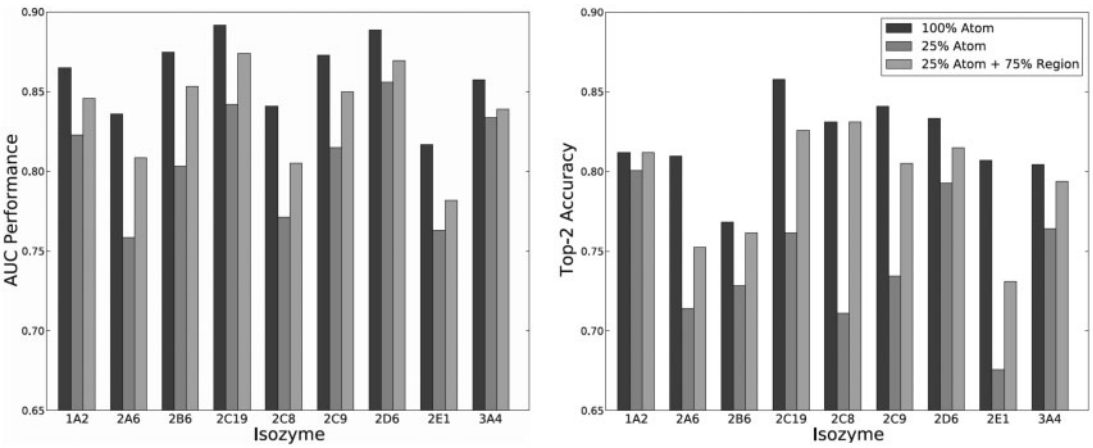

Fig. 6. Extending atom-trained models with region-resolution data. We compared the performances of models trained using 100% of atom-resolution data, 25% of atom-resolution data and 75% region-resolution data. According to both the Average AUC (left) and Top-2 (right), there is a performance drop across all isozymes when only 25% of the atom-resolution data is used. This performance drop is substantially reduced by adding back in the region-resolution data for the removed molecules. The figures show the results only using exact regions of size 5, but very similar results are observed for all regions datasets

region size increases. Nonetheless, the performance for regions of size is very close to atom-trained data. Similar results were observed using alternative performance metrics, including AUC and the Top-3 metric (results not presented). This is an exciting result, demonstrating that it is possible to train models on the region-resolution data.

3.4 Extending models with region-resolution data

The results presented so far suggest that we can, in fact, train site-of-metabolism models on region-resolution data alone. In practice, however, there is often atom-resolution data for some molecules and region-resolution data for other molecules. We hypothesize that using both types of data simultaneously is better than training on the atom-resolution data alone. Moreover, we hypothesize that using the region-resolution data can improve the model's accuracy on molecules structurally distant from the atom-resolution data.

Pharmaceutical companies want to predict the metabolism their molecules. They have atom-resolution metabolism data from publicly available sources and a much larger amount of private, region-resolution metabolism data, generated in-house, for molecules similar to theirs. Currently, the region-resolution data are ignored in modeling, but could including the region-resolution data improve the models? We simulate this scenario using a modified version of leave-one-molecule-out cross validation. For each cross-validation fold, there is one molecule in the test set, and the remaining molecules are used as the training set. Using fingerprints similar to Daylight fingerprints (Azencott *et al.*, 2007; Swamidass *et al.*, 2005), the molecules in the training set are sorted by their Tanimoto similarity to the test molecule. During training, the atom-resolution data are available for only the 25% most dissimilar molecules from the test molecule. Region-resolution data are available for the remaining 75% of molecules. Now, we train and test two models. One model trains on the atom-resolution data alone, on just 25% of the training set. The other model trains on both the atom- and region-resolution data together.

For this experiment, we measured performance with two metrics, Top-2 accuracy and Average AUC. The Average AUC is a more sensitive metric that also assesses how well sites of metabolism are picked out by the model. An ROC curve is constructed for each molecule, and the AUC is computed for curve. The average of these AUCs across all molecules in the training set is the Average AUC.

We expect that training on a reduced set of atom-resolution data will reduce the models' performance, but we also expect that some of this performance drop will be reversed by adding back the missing molecules as region-resolution data. This is exactly the behavior we

observe (Fig. 6). Training on just 25% of the atom-resolution data causes the performance to drop (by the Average AUC metric) by an average of 6% across enzymes. Adding the missing molecules as region-resolution data reduces the performance loss to 1, 2, 3 and 4% using regions of size 3, 5, 7 and 9, respectively. Very similar results are observed with the Top-2 metric. We see an average performance loss of 8% for the atom-only model across isozymes. Adding the region-resolution data back reduces this performance loss to 2, 3, 4 and 5% for regions of size 3, 5, 7 and 9, respectively.

When available, atom-resolution for all molecules is best, but this data are not always available. These results indicate that using region-resolution data where atom-resolution data are not available can train site-of-metabolism models nearly as accurately as if atom-resolution data for all molecules were available.

4 Conclusion

In this study, we present the first site-of-metabolism model that uses region-resolution data. Moreover, it appears that extending atom-resolution training data with region-resolution data improves the accuracy of site-of-metabolism models. This is an exciting result, because there is often much more region-resolution site of metabolism data available in pharmaceutical companies, much of which covers chemical space currently underrepresented in public atom-resolution literature-derived data. By using data already collected in drug development, but unused for modeling, Xenosite-Region may prove to be a powerful way of extending site of metabolism models making them more accurate. Furthermore, the approach we describe here is general and could extend other SOM modeling approaches to use region-resolution data as well, and we leave this for future work.

5 Supporting Information

Isozyme-specific substrate sets are provided as SDF files in the supporting information. These files indicate the specific atoms that are oxidized when the substrate is metabolized by the given CYP isozyme. Additional files are provided that indicate, for each atom, whether it is or is not metabolized, it's topological and multi-atom site information, and, in separate files, the substrate region to which the atom has been assigned.

Acknowledgements

The idea of building site-of-metabolism models that can train from region-resolution data was formed in conversations with Curt Breneman. This recognition of this problem within industry was confirmed in private, and very helpful, communications with Prashant Desai at Eli Lilly. Figure 1 created with OEDepict, version 1.7.4.5, OpenEye Scientific Software, Inc., Santa Fe, NM, USA, www.eyesopen.com, 2014. We thank the Department of Immunology and Pathology at the Washington University School of Medicine for its generous support of this work.

Conflict of Interest: none declared.

References

Azencott,C. *et al.* (2007) One- to four-dimensional kernels for small molecules and predictive regression of physical, chemical, and biological properties. *J. Chem. Inf. Model,* **47**, 965–974.

Baldi,P. and Brunak,S. (2001) *Bioinformatics: The Machine Learning Approach*. The MIT Press, Cambridge, MA.

Castro-Perez,J.M. (2007) Current and future trends in the application of HPLC-MS to metabolite-identification studies. *Drug Disc. Today,* **12**, 249–256.

Dapkunas,J. *et al.* (2009) Probabilistic prediction of the human cyp3a4 and cyp2d6 metabolism sites. *Chem. Biodivers.,* **6**, 2101–2106.

Dempster,A.P. *et al.* (1977) Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.. B,* **39**, 1–38.

Gerlich,M. and Neumann,S. (2013) Metfusion: integration of compound identification strategies. *J. Mass Spectrom.,* **48**, 291–298.

Guengerich,F.P. (2006) Cytochrome P450s and other enzymes in drug metabolism and toxicity. *AAPS J.,* **8**, E101–E111.

Heinonen,M. *et al.* (2012) Metabolite identification and molecular fingerprint prediction through machine learning. *Bioinformatics,* **28**, 2333–2341.

Hop,C.E. *et al.* (2008) High throughput ADME screening: practical considerations, impact on the portfolio and enabler of in silico ADME models. *Curr. Drug Metab.,* **9**, 847–853.

Huang,T.-W. *et al.* (2013) Dr-predictor: incorporating flexible docking with specialized electronic reactivity and machine learning techniques to predict CYP-mediated sites of metabolism. *J. Chem. Inf. Model.,* **53**, 3352–3366.

Karypis,G. and Kumar,V. (1999) A fast and highly quality multilevel scheme for partitioning irregular graphs. *SIAM J. Sci. Comput.,* **20**, 359–392.

Kerber,A. *et al.* (2001). Molgen-ms: evaluation of low resolution electron impact mass spectra with ms classification and exhaustive structure generation. *Adv. Mass Spectrom.,* **15**, 939–940.

Kirchmair,J. *et al.* (2012) Computational prediction of metabolism: sites, products, SAR, p450 enzyme dynamics, and mechanisms. *J. Chem. Inf. Model.,* **52**, 617–648.

Korolev,D. *et al.* (2003) Modeling of human cytochrome p450-mediated drug metabolism using unsupervised machine learning approach. *J. Med. Chem.,* **46**, 3631–3643.

Lawrence,C.E. and Reilly,A.A. (1990) An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins,* **7**, 41–51.

Nebert,D.W. and Russell,D.W. (2002) Clinical importance of the cytochromes p450. *Lancet,* **360**, 1155–1162.

Optibrium Ltd. (2009). Stardrop, version 4.3.

Rasche,F. *et al.* (2012) Identifying the unknowns by aligning fragmentation trees. *Anal. Chem.,* **84**, 3417–3426.

Redner,R.A. and Walker,H.F. (1984) Mixture densities, maximum likelihood and the EM algorithm. *SIAM Rev.,* **26**, 195–239.

Rudik,A.V. *et al.* (2014) Metabolism site prediction based on xenobiotic structural formulae and pass prediction algorithm. *J. Chem. Inf. Model.,* **54**, 498–507.

Rydberg,P. *et al.* (2010) SMARTCyp: a 2D method for prediction of cytochrome P450-mediated drug metabolism. *ACS Med. Chem. Lett.,* **1**, 96–100.

Scheubert,K. *et al.* (2013) Computational mass spectrometry for small molecules. *J. Cheminform.,* **5**, 1–24.

Schrödinger L.L.C. (2011). P450 SOM prediction, version 1.0.

Siegel,D. *et al.* (2013) Chemical and technical challenges in the analysis of central carbon metabolites by liquid-chromatography mass spectrometry. *J. Chromatogr. B.* **966**, 21–33.

Stein,S.E. (1995) Chemical substructure identification by mass spectral library searching. *J. Am. Soc. Mass Spectrom.,* **6**, 644–655.

Swamidass,S. *et al.* (2005) Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics,* **21**(Suppl. 1), i359.

Wolf,S. *et al.* (2010) In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics,* **11**, 148.

Xiao,J.F. *et al.* (2012) Metabolite identification and quantitation in LC-MS/MS-based metabolomics. *TrAC Trends Anal. Chem.,* **32**, 1–14.

Zaretzki,J. *et al.* (2011) RS-predictor: a new tool for predicting sites of cytochrome P450-mediated metabolism applied to CYP 3A4. *J. Chem. Inf. Model.*, **51**, 1667–1689.

Zaretzki,J. *et al.* (2012) Rs-predictor models augmented with smartcyp reactivities: robust metabolic regioselectivity predictions for nine CYP isozymes. *J. Chem. Inf. Model.,* **52**, 1637–1659.

Zaretzki,J. *et al.* (2013) Xenosite: accurately predicting CYP-mediated sites of metabolism with neural networks. *J. Chem. Inf. Model.,* **53**, 3373–3383.