Computationally Assessing the Bioactivation of Drugs by N-Dealkylation

Na Le Dang,[†] Tyler B. Hughes,[†] Grover P. Miller,[‡] and S. Joshua Swamidass[*,†]

[†]*Department of Pathology and Immunology, Washington University School of Medicine, Campus Box 8118, 660 S. Euclid Ave., St. Louis, Missouri 63110, United States*

[‡]*Department of Biochemistry and Molecular Biology, University of Arkansas for Medical Sciences, Little Rock, Arkansas 72205, United States*

E-mail: swamidass@gmail.com

Phone: 314.935.3567

Contents

Training Data Set

The list of unique molecules IDs, associated molecule and reaction registry numbers, and their metabolic status is provided in "AMD_Registry_Numbers.csv" file. This is a comma-separated values text file with the "UniqueDatasetMoleculeID" containing the assigned ID in the training data set. The "MOLREGNO" and "RXNREGNO" columns provide Accelrys Metabolite Database registry number. Columns "1A2", "2A6", "2B6", "2C19", "2C8", "2C9", "2D6", "2E1", "3A4", and "HLM" indicate whether the molecule is metabolized by the enzymatic entity. The histograms describing molecular properties of the combined data set are provided (Figures S1, S2, and S3).

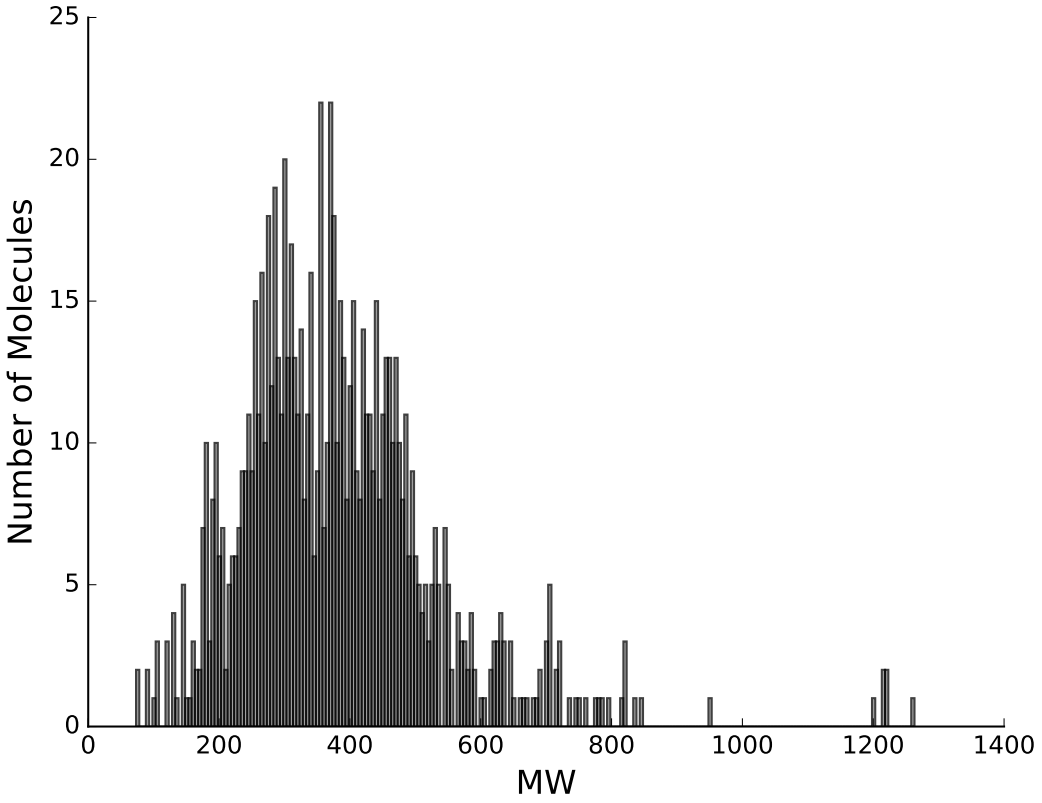

Figure S1: **Molecular Weight (MW) Distribution of the Combined Data Set.**

Descriptors

The following tables detail all the descriptors used by the model in this study.

Table S1: Molecule-level descriptors used by the XenoSite N-Dealkylation Model.

Name	Descriptions
atoms	number of atoms
bonds	number of bonds
TPSA	topological polar surface area

logP	octanol/water partition coefficient
HBD	number of hydrogen bond donors
HBA_1	number of hydrogen bond acceptors Pybel SMARTS string 1
HBA_2	number of hydrogen bond acceptors Pybel SMARTS string 2
MR	molar refractivity
MW	molecular weight
sbonds	number of single bonds
dbonds	number of double bonds
tbonds	number of triple bonds
abonds	number of aromatic bonds
heavy_atoms	number of heavy atoms
hydrogens	number of hydrogens
NumRings	number of rings

Table S2: Bond-level descriptors used by the XenoSite N-Dealkylation Model.

Name	Descriptions
Single	whether the bond is a single bond
Double	whether the bond is a double bond
Triple	whether the bond is a triple bond
Aromatic	whether the bond is an aromatic bond
Length	length of the bond
NTopologicalEquivalent	number of topological equivalent of the bond within the molecule

Table S3: Atom-derived bond-level descriptors used by the XenoSite N-Dealkylation Model.

Ne_d	number of atoms depth d (0,1,2,3,4) bonds away of type element e (C, O, N, S, P, F, Cl, Br or I)
Pe_d	percentage of atoms depth d (1,2,3,4) bonds away of type element e (C, O, N, S, P, F, Cl, Br or I)
Ne_sp_i_d	number of atoms depth d (0,1,2,3) bonds aways of type element e (C,O,N,S)with sp_i hybridization
sp_i_d	number of sp_i hybridization depth d (0,1,2,3) bonds aways
TotalBondOrder	total bond order
Span	(maximum path length from current atom)/(maximum path length from all atoms within the molecule)
InvertedSpan	$1/(1 + \text{Span})$
Normalized Span	Span/(maximum span within molecule)
$Ring_n$	within ring of size n
MaxInvRingSize	size of the largest ring containing the atom
NRings	total number of rings containing atom
SRing	smallest ring containing atom
HBonded	total number of hydrogens bound to atom
NHBonded	total number of non-hydrogens bound to atom
FarthestBondedHydrogen	distance to the farthest hydrogen bound to atom
FarthestBondedHydrogenIndicator	indicates whether or not an atom is bound to a hydrogen
RB	number of rotatable bonds for atom
PT_ElectronNeg	electron negativity
PT_ElectronAffinity	electron affinity
PT_Ionization	ionization state
PT_BondRad	Bond radius
PT_VdwRad	Vdw radius
Aromatic	binary value indicating whether atom is aromatic
SP^1	binary value indicating whether atom is sp^1 hybridized
SP^2	binary value indicating whether atom is sp^2 hybridized
SP^3	binary value indicating whether atom is sp^3 hybridized
HybX	binary value indicating whether atom is non-sp hybridized
Lone_Pair_Depth_d	Number of lone pair depth d (0,1,2,3) bonds away
AromaticNeighbors	Number of aromatic neighbors
BN_t_d	number of type t (single, double, triple, aromatic) bond neighbors of depth d away

Within_*substructure*	whether the atom is in a *substructure* (epoxide, α-β unsaturated ketone, carboxyl, sulfate,phosphate, nitro, amide)

Table S4: Descriptor groups in the modular multi-target neural network.

Group	Input Nodes	Number of ^1H nodes
Neighborhood 0	Ne_d with $d = 0$	3
Neighborhood 1	N$_d_e$, P$_d_e$, and sp$_i_d$ with $d = 0$	3
Neighborhood 2	N$_d_e$, P$_d_e$, and sp$_i_d$ with $d = 1$	3
Neighborhood 3	N$_d_e$, P$_d_e$, and sp$_i_d$ with $d = 2$	3
Neighborhood 4	N$_d_e$, P$_d_e$, and sp$_i_d$ with $d = 3$	3

Table S5: Descriptor groups used for sensitivity analysis.

Atom Element	Ne_d with $d = 0$
Atoms One Bond Away	N$_d_e$ and P$_d_e$ with $d = 1$
Atoms Two Bonds Away	N$_d_e$ and P$_d_e$ with $d = 2$
Atoms Three Bonds Away	N$_d_e$ and P$_d_e$ with $d = 3$
Atoms Four Bonds Away	N$_d_e$ and P$_d_e$ with $d = 4$
Size of Ring Containing Atom	Ring$_n$, NRings, and SRing
Hybridization State	SP^1, SP^2, SP^3, and HybX

Heuristic Python Script

The python script "Heuristic.py" can be used to make Heuristic N-dealkylation predictions on molecules in SDF format.

Metabolite Structure Prediction Python Script

The python script "get_Dealkylation_metabolite_structure.py" can be used to generate N-Dealkylation metabolite structure based on molecules in SDF format.

Accuracy Comparison

Alpha-Beta Unsaturated Aldehyde Reactivity Scores

To further assess the mode's performance in modeling reactivity of aldehydes, we identified 313 alpha-beta unsaturated aldehyde containing molecules and 528 alpha-beta unsaturated aldehyde sites from the reactivity model's training data set and used their cross-validated predictions to access the model performance on this subset of data. First, we assessed the ability of model to identify the correct atom in the molecule as reactive. The average site AUC is computed by averaging the AUC of sites computed within each molecule separately. The reactivity model predicted reactive atoms of 313 aldehyde containing molecules with average site AUC accuracies of 100.0% (1 molecule), 90.4% (21 molecules) , 96.1% (265 molecules), and 97.5% (23 molecules) for cyanide, DNA, GSH, and protein, respectively.

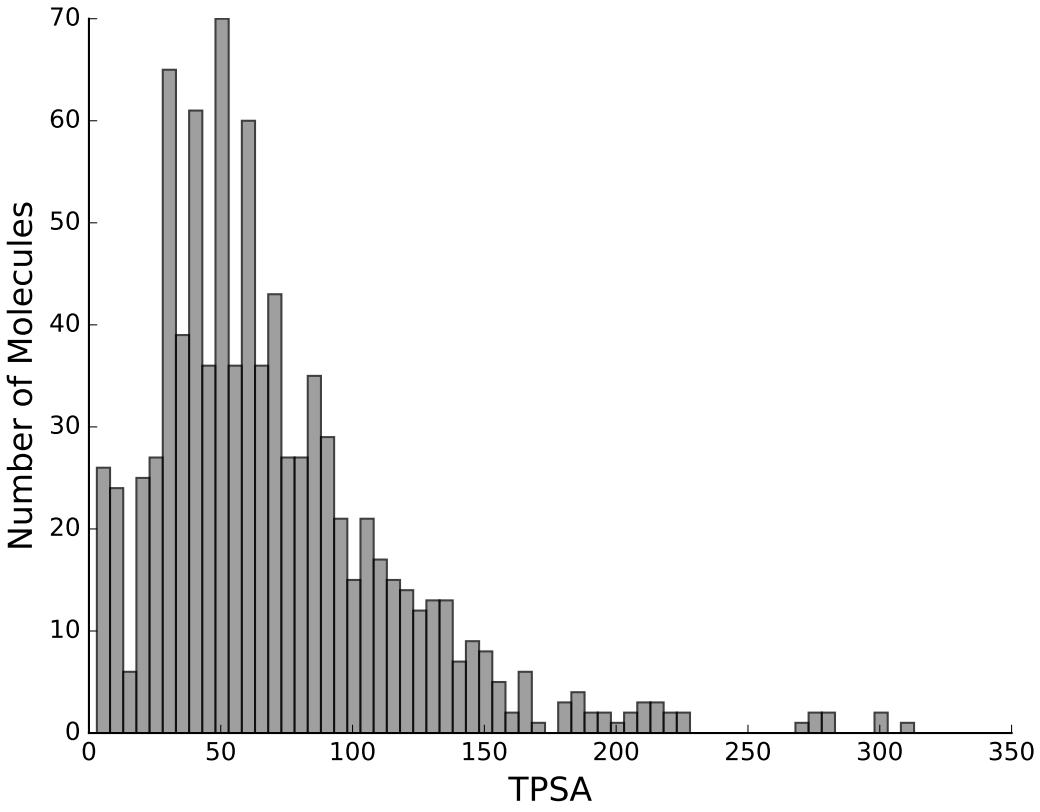

Figure S2: **Topological Polar Surface Area (TPSA) Distribution of the Combined Data Set.**

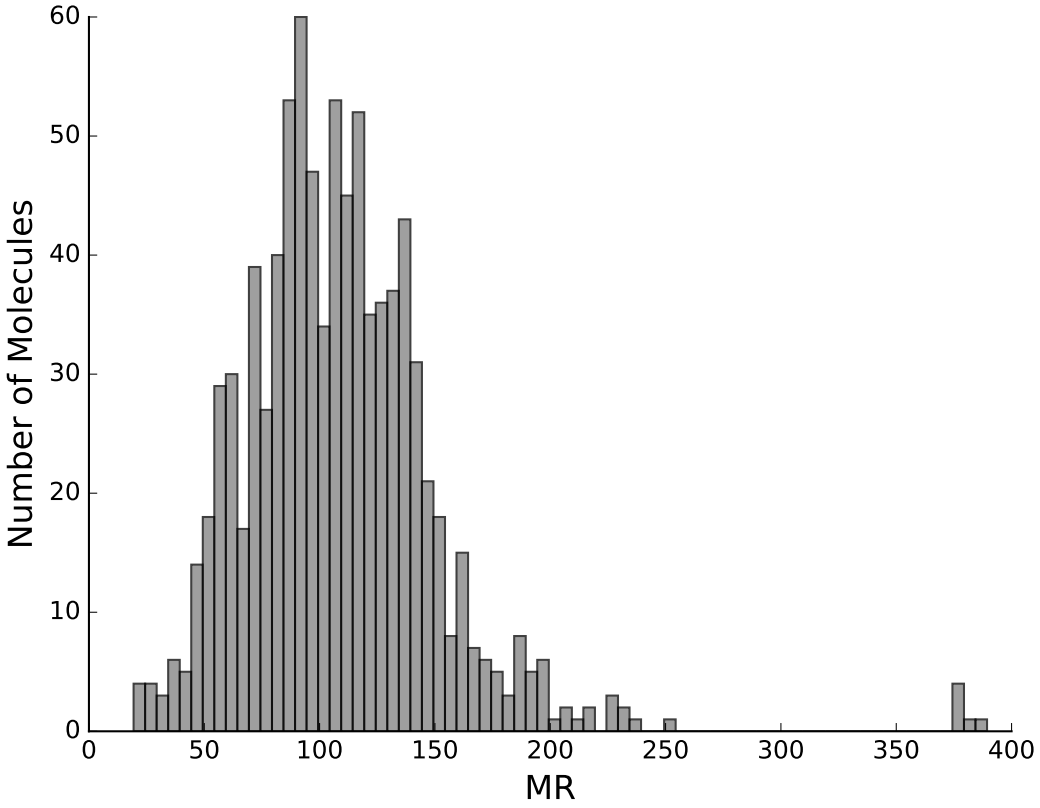

Figure S3: **Molar Refractivity (MR) Distribution of the Combined Data Set.**

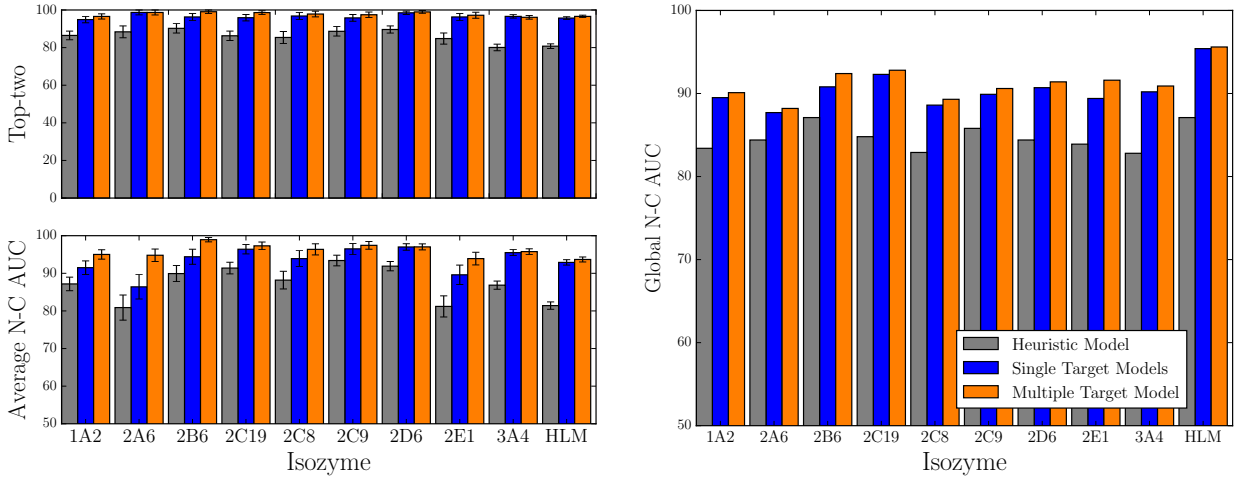

Figure S4: **The model accurately identifies sites of N-dealkylation.** Top left panel shows top-two performance metric for 883 molecules of the combined data set, by which a molecule was considered correctly predicted if any of its observed SNDs were predicted in the first- or second-rank position. Bottom left panel, the AUC for predictions of nitrogen-carbon bonds within each molecule is computed and then averaged across the whole data set, measuring the per-molecule performance. Most performance differences between two models are not statistically significant except for the CYP2A6 and CYP2B6 average nitrogen-carbon AUCs (Welch's t-test p-values cutoff of 0.05). By both molecule-level metrics, the cross-validated predictions generated by the multi-target modular neural network perform equally well in comparison to the predictions of a single target neural network. Right panel, for each isozyme and HML target, global nitrogen-carbon AUC was computed across all 4071 nitrogen-carbon bonds of the combined data set. This quantifies how often SNDs were ranked above other nitrogen-carbon bonds in the entire data set. The ten-fold cross-validated predictions generated by a multi-target modular neural network outperformed the predictions of a single target neural network. The performance difference between the two models is statistically significant (paired permutation test p-values cutoff of 0.05). Across all metrics, the heuristic model perform worst, with average top-two, average nitrogen-carbon AUC, and global nitrogen-carbon AUC accuracies of 83.8%, 85.8%, and 85.2%, respectively.

Next, we assessed the ability of the model to separate reactive and non-reactive aldehyde molecules. Across the full database 528 aldehydes, the model can accurately separate reactive and non reactive substructures with AUCs of 93.7%, 89.8% and, 98.7%, for DNA, GSH, and protein, respectively. These assessments demonstrate that the reactivity model can accurately model the reactivity of alpha-beta-reactive aldehyde containing compounds.

Descriptors Driving Performance

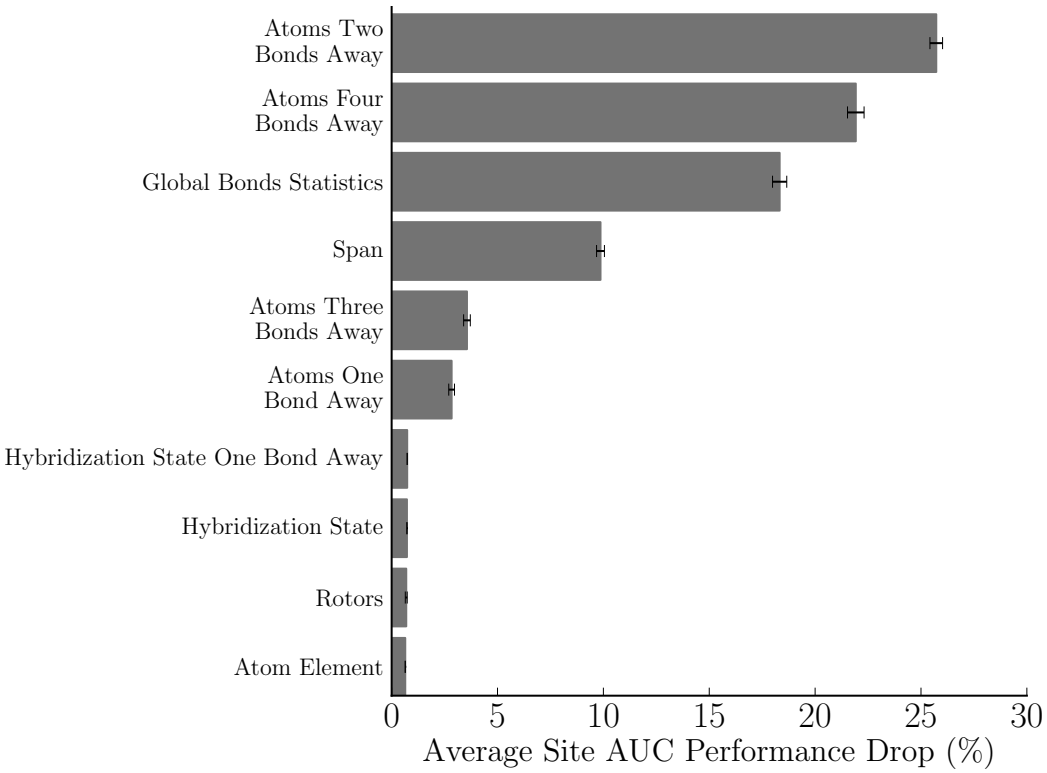

Figure S5: **The importance of specific descriptors to the bond N-dealkylation model.** A permutation sensitivity analysis quantified the importance of descriptors for the final trained site of N-dealkylation model. The ten most important individual descriptors are listed in decreasing order of importance from top to bottom. The graph shows the model performance drop associated after permuting the associated descriptor values, averaging over ten iterations.

A permutation sensitivity analysis identified the descriptors driving model accuracy.[1] We started with the trained model, and calculated its training accuracy. Next, we randomly permuted each descriptor column (or group of closely related descriptors) in the input data for these molecules. We applied the trained model to the permuted data, and recorded the performance drop across all molecules. The higher the performance drop, the more important the descriptor(s) to the model's performance. We saw similar results using all performance metrics. For example, using the average site AUC performance, this analysis identified the most important topological and molecular descriptors for differentiating metabolized sites

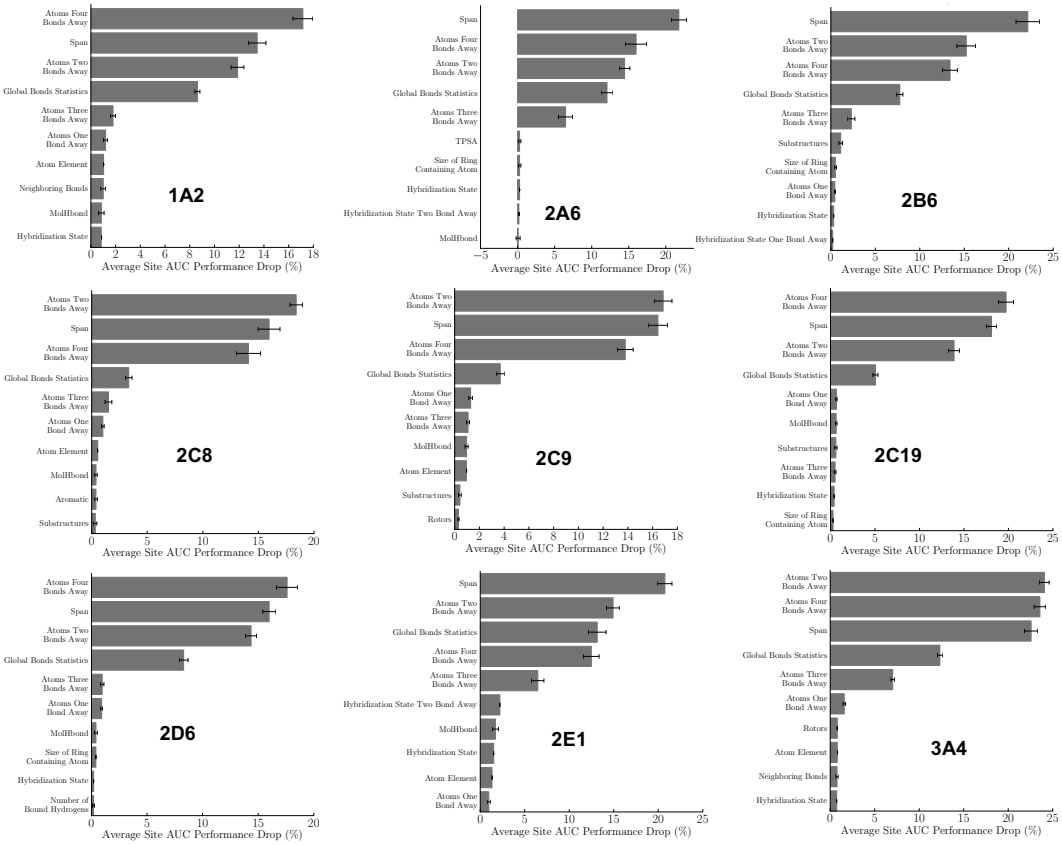

Figure S6: **The importance of specific descriptors to each isozyme N-dealkylation target.** A permutation sensitivity analysis quantified the importance of descriptors for the final trained site of N-dealkylation model. The ten most important individual descriptors are listed in decreasing order of importance from top to bottom. The graph shows the model performance drop associated after permuting the associated descriptor values, averaging over ten iterations.

from non-metabolized sites (Figure S5). The key topological descriptors are the identities and hybridization states of atoms on either sides of the bond and their neighbors (atoms one, two, three and four bond away), the relative distance from the bond to the center of the molecule (Span) and the number of rotatable neighbor bonds (Rotors). The molecule's number of single, double, triple and aromatic bonds (Global Bonds Statistics) are the most important group of molecular descriptors. This result revealed that the model heavily relies on local topology. Similar results were seen in the individual isozyme sensitivities (Figure S6).

Loss Functions Dependency on Iteration Cycles

The values of loss function at each iteration were tracked for 10 training experiments where targets were permuted (Figure S7). Similar results are seen for non-permuted data, but this data was excluded for brevity.

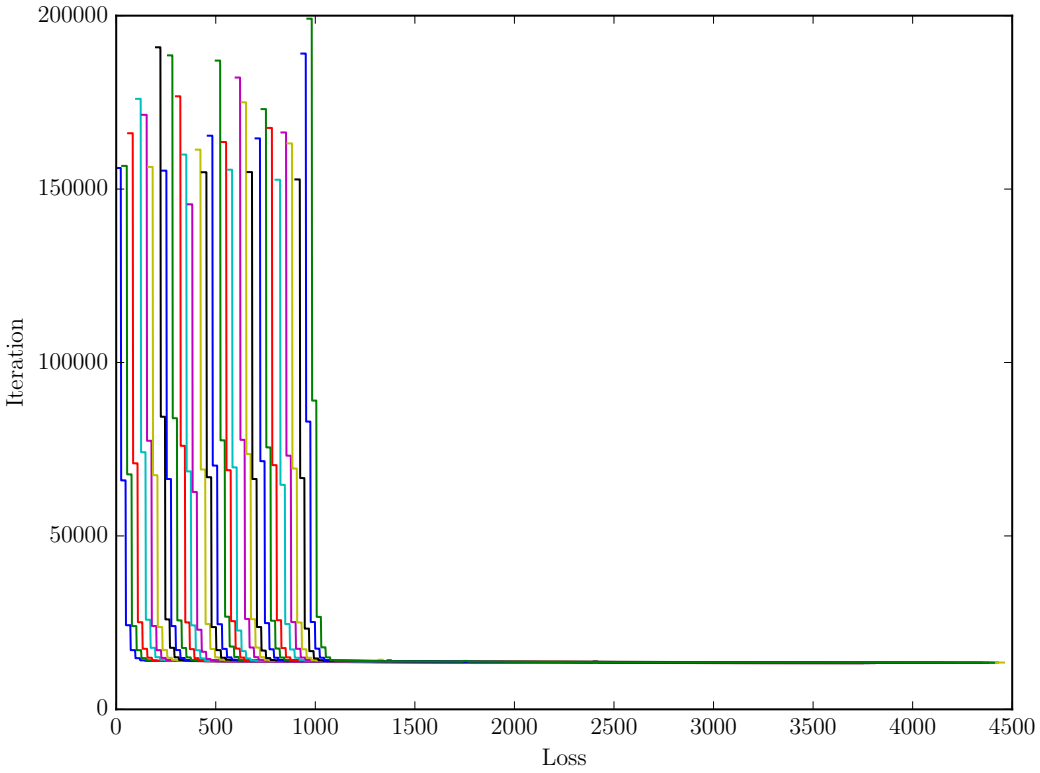

Figure S7: **Loss functions dependency on Iteration Cycles** The values of loss function at each iteration were tracked for 10 training experiments where targets were scrambled. Very quickly, the loss reaches a minimum. Similar results are seen for non-permuted data, and excluded for brevity.

Reliability Plots

Reliability plot for each CYPs isozymes are shown in Figure S8.

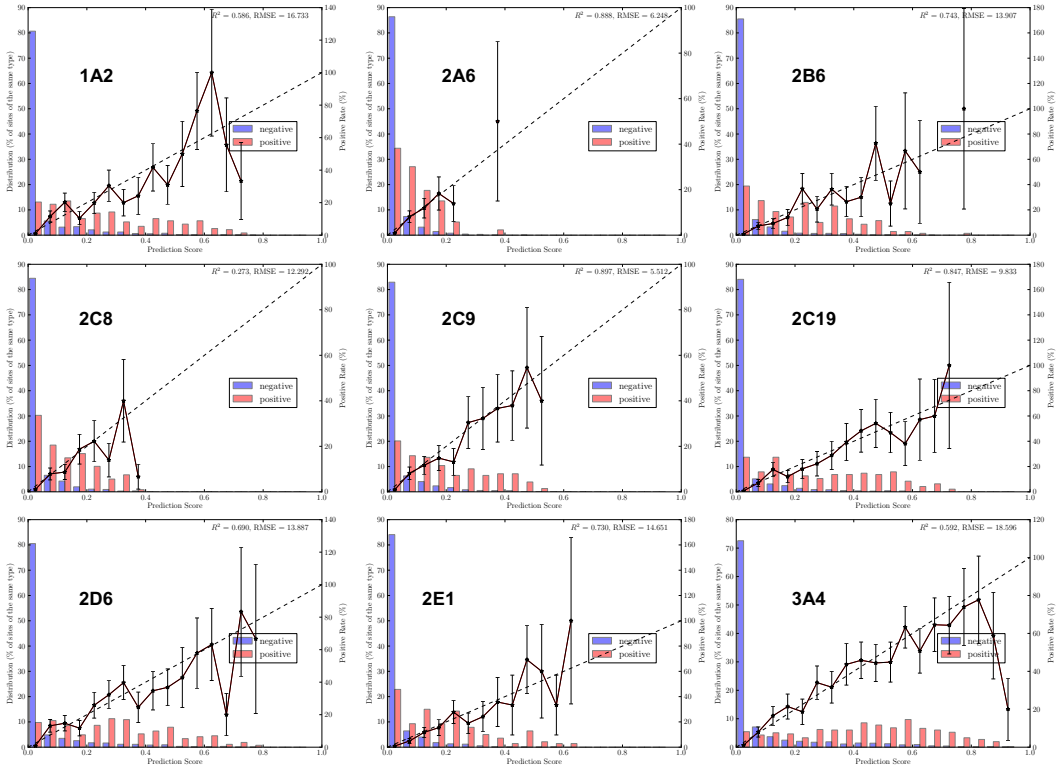

Figure S8: **XenoSite offers well-scaled probabilistic prediction scores of 1A2 N-dealkylation.** The bar graphs plot the normalized distributions of NDS across 4071 dealkylated and non-dealkylated N-C bonds. The solid lines plot the percentage of N-C bonds that are dealkylated (using non-normalized frequencies) in each bin. The diagonal dashed lines indicate a hypothetical perfectly scaled prediction.

References

(1) Hunter, A.; Kennedy, L.; Henry, J.; Ferguson, I. Application of neural networks and sensitivity analysis to improved prediction of trauma survival. *Computer methods and programs in biomedicine* **2000**, *62*, 11–19.