Supplementary Information: The Metabolic Rainbow: Deep Learning Phase I Metabolism in Five Colors

Na Le Dang, Matthew K. Matlock, Tyler B. Hughes, 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

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" column containing the assigned ID in the training data set. The "MOLREGNO" and "RXNREGNO" columns provide Accelrys Metabolite Database registry number. Columns "Stable Oxygenation", "Unstable Oxygenation", "Dehydrogenation", "Hydrolysis" and "Reduction" indicate whether the molecule is metabolized by the enzymatic entity.

Descriptors

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

Table S1: Atom-derived bond-level descriptors used by the XenoSite Human Phase I Metabolism 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 (3, 4, 5, 6, 7, 8)
NRings	total number of rings containing atom
MaxInvRingSize	size of the largest ring containing the atom
SRing	smallest ring containing atom
HBonded	total number of hydrogens bonded to atom
NANbrs	total number of atoms bonded to atom
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* (α-β unsaturated ketone, carboxyl, sulfate, phosphate, nitro, amide)

Table S2: Bond-level descriptors used by the XenoSite Human Phase I Metabolism 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
In Ring	whether the bond is part of a ring
Connected to Hydrogen	whether the bond is between a heavy atom and a hydrogen
Lone Pair	whether this is a lone pair (not a bond)
Ester	whether the bond is an ester bond
Amide	whether the bond is an amide bond
NTopologicalEquivalent	number of topological equivalent of the bond within the molecule

Table S3: Molecule-level descriptors used by the XenoSite Human Phase I Metabolism 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 S4: 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

Comparison between "Color Bond" and Generic Metabolism Labeling

Differences between our "Color Bond" and Generic Metabolism Labeling schemes are shown in Table S5.

Table S5: "Colored Bond" and Generic Site of Metabolism Labeling Schemes

		Labeled Site of Metabolism	
Color	Reaction Type	"Colored Bond"	Generic
Red	Epoxidation	the double/ aromatic bond between two heavy atoms	the two heavy atoms
	Hydroxylation	the bond between a heavy atom and a hydrogen	the heavy atom
	S-oxidation	the lone pair on a sulfur atom	the sulfur atom
	N-oxidation	the lone pair on an nitrogen atom	the nitrogen atom
Orange	N-dealkylation	the bond between a nitrogen and a carbon	the carbon atom
	O-dealkylation	the bond between an oxygen and a carbon	the carbon atom
	S-dealkylation	the bond between a sulfur and a carbon	The carbon atom
	C-dealkylation	the bond between two carbons	the carbon atom that the oxygen attaches to
	P-dealkylation	the bond between a phosphorus and a carbon atom	the carbon atom
	Oxidative Dehalogenation	the bond between a halogen and a carbon	the carbon atom
DH	Double/triple bond formation	the bond between the abstracted hydrogen and its connected heavy atom	the heavy atom
	Quinone/Imine/ Methide formation	the bond between heavy atom and hydrogen	the heavy atom
RD	Nitro reduction	the bond between nitrogen and attached oxygen	the nitrogen atom
	Carbonyl reduction	the carbonyl bond	the carbon atom
	Sulfo reduction	the bond between sulfur and attached oxygen	the sulfur atom
	Reductive dehalogenation	the bond between halogen and attached carbon	the carbon atom
	Hydrogenation	the double, triple bond between pair of hydrogenated atoms	the heavy atom
HD	Amide hydrolysis	amide bond	the heavy atoms on either sides of bond breakage
	Ester hydrolysis	ester bond	the heavy atoms on either sides of bond breakage
	Ether hydrolysis	ether bond	the heavy atoms on either sides of bond breakage

Identification of Potential Sites of Metabolism

Reaction type-specific potential sites are defined using SMARTS patterns (Table S6).

Table S6: SMARTS Strings used to Identify Reaction Type-Specific Potential Sites

Reaction Type	SMARTS
epoxidation	$([#6,#7,#16,#15]=[#6,#7,#16,#15]), $([c,n,s,p][c,n,s,p])
hydroxylation	$([#6;H1,H2,H3])
S-oxidation	!$([#16X4](=[OX1])(=[OX1])([OX2H,OX1H0-])[OX2][#6]); !$([#16X4+2]([OX1-])([OX1-])([OX2H,OX1H0-])[OX2][#6]); $([#16])
N-oxidation	!$([NX3](=O)=O);!$([NX3+](=O)[O-]);$([#7])
N-dealkylation	[#7][#6]
O-dealkylation	[#8][#6]
S-dealkylation	[#16][#6]
C-dealkylation	[#6][#6]
oxidative dehalogenation	[#9,#17,#35,#53][#6]
double-, triple- bond formation	$([#6;H1,H2,H3]), $([#7;H1,H2,H3]), $([#8;H1]), $([#16;H1])
quinone formation	$([#8H]cccc[#8H]), $([#8H]cc[#8H])
imine formation	$([#7H]cccc[#7H]), $([#7H]cc[#7H])
quinone imine formation	$([#7H]cccc[#8H]), $([#7H]cc[#8H]), $([#8H]cccc[#7H]), $([#8H]cc[#7H])
quinone methide formation	$([#6H]cccc[#8H]), $([#6H]cc[#8H]), $([#8H]cccc[#6H]), $([#8H]cc[#6H])
imine methide formation	$([#7H]cccc[#6H]), $([#7H]cc[#6H]), $([#6H]cccc[#7H]), $([#6H]cc[#7H])
nitro reduction	$([NX3](=O)=O), $([NX3+](=O)[O-]), $([#7][O]), $([#7+][O-])
carbonyl reduction	$([#6X3]=[OX1]),$([#6X3+][OX1-])
sulfo reduction	$([#16]O)
reductive dehalogenation	[#9,#17,#35,#53][#6]
hydrogenation	[#6,#7,#8]~#[#6,#7,#8]
amide hydrolysis	$([#7][CX3,P,S](=[OX1])),$([CX3,P,S](=[OX1])[#7])
ester hydrolysis	$([#8X2H0,#16X2H0]([#6,#7,#15])[C,P,S,N](=[#8X1,S])), $([C,P,S,N](=[#8X1,#16])[#8X2H0,#16X2H0][#6,#7,#15]), $([OX2H0,SX2H0]([#6,#7])P1SO1), $(P1(SO1)[OX2H0,SX2H0][#6,#7]), $([#9,#17,#35,#53][CX3,P,S](=[OX1])), $([CX3,P,S](=[OX1])[#9,#17,#35,#53])
ether hydrolysis	$([OD2]([#6])[#6]);!$([OX2H0]([#6])[CX3](=O))

Probabilisitic Output at Site- and Molecule-Levels

The model output can be interpreted as probabilities. When we binned class-specific sites by the Phase I prediction score, the proportion of class-specific SOMs in each bin closely correlated with the bin's score (Figure S1). Likewise, when we binned molecules by the Phase I molecule score, the proportion of class-specific metabolized molecules in each bin also correlates with the bin score (Figure S2). Quantitatively, Pearson regression coefficients of site and molecule levels are 0.996 and 0.947, respectively.

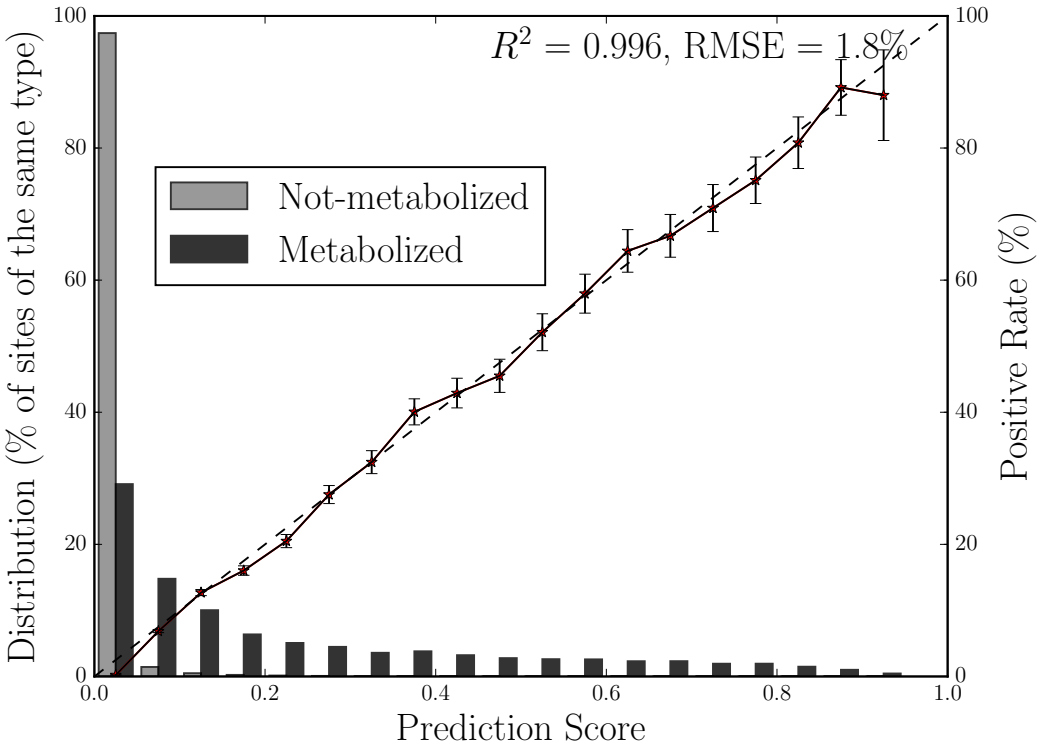

Figure S1: The model makes well-scaled predictions, corresponding to probabilities. The bar graphs plot the distributions of scores across class-specific 691349 metabolized and non-metabolized bonds and lone pair. The solid lines plot the percentage of bonds and lone pairs that are metabolized via a specific Phase I reaction class (using non-normalized frequencies) in each bin. The diagonal dashed lines indicate a hypothetical perfectly scaled prediction. Rainbow XenoSite score has a strong correlation to a perfectly scaled prediction (R^2 value of 0.996 and RMSE of 1.8%). This means that the score is interpretable as the probability that a bond or lone-pair is metabolized via a the corresponding Phase I reaction class.

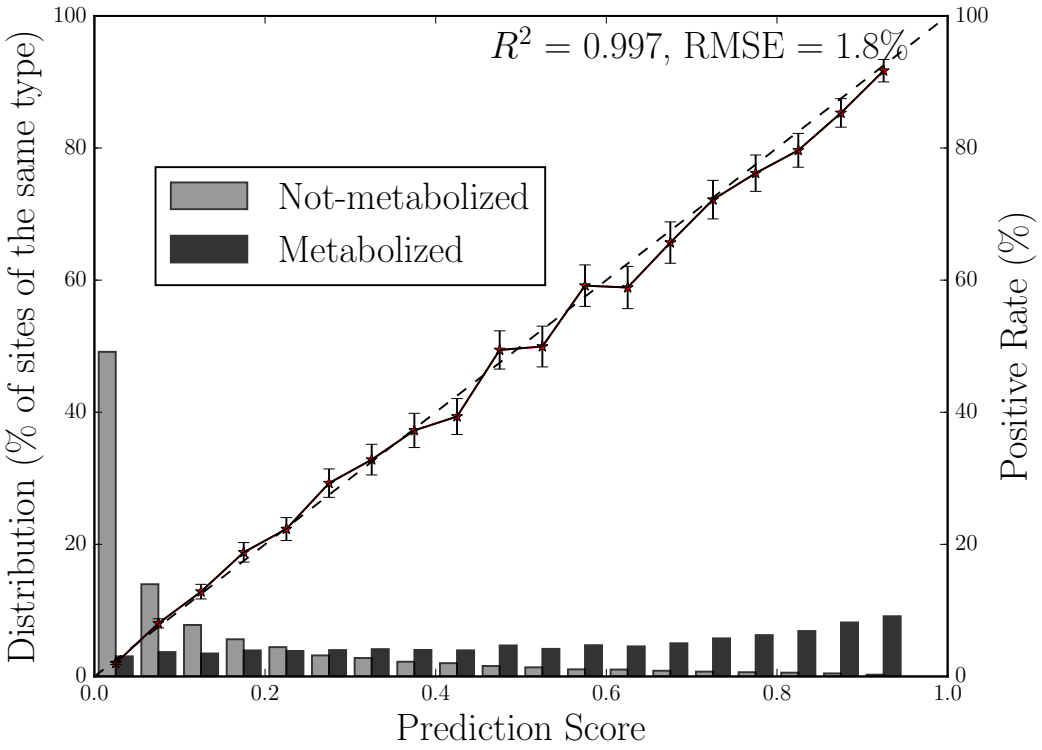

Figure S2: The model makes well-scaled predictions, corresponding to probabilities. The bar graphs plot the distributions of scores across 9674 class-specific metabolized and non-metabolized molecules. The solid lines plot the percentage of bonds and lone pairs that are metabolized via a specific Phase I reaction class (using non-normalized frequencies) in each bin. The diagonal dashed lines indicate a hypothetical perfectly scaled prediction. Rainbow XenoSite score has a strong correlation to a perfectly scaled prediction (R^2 value of 0.997 and RMSE of 1.8%). This means that a class-specific molecule score is interpretable as the probability that molecule is metabolized via the corresponding Phase I reaction class.

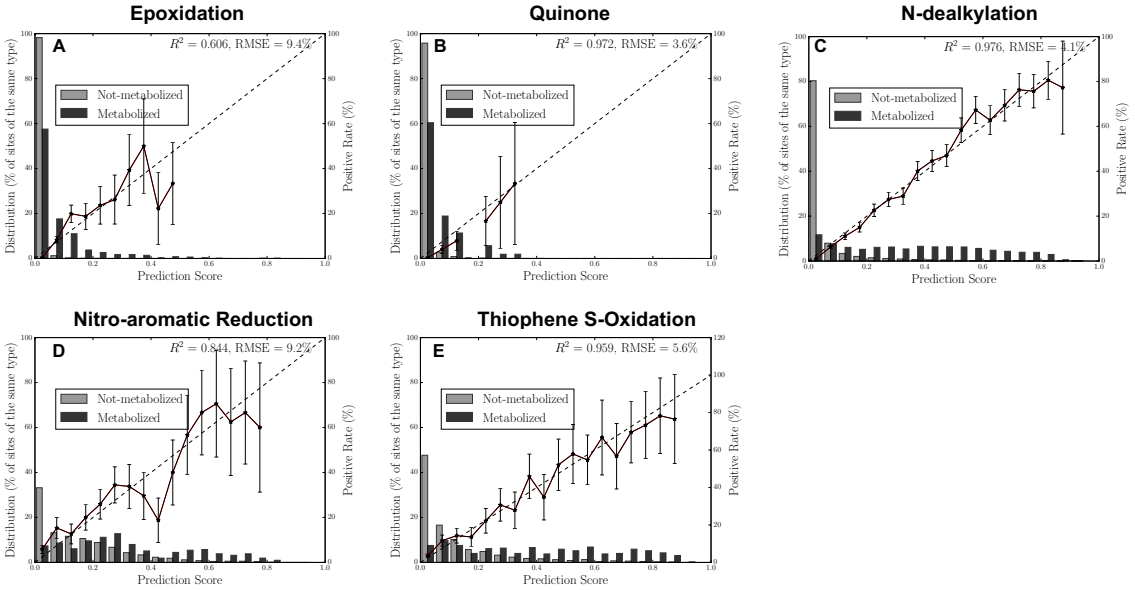

Figure S3: The model makes well-scaled predictions, corresponding to probabilities, for five key reactive metabolite formation reactions: epoxidation (A), one-step quinone formation (B), N-dealkylation (C), nitroaromatic reduction (D), and thiophene S-oxidation (E). The bar graphs plot the distributions of scores across metabolized and non-metabolized potential sites for each reaction type. The solid lines plot the percentage of bonds and lone pairs that are metabolized via the corresponding reaction (using non-normalized frequencies) in each bin. The diagonal dashed lines indicate a hypothetical perfectly scaled prediction. Rainbow XenoSite score has a strong correlation to a perfectly scaled prediction. This means that the score is interpretable as the probability that potential sites is metabolized via the corresponding bioactivation reaction.

Molecule Level Model Selection

Following site-level training, we investigated several methods of discriminating between type-specific metabolized and non-metabolized molecules (Table S7). Each of the tested model takes as input the top N site-prediction scores for each reaction class and molecule-level descriptors and outputs five prediction scores for each molecule. The number of molecule hidden layers (MHL) and L2 regularization coefficient (λ) also vary (Table S7). The best performing model (model 35 in Table S7) was a neural network that takes as input the top five site-prediction scores for each reaction class (25 descriptors in total), two hidden layers, each with 5 hidden nodes, and an L2 regularization of 0.3. This model was chosen as our final model.

Table S7: Parameter Sweep Class Targets (5) Model Structure.

ID	TopN	MHL	λ	SO	UO	DH	RD	HD
0	3	4	1.0	0.764775	0.821044	0.727566	0.879582	0.918844
1	3	4	0.3	0.771873	0.827215	0.704956	0.884103	0.921337
2	3	4	0.1	0.775550	0.831167	0.676844	0.882129	0.919762
3	3	3	1.0	0.773287	0.827940	0.752701	0.877028	0.922459
4	3	3	0.1	0.772452	0.830162	0.702973	0.887219	0.920524
5	5	4	0.3	0.776244	0.833654	0.739328	0.888282	0.922951
6	4	4	1.0	0.776654	0.830816	0.709263	0.882751	0.919626
7	5	4	0.1	0.780520	0.833027	0.767344	0.892992	0.923390
8	3	3	0.3	0.774295	0.834450	0.732586	0.886827	0.923245
9	4	4	0.3	0.782232	0.832007	0.742844	0.883918	0.925234
10	3	2	0.1	0.776196	0.835362	0.735300	0.886593	0.927163
11	6	4	1.0	0.778282	0.835395	0.749546	0.891311	0.923514
12	6	4	0.3	0.782201	0.834775	0.717160	0.894165	0.923016
13	4	3	1.0	0.775083	0.832192	0.739706	0.892056	0.921543
14	4	3	0.1	0.773861	0.834953	0.767055	0.894574	0.923560
15	4	4	0.1	0.772906	0.832040	0.759228	0.889462	0.922763
16	5	3	0.1	0.780201	0.836226	0.754724	0.900464	0.923731
17	3	2	0.3	0.775407	0.836006	0.740295	0.896264	0.922981
18	4	3	0.3	0.778681	0.835636	0.748597	0.894405	0.926030
19	3	2	1.0	0.772494	0.837126	0.735795	0.893673	0.925630
20	5	4	1.0	0.777216	0.835348	0.764207	0.895315	0.922364
21	5	2	0.1	0.777104	0.837707	0.749111	0.899152	0.928100
22	4	2	1.0	0.781098	0.836724	0.765054	0.900211	0.928151
23	6	4	0.1	0.778552	0.836661	0.757454	0.897818	0.927834
24	5	3	0.3	0.778268	0.837809	0.751910	0.900016	0.926298
25	6	3	0.3	0.784404	0.839816	0.720174	0.899714	0.927802
26	5	2	1.0	0.780069	0.835808	0.771889	0.899297	0.925255
27	6	3	0.1	0.781495	0.837520	0.775721	0.894807	0.925889
28	5	3	1.0	0.786632	0.840975	0.770587	0.896965	0.927352
29	4	2	0.1	0.779813	0.838217	0.773614	0.902970	0.923578
30	6	3	1.0	0.781348	0.836329	0.750546	0.900233	0.926382
31	6	2	0.3	0.778025	0.838190	0.747801	0.904363	0.928374
32	6	2	1.0	0.785288	0.841007	0.771881	0.903085	0.929299
33	6	2	0.1	0.779683	0.842008	0.747395	0.906530	0.925378
34	4	2	0.3	0.783752	0.837626	0.766873	0.898889	0.925374
35	5	2	0.3	0.783742	0.839015	0.773277	0.903848	0.927057

Following site-level training, we investigated several methods of discriminating between type-specific metabolized and non-metabolized molecules. Each of the tested model takes as input the topN site-prediction scores for each reaction class and molecule-level descriptors and outputs five prediction scores for each molecule. The number of molecule hidden layers (MHL) and L2 regularization coefficient (λ) also vary. We calculated the molecule AUCs for stable oxygenation, unstable oxygenation, dehydrogenation, reduction, and hydrolysis (SO, UO, DH, RD, and HD) for each model.