Supporting Information: Message Passing Neural Networks Improve Prediction of Metabolite Authenticity

Noah R. Flynn 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@wustl.edu

Contents

Example XenoNets

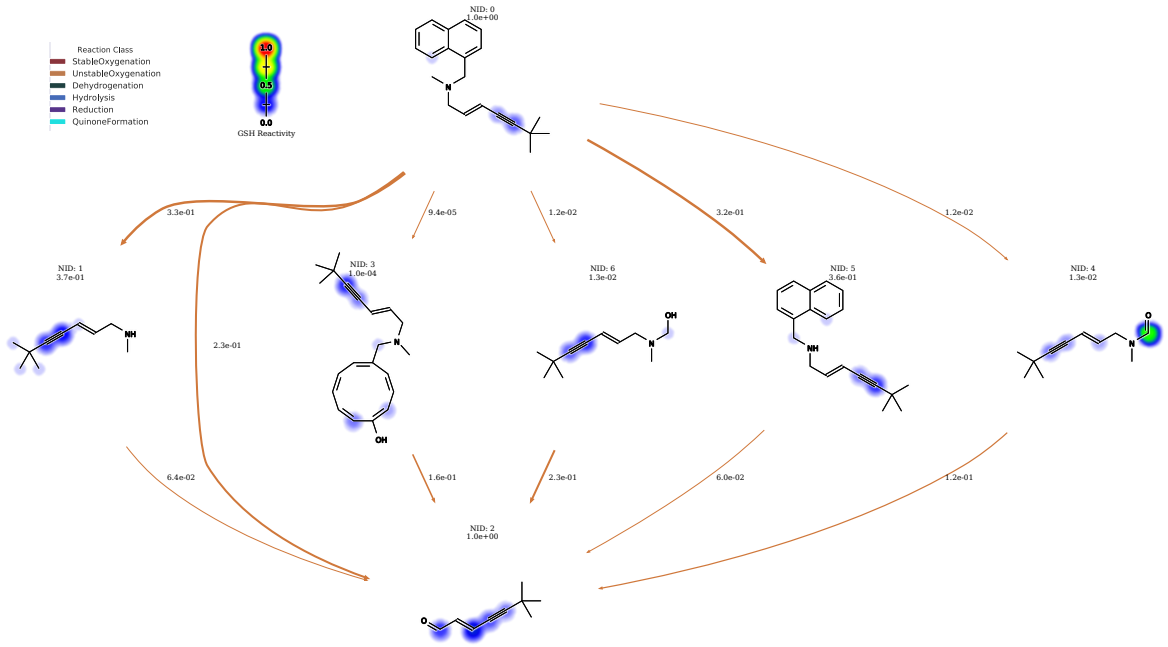

Figure S1. XenoNet identifies pathways between terbinafine (TBF) and TBF-A at a depth limit of 2 and beam width of 5. Several identified intermediate metabolites are consistent with those reported in literature.[1,2] In this example, XenoNet is able to restrict its search space to the N-dealkylation rule set. Scores on each edge correspond to predictions from the phase I SoM model. The heat map overlay corresponds to the site of reactivity model with respect to GSH. Note that metabolite scores assigned to each node were computed by the XenoNet algorithm and not the bidirectional MPNN.

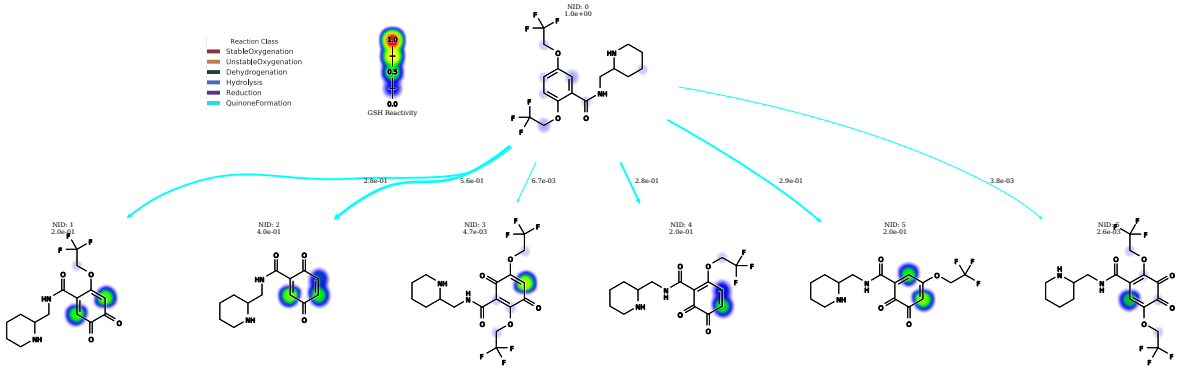

Figure S2. In this example, XenoNet demonstrates its ability to restrict the search space by both a reaction class (quinone formation) and a substructure (sites on or immediately adjacent to the benzene ring) for Flecainide (depth limit 1, beam width 10. Scores on each edge correspond to predictions from the phase I SoM model. The heat map overlay corresponds to the site of reactivity model with respect to GSH. Note that metabolite node scores were computed by the XenoNet algorithm and not the bidirectional MPNN.

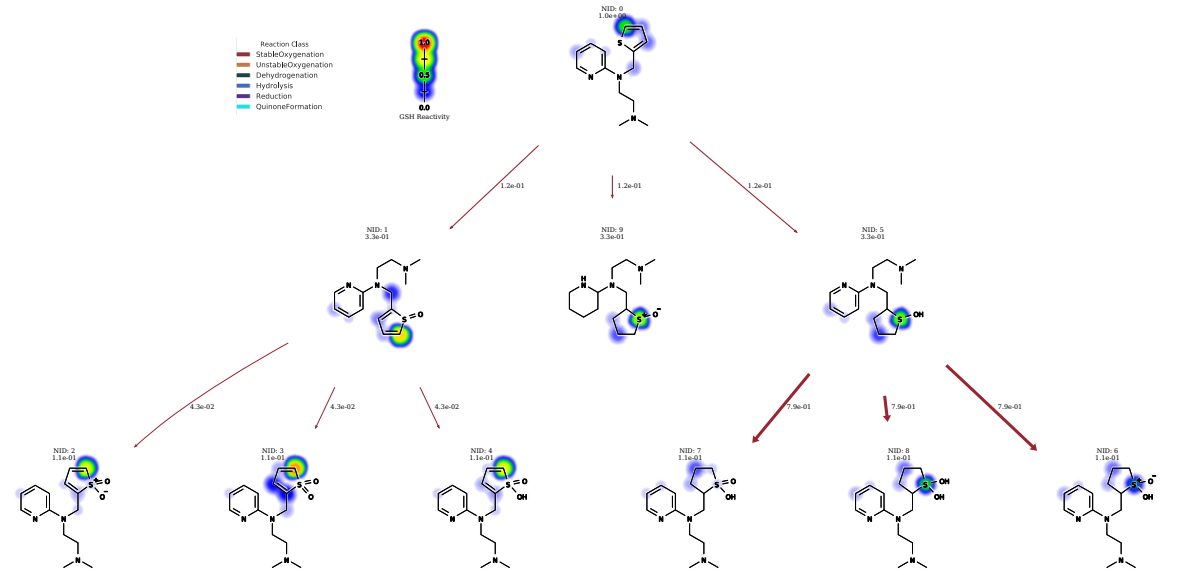

Figure S3. In this example, XenoNet demonstrates its ability to restrict the search space by both a reaction type (S-oxidation) and a substructure (sites on thiophene) for Methapyrilene (depth limit 2, beam width 10. Scores on each edge correspond to predictions from the phase I SoM model. The heat map overlay corresponds to the site of reactivity model with respect to GSH. Note that metabolite node scores were computed by the XenoNet algorithm and not the bidirectional MPNN.

Examples of Intermediates that XenoNet Found and Missed during Network Generation

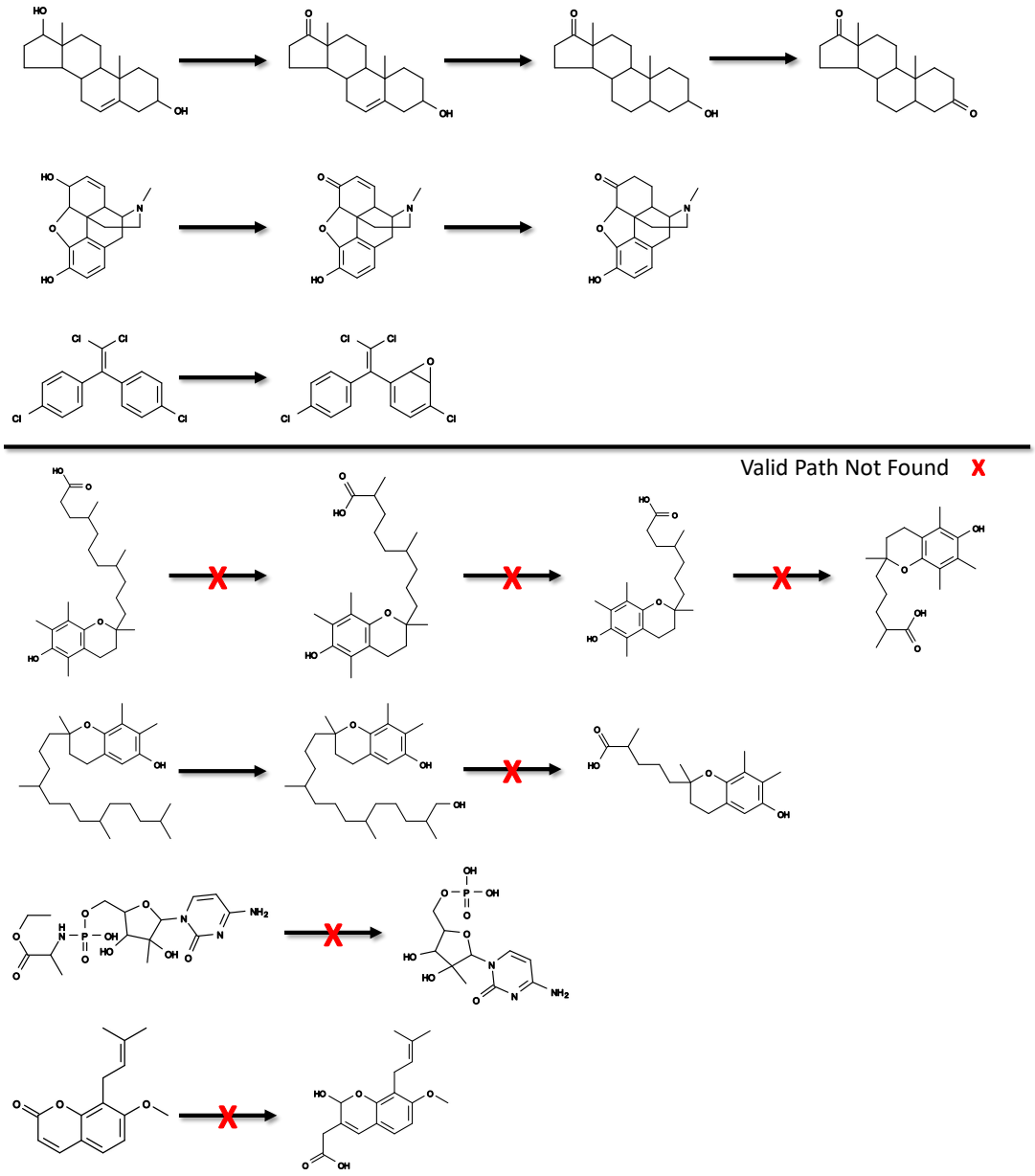

Figure S4. Examples of annotated intermediates that XenoNet succeeded and failed in finding during network generation. The example intermediates were randomly sampled. XenoNet has an easier time of finding intermediates with distinct, explicit steps because it strictly follows a set of transformation rules. The higher the granularity at which an annotated intermediate was recorded in the AMD, the greater the difficulty in XenoNet's ability to infer that intermediate if limiting search to a depth of 3.

Rainbow Model Reaction Class Calibration

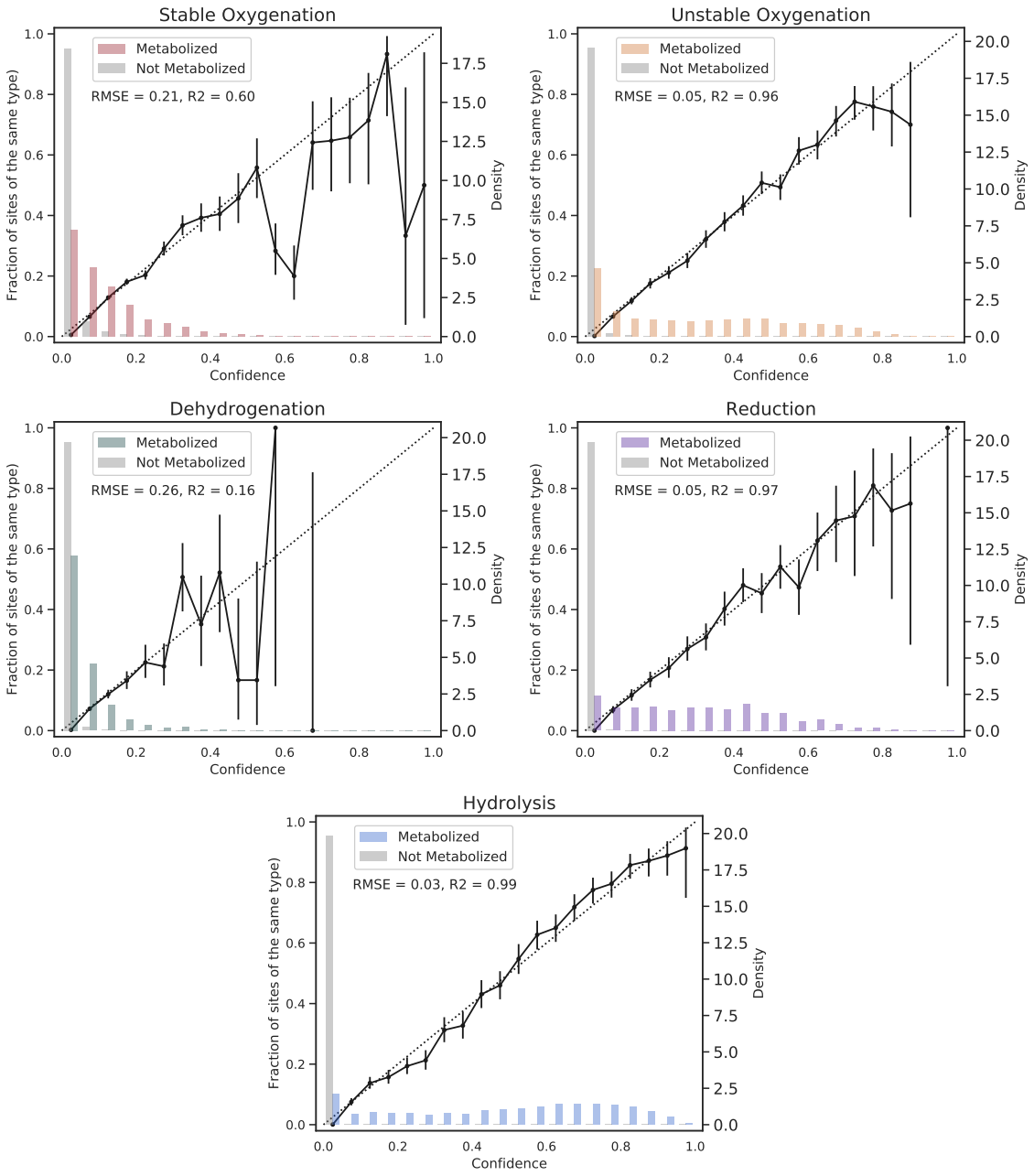

Figure S5. The model has differing levels of confidence in its reaction type specific predictions and incorporation of reaction type information may assist learning relative comparisons between predictions across different reaction types. Individually, site-level predictions specific to unstable oxygenation, reduction, and hydrolysis reaction types are well calibrated. Site-level predictions specific to stable oxygenation and dehydrogenation reaction types are not well calibrated. However, only the hydrolysis site-level predictions are probabilistic, as the unstable oxygenation and reduction site-level predictions do not range from 0 to 1.

Rainbow Model Site-level Reaction Class Thresholds

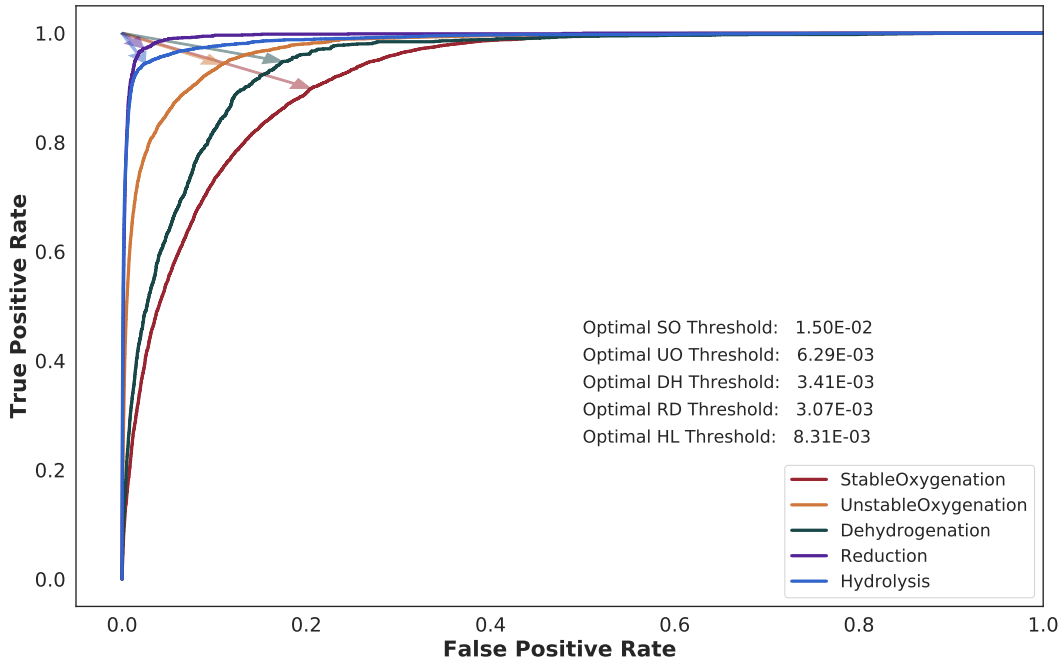

Figure S6. The optimal thresholds for binarizing the Rainbow model's site-level predictions are different for each reaction type. Thus, encoding reaction type information allows the model to adjust influence of predictions based on the Rainbow model's variability in confidence between different reaction types. To define the binarization thresholds, we computed the Youden index as the optimal point on the cross-validated ROC curves computed from the Rainbow model's site-level reaction predictions on its training data for each reaction type.[3,4] The site-level ROC curve is displayed across 10,280, 5,811, 2,794, 1,590, and 3,869 sites of stable oxygenation, unstable oxygenation, dehydrogenation, reduction, and hydrolysis respectively. The light colored arrows for each reaction type point out the corresponding location of the optimal site-level cutoff on that reaction type's respective ROC curve. The optimal cutoff for binarizing site-level predictions was 0.015, 0.0063, 0.0034, 0.0031, and 0.0083 sites of stable oxygenation, unstable oxygenation, dehydrogenation, reduction, and hydrolysis, respectively.

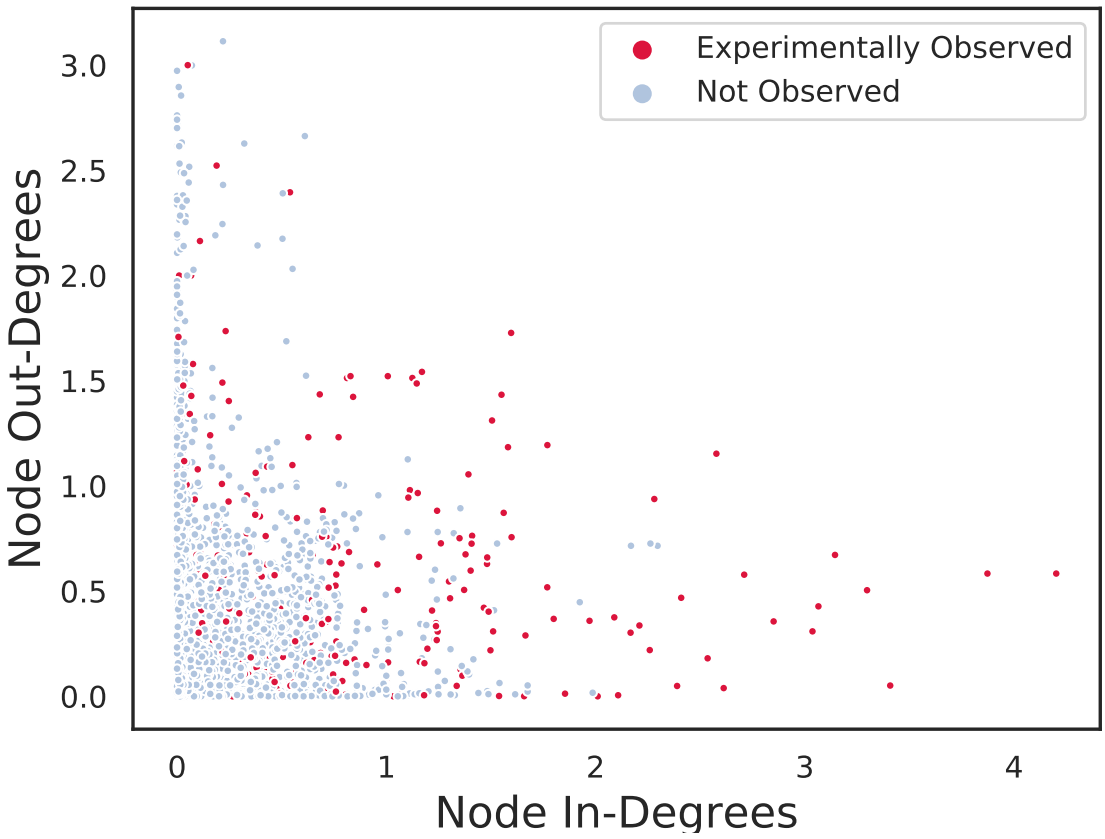

Figure S7. In the absence of node features, node in- and out-degree may be exploited as an informative feature with no additional computational overhead to compute. Though there is no clear demarcation, nodes with high in-degree and low out-degree tend to be experimentally observed and nodes with low in-degree and high out-degree tend to not be experimentally observed. Intuitively, a node with high in-degree is likely to be visited during metabolism and have a better chance of being observed. In contrast, a node with a high out-degree is likely to be transient or not as stable and have a lesser chance of being observed. This instability may cause the intermediate metabolite to be missed during experimental study, but it is possible the intermediate metabolite is short-lived enough to not be pernicious. The in- and out-degrees are calculated for each intermediate node using the weighted paths that enter and exit from the node. If multiple edges exist of same directionality between a pair of nodes, only the highest weighted edge is considered. Start and target nodes are left out as they are not being predicted on and significantly skew the distribution to high out-degrees and in-degrees, respectively.

Distribution of Information Silos for unidirectional MPNN

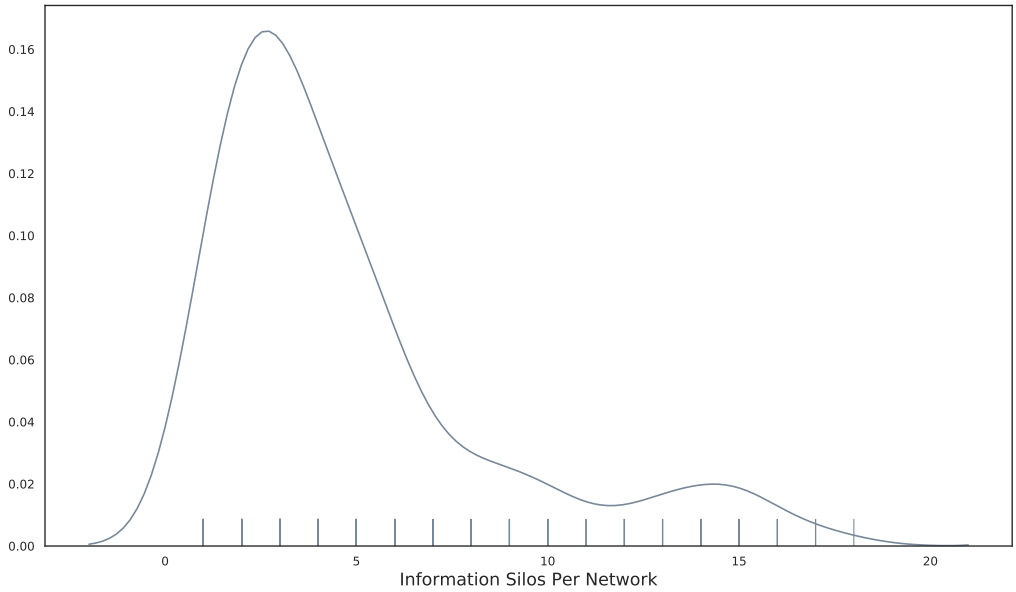

Figure S8. The first variant of the proposed MPNN architecture only aggregates messages from a node's parents and results in an inability for many competing pathways to be considered. Assuming the first variant's strategy, we compute the number of disconnected components that form information silos per network. In this case, we define an information silo as a group of intermediate nodes for which none of the intermediate nodes can be influenced by the hidden embedding state of any other intermediates nodes in the network, regardless of the k-hop neighborhood considered. On average, a network contains 5 information silos (minimum 1, maximum 18). Comparatively, the second variant's strategy results in 0 information silos for all networks.

Hyperparameters for Model Selection

Table S1. Hyperparameters used for model selection during nested cross validation inner loop. For the bidirectional variant used in follow-up case studies, the optimal hyperparameter values are bolded.

	Unidirectional Variant	**Bidirectional Variant**
Number of graph convolutional layers	[1, 2, 3, 4]	[2, **4**, 6]
Graph convolution output embedding size	[8, 16, 32, 64]	[8, **16**, 32, 64]
MLP hidden units	[8, 16, 32]	[8, **16**, 32]
DropOut	[0.0, 0.15, 0.20, 0.25]	[0.0, 0.15, **0.20**, 0.25]
Epochs	200	200
Optimizer	Adam	Adam
Learning rate	[1e-1, 1e-2, 1e-3]	[1e-1, 1e-2, **1e-3**]
Weight decay	[0, 5e-1, 5e-2, 5e-3]	[**0**, 5e-1, 5e-2, 5e-3]
Neighbors aggregation	[mean, max, sum]	[mean, max, **sum**]
JumpingKnowledge aggregation	[–, concatenation, max, LSTM]	[–, **concatenation**, max, LSTM]

Summary of Start-Multitarget Data Set Variants

Table S2. We evaluated the bidirectional variant, which was trained on networks constructed with a depth limit of 3 and no beam width constraint, for inference on networks constructed using different parameter values. The parameters for each variant are a depth limit of 2 with no beam width constraint, a depth limit of 3 and beam width of 5, and a depth limit of 2 and beam width of 5. A depth limit of 2 is selected as this is the lowest depth limit setting than can result in intermediate metabolites. A beam width of 5 is the lowest recommended beam width setting, as any lower does not result in meaningful decreases in time cost while increasing the chance that paths of interest are missed.

	Depth Limit 3	Depth Limit 3 Beam Width 5	Depth Limit 2	Depth Limit 2 Beam Width 5
Number of Networks	311	253	155	151
Number of Experimentally Observed Intermediates	516	302	191	183
Number of Experimentally Unobserved Intermediates	6090	1077	229	176
Mean Number of Nodes per Network	24	8	5	5
Max Number of Nodes per Network	102	16	16	14
Min Number of Nodes per Network	3	3	3	3

Influence of Important Intermediate Cut-off on Extracted Subnetworks

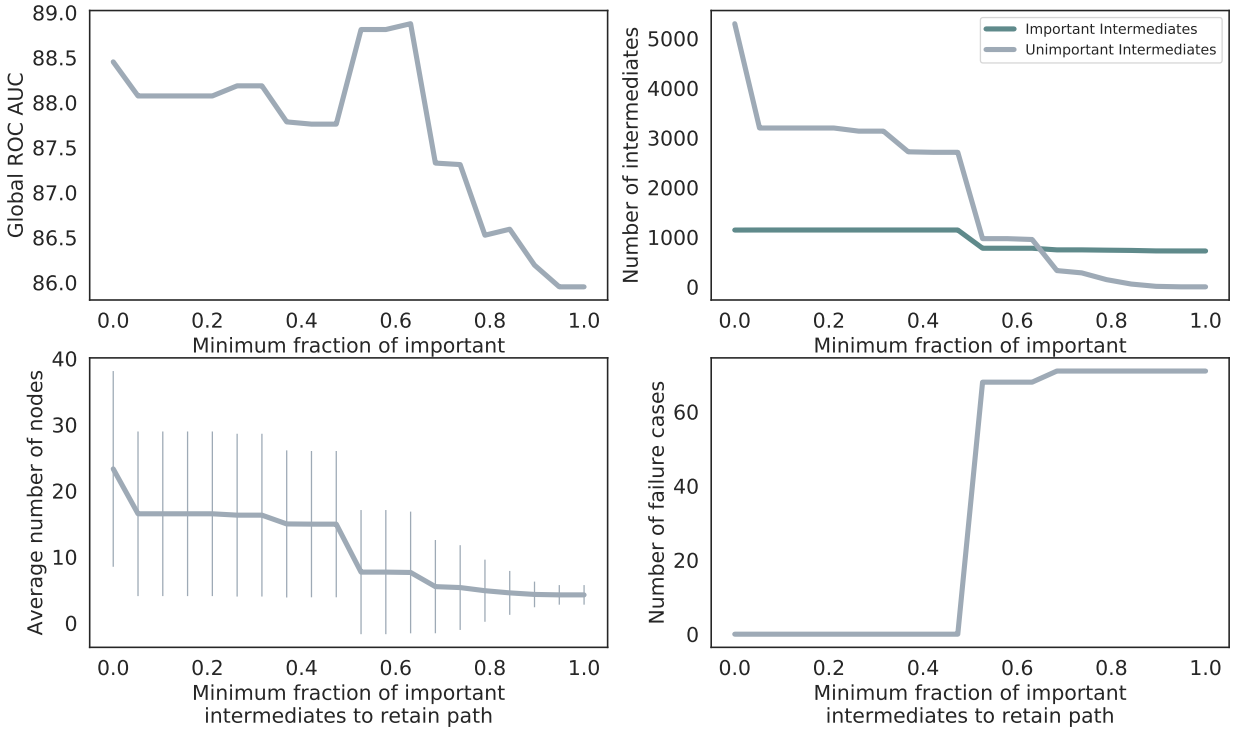

Figure S9. Optimal cut-off value depends on application context, sensitivity to possible false positives, and whether the cut-off results in a failure case. A cut-off of 1 minimized the number of unimportant intermediates to 0 and kept 722 important intermediates. The average network size was reduced from 24 ± 16 nodes to 4 ± 2 nodes. However, 421 important intermediates were excluded and 71 failure cases resulted. Alternatively, a cut-off of 0.5 was the maximum cut-off that kept all $1{,}143$ important intermediates and did not cause a single failure case. The average network size was reduced to 15 ± 11 nodes. The error bars for the average number of nodes represents the standard deviation. Though a 0.5 cut-off results in 0 failure cases, $2{,}664$ unimportant intermediates are retained.

MPNN Training Architectures

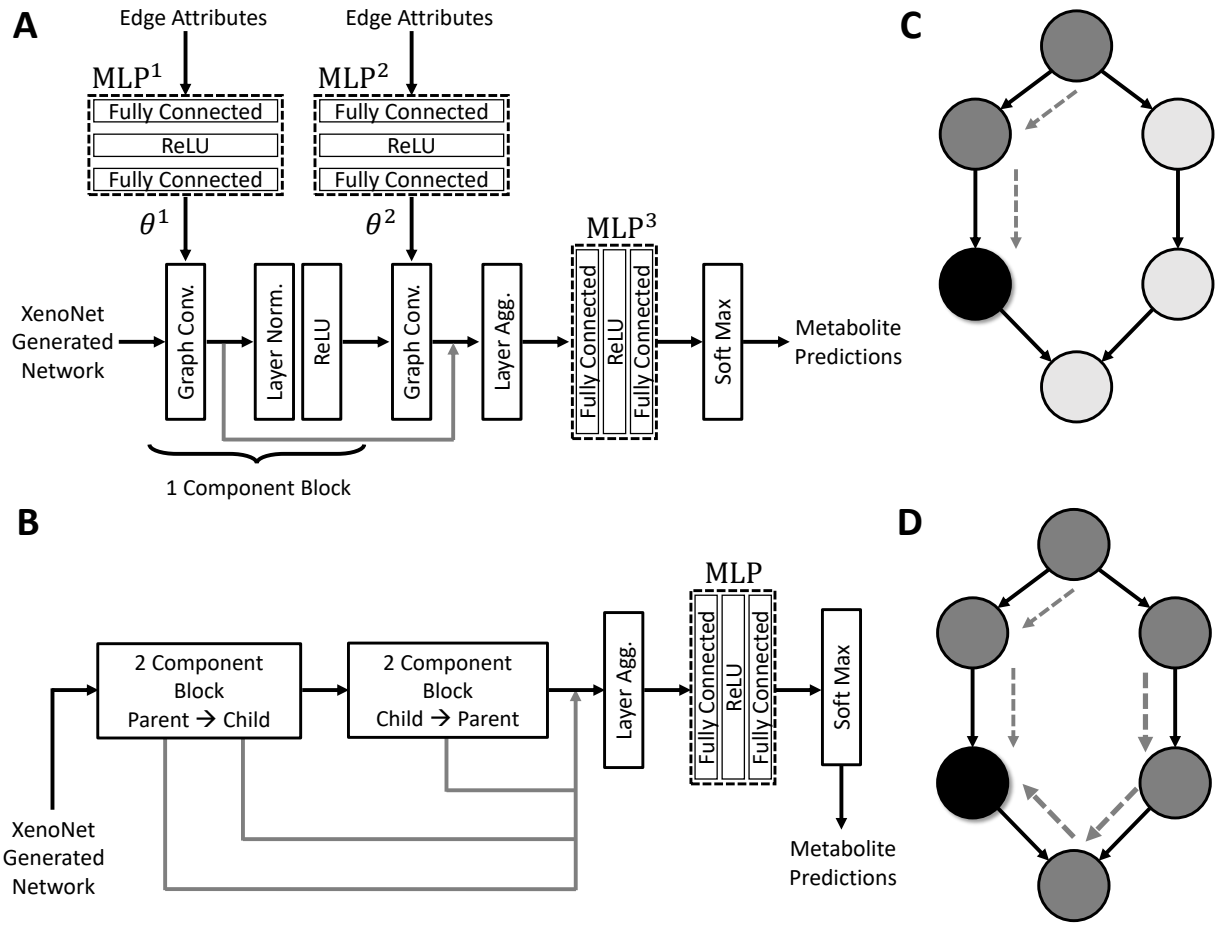

Figure S10. **A)** Example unidirectional metabolite scoring model with 2 graph convolutional layers. The message passing update for each node is calculated using an edge conditioned convolutional layer.[5,6] The output hidden embedding vector is processed by subsequent layer normalization and ReLU activation before being submitted to the next graph convolution layer. Graph convolution layers are aggregated and processed by a 2-layer MLP. The final layer output ranges from zero to one and reflects the metabolic importance of the metabolite. **B)** Example bidirectional metabolite scoring model with 4 graph convolutional layers, in which the second block of components reverses flow of message passing from child to parent nodes. **C)** We consider differences in information flow to the black node when using the unidirectional or bidirectional variants. Assuming the architecture is deep enough to learn from node's in a 2-hop neighborhood, the unidirectional variant allows for information flow from the dark grey nodes to the black node. In the unidirectional variant, the intermediate nodes on the left and right side of the network remain disconnected from each other. **D)** In the bidirectional variant, information is able to flow back upwards to the black node from its child metabolite, which would have been previously exposed to information from its 2-hop parent neighborhood. Thus, the black node is not disconnected from the greater network context and may be able to weigh its likelihood of occurring relative to competing pathways.

References

(1) Barnette, D. A.; Davis, M. A.; Flynn, N.; Pidugu, A. S.; Swamidass, S. J.; Miller, G. P. Comprehensive kinetic and modeling analyses revealed CYP2C9 and 3A4 determine terbinafine metabolic clearance and bioactivation. *Biochem. Pharmacol.* **2019**, *170*, 113661.

(2) Davis, M. A.; Barnette, D. A.; Flynn, N. R.; Pidugu, A. S.; Swamidass, S. J.; Boysen, G.; Miller, G. P. CYP2C19 and 3A4 Dominate Metabolic Clearance and Bioactivation of Terbinafine Based on Computational and Experimental Approaches. *Chem. Res. Toxicol.* **2019**, *32*, 1151–1164.

(3) Youden, W. J. Index for rating diagnostic tests. *Cancer* **1950**, *3*, 32–35.

(4) Dang, N. L.; Matlock, M. K.; Hughes, T. B.; Swamidass, S. J. The Metabolic Rainbow: Deep Learning Phase I Metabolism in Five Colors. *J. Chem. Inf. Model.* **2020**, *60*, 1146–1164.

(5) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. Proceedings of the 34th International Conference on Machine Learning. International Convention Centre, Sydney, Australia, 2017; pp 1263–1272.

(6) Simonovsky, M.; Komodakis, N. Dynamic Edge-Conditioned Filters in Convolutional Neural Networks on Graphs. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2017; pp 29–38.