Supporting Information: XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

Noah R. Flynn,[†] Na Le Dang,[†] Michael D. Ward,[‡] 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 Biophysics, 660 S Euclid Ave, St. Louis, Missouri 63110, United States*

E-mail: swamidass@wustl.edu

Contents

Metabolic Network Data Set

The "Metabolic_Network_Dataset.json" file contains the 17,054 metabolic networks that were derived from initial data reported in the Accelrys Metabolite Database (AMD). The metabolic networks are stored in JSON format and each network is most easily parsed via the NetworkX library in Python. Each metabolic network has its edges annotated with their corresponding reaction AMD registry number, which is denoted by the "RXNREGNO" field. AMD is a licensed database product and this approach is a standard practice in this field because the AMD license does not allow publication of structures.

Metabolite AUC Scores Evaluated on the Full Metabolic Network Dataset

For each network with at least on intermediate metabolite, we compute the AUC using the metabolite scores and their respective labels that indicate whether or not the metabolite was observed in the annotated set. The distribution of the AUCs for XenoNet and a random variant across 710 predicted networks are compared in Figure S1.

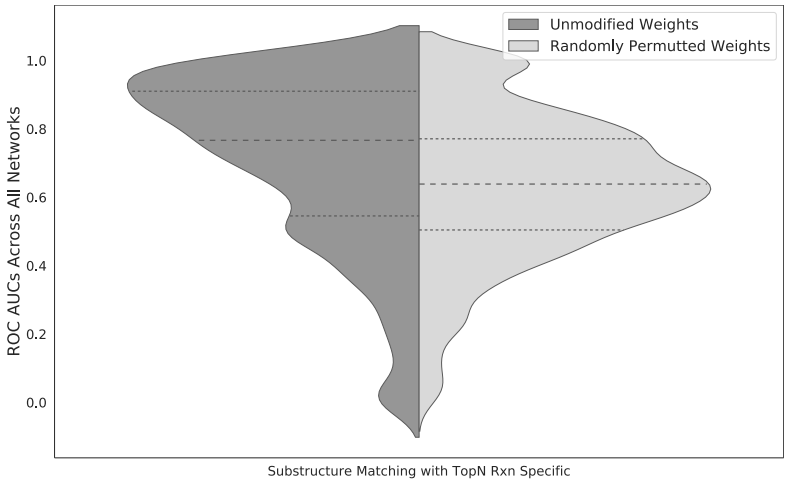

Figure S1: XenoNet's distribution of metabolite AUC scores evaluated on the full metabolic network data set and compared to network's with randomly permuted weights. Each AUC is calculated for an individual network in the data set. Both distributions cover 710 predicted networks that are not empty and have at least one intermediate metabolite in their corresponding annotated network. The middle-dashed line of each distribution designates the mean of the distribution.

Identification of Pathways Linking Problematic Substrate and Metabolite Pairs from the GLORY Test Set

Comparisons were made between XenoNet, GLORY, BioTransformer, and SyGMa using the GLORY test set. During this comparison, XenoNet was initially unable to predict 9 child metabolites of 8 parent molecules. However, XenoNet was able to later recover 5 of the child metabolites when given each child metabolite as a target to explicitly search for. The networks XenoNet produced that link each child metabolite to its corresponding parent molecule are shown in Figures S2, S3, S4, S5, and S6.

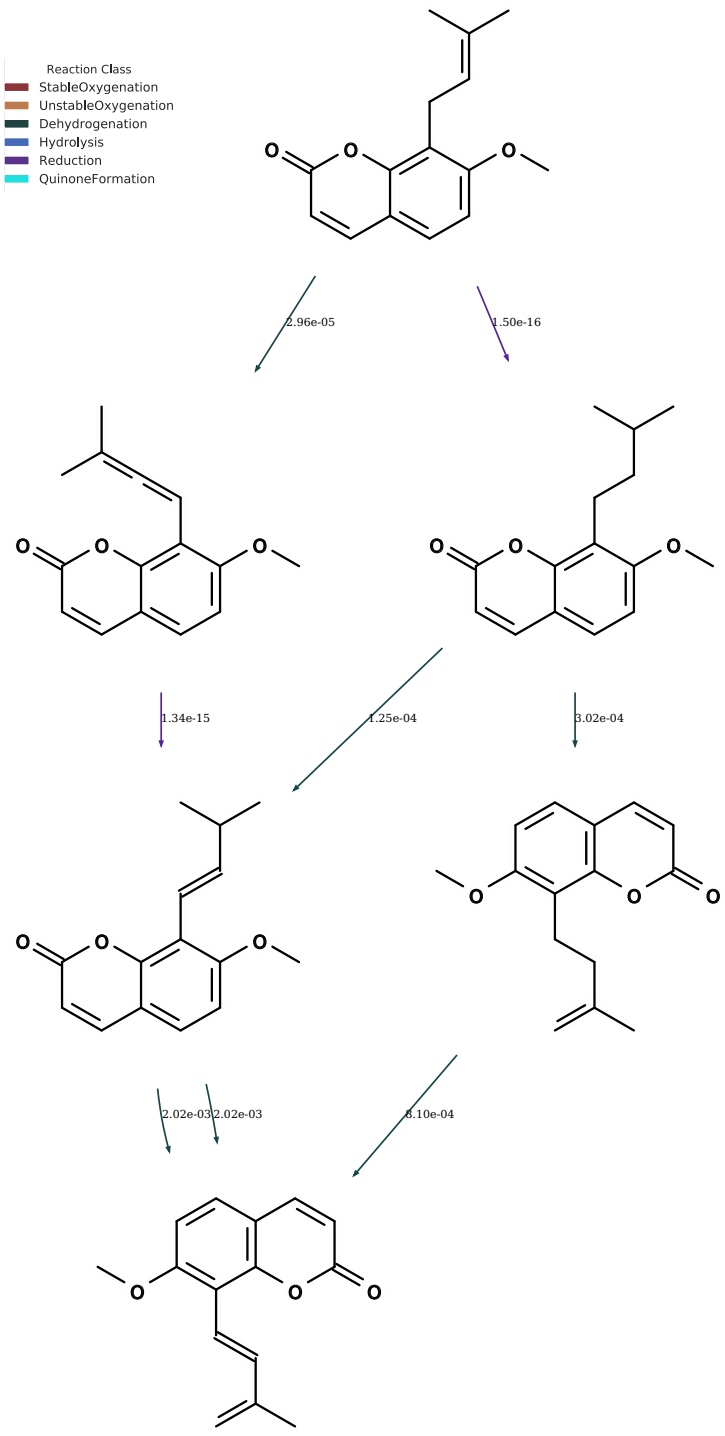

Figure S2: XenoNet identifies pathways between Osthole and cis-Dehydroosthol. Pathways linking the start and target metabolites for this network were not found by XenoNet or the other prior methods during initial comparison.

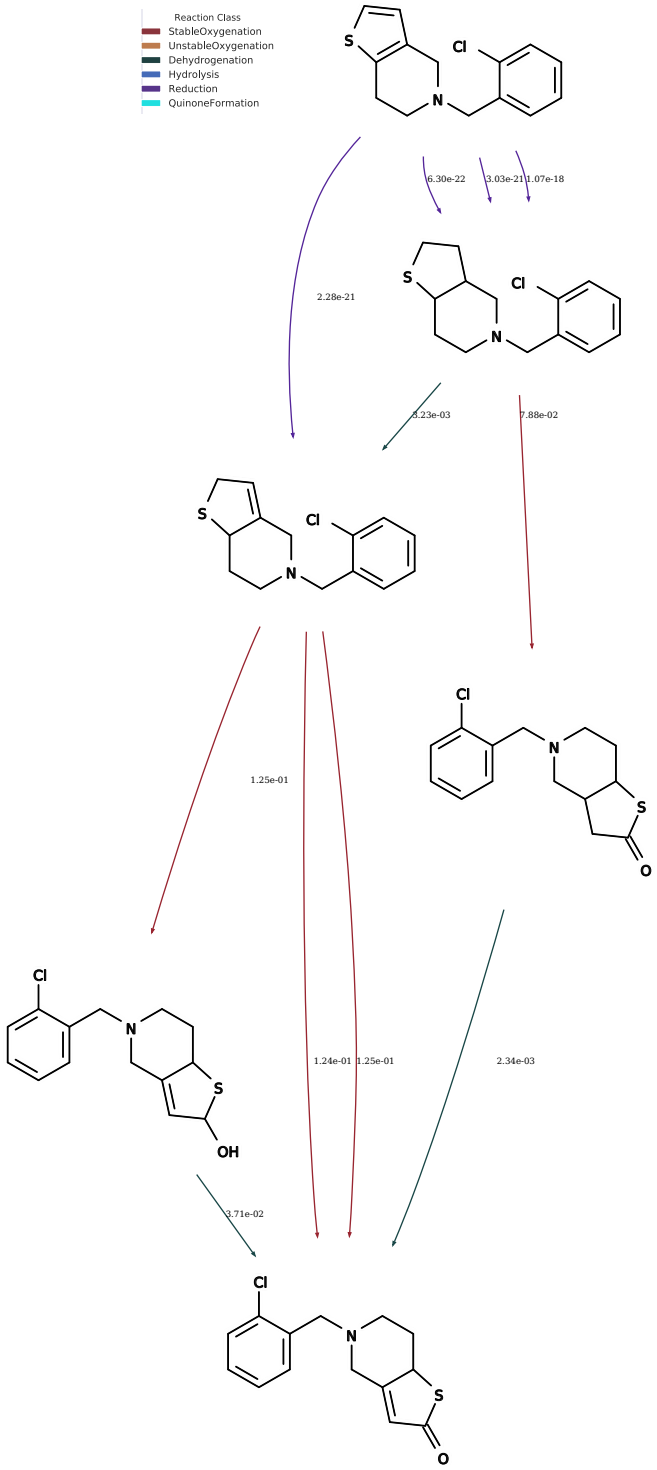

Figure S3: XenoNet identifies pathways between Ticlopidine and 2-Oxoticlopidine. Pathways linking the start and target metabolites for this network were not found by XenoNet or the other prior methods during initial comparison.

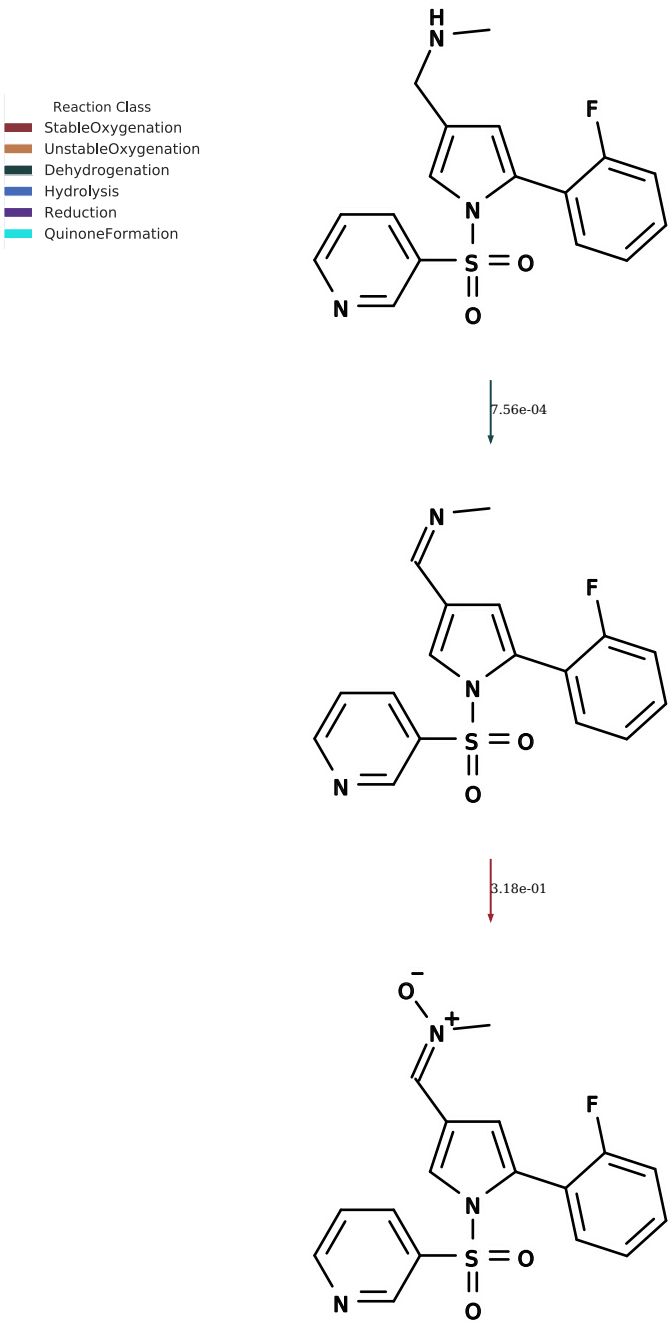

Figure S4: XenoNet identifies a pathway between Vonoprazan and 1-[5-(2-fluorophenyl)-1-(pyridine-3-sulfonyl)-1H-pyrrol-3-yl]-N-methylmethanimine oxide. A pathway linking the start and target metabolites for this network were not found by XenoNet or the other prior methods during initial comparison.

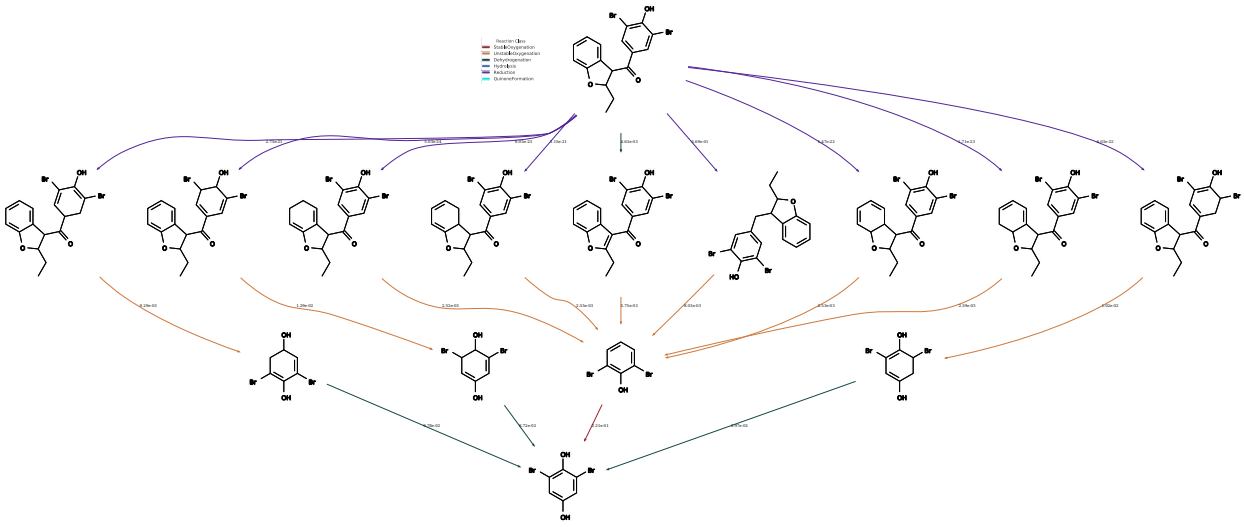

Figure S5: XenoNet identifies pathways between (3,5-dibromo-4-hydroxyphenyl)-(2-ethyl-2,3-dihydro-1-benzofuran-3-yl)methanone and 2,6-Dibromohydroquinone. Pathways linking the start and target metabolites for this network were not found by XenoNet or the other prior methods during initial comparison.

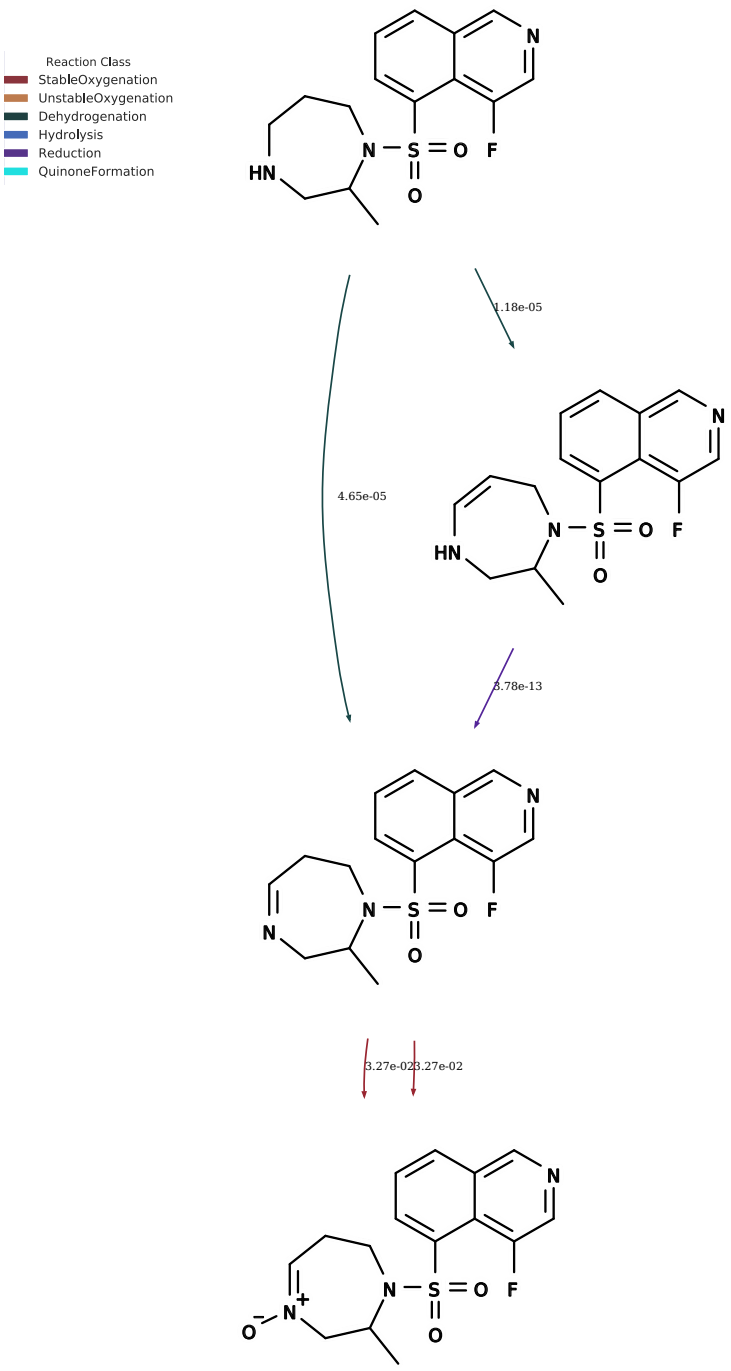

Figure S6: XenoNet identifies pathways between 4-Fluoro-5-[(2-methyl-1,4-diazepan-1-yl)sulfonyl]isoquinoline and 1-[(4-fluoroisoquinolin-5-yl)sulfonyl]-2-methyl-2,3,6,7-tetrahydro-1H-1,4-diazepin-4-ium-4-olate. Pathways linking the start and target metabolites for this network were not found by XenoNet or the other prior methods during initial comparison.

Problematic Substrate and Metabolite Pairs from the GLORY Test Set that no Method Found Valid Pathways for

Comparisons were made between XenoNet and GLORY, BioTransformer, and SyGMa using the GLORY test set. The 4 pairs of parent molecules and their missed metabolites that none of the aforementioned methods found a valid path for are shown in Figure S7.

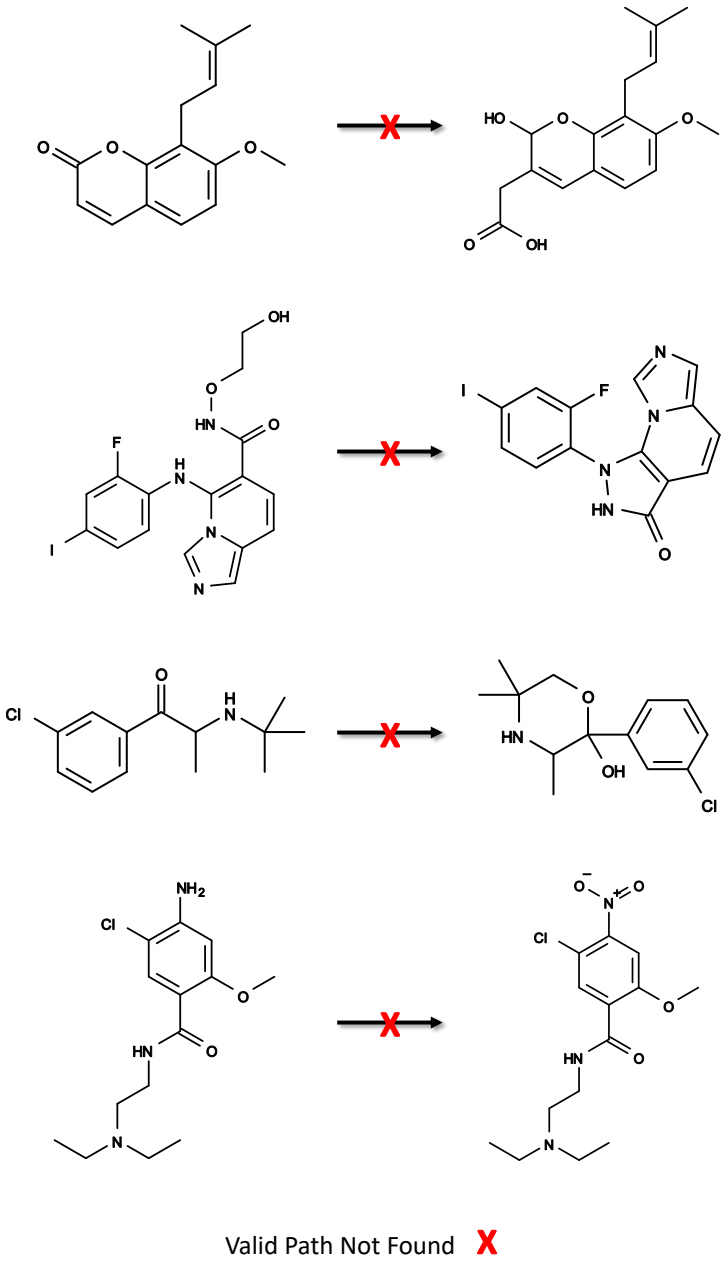

Figure S7: XenoNet, GLORY, SyGMa, and BioTransformer are not able to infer a pathway that links the substrate and metabolite for the 4 cases shown.

Frequency of Intermediate Metabolites Across the Full Metabolic Network Dataset

The number of annotated networks in the Metabolic Network Dataset are plotted against the exact number of intermediate metabolites they contain in Figure S8.

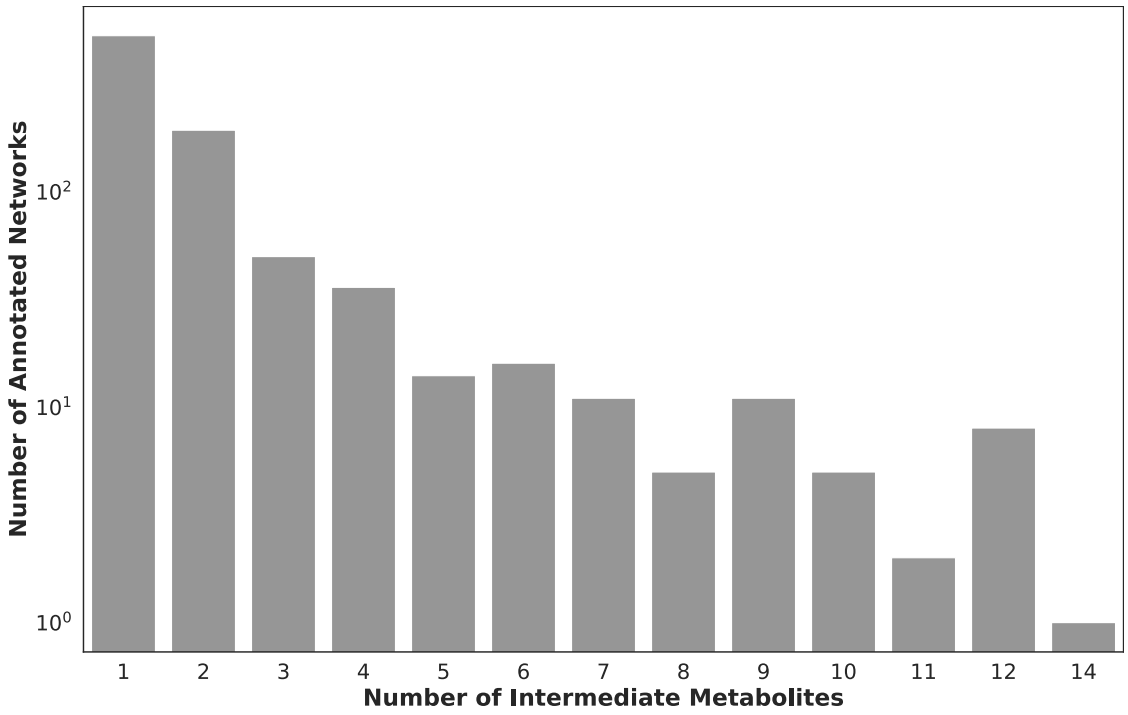

Figure S8: The number of annotated networks in the Metabolic Network Dataset with a specific number of intermediate metabolites. The majority of networks with an intermediate metabolite have either 1 or 2 intermediate metabolites. As the number of intermediate metabolites in a network increases, a smaller number of annotated networks are represented.

Examples of Found and Missed Paths by XenoNet

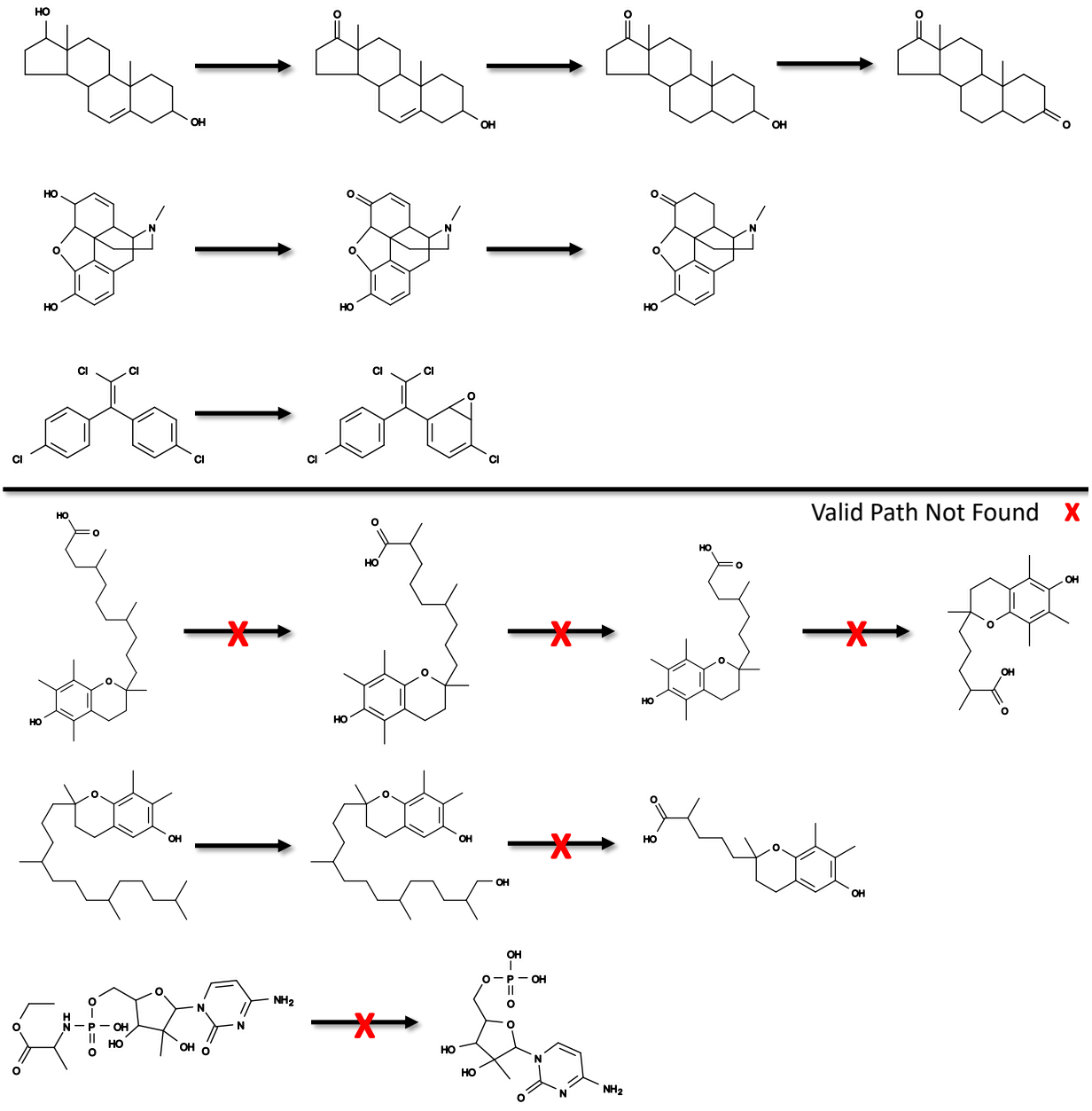

Figure S9: Examples of paths that XenoNet succeeded in finding and failed in finding for paths of length 1, 2, and 3. The example paths for each case were randomly sampled from the total set of paths. XenoNet has an easier time of finding paths with distinct, explicit steps.

Radius Hyperparameter Comparison for Substructure Matching Heuristic

Table S1: Comparison of results on metrics of path recall, intermediate recall, and time cost between substructure matching heuristic variants with radii of 1, 2, or 3. Increasing radius increases the network generation time cost without increasing performance across the path recall and intermediate recall metrics. In general, the weakness of the substructure matching heuristic is that it misses paths where any single step in the path requires a metabolic transformation to occur at a site that does not pass the substructure matching filter. In theory, increasing the radius hyperparameter would increase the number of viable sites of metabolism for a given molecule being processed and diminish the number of missed paths. In practice, the exchange in greater time cost may be the factor that mitigates the advantage that a greater radius hyperparameter would otherwise bring.

	Hyperparameter Value		
Method	Radius 1	Radius 2	Radius 3
Path Recall (Path Length 1)	0.86	0.86	0.86
Path Recall (Path Length 2)	0.39	0.39	0.36
Path Recall (Path Length 3)	0.22	0.21	0.17
Intermediate Recall (Minimum Depth 2)	0.36	0.36	0.34
Intermediate Recall (Minimum Depth 3)	0.31	0.31	0.30
Intermediate Recall (Minimum Depth 4+)	0.22	0.20	0.19
Average Network Generation Time (Minutes)	6	10	18
Networks Fully Generated	722	593	352