JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

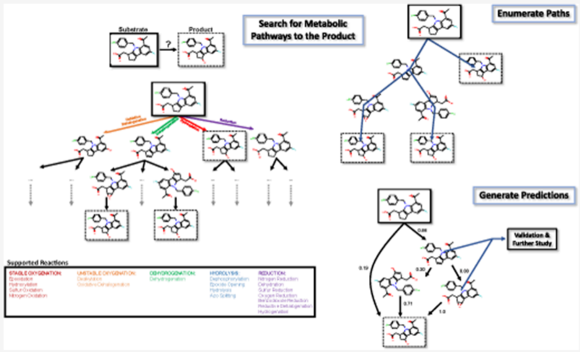
XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

Noah R. Flynn, Na Le Dang, Michael D. Ward, and S. Joshua Swamidass*

Cite This: *J. Chem. Inf. Model.* 2020, 60, 3431−3449 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Drug metabolism is a common cause of adverse drug reactions. Drug molecules can be metabolized into reactive metabolites, which can conjugate to biomolecules, like protein and DNA, in a process termed bioactivation. To mitigate adverse reactions caused by bioactivation, both experimental and computational screening assays are utilized. Experimental assays for assessing the formation of reactive metabolites are low throughput and expensive to perform, so they are often reserved until later stages of the drug development pipeline when the drug candidate pools are already significantly narrowed. In contrast, computational methods are high throughput and cheap to perform to screen thousands to millions of compounds for potentially toxic molecules during the early stages of the drug development pipeline. Commonly used computational methods focus on detecting and structurally characterizing reactive metabolite−biomolecule adducts or predicting sites on a drug molecule that are liable to form reactive metabolites. However, such methods are often only concerned with the structure of the initial drug molecule or of the adduct formed when a biomolecule conjugates to a reactive metabolite. Thus, these methods are likely to miss intermediate metabolites that may lead to subsequent reactive metabolite formation. To address these shortcomings, we create XenoNet, a metabolic network predictor, that can take a pair of a substrate and a target product as input and (1) enumerate pathways, or sequences of intermediate metabolite structures, between the pair, and (2) compute the likelihood of those pathways and intermediate metabolites. We validate XenoNet on a large, chemically diverse data set of 17 054 metabolic networks built from a literature-derived reaction database. Each metabolic network has a defined substrate molecule that has been experimentally observed to undergo metabolism into a defined product metabolite. XenoNet can predict experimentally observed pathways and intermediate metabolites linking the input substrate and product pair with a recall of 88 and 46%, respectively. Using likelihood scoring, XenoNet also achieves a top-one pathway and intermediate metabolite accuracy of 93.6 and 51.9%, respectively. We further validate XenoNet against prior methods for metabolite prediction. XenoNet significantly outperforms all prior methods across multiple metrics. XenoNet is available at https://swami.wustl.edu/xenonet.

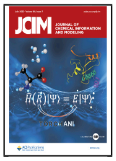


■ INTRODUCTION

Adverse drug reactions are a limiting factor in the development of drugs and their distribution. Approximately 10−15% of adverse drug reactions are categorized as dose-independent, idiosyncratic (idiosyncratic adverse drug reactions—IADRs) and occur at very low frequencies of 1 in 10 000 to 1 in 100 000.[1] As a result, IADRs are notoriously difficult to plan for because they cannot always be treated by simple dose adjustments and can evade detection until the drug has gained increased exposure at large population levels that are not tractable in clinical trials. At least 17% of liver transplant cases and 50% of acute liver failure cases can be traced to IADRs.[2,3] Even with regards to drugs that have been approved, drug toxicity such as those caused by IADRs can result in significant health consequences, including hepatotoxicity and drug-induced liver injury, which can lead to drug withdrawal.[4,5]

Many IADRs are associated with the production of electrophilic reactive metabolites, which may result in harm by conjugating to nucleophilic sites within DNA and protein through a process known as bioactivation. Conjugation to proteins can lead to deleterious alterations in protein structure and folding that provoke immune response,[6−9] and interactions with nucleic acids can alter DNA structure or gene expression in ways that invoke carcinogenicity and teratogenicity.[10,11] Though there are many ways that drug toxicity can arise following bioactivation, most studies focus on detecting precursors to and understanding the bioactivation process itself.

Received: April 11, 2020
Published: June 11, 2020



https://dx.doi.org/10.1021/acs.jcim.0c00361

The risk of bioactivation can be allayed by the early identification of reactive metabolites. Current experimental approaches, such as trapping studies or covalent binding studies, are geared toward detecting reactive metabolites or their conjugated forms.[12,13] However, such methods are resource intensive, time consuming, and can be expensive and biased by experimental design choices, such as the type of trapping agent used. *In silico* methods can operate at a higher throughput and under shorter time scales. A common computational approach is to try to predict the site on a drug molecule where a reactive metabolite might form.[11] Reactive metabolite identification can then be followed up by rational drug modification that avoids potential reactive metabolite formation while retaining the desired pharmaceutical effect of the now altered drug.[9,10,14]

However, there is a dearth of commonly used techniques that can monitor sequential metabolic transformations, and resulting intermediate metabolites, that are required to form a reactive metabolite. This is an important task since the majority of drug molecules require more than one metabolic transformation, and therefore an intermediate metabolite, to form a reactive metabolite.[15] The fulfillment of such a goal would reveal important intermediate metabolites that are required to form reactive metabolites. Consequently, this would open up additional avenues for preventing reactive metabolite formation and subsequent IADRs, as one could modify drug molecules to avoid these intermediates.

A recently published study highlights the need and the ability to use *in silico* methods to identify intermediate metabolites that are necessary to form reactive metabolites.[16] The authors identified an important intermediate metabolite by combining a site of metabolism (SOM) model with a model than can infer metabolite structures. Specifically, they discovered that terbinafine (TBF), an antifungal drug known to cause toxicity,[17] forms a previously unidentified intermediate, desmethyl terbinafine (TBF-D), which has since been shown to be an important precursor in the formation of the reactive metabolite TBF-A.[18] Further experimental investigations of TBF and TBF-D metabolism into TBF-A by P450 isozymes were later able to reveal the degree of involvement of CYP2C9 and 3A4 in TBF's metabolic clearance and bioactivation potential.[19,20] Knowledge of the intermediate, TBF-D, provides a better mechanistic understanding of reactive metabolite formation and could inform potential modifications that would reduce the bioactivation potential of TBF. However, this approach has not been generalized. Instead, the author's approach involved manual application of separate models in an *ad hoc* manner to understand how TBF is bioactivated.

In this work, we explore a generalized approach that combines two types of models: site of metabolism (SOM) models, which identify atoms, or bonds, which are liable to be metabolized, and metabolite structure inference models, which can infer the structures of potential metabolites formed during the metabolism of a given molecule. With regard to SOM prediction, there are several previously published phase-I SOM prediction models that are freely available. SMARTCyp, RSPredictor, SOMP, MetaPrint2D, FAME 3, Site of Metabolism Estimator (SOME), and He et al. are a few examples of the methods available.[21−28] We employ our Rainbow XenoSite model, which was developed previously as part of our ongoing effort to develop a collection of free, usable metabolism and reactivity models, referred to collectively as XenoSite.[29] Our motivation for doing so is that Rainbow XenoSite, compared to the aforementioned models, is the only model that produces well-scaled, probabilistic outputs in a reaction-type specific manner that, in conjunction with knowledge of the site of metabolism, allow for unambiguous inference of the metabolite structures that a molecule is likely to form.[30] Furthermore, our prior work has also shown that Rainbow XenoSite covers the largest proportion of known human phase-I metabolic reactions and its coverage of reaction types includes many important bioactivation reactions not covered by the other models.

Comparatively, metabolite structure inference models are a less explored space than SOM models. Previously published and freely available methods worth noting include BioTransformer, RD-Metabolizer, SyGMa, and GLORY.[31−34] The earliest approach of those listed, SyGMa, presents a rule-based method for predicting the potential child metabolites produced from a given parent metabolite.[33] SyGMa's reaction rules cover both phase 1 and phase 2 metabolism and are augmented by probability scores to allow ranking of the predicted metabolites. SyGMa's probability scores for reaction rules are derived from statistical analysis on a large data set of experimental metabolic reactions. A second method, BioTransformer, combines a knowledge-based approach with a machine learning approach to predict small molecule metabolism across a wide range of contexts, including CYP450-mediated metabolism.[31] The former includes the use of MetXBioDB, a biotransformation database of annotated, experimentally derived metabolic reactions that informs a reaction knowledge base for metabolite prediction. The latter involves use of CypReact for CYP specificity prediction.[35] In contrast, GLORY accomplishes metabolite prediction without reliance on metabolic reaction data sets and instead implements rules derived from scientific literature and chemistry knowledge.[34] GLORY also applies FAME 2 to enable filtering out potential false-positive predictions and ranking of predicted metabolites by their likelihood of occurrence.

We employ our own metabolite structure inference model, called the Metabolic Forest, which we have previously established and validated as an accurate tool for predicting metabolic structures in comparison against RD-Metabolizer.[36] Similar to the aforementioned methods, Metabolic Forest uses a rule-base approach for metabolite inference. However, an elusive problem common to all of the above methods is the generation of large numbers of false-positive metabolite structures. Use of an SOM model in tandem with a metabolite structure inference model can remedy this by allowing early filtration of metabolites that would result from low-value predictions and a basis for ranking the remaining predicted metabolites in accordance to their likelihood of being legitimate.

In this work, we propose an approach that combines our Rainbow XenoSite and Metabolic Forest models to build networks of metabolic transformations that include reaction-specific SOM predictions and metabolism structure inference across a set of phase-I metabolism rules. We also validate how this method, referred to as XenoNet, can be used to infer intermediate metabolites precluding formation of a given target metabolite known to eventuate from a given parent molecule. Furthermore, we can use Rainbow XenoSite to determine the probability of each metabolic transformation and then calculate the likelihood of observing a given metabolic pathway. Briefly,

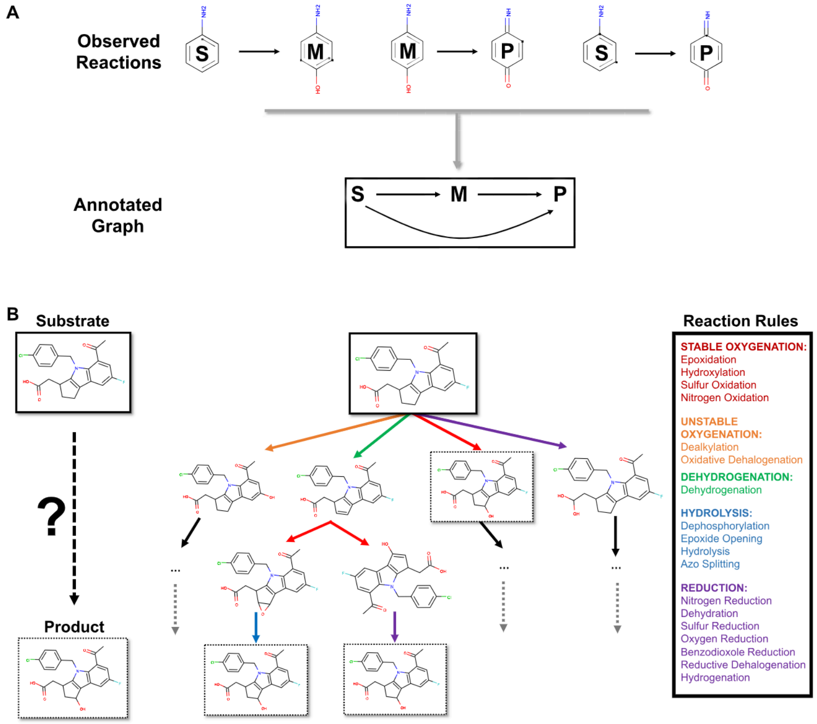


Figure 1. Metabolic network data set construction and overview of XenoNet. (A) Multiple experimentally observed reactions from the AMD can be linked to an annotated network. Using these annotated networks, we can evaluate how well different metabolic network-generating algorithms can infer observed intermediate metabolites. (B) XenoNet is a metabolic network predictor that, given a substrate and a target product as inputs, can infer the metabolic pathways connecting two input molecules and the corresponding likelihood of each pathway. In XenoNet, the Metabolic Forest algorithm is applied iteratively to generate a tree of potential pathways that span multiple metabolic transformations. During construction of this tree, pathways between the starting molecule and a target molecule can be enumerated. The likelihood of each step in a pathway can then be computed from the five-color predictions by Rainbow XenoSite. The five colors are used to allow for ease of use when visualizing the networks that XenoNet generates. Each of the major groups corresponding to one of the five colors is further subdivided into more nuanced reaction-type classes that are used to annotate the edges in the generated network. These more detailed edge annotations are accessible through the XenoNet network object. XenoNet's predicted metabolic pathways are stored using a graph-based data structure where each molecule is a node and each metabolic transformation is an edge.

we also compare XenoNet's ability to infer known metabolites in comparison to GLORY, SyGMa, and BioTransformer.

■ DATA AND METHODS

Metabolic Network Data Set. To construct our metabolic network data set, we used a previously described data set that contains 20 736 *in vitro* and *in vivo* human phase-I reaction records filtered from the Accelrys Metabolite Database (AMD), a literature-derived database.[30] Each reaction record was classified into one of the five phase-I metabolism classes: stable oxygenation, unstable oxygenation, dehydrogenation, hydrolysis, and reduction reactions and manually labeled with the site of metabolism (SOM). The final data set contains 10 280, 5811, 2794, 3869, and 1590 stable oxygenation, unstable oxygenation, dehydrogenation, hydrolysis, and reduction reactions, respectively. Due to their low prevalence, some phase-I reaction types such as tautomerization, isomerization, rearrangement, radical formation, hydration, deacylation, denitrogenation, and decarbonylation were not included in our data set.

These reaction records were converted into graph representations where the nodes represent the substrates and the products and the edges represent metabolic transformations. These graphs were then linked together to construct networks based on shared nodes—identical chemical structures. Figure 1A illustrates an example of how these records are collated into an annotated graph. In this example, one record may indicate that a substrate molecule (S) undergoes a reaction that generates a metabolite (M). A second record may indicate that (M) undergoes a reaction to generate another metabolite (P). A third record indicates that (S) is directly metabolized into (P). Therefore, these three records can be linked to show that the substrate (S) is connected to the downstream metabolite (P) through the metabolite (M), serving as an intermediate node. At the end of the network collation process, all networks that are induced subgraphs of another network are removed. The final data set contains 17 054 annotated metabolic networks with at least one direct path connecting each substrate molecule to its recorded product. Each of the networks represents a unique substrate−product pair. Approximately 91% of the paths across all annotated networks require three or fewer metabolic transformations. Although we cannot share the exact chemical structures from the proprietary AMD, we provide the AMD reaction records for each metabolic network in our data set in the "Metabolic_Network_Dataset.json" file.

Metabolic Network Generator XenoNet. We built XenoNet, a metabolic network predictor that, given a substrate and a target product as inputs, can infer the metabolic pathways connecting two input molecules and the correspond-

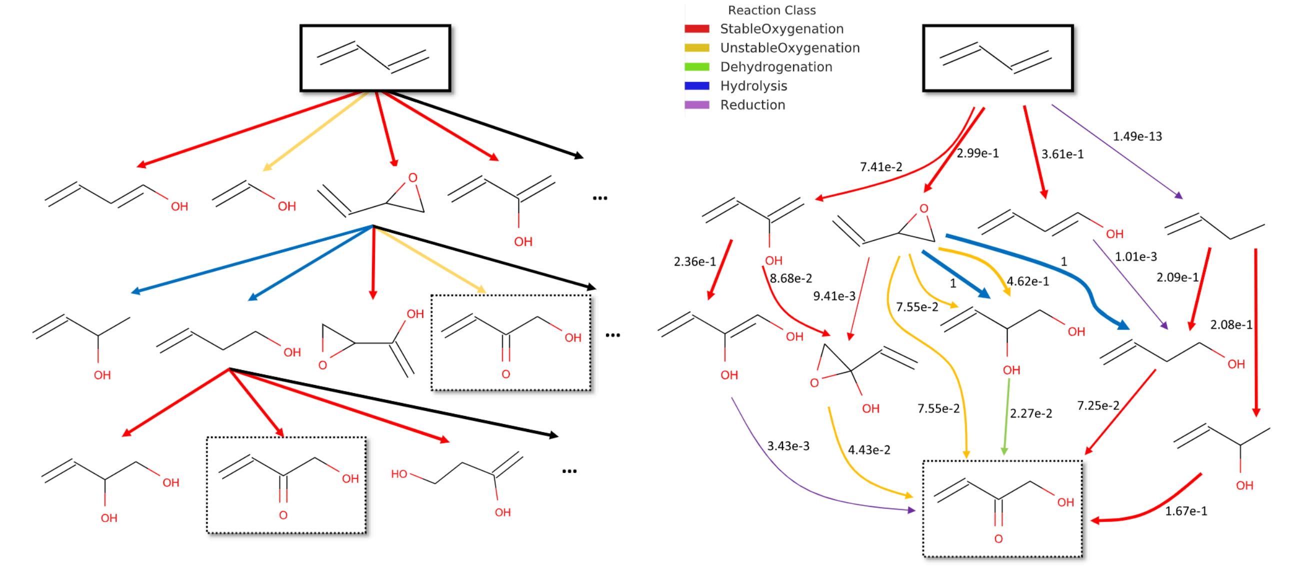


Figure 2. XenoNet can infer the metabolic pathways connecting a substrate and a product from their structures in three steps. First, using Metabolic Forest,[36] a depth-first search scans across a space of possible metabolite structures for paths that terminate at the given product metabolite. Second, the discovered paths are used to construct a new metabolic network with the input substrate and product as terminal nodes. Third, Rainbow XenoSite[30] yields predictions on the metabolic transformation edges of the constructed metabolic network. As an example, we show here how the process works when the model is given 1,3-butadiene (in solid-line box) and hydroxymethylvinyl ketone (in dash-line box) as its substrate and product pair inputs. The search starts at 1,3-butadiene and explores as far as possible along a branch of metabolic transformations up to a limited depth, before backtracking to continue the search along other branches (left). Only pathways that go from 1,3-butadiene to hydroxymethylvinyl ketone are retained to construct a new metabolic network (right). Once all metabolites and directed edges linking them are in the network, the network is passed into Rainbow XenoSite to compute the likelihood for each metabolic transformation, shown as the numbers next to the edges.

ing likelihood of each pathway. Two previously developed, in-house models are combined and augmented to develop XenoNet. The first is Rainbow XenoSite, a deep learning phase-I metabolism model that, given a molecule as input, can accurately predict sites of metabolism (SOMs) for each of the 20 phase-I reaction types via five reaction classes.[30] In the current work, we utilize the same set of transformation rules that define the reaction classes and reactions types in Rainbow XenoSite. The second is the Metabolic Forest, which takes the substrate molecule as input and infers possible metabolite structures that are one metabolic step away.[36] In addition to inferred metabolite structures, the Metabolic Forest also outputs the specific SOMs and the corresponding metabolic transformations that act on those SOMs to produce the inferred metabolite. Metabolic Forest infers metabolite structures of a given substrate using breadth-first searches and transformation rules encoded through a combination of reaction SMARTS, resonance pair rules, and resonance structure rules.

Briefly, in XenoNet, the Metabolic Forest algorithm is applied iteratively to generate a tree of potential pathways that span multiple metabolic transformations (Figure 1B). During construction of this tree, pathways between the starting molecule and a target molecule can be enumerated. Through the application of Rainbow XenoSite, the likelihood of each step in a pathway can then be computed. XenoNet's weighted metabolic pathways are stored using a graph-based data structure where each molecule is a node and each metabolic transformation is an edge. Thus, unlike the Metabolic Forest, XenoNet not only infers metabolite structures but also predicts the likelihood of the metabolic transformation that the input molecule undergoes to form the inferred metabolite structures, whereas the Metabolic Forest outputs a list of the inferred metabolite structures it enumerates, XenoNet encodes its inferred metabolite structures, along with edge-level data such as SOM predictions, in a network object. This network object supports additional functionality, such as the ability to compute the likelihoods of paths and metabolites in the network. In addition, the network object readily supports conversion to a NetworkX MultiDiGraph class object for further utility. Finally, XenoNet incorporates a depth-first search strategy that yields much lower memory requirements compared to the breadth-first search strategy employed by Metabolic Forest. XenoNet was implemented in Python with the 2018.09.03 release of RDKIT.[37] Seven different variants of XenoNet were implemented, and the best variant was selected as the final version. The detailed descriptions of each step are in the following sections.

The Naive XenoNet Variant. The first XenoNet variant is a naive model that employs a brute-force approach to network construction, which enumerates all possible paths between the starting molecule and the target molecule. Given a query consisting of a substrate–product pair of molecules, Metabolic Forest is iteratively applied to construct a metabolic tree of successive intermediates in a depth-first manner until a pathway connecting the queried pair is found or the depth limit is reached. Next, all discovered paths from the substrate to the target product are used to construct a new metabolic network. Once the search completes, and the structures in the network are known, the whole network is given as input to Rainbow Xenosite. In this step, each transformation edge is given a prediction score depending on the reaction class that it belongs to. As an example, we show how the process works

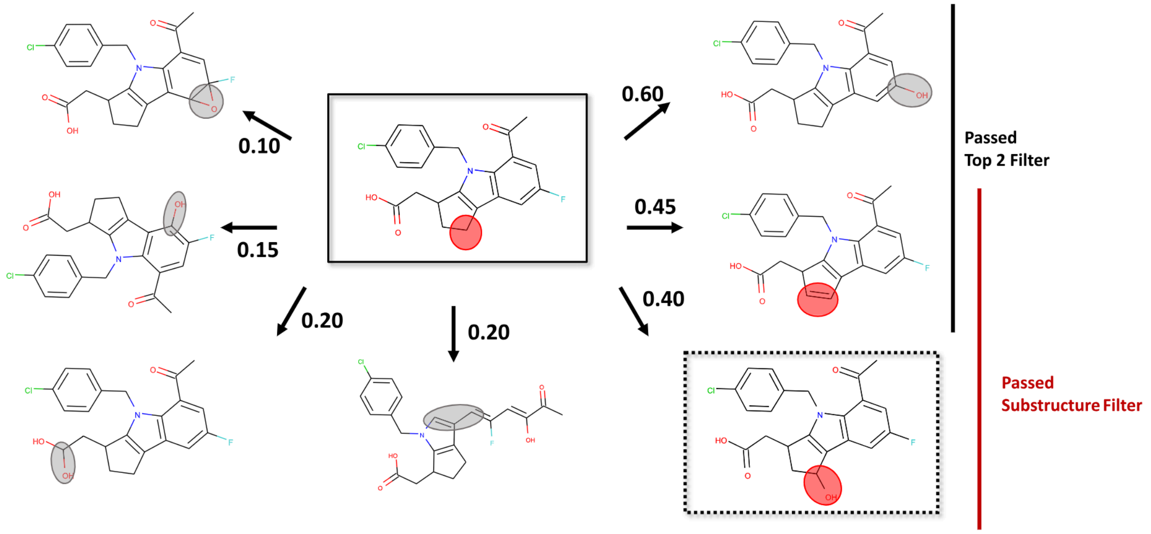


Figure 3. Top-*N* and substructure matching heuristics. To limit the branching factor of potential metabolic trees, we use top-*N* and substructure matching heuristics. As an example, let us consider a pair of a substrate (in solid-line box) and a target product (in dash-line box), as shown above. For the first generation of metabolites, XenoNet discovers seven structures, one of which is our target product. However, only the target product and structures that meet top-*N* and/or substructure heuristics would be considered in the next step of the search and the remaining metabolites would be ignored. The top-*N* heuristic filters for metabolite structures that receive *N* highest scores. The numbers next to the arrows are scores assigned by Rainbow XenoSite for the corresponding metabolic transformation. Only two metabolites with the scores of 0.60 and 0.45 pass the Top-2 filter. The substructure matching heuristic filters for metabolite structures produced via transformation at sites expected to lead to the target product. In the substrate structure, the site of the structural difference between the substrate and target compound, i.e., the site of metabolism, is circled in red. In each metabolite structure, the site of structural difference between the metabolite and the substrate is circled in red if it contains the site of metabolism and gray if it does not. Only two metabolites with red circles pass the substructure matching filter. Different combinations of top-*N* and substructure matching heuristics are used to construct six XenoNet variants.

when the model is given 1,3-butadiene and hydroxymethylvinyl ketone as its substrate and product pair inputs (Figure 2).

In addition to detecting and removing cycles, such as repeatedly adding and removing a hydroxyl group, from the network, we prevent wasteful computations by using two parameters: the depth limit and the time limit. The depth limit parameter specifies the maximum length of the searched paths. The time limit parameter specifies the maximum amount of time that the network generation cannot exceed. During path enumeration, we construct a separate metabolic tree to keep track of all pathways that connect the queried pair. If the metabolic network exceeds its allotted time, the current state of the partially generated network is saved. This brute-force approach is only a baseline to demonstrate the necessity of clever algorithms to make the problem tractable.

Notably, network generation does not require a defined product molecule for the search to terminate. Instead, it can also function in cases where only a starting molecule is given. We can input only a substrate molecule and XenoNet will still generate all paths of metabolic transformations within the limits of its phase-I rule set and defined user parameters such as the depth limit. Thus, we can also compare our method to other tools that only try to infer metabolites yielded from a starting molecule and do not try to enumerate paths with a defined end state.

Heuristic XenoNet Variants. In addition to the naive variant, we develop six other XenoNet variants. To limit the branching factor of potential metabolic trees, we use different combinations of a top-*N* heuristic and a substructure matching heuristic in six other variants of XenoNet (Figure 3).

The top-*N* heuristic drives the metabolic network to explore only metabolic transformations whose probability of occurrence is among the highest. Using Rainbow XenoSite, we calculate the probabilities of all metabolic transformations that the parent metabolite may undergo and only generate the child metabolite structures that correspond to the top-*N* metabolic transformations.

The top-*N* heuristic limits the tree to grow as x^N, where x is the number of steps in a pathway and N is the number of metabolic transformations that can be explored for each molecule. Without this heuristic, N can effectively be on the order of hundreds of metabolic transformations. Furthermore, the top-*N* heuristic has two forms: reaction-agnostic top-*N* and reaction-specific top-*N*. In reaction-agnostic top-*N*, the top-*N* child metabolites are selected to explore in the next step, as previously discussed. However, reaction-agnostic top-*N* assumes that the probabilities generated by the Rainbow XenoSite for each of the 20 reaction types are comparable, which is an oversimplification. Therefore, we also employ a reaction-specific top-*N*, where each of the 20 reaction types is considered separately. In the reaction-specific top-*N*, the child metabolites produced from the top-*N* transformations within each of the reaction types are eligible for further exploration. For example, when N is set equal to 3, the top-3 subsequent metabolites formed via each reaction type would be considered for the next step, allowing for a maximum of 60 child metabolites to be considered in the next step.

One final variant of the top-*N* heuristic, hereon referred to as the optimal thresholds heuristic, limits the acceptable child metabolites based on whether the metabolic transformation required to produce them is above a probability threshold. We used a separate threshold for each of the five major reaction classes—stable oxygenation, unstable oxygenation, dehydrogenation, hydrolysis, and reduction. To define the thresholds, we used the optimal point on the cross-validated receiver-operating curves (ROC) curves computed from Rainbow XenoSite's atom-level reaction predictions on its training data for each reaction class.[30] The threshold was computed to optimize both sensitivity and specificity using the Youden index.[38] For example, a metabolic transformation with the epoxidation reaction type must have a score greater than the stable oxygenation reaction class threshold for the produced child metabolite to be retained for further processing at the next step of the search.

In the substructure matching heuristic, XenoNet only generates intermediate metabolite structures that result from applying reaction rules to sites expected to lead to the product metabolite. Concretely, metabolic transformations are only computed for sites where the current intermediate metabolite being assessed differs from the product metabolite, as well as their immediately neighboring sites. For instance, if a substrate−product pair only differs by hydroxylation at a single site, then the reaction rules will only be applied to that differing site and its adjacent sites, rather than every site in the substrate. Substructure matching can be enhanced by allowing a parameter for the radius from sites of differences. As described previously, the radius is one, i.e., we find the sites where differences are present and only run the SOM model on those sites and their adjacent sites. The radius hyperparameter for substructure matching was also evaluated for radii of 2 and 3 (Table S1), but these variants did not perform better than substructure matching with a radius of 1. Ultimately, substructure matching was evaluated with respect to the other heuristics with a radius of 1 only. Some reaction rules, such as dehydrogenation, can operate on two distinct sites in a molecule. If the reaction rule specifies two sites, then both sites must be in the set of valid sites derived from the substructure matching heuristic.

However, both the top-*N* and the substructure matching heuristics have weaknesses. The top-*N* heuristic is greedy. If a high probability transformation follows a low probability transformation, the pathway would be missed because the first low probability transformation would be skipped. On the other hand, substructure matching treats every child metabolite as equally likely to form in the subsequent transformation. Additionally, it does not account for initial reactions that occur at a site that lies outside the common substructure, even if a reaction at that site has a high probability and could eventually lead to the target via further downstream reactions. Combining both approaches could help ameliorate their specific deficiencies. In the combination approach, the set of child metabolites that is eligible for further exploration is derived from the union between the set of children produced by individual application of the substructure heuristic and the set of children produced by individual application of the top-*N* heuristic.

In summation, the six additional XenoNet variants include top-*N* reaction-agnostic, top-*N* reaction-specific, optimal thresholds, substructure matching, substructure matching in combination with top-*N* reaction agnostic, and substructure matching in combination with top-*N* reaction specific.

It should be noted that there is a speed optimization for variants of the model that employ the top-*N* heuristic or optimal thresholds heuristic. Since evaluating the top-*N* child metabolites for further search exploration involves predictions from Rainbow XenoSite, those predictions can be cached. Though at the expense of increased memory cost, saving the predictions for a given metabolite in anticipation of that metabolite potentially involved in a valid pathway prevents having to execute Rainbow XenoSite more than once on each metabolite in the network. If a path is found between the start and target molecules, then the path, along with its cached predictions, can be stored in the network.

Metabolite Scoring Algorithm. Several of the metrics used for evaluating our models require a way to rank the relative importance of metabolites in a given network. Thus, except for the substrate that is always assigned a score of 1.0, we calculate metabolite scores for all other compounds in the network through three steps. First, we start with the raw score $w_{M_j \to M_k}$ that Rainbow XenoSite assigns to the metabolic transformation between a metabolite M_j to one of its children M_k (Figure 4a). Second, the raw score $w_{M_j \to M_k}$ is normalized over all metabolic transformation from M_j to its children (Figure 4b)

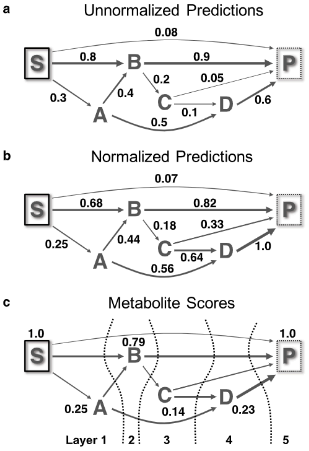


Figure 4. Metabolite score. To rank the relative importance of metabolites in a given network, we calculate metabolite scores. There are three steps in this process. (a) First, each metabolic transformation in the network is assigned with a raw prediction by Rainbow XenoSite. (b) Second, the raw prediction is normalized using eq 1. (c) Third, the substrate is assigned a score of 1.0, and downstream metabolites are assigned with scores computed using eq 2. Because a node can only be scored after all of its parents have been scored, the third step is carried out in multiple layers. The dashed-line dividing the network into segments corresponds to the computation of the metabolite scores for subsequent layers over the network. On the first layer, for example, scores could be computed for A. The computed score for A, displayed as numbers adjacent to each metabolite, is 0.25. The number of layers required to compute all metabolite scores is equal to the path of maximum length. The maximum length path in this network requires five steps. Computing metabolite scores for all metabolites in the network shown required five layers.

$$W_{M_j \to M_k} = \frac{w_{M_j \to M_k}}{\sum_{M_x \in M_j^{\text{children}}} w_{M_j \to M_x}} \tag{1}$$

Third, the metabolite score F_{M_j} is a weighted sum of the normalized $W_{M_i \to M_j}$ where M_i is one of the parents of M_j

$$F_{M_j} = \sum_{M_i \in M_j^{\text{parents}}} F_{M_i} \times W_{M_i \to M_j} \tag{2}$$

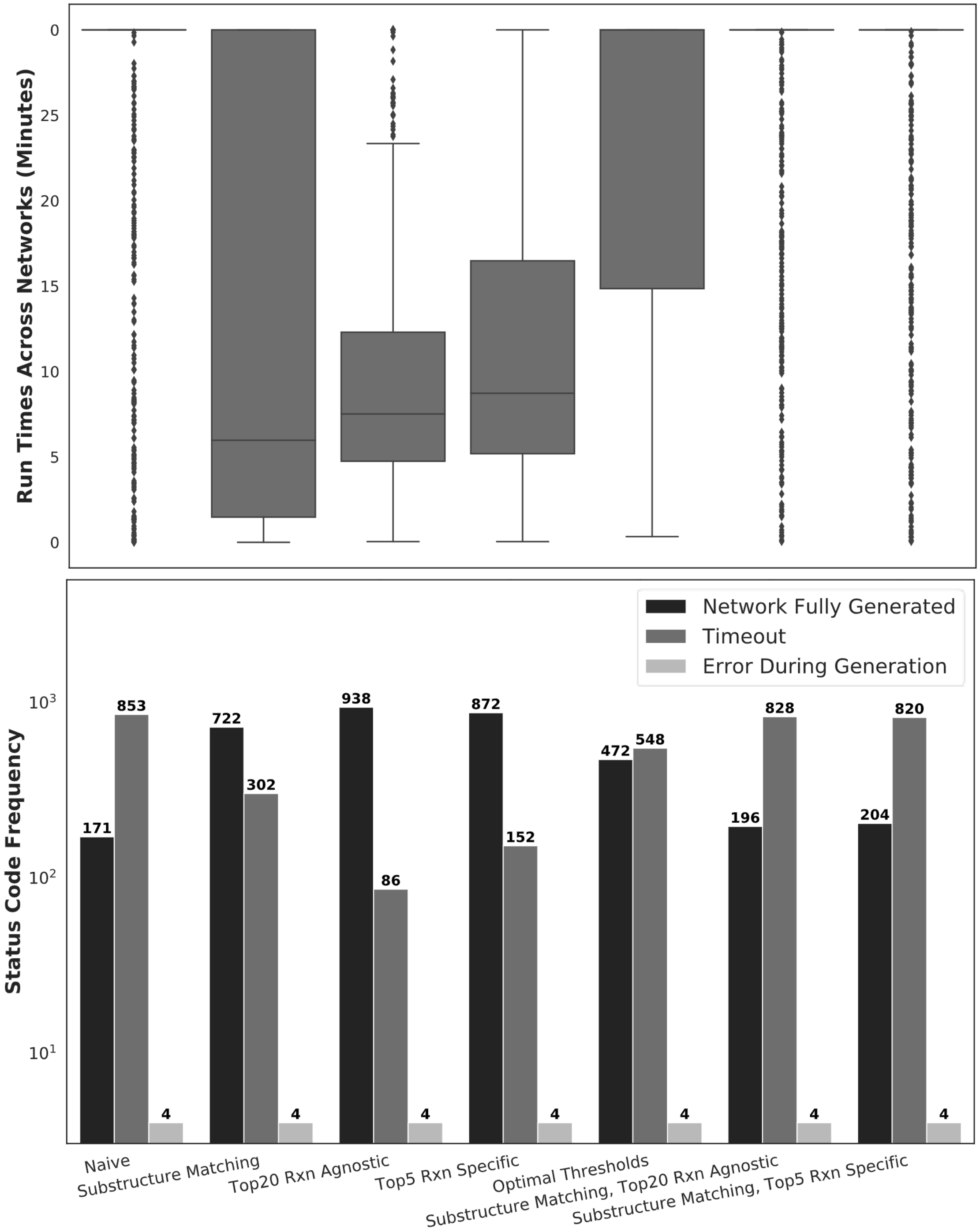
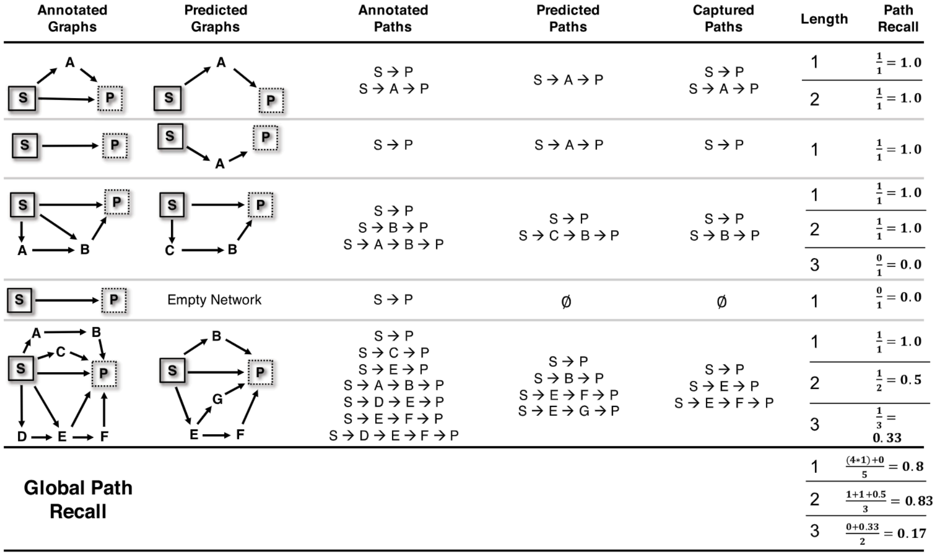


Figure 5. Time cost varies greatly between the seven XenoNet variants. In this comparison, each method was allotted 30 min for network generation of each substrate−product pair at a depth limit of three steps from the substrate. (Top) Across 1024 substrate−product pairs, only the substructure matching, top-*N* specific, and top-*N* agnostic variants can produce above 70% fully generated networks within the 30 min allotted time. In the same allotted time limit, the optimal thresholds heuristic fully generated almost 50% of the networks. In contrast, while the naive and the combination variants are the least restricted in terms of the metabolite structures generated, they are only able to produce at most 20% fully generated networks within 30 min. (Bottom) A similar trend is found when comparing the run time distributions. The substructure matching, top-*N* specific, and top-*N* agnostic variants, on average, take less than 10 min to generate a network. The time distribution for the naive and the combination variants is much broader, and many runs hit the 30 min timeout before producing a fully generated network.

Because a node can only be scored after all of its parents have been scored, the third step is carried out in multiple layers (Figure 4c).

■ RESULTS AND DISCUSSION

In the following sections, we examine the inner workings of XenoNet, our metabolic network predictor. First, we compared the seven variants of XenoNet on a randomly sampled subset



Annotated Graphs	Predicted Graphs	Annotated Paths	Predicted Paths	Captured Paths	Length	Path Recall
		S → P S → A → P	S → A → P	S → P S → A → P	1	$\frac{1}{1} = 1.0$
					2	$\frac{1}{1} = 1.0$
		S → P	S → A → P	S → P	1	$\frac{1}{1} = 1.0$
		S → P S → B → P S → A → B → P	S → P S → C → B → P	S → P S → B → P	1	$\frac{1}{1} = 1.0$
					2	$\frac{1}{1} = 1.0$
					3	$\frac{0}{1} = 0.0$
	Empty Network	S → P	∅	∅	1	$\frac{0}{1} = 0.0$
		S → P S → C → P S → E → P S → A → B → P S → D → E → P S → E → F → P S → D → E → F → P	S → P S → B → P S → E → F → P S → E → G → P	S → P S → E → P S → E → F → P	1	$\frac{1}{1} = 1.0$
					2	$\frac{1}{2} = 0.5$
					3	$\frac{1}{3} = 0.33$
Global Path Recall					1	$\frac{(4*1)+0}{5} = 0.8$
					2	$\frac{1+1+0.5}{3} = 0.83$
					3	$\frac{0+0.33}{2} = 0.17$

Figure 6. Path recall calculation. Path recall is a metric designed to measure how well a model captures annotated paths. Here, we show how path recall is calculated for a hypothetical data set of five substrate−product pairs. For each substrate−product pair, a predicted graph is generated within a depth limit of 3. An annotated path from the substrate to the target product is considered as being captured if (1) all of its nodes are contained in a predicted path and (2) the order of traversal through these nodes is the same in the annotated and predicted paths. For each pair of substrate and product molecules, the proportions of annotated paths of a certain length that were captured are calculated. For each length classification, the path recall of a test set is the average of the captured proportions at that specific path length across the entire test set.

of our metabolic network data set to identify the optimal model. The most efficient XenoNet variant was chosen as the final XenoNet model. Second, we assess the performance of the final XenoNet model on the full metabolic network data set. Third, we compare XenoNet to prior work on the task of predicting metabolite formation with respect to phase-I metabolism.

Comparison Between Metabolic Network Variants. The main goal of XenoNet is to enumerate and assign prediction scores to sequences of metabolic transformations (edges) and intermediates (nodes) in the pathways that connect a given input substrate−product pair. We generated seven XenoNet variants and wanted to select the best variant as our final model. However, the time cost of producing networks poses a significant computational challenge when running each method across the full metabolic network data set. Inferring pathways of depth 3 or higher, along with model predictions, can take a long time. As such, it is infeasible to run each XenoNet variant on the full data set. Instead, we randomly sample from our full data set a subset of 1024 substrate−product pairs to compare the model variants. In addition to the naive XenoNet that uses no heuristics, the other six variants employ either substructure matching, top-*N* heuristics, or a combination of both. The choice of heuristic influences the paths that are enumerated from the substrate molecule to the product molecule. Consequently, heuristic choice influences the presence of an edge in the predicted graph but does not influence the scoring of the edges provided by the Rainbow model. As such, metrics for comparing performance between heuristics focused on the time cost of network generation, path recall, intermediate metabolite recall, and intermediate metabolite recovery. Metrics for comparing the relevance of edge predictions during metabolic network construction were not compared. The model variant with the best performance on this subset would be the final XenoNet model for analysis on the full data set.

Prior to comparing distinct variants, we first tuned the *N* hyperparameter for the top-*N* variants. As *N* increases, the time cost increases. Preferably, we want the lowest value of *N* that does not result in a significant performance decrease. We increased the value of *N* until we hit a value whose performance increase based on metrics of path recall and intermediate recall, which will be discussed in detail in the following sections, was no longer significant. We then selected the value of *N* preceding this drop in significance. Significance between the current top-*N* variant and its immediately preceding variant was evaluated using a paired *t*-test. For the reaction-specific top-*N*, the hyperparameter *N* was incremented by 1 for each evaluation. The full range of assessed values was for *N* equal to 2, 3, 4, 5, and 6. The value of *N* chosen for this heuristic was *N* equal to 5. For reaction-agnostic top-*N*, the hyperparameter *N* was incremented by 5 for each evaluation. The full range of assessed values was for *N* equal to 5, 10, 15, 20, and 25. The value of *N* chosen for this heuristic was *N* equal to 20. The optimal value of *N* for each of the top-*N* variants was used for both of their corresponding heuristic combination variants.

Time cost. An effective algorithm would be able to quickly identify a pathway to known metabolites in all cases and identify known intermediate metabolites. A brute-force, depth-first search alone is not tractable since the time complexity of searching the metabolic forest is $O(n^d)$, where n is the number of metabolites for a given molecule and d is the depth of the tree—the number of metabolic steps that is allowed between the starting metabolite and the target metabolite. In some instances, n can be on the order of 10^3. In this comparison, each method was allotted 30 min for network generation of each substrate−product pair at a depth limit of three steps from the substrate. In the same allotted time limit, the optimal thresholds heuristic fully generated almost 50% of the networks. Across 1024 substrate−product pairs, only the substructure matching, top-*N* specific, and top-*N* agnostic

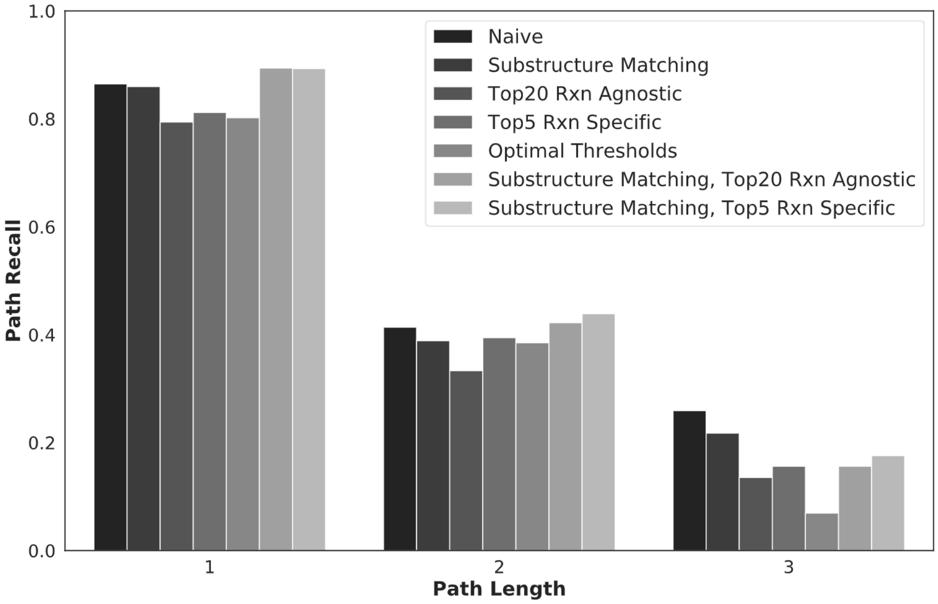


Figure 7. Path recall performance is dependent on the path length. While heuristic approaches capture shorter-length annotated paths better, the naive model is superior in capturing longer-length annotated paths. The performance trend highlights the trade-off between employing heuristics to speed up network generation at the cost of constraining the set of possible child metabolites. The constraining factor grows exponentially as the path length grows.

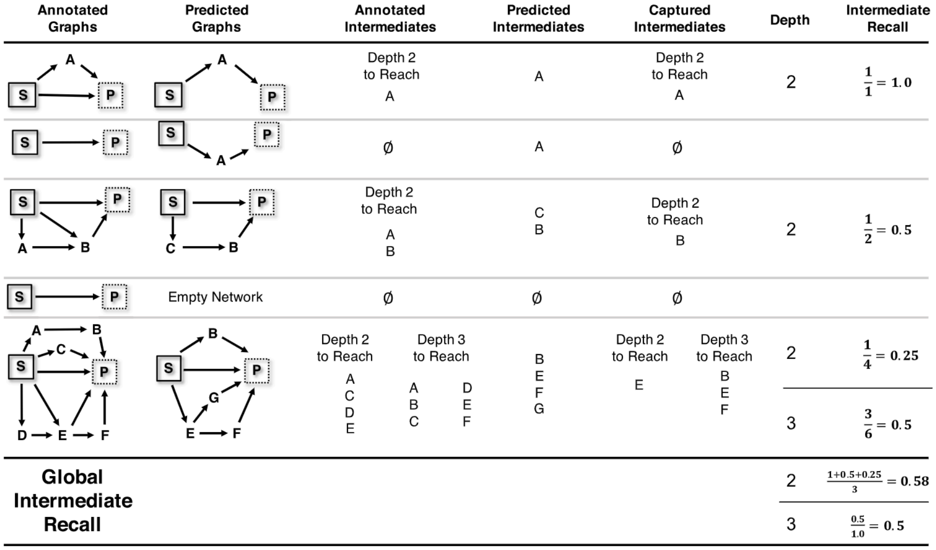
variants can produce above 70% fully generated networks within the 30 min allotted time (Figure 5, top panel).

In contrast, while the naive and the combination variants are the least restricted in terms of the metabolite structures generation, they are only able to produce at most 20% fully generated networks within 30 min. A similar trend is found when comparing the run time distributions (Figure 5, bottom panel). The substructure matching, top-*N* specific, and top-*N* agnostic variants, on average, take less than 10 min to generate a network. The time distributions for the naive and the combination variants are much broader, and many runs hit the 30 min timeout before producing a fully generated network. For applications where run time cost needs to be conserved, the substructure matching, top-*N* specific, and top-*N* agnostic XenoNet variants may be the most optimal. Nevertheless, a fast method is not useful if it cannot correctly identify known intermediate metabolites. We hypothesize that the combination methods traverse more of the metabolite space for each partially generated network than the naive method. We test this hypothesis in the following sections using path recall, intermediate metabolite recall, and intermediate metabolite recovery metrics.

Path Recall. Path recall is a metric designed to measure how well a model captures annotated paths (Figure 6). Given an input substrate−product pair, a model would output a graph with paths that lead from the substrate to the product. Ideally, these predicted paths would be the same as the annotated paths that were derived from the literature. However, depending on the various experimental details, annotated paths may miss some intermediates, especially short-lived and reactive metabolites like epoxides or quinones. In contrast, predicted paths consistently include even short-lived metabolites because they are built with fixed and comprehensive reaction-rule sets. Consequently, an annotated path from the substrate to the target product is considered as being captured if (1) all of its nodes are contained in a predicted path and (2) the order of traversal through these nodes is the same in the annotated and predicted path. The paths vary from a length of one, i.e., a direct path linking the substrate to the product, up to the depth limit specified during the generation of the predicted network. For each pair of substrate and product molecules, the proportions of annotated paths of a certain length that were captured are calculated. For each length classification, the path recall of a test set is the average of the captured proportions at that specific path length across the entire test set.

The length-specific path recalls were evaluated over all metabolic networks in the test set (Figure 7). All 1024 networks contain annotated paths with a length of 1, but only 60 and 24 of these networks contain annotated paths with a length of 2 and 3, respectively. For all methods, the path recall decreases as the path length increases. This is expected since the longer a path gets, the probability that a model generates all of its intermediates and places them in the correct order diminishes.

No single method has the highest length-specific path recall across all three path lengths. While heuristic approaches capture shorter-length annotated paths better, the naive model is superior in capturing longer-length annotated paths. For example, at a path length of 1, both combination approaches perform best with a path recall of 0.89. For a path length of 2, the top-*N* reaction-specific combination approach performs best with a path recall of 0.44. For a path length of 3, the naive method performs best with a path recall of 0.26. The performance trend highlights the trade-off between employing heuristics to speed up network generation at the cost of constraining the set of possible child metabolites. The constraining factor grows exponentially as the path length grows. Though the naive method generally takes the longest time to iterate over all possible child metabolites branching out from a given parent metabolite, it has the highest expressivity—the greater the variety and quantity of metabo-



Annotated Graphs	Predicted Graphs	Annotated Intermediates	Predicted Intermediates	Captured Intermediates	Depth	Intermediate Recall
		Depth 2 to Reach A	A	Depth 2 to Reach A	2	$\frac{1}{1} = 1.0$
		∅	A	∅		
		Depth 2 to Reach A B	C B	Depth 2 to Reach B	2	$\frac{1}{2} = 0.5$
	Empty Network	∅	∅	∅		
		Depth 2 to Reach A C D E Depth 3 to Reach A B C D E F	B E F G	Depth 2 to Reach E Depth 3 to Reach B E F	2	$\frac{1}{4} = 0.25$
					3	$\frac{3}{6} = 0.5$
Global Intermediate Recall					2	$\frac{1+0.5+0.25}{3} = 0.58$
					3	$\frac{0.5}{1.0} = 0.5$

Figure 8. Intermediate metabolite recall calculation. The intermediate recall is a metric designed to measure how well the model can infer experimentally observed intermediates, compounds on the paths from the substrate to the target product. Here, we show how the intermediate recall is calculated for a hypothetical data set of five substrate−product pairs. For each substrate−product pair, a predicted graph is generated within a depth limit of 3. For each input substrate−product pair, the proportions of experimentally observed intermediates that can be inferred in the predicted graph of certain minimal depth are calculated. For each minimal depth classification, an intermediate recall of a test set is the average of these depth limit-specific proportions across the entire data set.

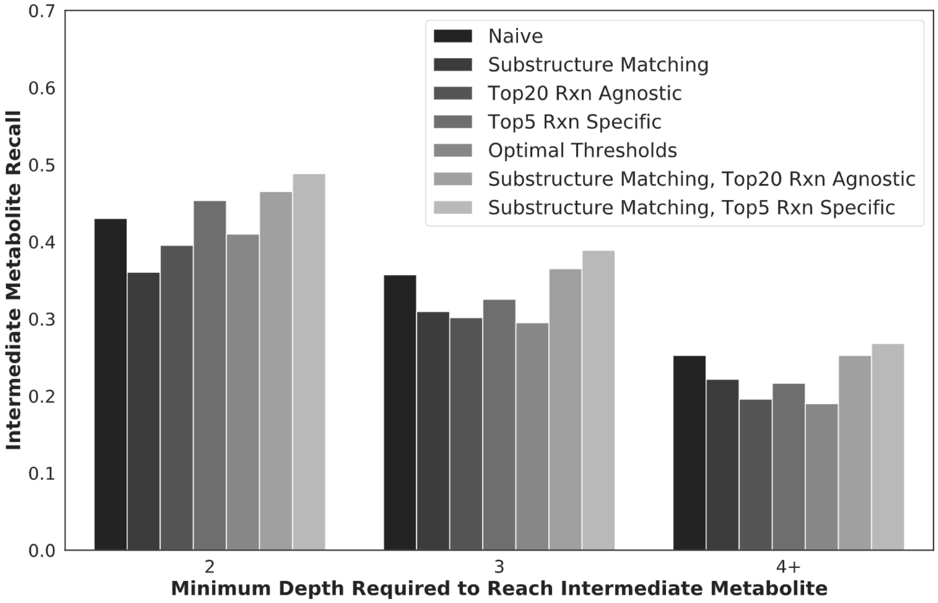


Figure 9. Combined substructure matching and top-*N* specific heuristic model has the highest intermediate metabolite recall across all depths.

lites it can infer. If network generation were allowed to run without a time limit, the naive method would be expected to have the best performance. In contrast, each heuristic makes certain assumptions about the importance of child metabolites to eventuating into the product metabolite. As the path length increases, the constraints imposed by the heuristic compound over each set of child metabolites produced at each step in the path. Eventually, the heuristic's trade-off worsens performance relative to the naive method.

It is not yet apparent which method achieves an optimal trade-off. However, the methods that have the highest constraints—both top-*N* methods, the substructure matching method, and the optimal thresholds method—tend to have lower path recall. Finally, each pair of approaches that use some form of the top-*N* heuristic perform better or equivalent when the heuristic uses the top-*N* specific form compared to the top-*N* agnostic form.

Intermediate Metabolite Recall. Since most current experimental approaches are designed to identify and characterize only the reactive metabolite, they are liable to miss important intermediate metabolites. Identification of intermediate metabolites could help in the generation of hypotheses for how a given drug molecule could be modified to be less likely to form the intermediate metabolite. In effect, such a modification could prevent the further formation of the reactive metabolite.

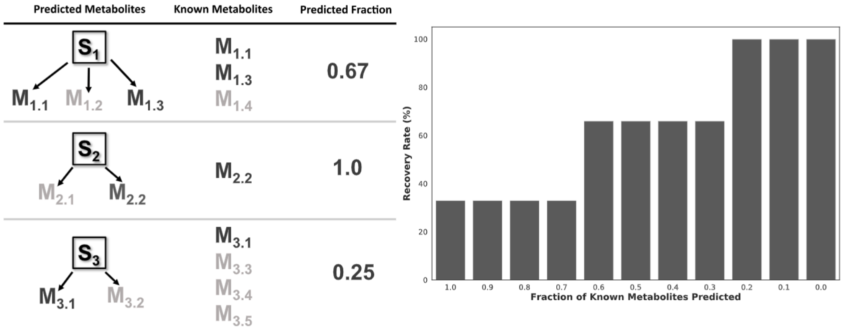


Figure 10. Recovery rate calculation. Here, we show how the recovery rate is calculated for a set of three hypothetical metabolic networks. (Left) Predicted metabolic networks with a depth limit of one of three different substrates, S_1, S_2, and S_3, and their known metabolites are shown. Metabolites M that are both predicted and experimentally observed are in darker gray. Metabolites M that are predicted but not experimentally observed or vice versa are in lighter gray. (Right) The proportion of networks that have fractions of known metabolites predicted by the model above a certain threshold is computed. In our hypothetical database of three metabolic networks, 100% of substrates have at least 0−20% of their known metabolites predicted, and 66.7% of the substrates (S_1 and S_2) have at least 60% of their known metabolites predicted. Only 33.3% of the substrates (S_1) have at least 70% of their known metabolites predicted.

Table 1. Intermediate Metabolite Recovery Rate of XenoNet Variants. The Combined Substructure Matching and Top-*N* Reaction-Specific Variant is the Best Model across Thresholds[a]

	fraction of known intermediates predicted									
method	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
naive	0.57	0.57	0.57	0.54	0.54	**0.49**	**0.49**	**0.49**	**0.49**	**0.44**
substructure matching, top-*N* reaction specific	**0.62**	**0.62**	**0.62**	**0.59**	**0.59**	**0.49**	**0.49**	**0.49**	**0.49**	**0.44**
substructure matching, top-*N* reaction agnostic	0.60	0.60	0.60	0.57	0.57	0.46	0.44	0.44	0.44	0.40
substructure matching	0.57	0.55	0.55	0.51	0.51	0.41	0.40	0.40	0.29	0.27
top-*N* reaction specific	0.57	0.57	0.57	0.57	0.56	0.46	0.41	0.38	0.35	0.35
top-*N* reaction agnostic	0.52	0.51	0.49	0.44	0.41	0.35	0.30	0.30	0.30	0.30
optimal thresholds	0.58	0.50	0.44	0.44	0.44	0.40	0.37	0.37	0.37	0.37

[a] The naive model is also competitive at thresholds ranging from 0.6 to 1. The highest values in each threshold are bolded.

The intermediate recall is a metric designed to measure how well the model can infer experimentally observed intermediates—compounds on the paths from the substrate to the target product (Figure 8). The intermediate recall is depth limit-dependent. To infer a certain annotated intermediate, a predicted network needs a minimal depth limit. As an example, for an intermediate with a specified minimum depth of 2, any method for predicting a network that potentially contains a path that both infer the intermediate and terminate at the target product would need to have the depth limit set to 2 at minimum. Thus, for each input substrate−product pair, the proportions of experimentally observed intermediates that can be inferred in the predicted graph of certain minimal depth are calculated. For each minimal depth classification, an intermediate recall of a test set is the average of these depth limit-specific proportions across the entire data set.

Among the randomly sampled 1024 test set substrate−product pairs, there are 60, 82, and 97 annotated networks that contain intermediates that require XenoNet's depth limit to be set to, at minimum, 2, 3, and 4 or more, respectively. There are 516, 756, and 1164 metabolites in the 60, 82, and 97 annotated networks, respectively. The depth-specific intermediate recalls were computed over these 60, 82, and 97 annotated networks (Figure 9).

For all methods, the intermediate metabolite recall decreases as the minimum depth required to reach the intermediate metabolite increases. This phenomenon is consistent with the fact that while the chemical diversity of annotated graphs would vary greatly depending on experimental conditions, predicted graphs strictly follow a set of transformation rules. Because of such difference in granularity, the discrepancy between the annotated graphs and predicted graphs grows exponentially with the depth of the networks.

The combined substructure matching and top-*N* specific heuristic model has the highest intermediate metabolite recall across all depths. For this variant, the intermediates metabolite recalls are 0.49, 0.39, and 0.27 for networks of depths 2, 3, and 4 or more, respectively. The naive model, in contrast, only achieves intermediate metabolite recalls of 0.47, 0.36, and 0.25 for networks of depth 2, 3, and 4 or more, respectively. Other models have even lower intermediate metabolite recalls (Figure 9). The result is further evidence that the combined substructure matching and top-*N* specific heuristic mode are the optimal XenoNet variant that can decrease its time cost for inferring metabolites without hampering its ability to infer known metabolites.

Intermediate Metabolite Recovery. To measure the ability of a model to capture known metabolites, we use the recovery rate metric. The recovery rate is the proportion of networks that have fractions of known metabolites predicted by the model above a threshold (Figure 10).

We assess the intermediate metabolite recovery rate of all seven XenoNet variants on the randomly sampled 1024 network subset (Table 1). The combined substructure matching and top-*N* reaction-specific model is the best model across all thresholds. The naive model is also competitive at thresholds ranging from 0.6 to 1. Both of

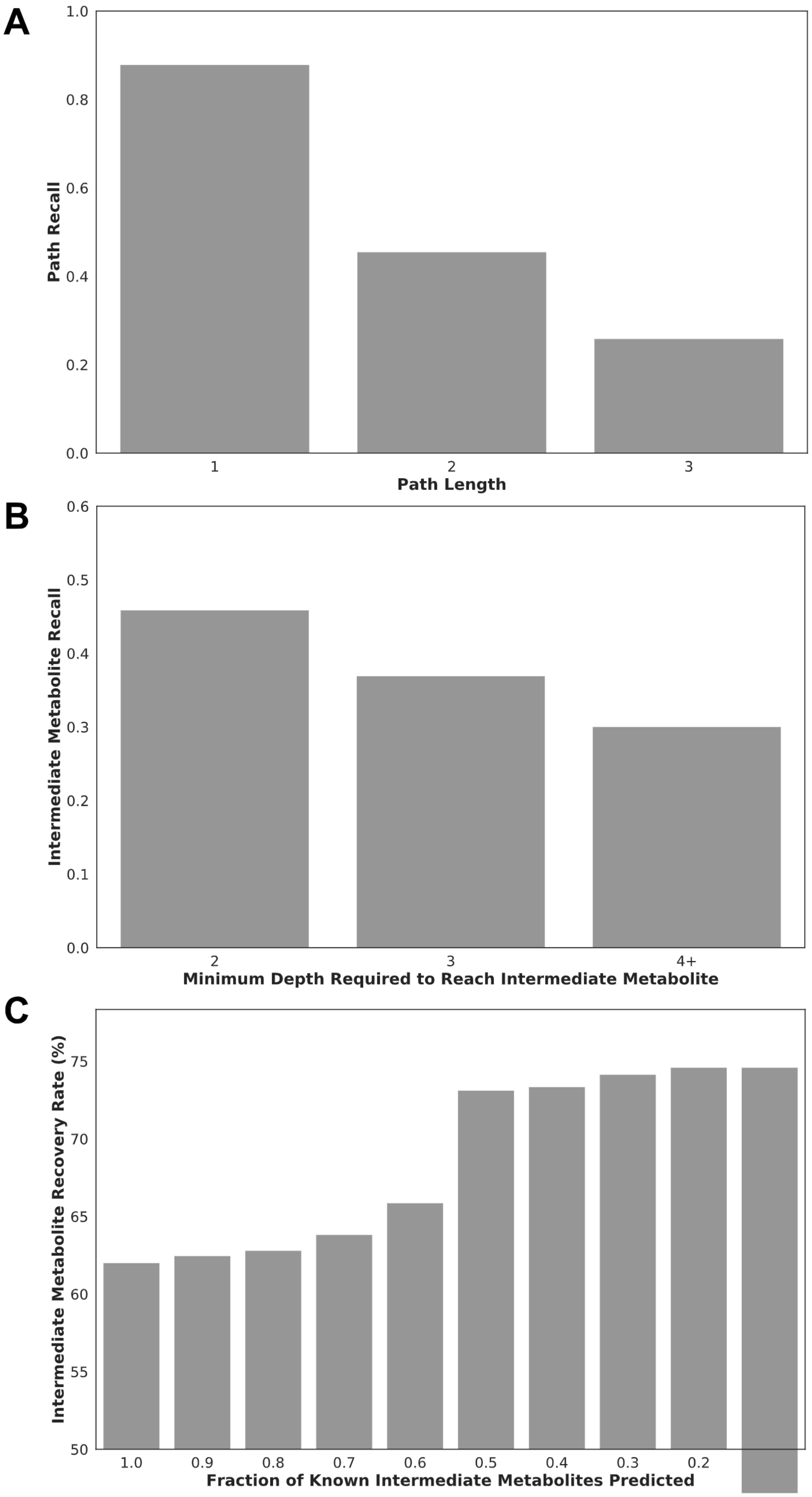
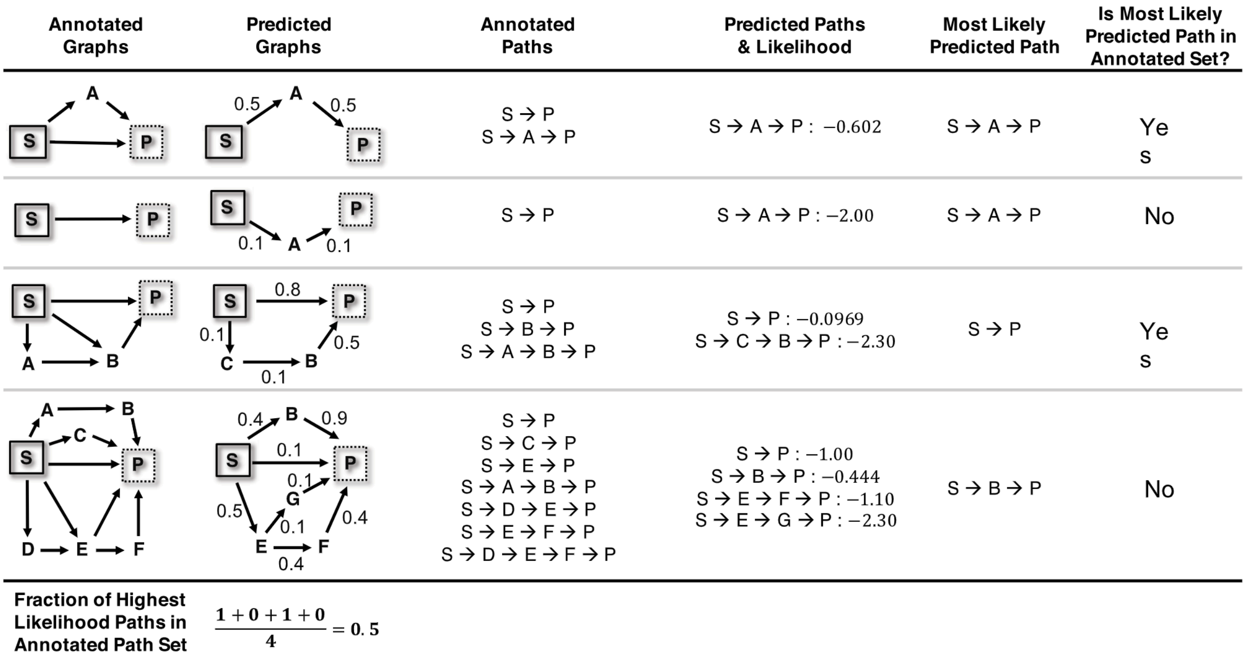


Figure 11. XenoNet accurately predicts metabolic networks. Performance of XenoNet across the full metabolic network data set on the same metrics assessed during the comparison between all heuristics on the subset of 1024 networks. (A) Path recall stratified across paths of length 1, 2, and 3. (B) Intermediate metabolite recall stratified across minimum network depth limit required to reach the intermediate metabolite of 2, 3, and 4 or more. (C) Intermediate metabolite recovery across recovery rate thresholds of 0.1 to 1. The results for all three plots are close to the initial results for each metric when validated on the subset of 1024 samples and supports that the comparison between the XenoNet variants generalizes to performance on the full data set.



Annotated Graphs	Predicted Graphs	Annotated Paths	Predicted Paths & Likelihood	Most Likely Predicted Path	Is Most Likely Predicted Path in Annotated Set?
		S → P S → A → P	S → A → P : −0.602	S → A → P	Ye s
		S → P	S → A → P : −2.00	S → A → P	No
		S → P S → B → P S → A → B → P	S → P : −0.0969 S → C → B → P : −2.30	S → P	Ye s
		S → P S → C → P S → E → P S → A → B → P S → D → E → P S → E → F → P S → D → E → F → P	S → P : −1.00 S → B → P : −0.444 S → E → F → P : −1.10 S → E → G → P : −2.30	S → B → P	No

Fraction of Highest Likelihood Paths in Annotated Path Set $\frac{\mathbf{1+0+1+0}}{\mathbf{4}} = \mathbf{0.5}$

Figure 12. Path ranking calculation. This metric is designed to measure how well the predicted path with the highest likelihood of traversal corresponds to a known path in the annotated graph. Here, we show how path ranking is calculated for a hypothetical data set of five substrate−product pairs. The likelihood of each predicted pathway is the logarithmic sum over the scores that Rainbow XenoSite assigns to the metabolic transformations in that pathway. We then rank all predicted pathways in a given network by their likelihoods. If the highest-likelihood path of a given network has an exact match in the annotated path set, then that network is assigned a score of 1. Otherwise, that network is assigned a score of 0.

these methods can infer all intermediate metabolites in 44% of the networks.

Performance on the Full Metabolic Network Data Set. Comparisons between variants of metabolic network construction were drawn on a small fraction of our metabolic network data set. From these comparisons, we selected the best-performing variant, the combined substructure matching, and top-*N* reaction-type specific heuristic model, to run on our entire data set of 17 054 metabolic networks. Networks were generated with a depth limit of 3 and a max time limit of 30 min per network.

Network Coverage, Path Recall, Metabolite Recall, and Metabolite Recovery. In terms of coverage, 2804 of the networks reached completion within the 30 min allotted for network generation. The partial network states reached by the remaining 14 250 networks were also preserved for analysis. In total, 14 882 of the networks found at least one path between the substrate−product pair. One thousand four hundred and ninety networks timed out at the 30 min limit before finding any valid paths between the substrate and the target product. The remaining 673 of the networks reached completion within the time limit but did not find any valid paths. In total, the total time required to generate networks for the full metabolic network data set was approximately 468 000 min.

We also evaluate the final XenoNet's ability to infer known pathways and intermediates between a given substrate−product pair of molecules on the 17 054 pairs using path recall, intermediate metabolite recall, and intermediate metabolite recovery metrics (Figure 11).

First, we assess path recall. In our data set, all 17 054 networks contain annotated paths with a length of one, but only 817 and 240 of these networks contain annotated paths with a length of two and three, respectively. The path recall for path lengths of one, two, and three steps is 0.88, 0.46, and 0.26 across 17 054, 817, and 240 networks, respectively (Figure 11A). Examples of paths that were successfully elucidated are highlighted in Figure S9, along with the examples for which a valid path was not found. XenoNet's path recall for a path length of 1 is comparable to the accuracy of Metabolic Forest (88.4−88.8%).[36] This result is expected because, despite aiming at different tasks, the two models use the same rule set and their data was built using the same set of AMD reactions. XenoNet's path recall at a path length of 1 is only slightly lower than the path recall assessed for Metabolic Forest, which can be attributed to the heuristics limiting the extent of the chemical transformation space.

Second, we asses intermediate recall. In our data set, there are 817, 1041, and 1143 annotated networks that contain intermediates that require XenoNet's depth limit to be set to, at minimum, 2, 3, and 4 or more. XenoNet's depth-specific intermediate metabolite recalls at depth limits of 2, 3, and 4 or more are 0.46, 0.38, and 0.30, respectively (Figure 11B). Unsurprisingly, there is a drastic drop-off in intermediate metabolite recall when going to metabolites that would require a minimum depth limit of greater than 3 to reach. Since XenoNet's depth limit parameter was set to 3, we expect low recall for intermediates, which, based on their annotated graph, would require a minimal depth limit of 3 to locate.

Recall that the path recall for paths of length one, two, and three steps over the subset of 1024 networks was 0.89, 0.44, and 0.18 using the same heuristic. Furthermore, the intermediate recall for minimum network depths of two, three, and four or more metabolic steps over the subset of 1024 networks were 0.49, 0.39, and 0.27. It is reassuring that the recall metrics computed over the subset of 1024 networks translated to recall evaluation over the full data set. Most of the

Table 2. Comparison between XenoNet and a Random Model

	highest-likelihood path in observed paths (%)	top-one intermediate metabolite (%)	top-two intermediate metabolite (%)	top-three intermediate metabolite (%)	average AUC (%)
unmodified weights	93.6	51.9	66.2	77.5	78.7
randomly permuted weights	61.8	31.9	46.4	60.7	62.5

recall metrics are similar in value, with more notable increases in performance on the full data set for depth 3 path recall and depth 2 intermediate metabolite recall. Regardless, the results on the whole metabolic network data set are broadly consistent with the results across the subset of data used to compare various heuristics in the previous section.

Last but not least, we assess XenoNet's performance in regard to intermediate metabolite recovery on the full data set. The recovery rate only drops ∼13% from ∼75% at 0.1 threshold to ∼62% at 1.0 threshold (Figure 11C). Inversely, 25% of networks (2163) found 0 of the known intermediate metabolites. About half of these 25% networks are a result of networks that could not find any valid paths.

Path and Metabolite Rankings. The ability to enumerate multiple metabolites and metabolic pathways alone is useful, but not enough for triage purposes. Knowing the likelihood of each metabolite or pathway would help differentiate between true positives and false positives. Ideally, an excellent metabolic network generator would always be able to assign experimentally observed pathways or metabolites with higher likelihood than not-observed pathways or metabolites. The more frequently the model assigns the highest probability to pathways/metabolites that have been observed, the more confident users would be that, when applied to new data, the highest-likelihood pathway and metabolite would be designated to a pathway and metabolite with a high chance of existing. In this section, we explore how well the XenoNet model would predict experimentally observed metabolites and pathways with higher likelihood than not-observed metabolites and pathways.

The path ranking metric is designed to measure how well the predicted path with the highest likelihood corresponds to an annotated path (Figure 12). Here, the likelihood of each predicted pathway is the logarithmic sum over the scores that Rainbow XenoSite assigns to the metabolic transformations in that pathway. We then rank all predicted pathways in a given network by their likelihoods. If the highest-likelihood path of a given network has an exact match in the annotated path set, then that network is assigned a score of 1. Otherwise, that network is assigned a score of 0. The path ranking of a set is the average score across all networks in the set (Figure 12). Note that, unlike in the assessment of the path recall metric, the predicted path with the highest likelihood is assessed for an exact match—containing the same set of metabolites in identical order—in the annotated path set. This is because every annotated network contains a one-step path that directly links the substrate to the product, and every captured path is an approximate match to this one-step path. Allowing approximate matches would inflate the path ranking.

Top-one, -two, and -three metabolite ranking metrics are designed to measure how well the predicted metabolite with the highest metabolite scores (Figure 4) corresponds to an observed metabolite. Top-*N* accuracy of a set is the fraction of predicted networks that have experimentally observed metabolites among the *N* predicted metabolites with the highest metabolite scores. Note that an obstacle can arise in computing metabolic scores for networks that contain a cycle. Though an individual enumerated path is not allowed to have a cycle, a cycle can still arise in the global network structure. Consider the case of a network with the following two paths: S → M_1 → M_2 → P and S → M_2 → M_1 → P. The method for computing metabolite scores, as initially described, would result in an infinite loop due to the one-step cycle between M_1 and M_2. Since networks with a cycle show up in under 1% of the networks predicted during our experiments, they are ignored when computing metabolite scores for the following results over the full metabolic network data set.

We also compute the area under the receiver-operating characteristic curve (AUC) using metabolite scores and their respective labels, 1 or 0, indicating whether the metabolite was experimentally observed. We then calculate an average AUC across all of the networks to measure the performance of this task across the whole data set.

Among the 817 metabolic networks with at least one intermediate in our data set, XenoNet was able to complete 710 networks. The path ranking, top-*N*, and average AUC metrics calculated over these 710 networks (Table 2). Overall, XenoNet was able to accurately predict experimentally observed pathways with 93.6% path-rank accuracy. The model also accurately predicted experimentally observed metabolites with 51.9, 66.2, 77.5, and 78.7% top-one, -two, -three, and average AUC accuracies, respectively.

To assess the value of scoring pathways and metabolites using predictions from the Rainbow XenoSite model, we compare XenoNet's performance on path ranking, Top-*N*, and average AUC metrics to a model where these predictions are permuted (Table 2). Specifically, for each of the predicted networks, while the nodes and edges are kept in the original order, the scores assigned by Rainbow XenoSite to the edges are randomly permuted. The values computed for the "randomly permuted weights" case are computed over 10 trials. Overall, the randomly permuted model predicts experimentally observed pathways with 61.8% path-rank accuracy. The randomly permuted model also predicts experimentally observed metabolites with 31.9, 46.4, 60.7, and 62.5% Top-one, -two, -three, and average AUC accuracies, respectively.

We use two different statistical tests to assess the statistical significance of the difference in the performance of XenoNet and the randomly permuted model. For path ranking and top-*N* metrics, we use McNemar's test. McNemar's test is a paired nonparametric statistical hypothesis test for evaluating the disagreements between two cases.[39] In the context of classification, McNemar's test can be used to interpret whether both models make different errors and the difference in the relative proportions of those errors. The two cases are the unmodified XenoNet and the randomly permuted XenoNet, and a contingency table is constructed using both cases based on the numbers of highest-ranked pathways or metabolites that are experimentally observed versus those that are not. The null hypothesis of marginal homogeneity would mean that there is no effect in regards to where the edge weights are shuffled in

Table 3. XenoNet Outperforms Published Methods on the GLORY Test Set[a]

	GLORY, MaxCoverage Mode	GLORY, MaxEfficiency Mode	SyGMa	BioTransformer	XenoNet
precision	0.08	0.16	0.15	**0.17**	0.06
recall	0.83	0.64	0.74	0.72	**0.89**
total number of predicted metabolites	793	327	406	344	1179
number of successfully predicted reported metabolites	67	52	60	58	**72**
AUC	67.6%		50.1%		**73.3%**
Top-1	68.97%	68.97%	0%		**72.41%**
Top-2	72.41%	72.41%	48.28%		**75.86%**
Top-3	75.86%	75.86%	68.97%		**79.31%**

[a]For each metric, the best method's value is displayed in bold. Values for both GLORY variants, SyGMa, and BioTransformer were extracted from the initial comparisons made in de Bruyn Kops et al.[34]

the network, i.e., the two cases should have the same error rate.[40] McNemar's test yielded a p-value less than 0.001 for both path ranking and Top-N metrics, so there is strong evidence to reject the null hypothesis. For the average AUC metric, we use a paired t-test. Comparing the performance between the two cases via a paired t-test yielded a p-value less than 0.001. In summary, XenoNet performs better than a randomly permuted model across all considered metrics.

Comparison to Prior Work. No published model does the exact main task that XenoNet is designed for: to take a pair of start-target molecules as inputs and output a network of metabolic pathways between them. The closest comparable works in the literature include GLORY, BioTransformer, and SyGMa.[31,33,34] However, the main task of these methods is to receive a molecule as input and output computationally predicted metabolites. Though these prior works do not readily support XenoNet's main task of interest, XenoNet does support the main task of these prior works and so we can compare XenoNet to them in some capacity.

To enable this comparison, we first ran the naive XenoNet on GLORY's reference set of 848 parent molecules with 1588 known metabolites.[34] This reference set was used by GLORY's MaxEfficiency mode to decide upon a SOM probability cutoff that could be used as a preliminary filter and by GLORY's MaxCoverage mode to develop a priority score for ranking predicted metabolites. After XenoNet produced networks for each of the 848 parent molecules in the reference set, precision-recall curves for each of the 20 phase-I reaction types were used to define an optimal, reaction-type-specific threshold for filtering predicted metabolites. For each rule, the filtering threshold was set to the lowest threshold that did not reduce the original recall across the reference set's known metabolites. Similar to how the thresholds were used by GLORY, they are used to filter predicted metabolites for a test substrate after its metabolic network is generated by XenoNet.

As stated earlier, the extent to which probabilities generated by XenoNet may be compared across different reaction rules is unknown. The scores emitted for each reaction type are not proven to be well scaled across different reaction rules. Keeping the filtering threshold constant for each reaction class could result in a threshold that is high enough to completely filter out certain reaction rules from the network. As a result, we individually set the filtering threshold for each rule. The effect of the filtering step served mostly to filter out those reaction rules that were not being tested in the reference set. Ultimately, the only reaction rules relevant to yielding the observed metabolites in the reference set were hydroxylation, nitrogen reduction, dehydrogenation, nitrogen oxidation, oxidative dehalogenation, hydrolysis, sulfur oxidation, dealkylation, and epoxidation. Networks were still generated using the full set of reaction rules, but metabolites resulting from reaction rules outside of those listed above were filtered out from the final output.

Next, we tested the naive XenoNet variant on the curated test set of 29 substrates and 81 products that was used in GLORY[34] to compare against BioTransformer[31] and SyGMa.[33] None of the 29 substrates in this test set were contained within the GLORY reference set. In this test, XenoNet was set to generate metabolic networks using only the substrates without any secondary target molecules to enforce paths to terminate at. We allotted a 5 min time limit for XenoNet to generate each metabolic network and set the network's depth limit to 1. The generation of networks included both the inference of metabolite structures as well as assigning edge weights via SOM predictions generated by Rainbow XenoSite. The final list of predicted metabolites is filtered using the thresholds described in the previous paragraph. Production of the metabolites from the 29 test set substrates by XenoNet was computed on a single Intel Xeon Processor E5-2630 v3 CPU using a Linux operating system. The total run time using a single core was 80.5 min, and the average run time per parent molecule was 2.78 min.

XenoNet outperformed both GLORY variants, SyGMa, and BioTransformer at multiple metrics (Table 3). A metabolite is considered to be a true positive when it is both predicted and experimentally observed. Precision is the proportion of true positives among all predicted metabolites, and recall, also known as sensitivity, is the proportion of true positives among all experimentally observed metabolites. The precision values on the test set for XenoNet, GLORY MaxCoverage, GLORY MaxEfficiency, SyGMa, and BioTransformer are 0.06, 0.08, 0.16, 0.15, and 0.17, respectively. The recall values on the same test set for XenoNet, GLORY MaxCoverage, GLORY MaxEfficiency, SyGMa, and BioTransformer are 0.89, 0.83, 0.64, 0.74, and 0.72, respectively. It should be noted that while XenoNet has lower precision than other models, it has a higher recall. The trade-off between precision and recall is dependent on the chosen threshold, so either metric is not reliable to compare between models. The receiver-operating curves (ROC) and their corresponding areas under the curve (AUC) are more reliable metrics than precision and recall. The AUC of XenoNet, GLORY MaxCoverage, and SyGMa are 73.3, 67.6, and 50.1%, respectively (Table 3). We were unable to construct a ROC curve and calculate the AUC for BioTransformer based on their publication, but their recall of 0.72 is lower than other models if all of them were set to have

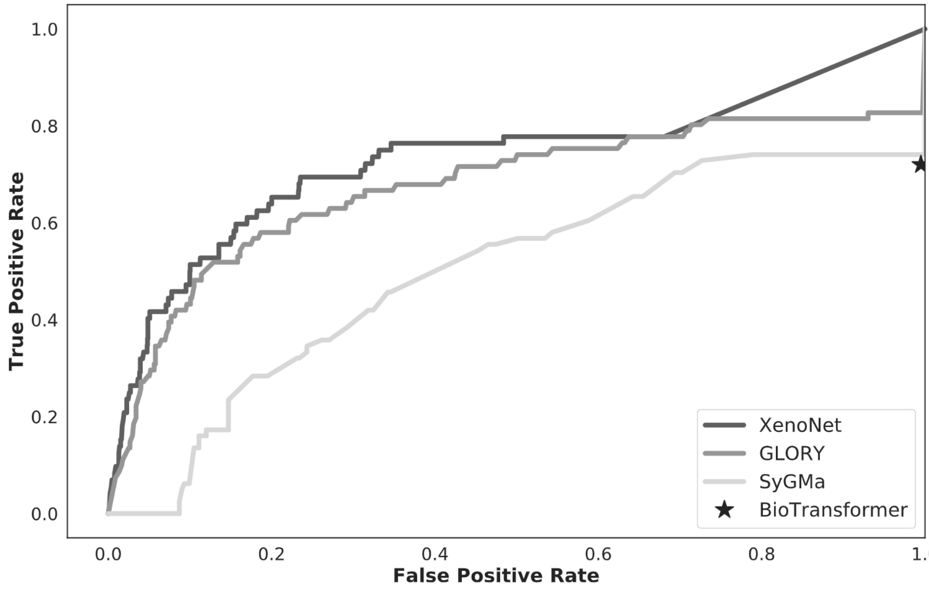


Figure 13. XenoNet is superior to published metabolite prediction models. The receiver-operating characteristic curves (ROC) and their corresponding areas under the curve (AUC) are reliable metrics to compare between the models. The AUC of XenoNet, GLORY MaxCoverage, and SyGMa are 73.3, 67.6, and 50.1%, respectively. We were unable to construct a ROC curve and calculate the AUC for BioTransformer based on their publication, but their recall of 0.72 is lower than other models if all of them were set to have the same precision of 0.17.

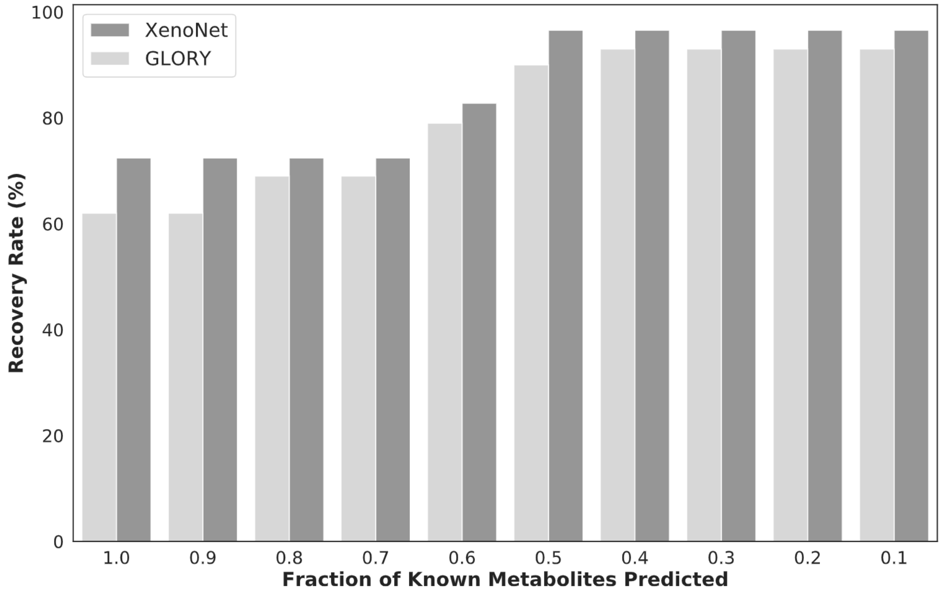


Figure 14. XenoNet is superior to GLORY MaxCoverage in terms of the recovery rate. Across all thresholds, XenoNet's recovery rate is higher than GLORY's.

the same precision of 0.17 (Figure 13). In addition, since each metabolic transformation was assigned with a score based on Rainbow XenoSite's prediction and the network depth limit was set to 1, we use these scores as the proxy for the likelihood that the corresponding metabolites would exist. Here, the top-*N* metric is the fraction of substrates that have at least one experimentally observed metabolite among their group of *N* predicted metabolites with the highest scores. The top-3 performance for XenoNet, GLORY, and SyGMa is 79.3, 75.9, and 69.0%, respectively.

Moreover, XenoNet predicts many more metabolite structures than other models because it has a broader chemical transformation rule set. XenoNet's rule set was derived from Rainbow XenoSite's database, which covered 92.3% of phase-I reactions. In contrast, the most extensive rule sets by previous models, GLORY, was based on FAME 2, which covered only 48.0% of phase-I reactions.[30] While this feature lowers XenoNet's precision, it could also discover metabolites that were missed by experimental methods. Depending on the experimental assay and conditions used, certain metabolites are not easily detected. Computational prediction could serve as a guide for future experiments.[16,18]

In terms of the recovery rate, XenoNet outperformed GLORY MaxCoverage Mode at all thresholds (Figure 14). GLORY MaxCoverage Mode was chosen for comparison because it previously outperformed BioTransformer, SyGMa, and GLORY MaxEfficiency Mode with regard to the recovery rate metric.[34] XenoNet can predict at least 50% of the known metabolites for 96.6% of the parent molecules in the test data set, while GLORY MaxCoverage did so for 90% of the parent molecules. The proportion of parent molecules that have all of their known metabolites predicted is 72.4% for XenoNet and 62.0% for GLORY MaxCoverage.

In terms of absolute numbers, XenoNet captured 72 of the 81 known metabolites in the test set. The method with the next highest number of captured metabolites, GLORY MaxCoverage, yields 67 of the known metabolites. There were nine metabolites of eight parent molecules that XenoNet was not able to predict when the models' depth limit was 1. These nine metabolites were also missed by published models. However, XenoNet is designed to infer metabolic structures from preceding known, potentially reactive, metabolites. We wanted to test whether XenoNet's ability to specify a given target metabolite in addition to a start metabolite could allow for the detection of pathways between each of the parent molecules and their missed metabolite(s). XenoNet was run at a depth limit of 3 on 9 substrate−product pairs of molecules representing the parent molecule and one of their missed metabolites. For 5 of the 9 networks, paths were found linking the parent molecule and their previously missed metabolite. The generated networks by XenoNet are included in Figures S2−S6. The four remaining pairs of parent molecules and their missed metabolites that no method found a valid path for are included in Figure S7. On manual inspection, most of the missed cases require metabolic transformations that are phase-II reaction types or rare phase-I reaction types that our Rainbow XenoSite model does not yet account for. We plan to expand our rule set to cover these reactions in future work.

■ MODEL LIMITATIONS AND FUTURE DIRECTIONS

Future development of XenoNet can be classified into improvements and extensions of the model and further exploration of the model's capabilities to real-world application. To begin with, a key assumption in the current implementation is that the probability of each metabolic transformation is memoryless or independent of the state it came from. This deficiency may be overcome by adjusting probabilities to take contextual dependencies into account. As an example, deeper metabolic steps are less likely since excretion probability is higher with each given transformation. One way to amend this would be to have a parameter that adjusts the metabolic transformation probabilities as a function of the depth at which the transformation takes place.

Besides, the current heuristics could be modified further to resolve weaknesses highlighted by the previous comparisons raised between the XenoNet variants. With regards to the top-*N* heuristics, the relationship between reaction-type specific and reaction-type agnostic variants can be explored further. Specifically, we need to find the best way to tune the value of *N* on large data sets of substrate−product pairs outside of the AMD data set.

If the rule sets from the Metabolic Forest model can be implemented in the opposite direction, from child to parent molecule, then a bidirectional search could also be employed. Consider a bidirectional search on a network with a depth limit of 4. The bidirectional search will aim to form two sets of paths. The first set of paths is formed by enumerating all paths of a depth limit of 2 that begin with the starting metabolite. The second set of paths is formed by enumerating all paths of a depth limit of 2 that begin with the target metabolite. Afterward, paths from the start to the target may be discovered by linking the two sets of paths at points where they have overlapping states.

As previously mentioned, since heuristics are deterministic, improvements may be noticed if a variant is developed that introduces stochasticity, which could escape the inference of spurious metabolites. The top-*N* heuristic could be modified to be a Monte−Carlo heuristic, where we generate *N* metabolites drawn according to their relative probabilities across all the possible metabolites. We can run several trials of Monte−Carlo sampling and concatenate all discovered pathways. Such a heuristic would increase the likelihood of discovering pathways with a low probability first step, but successive high probability steps.

An additional mode allowing for multiple targets to be specified in the input could also be implemented to improve the functionality of XenoNet. Furthermore, XenoNet infers metabolites and assigns predictions to the metabolic transformations that preclude them, but it does not capture the full bioactivation process. Previously, we have developed a SOM model for predicting the reactivity of a molecule with respect to DNA, protein, GSH, and cyanide.[41] A natural next step is to incorporate this reactivity model into XenoNet as a way to predict whether an inferred metabolite is likely to be reactive.

Finally, XenoNet is specifically designed to infer intermediate metabolites when drug molecules form reactive metabolites. XenoNet's capabilities could be further assessed by using it to screen for missing intermediates among known, withdrawn drugs that form reactive metabolites. XenoNet could be applied to drug-reactive metabolite pairs to find sequential metabolic transformations that lead to reactive metabolite formation and identify previously unknown intermediate metabolites for further experimental validation.

■ CONCLUSIONS

We have established a method, XenoNet, that combines a SOM model with a structure inference model for the enumeration of metabolic pathways between a known parent molecule and target molecule and the intermediate metabolite structures that link them. XenoNet can predict experimentally observed pathways and intermediate metabolites with high accuracies. Our method can also function in a similar capacity to prior methods, such as BioTransformer and GLORY, when only given a parent molecule as input. When given the task, XenoNet outperforms prior methods across multiple metrics. While we have yet to model the full bioactivation pathway potential between two molecules or starting from a single molecule, we anticipate the successful incorporation of reactivity models to XenoNet's workflow in the near future. Incorporation of reactivity into XenoNet is a natural extension of the current work and will hopefully cement XenoNet as an informative tool that experimentalists can use to generate specific, testable hypotheses for understanding reactive metabolite formation. Importantly, if it helps experimentalists discover otherwise unknown intermediates, they can then use that knowledge to modify drug molecules to prevent the formation of the intermediate(s) that are antecedent to reactive metabolite formation.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.0c00361.

> The "Metabolic_Network_Dataset.json" file contains the 17 054 metabolic networks that were derived from AMD database; metabolic networks are stored in JSON format and each network is most easily parsed via the NetworkX library in Python; in accordance with the AMD license agreement, we are not allowed to disclose the structures of molecules in the networks; as such, the nodes of each network do not contain structural information and are instead numbered beginning from 0 for each network; to enable rebuilding the complete metabolic network data set, each network has its edges annotated with their corresponding reaction AMD registry numbers (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

S. Joshua Swamidass − *Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, Missouri 63110, United States;* orcid.org/0000-0003-2191-0778; Email: swamidass@wustl.edu

Authors

Noah R. Flynn − *Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, Missouri 63110, United States;* orcid.org/0000-0002-8542-8887

Na Le Dang − *Department of Pathology and Immunology, Washington University School of Medicine, St. Louis, Missouri 63110, United States;* orcid.org/0000-0001-7458-1264

Michael D. Ward − *Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine, St. Louis, Missouri 63110, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.0c00361

Funding

Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award no. R01LM012222 and R01LM012482. Computations were performed using the facilities of the Washington University Center for High Performance Computing, which were partially funded by National Institutes of Health Grants nos. 1S10RR022984-01A1 and 1S10OD018091-01. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We also thank both the Department of Immunology and Pathology at the Washington University School of Medicine and the Washington University Center for Biological Systems Engineering for their generous support of this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge that they have no competing financial interests. We thank the developers of the open-source chemoinformatics tools Open Babel[42] and RDKit, of which we made extensive use. Specially thanks to Tyler B. Hughes for developing the Metabolic Forest model, which our current model employed heavily. Computations were performed using the facilities of the Washington University Center for High Performance Computing, which were partially funded by NIH Grants nos. 1S10RR022984-01A1 and 1S10OD018091-01.

■ ABBREVIATIONS

AMD, Accelrys Metabolite Database; AUC, area under the receiver-operating characteristic curve; IADR, idiosyncratic adverse drug reaction; XenoNet, metabolic network; ROC, receiver-operating characteristic; SoM, site of metabolism

■ REFERENCES

(1) Kalgutkar, A. S.; Didiuk, M. T. Structural Alerts, Reactive Metabolites, and Protein Covalent Binding: How Reliable Are These Attributes as Predictors of Drug Toxicity. *Chem. Biodiversity* **2009**, *6*, 2115−2137.

(2) Ostapowicz, G.; Fontana, R. J.; Schiødt, F. V.; Larson, A.; Davern, T. J.; Han, S. H.; McCashland, T. M.; Shakil, A. O.; Hay, J. E.; Hynan, L.; Crippin, J. S.; Blei, A. T.; Samuel, G.; Reisch, J.; Lee, W. M. Results of a prospective study of acute liver failure at 17 tertiary care centers in the united states. *Ann. Intern. Med.* **2002**, *137*, 947−954.

(3) Srivastava, A.; Maggs, J.; Antoine, D.; Williams, D.; Smith, D.; Park, B. *Adverse Drug Reactions*; Springer, 2010; pp 165−194.

(4) Watkins, P. B.; Seeff, L. B. Drug-Induced Liver Injury: Summary of a Single Topic Clinical Research Conference. *Hepatology* **2006**, *43*, 618−631.

(5) Babai, S.; Auclert, L.; Le-Louie, H. Safety data and withdrawal of hepatotoxic drugs. *Therapie*. 2018. DOI: 10.1016/j.therap.2018.02.004.

(6) Uetrecht, J.; Naisbitt, D. J. Idiosyncratic Adverse Drug Reactions: Current Concepts. *Pharmacol. Rev.* **2013**, *65*, 779−808.

(7) Aithal, G. P.; Ramsay, L.; Daly, A. K.; Sonchit, N.; Leathart, J. B. S.; Alexander, G.; Kenna, J. G.; Caldwell, J.; Day, C. P. Hepatic adducts, circulating antibodies, and cytokine polymorphisms in patients with diclofenac hepatotoxicity. *Hepatology* **2004**, *39*, 1430−1440.

(8) Robin, M. A.; Maratrat, M.; Roy, M. L.; Breton, F. P. L.; Bonierbale, E.; Dansette, P.; Ballet, F.; Mansuy, D.; Pessayre, D. Antigenic targets in tienilic acid hepatitis. Both cytochrome P450 2C11 and 2C11-tienilic acid adducts are transported to the plasma membrane of rat hepatocytes and recognized by human sera. *J. Clin. Invest.* **1996**, *98*, 1471−1480.

(9) Cribb, A. E.; Nuss, C. E.; Alberts, D. W.; Lamphere, D. B.; Grant, D. M.; Grossman, S. J.; Spielberg, S. P. Covalent Binding of Sulfamethoxazole Reactive Metabolites to Human and Rat Liver Subcellular Fractions Assessed by Immunochemical Detection. *Chem. Res. Toxicol.* **1996**, *9*, 500−507. PMID: 8839055.

(10) Skipper, P. L.; Kim, M. Y.; Sun, H. L.; Wogan, G. N.; Tannenbaum, S. R. Monocyclic aromatic amines as potential human carcinogens: old is new again. *Carcinogenesis* **2010**, *31*, 50−58.

(11) Wells, P. G.; Lee, C. J. J.; McCallum, G. P.; Perstin, J.; Harper, P. A. *Adverse Drug Reactions*; Springer: Berlin, 2010; pp 131−162.

(12) Kalgutkar, A. S.; Gardner, I.; Obach, R. S.; Shaffer, C. L.; Callegari, E.; Henne, K. R.; Mutlib, A. E.; Dalvie, D. K.; Lee, J. S.; Nakai, Y.; O'Donnell, J. P.; Boer, J.; Harriman, S. P. A Comprehensive Listing of Bioactivation Pathways of Organic Functional Groups. *Curr. Drug Metab.* **2005**, *6*, 161−225.

(13) Evans, D. C.; Watt, A. P.; Nicoll-Griffith, D. A.; Baillie, T. A. Drug-Protein Adducts: An Industry Perspective on Minimizing the Potential for Drug Bioactivation in Drug Discovery and Development. *Chem. Res. Toxicol.* **2004**, *17*, 3−16.

(14) Lewis, D. F.; Ito, Y. Human cytochromes P450 in the metabolism of drugs: new molecular models of enzyme-substrate interactions. *Expert Opin. Drug Metab. Toxicol.* **2008**, *4*, 1181−1186. PMID: 18721112.

(15) Testa, B.; Pedretti, A.; Vistoli, G. Reactions and Enzymes in the Metabolism of Drugs and Other Xenobiotics. *Drug Discovery Today* **2012**, *17*, 549−560.

(16) Dang, N. L.; Hughes, T. B.; Miller, G. P.; Swamidass, S. J. Computationally Assessing the Bioactivation of Drugs by N-Dealkylation. *Chem. Res. Toxicol.* **2018**, *31*, 68−80.

(17) Iverson, S. L.; Uetrecht, J. P. Identification of a Reactive Metabolite of Terbinafine: Insights Into Terbinafine-Induced Hepatotoxicity. *Chem. Res. Toxicol.* **2001**, *14*, 175−181.

(18) Barnette, D. A.; Davis, M. A.; Dang, N. L.; Pidugu, A. S.; Hughes, T.; Swamidass, S. J.; Boysen, G.; Miller, G. P. Lamisil (terbinafine) toxicity: Determining pathways to bioactivation through computational and experimental approaches. *Biochem. Pharmacol.* **2018**, *156*, 10−21.

(19) 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*, No. 113661.

(20) 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.

(21) Rydberg, P.; Gloriam, D. E.; Zaretzki, J.; Breneman, C.; Olsen, L. SMARTCyp: A 2D Method for Prediction of Cytochrome P450-Mediated Drug Metabolism. *ACS Med. Chem. Lett.* **2010**, *1*, 96−100.

(22) Zaretzki, J.; Rydberg, P.; Bergeron, C.; Bennett, K. P.; Olsen, L.; Breneman, C. M. RS-Predictor Models Augmented With SMARTCyp Reactivities: Robust Metabolic Regioselectivity Predictions for Nine CYP Isozymes. *J. Chem. Inf. Model.* **2012**, *52*, 1637−1659.

(23) Rudik, A. V.; Dmitriev, A. V.; Lagunin, A. A.; Filimonov, D. A.; Poroikov, V. V. Metabolism Site Prediction Based on Xenobiotic Structural Formulas and PASS Prediction Algorithm. *J. Chem. Inf. Model.* **2014**, *54*, 498−507.

(24) Rudik, A.; Dmitriev, A.; Lagunin, A.; Filimonov, D.; Poroikov, V. SOMP: Web-Server for in Silico Prediction of Sites of Metabolism for Drug-Like Compounds. *Bioinformatics* **2015**, *31*, 2046−2048.

(25) Adams, S. E. Molecular Similarity and Xenobiotic Metabolism. Ph.D. Thesis, University of Cambridge, 2010.

(26) Šícho, M.; Stork, C.; Mazzolari, A.; de Bruyn Kops, C.; Pedretti, A.; Testa, B.; Vistoli, G.; Svozil, D.; Kirchmair, J. FAME 3: Predicting the Sites of Metabolism in Synthetic Compounds and Natural Products for Phase 1 and Phase 2 Metabolic Enzymes. *J. Chem. Inf. Model.* **2019**, *59*, 3400−3412.

(27) Zheng, M.; Luo, X.; Shen, Q.; Wang, Y.; Du, Y.; Zhu, W.; Jiang, H. Site of Metabolism Prediction for Six Biotransformations Mediated by Cytochromes P450. *Bioinformatics* **2009**, *25*, 1251−1258.

(28) He, S.; Li, M.; Ye, X.; Wang, H.; Yu, W.; He, W.; Wang, Y.; Qiao, Y. Site of Metabolism Prediction for Oxidation Reactions Mediated by Oxidoreductases Based on Chemical Bond. *Bioinformatics* **2017**, *33*, 363−372.

(29) Matlock, M. K.; Hughes, T. B.; Swamidass, S. J. XenoSite Server: A Web-Available Site of Metabolism Prediction Tool. *Bioinformatics* **2015**, *31*, 1136−1137.

(30) 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.

(31) Djoumbou-Feunang, Y.; Fiamoncini, J.; Gil-de-la Fuente, A.; Greiner, R.; Manach, C.; Wishart, D. S. BioTransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *J. Cheminf.* **2019**, *11*, No. 2.

(32) Meng, J.; Li, S.; Liu, X. RD-Metabolizer: an integrated and reaction types extensive approach to predict metabolic sites and metabolites of drug-like molecules. *Chem. Cent. J.* **2017**, *11*, No. 65.

(33) Ridder, L.; Wagener, M. SyGMa: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites. *ChemMedChem* **2008**, *3*, 821−832.

(34) de Bruyn Kops, C.; Stork, C.; Šícho, M.; Kochev, N.; Svozil, D.; Jeliazkova, N.; Kirchmair, J. GLORY: Generator of the Structures of Likely Cytochrome P450 Metabolites Based on Predicted Sites of Metabolism. *Front. Chem.* **2019**, *7*, No. 402.

(35) Tian, S.; Djoumbou-Feunang, Y.; Greiner, R.; Wishart, D. S. CypReact: A Software Tool for in Silico Reactant Prediction for Human Cytochrome P450 Enzymes. *J. Chem. Inf. Model.* **2018**, *58*, 1282−1291.

(36) Hughes, T. B.; Dang, N. L.; Kumar, A.; Flynn, N. R.; Swamidass, S. J. The Metabolic Forest: Predicting the Diverse Structures of Drug Metabolites *Under Review*, 2020.

(37) Landrum, G. Open-Source Cheminformatics and Machine Learning, 2006, http://www.rdkit.org/ (accessed June 14, 2017).

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

(39) McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153−157.

(40) Dietterich, T. G. Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms. *Neural Comput.* **1998**, *10*, 1895−1923.

(41) Hughes, T. B.; Dang, N. L.; Miller, G. P.; Swamidass, S. J. Modeling Reactivity to Biological Macromolecules with a Deep Multitask Network. *ACS Cent. Sci.* **2016**, *2*, 529−537.

(42) OLBoyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminf.* **2011**, *3*, No. 33.