Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

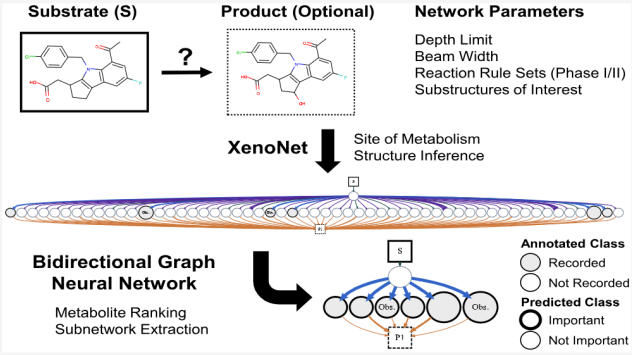
Message Passing Neural Networks Improve Prediction of Metabolite Authenticity

Noah R. Flynn and S. Joshua Swamidass*

Cite This: J. Chem. Inf. Model. 2023, 63, 1675−1694 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Cytochrome P450 enzymes aid in the elimination of a preponderance of small molecule drugs, but can generate reactive metabolites that may adversely react with protein and DNA and prompt drug candidate attrition or market withdrawal. Previously developed models help understand how these enzymes modify molecule structure by predicting sites of metabolism or characterizing formation of metabolite-biomolecule adducts. However, the majority of reactive metabolites are formed by multiple metabolic steps, and understanding the progenitor molecule's network-level behavior necessitates an integrative approach that blends multiple site of metabolism and structure inference models. Our previously developed tool, XenoNet 1.0, generates metabolic networks, where nodes are molecules and weighted edges are metabolic transformations. We extend XenoNet with a bidirectional message passing neural network that integrates edge feature information and local network structure using edge-conditioned graph convolutions and jumping knowledge to predict the authenticity of inferred Phase I metabolite structures. Our model significantly outperformed prior work and algorithmic baselines on a data set of 311 networks and 6606 intermediates annotated using a chemically diverse set of 20 736 individual in vitro and in vivo reaction records accounting for 92.3% of all human Phase I metabolism in the Accelrys Metabolite Database. Cross-validated predictions resulted in area under the receiver operating characteristic curves of 88.5% and 87.6% for separating experimentally observed and unobserved metabolites at global and network levels, respectively. Further analysis verified robustness to networks of varying depth and breadth, accurate detection of metabolites, such as D,L-methamphetamine, that are experimentally observed or unobserved in different network contexts, extraction of important metabolic subnetworks, and identification of known bioactivation pathways, such as for nimesulide and terbinafine. By exploiting network structures, our approach accurately suggests unreported metabolites for experimental study and may rationalize modifications for avoiding deleterious pathways antecedent to reactive metabolite formation.



INTRODUCTION

In silico prediction of reactive metabolites is an important cheminformatics problem. Reactive metabolite formation is an unfortunate consequence of drug-clearing defense mechanisms, i.e., metabolism.[1] Metabolism is generally beneficial, making drugs more hydrophilic and easier to excrete. However, metabolism can transform drugs into pernicious reactive metabolites that may conjugate to DNA or off-target proteins and result in adverse events.[2,3] In particular, P450s are highly relevant to bioactivation processes surrounding drugs, as they collectively have more substrates than any other enzymes and several of their reaction products have strong electrophilic properties. Reactive metabolites are significant drivers of drug candidate attrition and market withdrawal.[4−6] Detection of reactive metabolite formation during metabolism of a drug to known structural end points could be leveraged to engineer rational modifications that bypass formation of the reactive metabolite in favor of benign metabolic pathways.

Current in vitro and in vivo methods for reactive metabolite detection, such as metabolite trapping[7] and covalent binding studies,[8,9] are time-, labor-, and cost-intensive. In addition to prohibitive expenses of manual examination, in vivo study is difficult because reactive metabolites generally are transitory and do not circulate. In contrast, in silico approaches can confidently supply a fast, inexpensive method to triage studies and understand step-by-step formation of reactive metabolites, which helps design safer drugs and overcome experimental bottlenecks. Ideally, the model can classify computationally inferred intermediate metabolites as likely to be present or not given a small dataset of manually labeled metabolic networks.

We previously developed XenoNet, which receives an input substrate molecule and optional target metabolite(s) and

Received: November 1, 2022
Published: March 16, 2023

ACS Publications © 2023 American Chemical Society

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

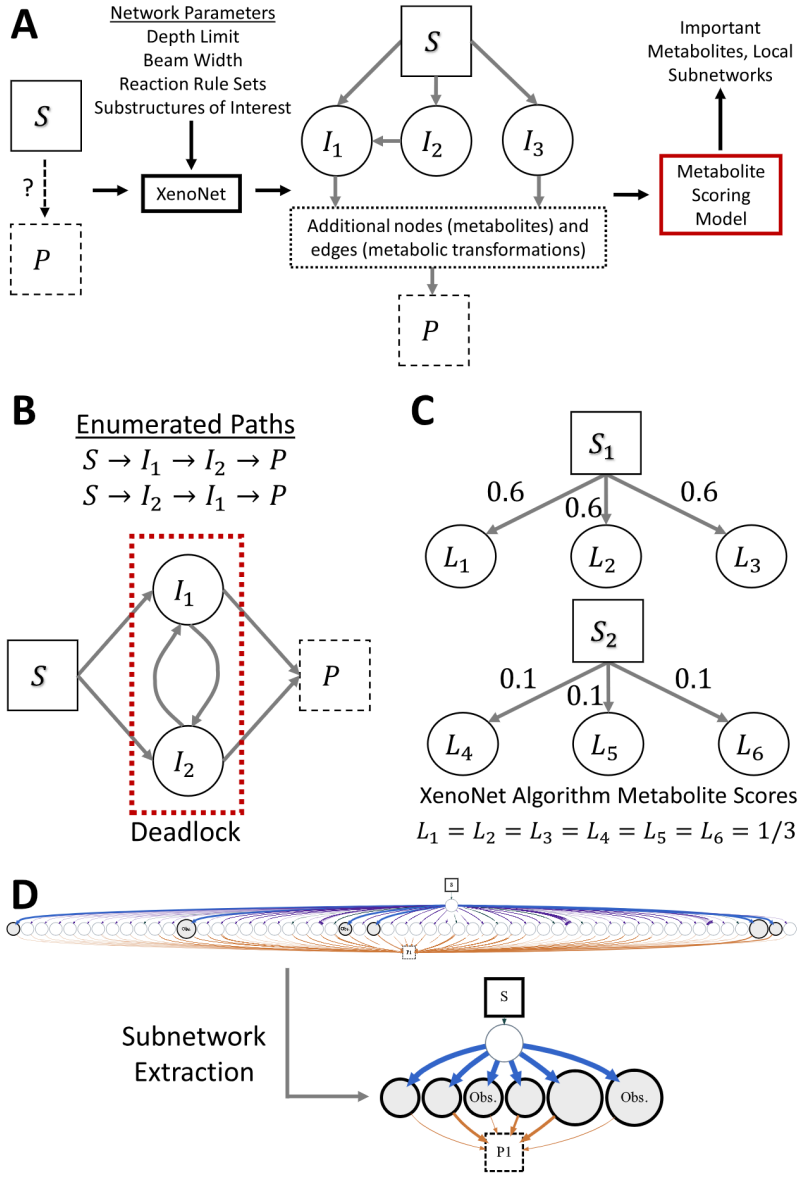


Figure 1. (A) Modified XenoNet workflow to incorporate metabolite predictions. Predicting the authenticity of inferred metabolites requires a "metabolite scoring model", such as the XenoNet algorithm, that processes a generated network. Important metabolites or their local subnetworks can be extracted for further study or to highlight processes pertinent to toxicity. (B) Neither path $S \rightarrow I_1 \rightarrow I_2 \rightarrow P$ nor path $S \rightarrow I_2 \rightarrow I_1 \rightarrow P$ individually contain a cycle. Addition of both paths to the network does result in a cycle between nodes I_1 and I_2. Cycles are problematic for the XenoNet metabolite scoring algorithm and provoke deadlock conditions. (C) Choice of normalization strategy results in loss of information for comparing metabolites and their formation scores across different networks. (D) Our proposed model is able to designate intermediate metabolites of greatest authenticity, as well as extract important connected subnetworks for focused review.

generates a metabolic network.[10] In the metabolic network, molecules are nodes and directional edges convey metabolic transformations. The metabolic network is generated by enumerating pathways, or sequences of intermediate metabolites, between the substrate and targets and predicting the probability of each metabolic transformation. XenoNet is not the only tool for predicting and inferring metabolites from a substrate molecule.[11,12] However, it remains the only tool for enumerating scored paths between a substrate and target metabolite(s) to form a network, allowing for identification of elusive intermediate metabolites that may be easily missed in experimental studies.

A metabolite's presence in the generated network may correspond to an authentic experimental observation or spurious model inference (Figure 1A). We previously proposed the XenoNet algorithm[10] to output a "metabolite formation score" for each metabolite as a proxy for its authenticity. The XenoNet algorithm propagates information from the substrate down each branch of the metabolic network to score every child node by a combination of the edge weights leading to the child node and the metabolite scores of the child's parent nodes.

The XenoNet algorithm has several deficiencies. It requires multiple passes across the network equivalent to the length of its longest path. Furthermore, the algorithm assumes a directed, acyclic graph. To score a node, that node's parents must have already been scored. In a cycle, each node is waiting on other nodes in the cycle to be scored and a deadlock occurs.

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

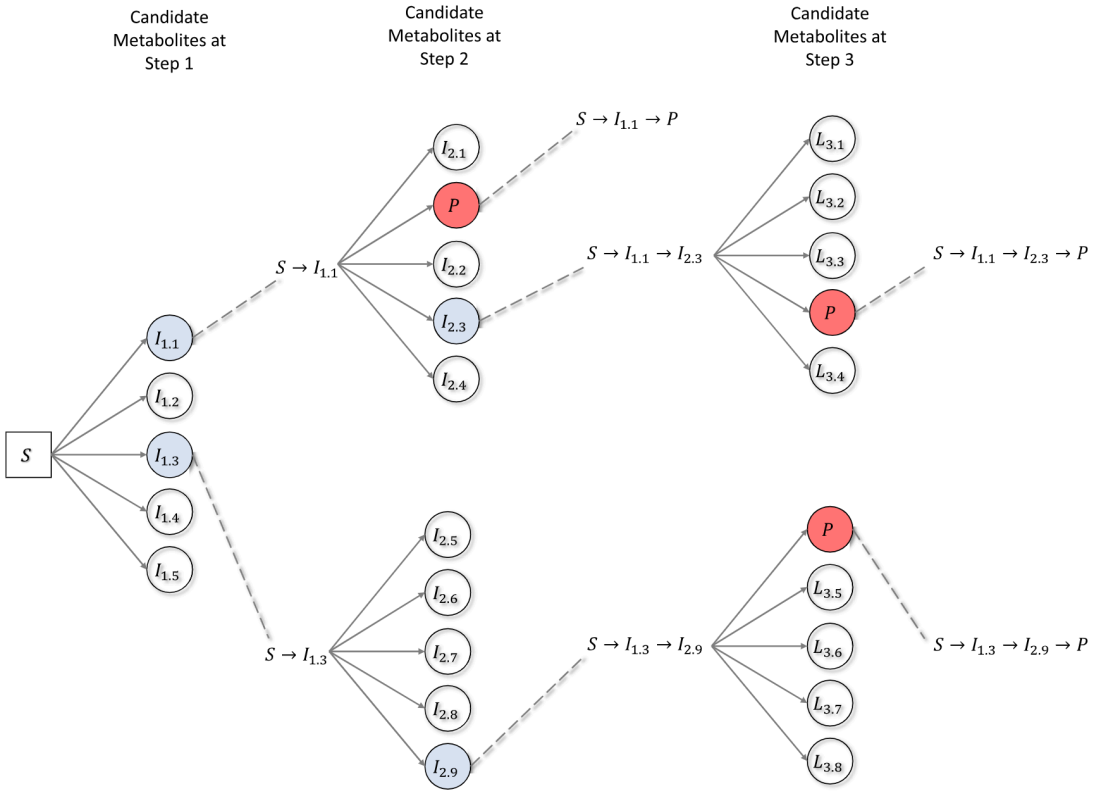


Figure 2. Example process of beam search using a beam width of 2 and depth limit of 3. *S*, *I*, and *P* signify the substrate (or start) molecule, the intermediate metabolites, and the product (or target) metabolite, respectively. At the first step, suppose that the transformations leading to $I_{1.1}$ and $I_{1.3}$ have the highest prediction. Then the second search step commences across all possible paths branching out from $I_{1.1}$ and $I_{1.3}$, where we suppose that the two-step pathways with highest probabilities that pass through either $I_{1.1}$ or $I_{1.3}$ are $S \rightarrow I_{1.1} \rightarrow I_{2.3}$ and $S \rightarrow I_{1.3} \rightarrow I_{2.9}$. Even if the two-step pathway $S \rightarrow I_{1.1} \rightarrow P$ is less probable than the aforementioned two-step pathways and does not satisfy the beam width criteria, we retain that pathway in the network because we know it terminates at *P*. The process continues until the depth limit of 3 is reached. The output metabolite sequences are $S \rightarrow I_{1.1} \rightarrow I_{2.3} \rightarrow P$, $S \rightarrow I_{1.3} \rightarrow I_{2.9} \rightarrow P$, and $S \rightarrow I_{1.1} \rightarrow P$.

XenoNet does not add individual cycle-containing paths, but this does not prohibit the formation of cycles in the network (Figure 1B), and so computing metabolite scores must be compatible with cyclic structures.

Moreover, XenoNet's metabolite scores are not well-scaled—the confidence of the model does not reflect its accuracy—and they do not generalize well when comparing metabolites across different networks (Figure 1C). Currently, normalization is applied across edges that share a common 1-hop predecessor node within the same network. A caveat of this strategy is loss of information necessary for relating nodes with membership in different networks. Thus, a metabolite's score only has meaning relative to other metabolites in the same network and is not comparable to metabolites outside of its network. Lastly, the XenoNet algorithm is inflexible to descriptors beyond the raw edge scores predicted by our Phase I site of metabolism (SoM) model.[13] For instance, XenoNet supports 20 metabolic transformations, but predictions from different transformations scale differently and are not directly comparable without considering the transformation's reaction type.

We propose a method that addresses the aforementioned deficiencies and increases performance across several metrics. Instead of manually developing an algorithmic approach, we train a graph neural network to learn an encoding function that maps nodes into a low-dimensional space, where their position in the embedding space corresponds to a measure of their authenticity within the context of Phase I metabolism. Rather than score a node based only on its features, a graph neural network is able to capture the node's local network structure and enrich its feature representation by borrowing information from neighboring nodes.

This modeling approach requires only a single pass across all nodes in the network, manages networks with cycles, results in increased accuracy and calibration, and accounts for numerical and categorical features, such as reaction type, beyond just edge weights. As the embedding space represents a learned representation of metabolites across all networks, the learned embeddings represent an optimal normalization strategy and allow comparison of metabolites within an individual and across multiple networks. Finally, the learned model is robust to network inputs of varying sizes, accurately adjusts predictions of the same metabolite structure in different network contexts, and highlights important subnetworks for focused study (Figure 1D).

DATA AND METHODS

XenoNet. XenoNet integrates multiple machine-learning approaches to modeling P450 metabolism and reactivity with a rule-based structure inference model.[10] XenoNet iteratively chains P450 SoM models with a structure inference model, enumerating acyclic sequences of metabolic transformations (edges) and metabolites (nodes) that are then stored in a directed, multiedge graph-based data structure. Specifically, we

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

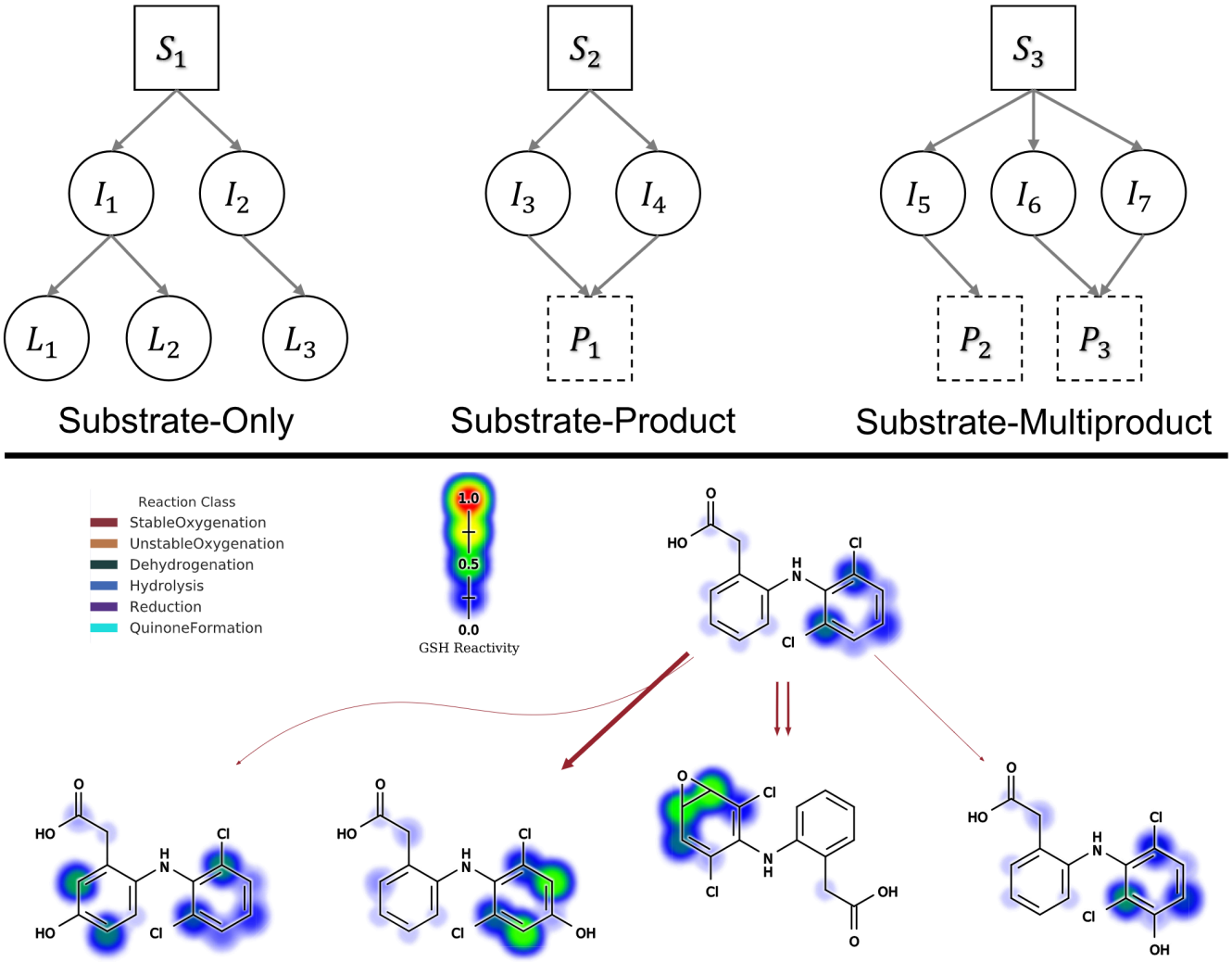


Figure 3. (Top) Example network structures for all three cases: no target metabolite, one target metabolite, and multiple target metabolites. *S* denotes a substrate, or start, molecule. *I* denotes an intermediate metabolite. *L* and *P* denote leaf metabolites and product, or target metabolites, respectively. (Bottom) Example of a network generated by XenoNet for diclofenac metabolism at a depth limit of 1, a beam width of 4 child metabolites, and with reactivity to GSH. Dilofenac is an anti-inflammatory drug that has been implicated in incidences of drug-induced liver injury. At a limited depth of 1, the generated network infers the presence of several experimentally reported intermediate metabolites, formed via epoxidation or hydroxylation, that are known precursors to reactive metabolites or known to readily form adducts with GSH.[16]

use our rule-based structure inference model, the Metabolic Forest, to infer possible SoMs on the parent molecule at which a metabolic transformation will initiate at and the resultant child metabolite structure(s) .[14] Multiple edges may occur between substrate and metabolite when the same reaction type initiates at different sites but forms the same metabolite. For example, epoxide formation might assign either of the two carbons as the site of metabolism. In tandem, our deep learning Phase I SoM model, Rainbow XenoSite, provides a probabilistic prediction of the corresponding metabolic transformation.[13] We then reapply this procedure on the inferred child metabolites to build out a widening tree of potential pathways across multiple metabolic transformations.

Additional network construction parameters further constrain the network's search space. Namely, the depth limit specifies the maximum number of edges a path can be constructed with and the beam width specifies the maximum number of candidate paths under consideration at any search step. Upon surpassing the maximum allotted time parameter, XenoNet terminates network generation and outputs the partially generated network. The user may also impose constraints on the reaction rule sets used and the valid sites to conduct a search across via substructure matching, resulting in a more focused and efficient search (see Figures S1−S3 in the Supporting Informatin).

Extensions to XenoNet. *Beam Search.* XenoNet now finds paths leading from a start molecule to target metabolites using beam search (Figure 2). At each search step, XenoNet may consider thousands of possible child metabolites and prioritizes finding the most likely paths. A greedy approach selects the most probable metabolites at each search step, but is problematic because it is impossible to know a priori which chosen metabolite will lead to the more promising future. The effect of choosing a metabolite may not be obvious until several search steps later and early choice of the wrong metabolite may cumulatively lead to a poorer overall path probability over time.

We perform an exhaustive search, retain all choices, and fork the network with every possible molecule under consideration at a given step. Since this approach rapidly blows up, we

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

instead retain only the top N scoring forks at each search step. We choose the top N based on the product of probabilities of all metabolites along the path constructed so far. For a given path, termination occurs when the number of metabolic transformations is equal to the depth limit parameter. As we conduct the search, we retain any paths that terminate at any user-defined target metabolites, regardless of the probability of those paths. The number of retained paths may exceed the beam width, as paths that terminate at a target no longer contribute to the beam width threshold once stored in the constructed network. At the penultimate search step, we enumerate all possible candidate child metabolites for each candidate path and retain all paths whose candidate child metabolite is a target metabolite.

Multitarget Network Construction. Initially, XenoNet only created networks where a starting substrate molecule and either 1 or 0 target metabolites are defined. We extended XenoNet to accommodate multiple target metabolites, allowing for quicker generation of overlapping networks (Figure 3). For example, our metabolic network dataset used to validate XenoNet had ∼17 000 substrate molecule and target metabolite pairs.[10] Merging networks that share the same substrate into a single network with one substrate and multiple targets decreased the number of networks to 9686 and reduced redundant computation.

Reactivity Model Integration. Originally, XenoNet was capable of inferring metabolite structures and assigning probabilistic predictions to the metabolic transformations that precede their formation. XenoNet did not provide reactivity predictions and so could not capture the full bioactivation process. Previously, we developed a site of reactivity model, with respect to DNA, protein, GSH, and cyanide.[15] We integrated this model into XenoNet, allowing for computation of atom-level and molecule-level reactivity predictions for each molecule in the generated network (Figure 3). Later, we demonstrate how this modification allows for screening of missing or unknown intermediate metabolites and assessing reactivity of inferred intermediates.

Single Start-Multitarget Metabolism Dataset. We extracted 17 054 annotated networks from the metabolic network dataset described in prior work.[10] The annotated networks are derived from 20 736 individual in vitro and in vivo human Phase I reaction records filtered from the literature-derived Accelrys Metabolite Database (AMD, 2019). Each annotated network is defined by a substrate molecule, a target metabolite, and zero or more experimentally observed intermediate metabolites. Multiple entries share the same substrate molecule but have different target metabolites. To prevent networks with the same substrate molecule contributing unequally to performance metrics, we merged annotated networks with the same substrate molecule, resulting in 9686 annotated networks.

Next, we retained annotated networks with at least one intermediate and used XenoNet to generate metabolic networks with a single substrate molecule, multiple target metabolites, a depth limit of 3, and no beam width constraint. We only retained generated networks that inferred the presence of at least one intermediate metabolite recorded in the corresponding annotated network. The final XenoNet start-multitarget network dataset consisted of 311 pairs of generated networks and their annotated counterparts.

Across all 311 annotated networks, there are 676 known intermediate metabolites, of which 467 are unique. Across all 311 generated networks, XenoNet inferred 516 of the 676 known intermediate metabolites. This discrepancy results because each generated network may not have inferred all its corresponding annotated network's known intermediate metabolites—it only needed to infer at least one known intermediate to be retained. The 96 intermediates in the annotated networks but not the generated networks were not inferred by XenoNet because they require a depth limit greater than 3 to discover (see Figure S4 in the Supporting Information). With respect to the 311 generated networks, the 516 known intermediates were labeled as observed (or 1), and the 6090 remaining intermediates were labeled as unobserved (or 0). Note that not all intermediates (both observed and unobserved) are unique, since the same metabolite can be present across multiple metabolic networks and we want the model to be identify different contexts under which the same inferred metabolite may be authentic or spurious. We provide the start-multitarget networks in the "Start_Multitarget_Training_Dataset.json" file.

DrugBank Dataset. We used the 5.1.6 release version of DrugBank, which contains records for 242 withdrawn, small-molecule drugs, as a source of external data for analysis of case studies. We extracted annotated networks with no reported intermediate metabolite, whose substrate molecule is one of the 242 withdrawn drugs, and who have at least one target metabolite defined in the metabolic network dataset. We removed withdrawn drugs that were present as substrate molecules in the training dataset. The final DrugBank dataset consisted of 70 annotated networks for each remaining withdrawn drug and 70 generated networks containing 2832 intermediate metabolites. We generated the DrugBank networks at a depth limit of 3 and a beam width of 15. We provide the start-multitarget networks used for each network in the DrugBank withdrawn drug dataset in the "Start_Multitarget_DrugBank_Withdrawn_Dataset.json" file.

XenoNet Start-Multitarget Network Structures and Features. We propose neural network models designed around the message passing neural network (MPNN) framework[17] and predicated on the structural properties of the start-multitarget networks. Each start-multitarget network is a directed, weighted network where every path originates from the start molecule and terminates at a target metabolite. Multiple metabolic transformations, of differing reaction types and sites of metabolism, can link a parent molecule to a child metabolite and so the network is a multidigraph. However, some methods used in later comparisons do not support multidigraphs. As a post-processing step, parent and child metabolites connected by multiple edges of same directionality are simplified by only considering the edge with maximum score.

Network edge scores were computed by the Phase I SoM model and can propagate information on which downstream metabolites are likely to be encountered. Intuitively, a higher edge score denotes a higher chance of a metabolic transformation occurring. Flow of information on incoming reactions into a child metabolite may indicate the likelihood of the metabolite occurring and being genuine, as opposed to being unlikely to form or spurious. Nevertheless, the predicted edge scores correspond to one of five reaction types—stable oxygenation, unstable oxygenation, dehydrogenation, reduction, and hydrolysis—and these reaction-type-specific predictions do not scale the same, differ in calibration (Figure S5 in the Supporting Information), and support different optimal

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

binarization thresholds (Figure S6 in the Supporting Information). We do not have an SoM model for epoxide opening and treat it as a separate, sixth class. To increase model capacity to compare predictions across reaction types, we represent each edge by a feature vector containing the Phase I model prediction and a one-hot encoding of its reaction type.

In the absence of node features to inform the metabolite scoring task, node in-degree and node out-degree are easily exploited with minimal computational overhead. Furthermore, we noticed that nodes with high in-degree and low out-degree (stable metabolites) values have a tendency to be experimentally observed and nodes with low in-degree and high out-degree (transient metabolites) have a tendency to be experimentally unobserved (Figure S7 in the Supporting Information). The in-degrees and out-degrees are calculated using the weighted paths entering and exiting each node.

We later use reactivity scores in conjunction with learned metabolite scores to assess the likelihood of reactive intermediate metabolite formation. Formally, each node is represented by a vector containing its predicted metabolite score and its molecule-level and site-level reactivity scores to cyanide, DNA, GSH, and protein. Site-level and molecule-level reactivity scores are not used as node feature information for learning metabolite scores.

Algorithmic Approaches. During model comparison, we assessed several baselines that depart from the MPNN paradigm. We describe baseline algorithmic approaches designed specifically for metabolite scoring (e.g., the XenoNet algorithm), in addition to approaches commonly used to compute node importance (e.g., PageRank and random walk with restarts).

XenoNet Metabolite Scoring Algorithm. The XenoNet metabolite scoring algorithm uses network edge predictions to score nodes (Figure 4). Each edge represents a metabolic transformation and is weighted by a raw score predicted by the Phase I SoM model. Formally, a metabolic transformation between metabolite M_j and one of its children, M_k, is weighted by a raw score $w_{M_j \to M_k}$. We normalize each raw score, $w_{M_j \to M_k}$, by the sum of raw scores over all metabolic transformations from M_j to its children. The normalization step is computed using eq 1 and results in a normalized score, $W_{M_j \to M_k}$, between a metabolite, M_j, and one of its children, M_k.

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

After edge normalization, we iteratively score each metabolite. First, the substrate is assigned a score of 1.0. Second, a downstream metabolite, M_j, is scored using eq 2.

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

The resultant metabolite score, F_{M_j}, is a weighted sum of the normalized scores, $W_{M_i \to M_j}$, where each M_i is one of the parents of M_j and the weight of the normalized score, F_{M_i}, is the score previously computed for M_i.

Scoring a node requires all of its parents to have been scored, so computing metabolite scores is carried out over multiple iterations. Each iteration commits a traversal over the network and checks whether a node can be scored and, if so, scores the node. Assuming a directed, acyclic network, the number of iterations required to score all metabolites is no more than the network's maximum length path.

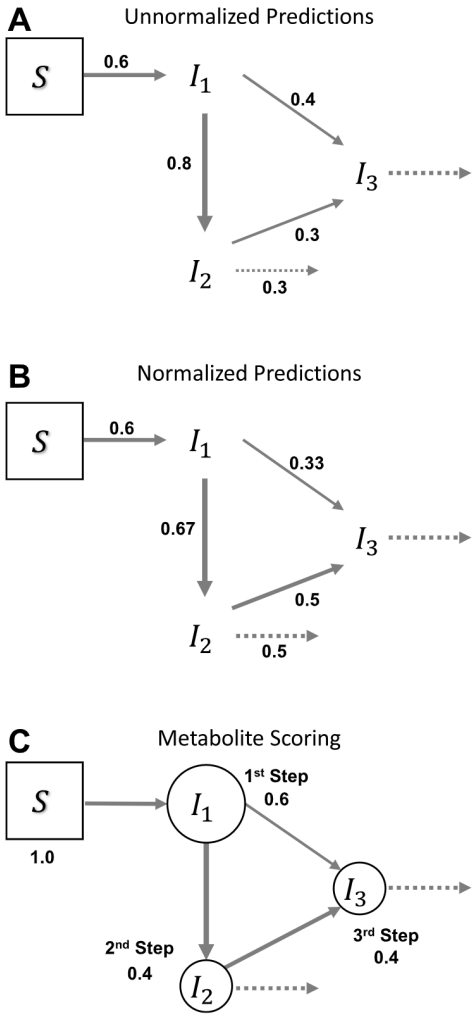


Figure 4. XenoNet algorithm calculates metabolite scores using a three step process. (A) First, edge predictions are computed by the Phase I SoM model. (B) Second, the predictions are normalized using eq 1. (C) Third, the substrate is assigned a score of 1.0 and downstream metabolites are scored using eq 2. A metabolite can be scored only if all its parents are scored. Once I_1 is scored on the first iteration, I_2 can be scored. Once I_2 is scored on the second iteration, I_3 can be scored. The final metabolite score is displayed adjacent to each metabolite. Dashed arrows indicate additional downstream structures in a larger network.

However, 73 (12%) of the start-multitarget XenoNets violate the acyclic assumption. To enable the XenoNet algorithm to handle cyclic networks, a preprocessing step is applied to identify and eliminate cycles by removing the minimum number of edges necessary, in order of edges with the smallest prediction values. This preprocessing step is only applied to the XenoNet algorithm and its random variant. The random model takes each network, randomly permutes the predicted edge weights, and then applies the XenoNet algorithm to score metabolites. Performance metrics for the random model are averaged over 10 trials.

Graph Analysis Algorithms. For comparison, we utilized PageRank, a common centrality measure used to compute the importance of web pages (nodes) in web graphs.[18] PageRank

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

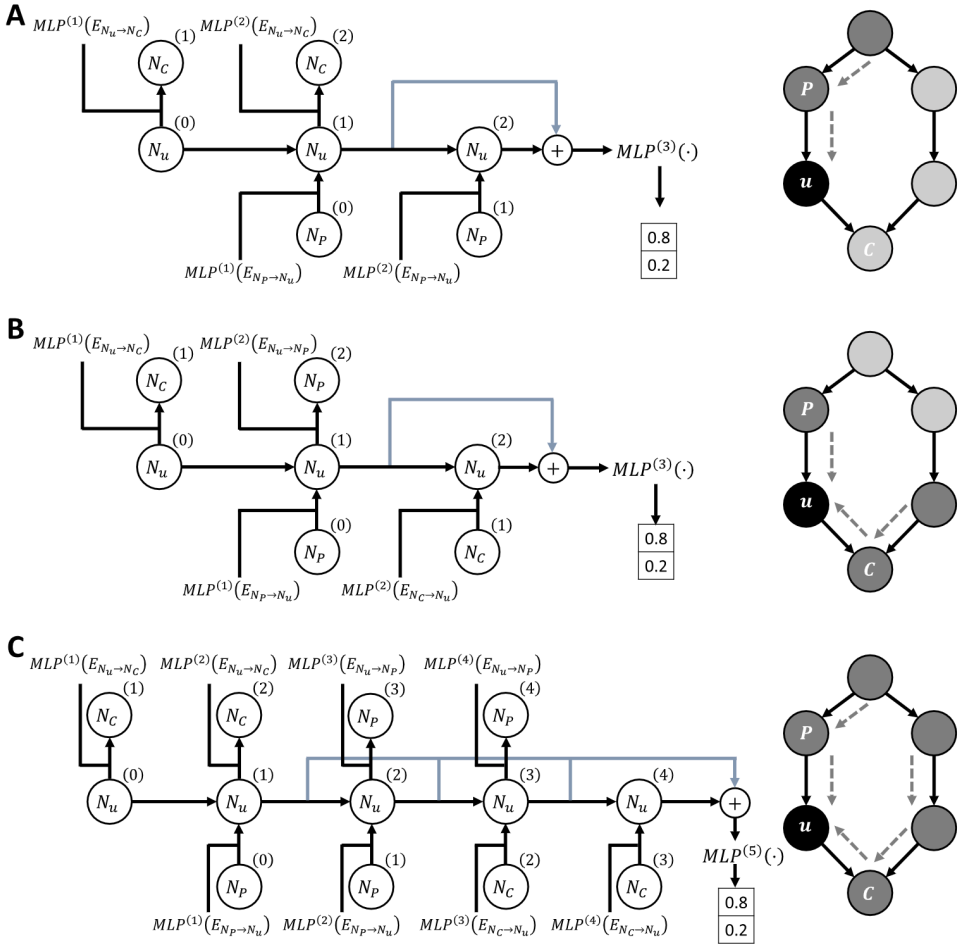


Figure 5. Bidirectional MPNN takes advantage of information flow in both directions, from parent to child and from child to parent nodes, to extend each node's receptive field and improve learning. (A) Example information flow for the 2-layer unidirectional MPNN. At each graph convolution layer, node u is influenced by messages from its parents node(s), N_P, and influences its child node(s), N_C. Messages are conditioned by edge information. For each message passing iterations, the edge features remain constant but are parametrized by distinct MLPs. Output embeddings from each layer are concatenated and fed through another MLP to produce class predictions. Node u is only exposed to, at most, information from the dark gray nodes and is disconnected from half of the network, specifically a competing pathway. (B) Example information flow for the 2-layer bidirectional MPNN. The only modification is to the second graph convolution layer, which reverses flow of message passing such that node u is influenced by messages from its child node(s), N_C, and influences its parent node(s), N_P. Node u is partially exposed to information in both its own pathway and the competing pathway. (C) Example information flow for the 4-layer bidirectional MPNN. The first two and last two layers transmit information flow from parent to child and from child to parent, respectively. Node u gathers information from all nodes in the network.

interprets incoming edges as votes toward a node's importance, where votes are further weighted by the importance of the parent node that is casting the vote and of the score of the edge itself. This process, whereby a node's state is transmitted as a signal through its connections to update its neighboring node's states, continues for multiple iterations until convergence or a stopping criterion is reached. A node's final score reflects the probability that the sequence of metabolic transformations will produce the metabolite. Because of the recursive nature of scoring node importance based on the importance of other nodes in the network, the PageRank implementation is nontrivial and we refer to primary literature for comprehensive description.[19]

We also compared model performance to random walk with restarts and betweenness centrality. For the former, we commence multiple random walks of variable path length starting from the substrate molecule. A node receives higher importance the more times we pass over it. We implemented random walk with restarts as a personalized PageRank where the personalization vector has a value of 1 for the substrate molecule and a value of 0 everywhere else.[20,21] Betweenness centrality considers the extent to which a node contributes to paths between all other nodes and we used the shortest-path variant.[22,23]

Message Passing Neural Network Architectures. We formalize relevant aspects of the neural message passing framework, which incorporates both structural features of the graph and node-, edge-, or graph-level feature information to learn node representations, or embeddings. Each node, u, in the graph is represented by an initial hidden embedding vector, $\mathbf{h}_u^{(0)}$, which is just the node's input features. At each iteration k of message passing, we transmit hidden embeddings between each node's 1-hop neighborhood based on the directionality of the message. The message node u receives, $\mathbf{m}_{N(u)}^{(k)}$, is defined by aggregating the hidden embeddings received from its neighborhood, $N(u)$. The node's hidden embedding for the

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

current iteration is updated based on its hidden embedding at iteration $k-1$, $\mathbf{h}_u^{(k-1)}$, and the aggregated message, $\mathbf{m}_{N(u)}^{(k)}$. After K iterations of message passing, we define each node u's output embedding as $\mathbf{h}_u^{(K)}$.

Intuitively, each message passing iteration exposes a node to information from further away in the graph. After k iterations, the updated node embedding has been influenced by structural information, such as node degrees, or feature information, such as edge scores or reaction types, within its k-hop neighborhood. Ideally, learned embeddings represent a projection of the nodes to a latent space where the distance between points corresponds to similarity in the structural and feature information that is relevant to predicting node importance.

The aggregation and update operators used for each message passing update, formalized in eqs 3 and 4, are derived from prior work, namely, the edge-conditioned convolutional layer.[17,24] Each multilayer perceptron (MLP) is parametrized by learnable network weights Θ that are not shared and are specific to each layer k. This enables the edge weight interpretations, as learned by the MLP, to be different for each graph convolution layer, regardless of the edge features remaining fixed throughout the learning process. For each graph convolution layer, the MLP is defined by 1 input layer, 1 hidden layer of variable size with ReLU activation, and 1 output layer. The input layer is a vector of length 7 that embodies edge feature information (i.e., the Phase I model prediction and one-hot encodings of six reaction types) and the output layer is a vector equivalent in length to the node hidden embedding size. The output layer represents a vector of learned edge-specific weights for the edge connecting node u and neighboring node v that are multiplied by the hidden embedding of node v. The following formalization defines the messages across the neighborhood as being summed, but we also employ alternate permutation invariant functions such as the max or mean.

$$\mathbf{m}_{N(u)}^{(k)} = \sum_{v \in N(u)} \mathbf{h}_v^{(k-1)} \cdot \mathrm{MLP}_{\Theta}^{(k)}(\mathbf{e}_{u,v}) \tag{3}$$

The hidden embedding of node u is updated by summing its hidden embedding from the previous iteration, parametrized by $\Theta_{\mathrm{self}}^{(k)}$, with the aggregated messages, $\mathbf{m}_{N(u)}^{(k)}$, and learnable bias terms $\mathbf{b}^{(k)}$.

$$\mathbf{h}_u^{(k)} = \Theta_{\mathrm{self}}^{(k)} \mathbf{h}_u^{(k-1)} + \mathbf{m}_{N(u)}^{(k)} + \mathbf{b}^{(k)} \tag{4}$$

We combined multiple-edge-conditioned convolutional layers to form a deeper architecture for node prediction (see Figure S10 in the Supporting Information). The first convolutional layer receives the start-multitarget network containing each node's initial hidden embedding. The output of the convolutional layer is followed by layer normalization, ReLU activation, and dropout. Regardless of the initial hidden embedding size, the first convolutional layer updates the node hidden embedding to a different size that is maintained for the input and output of all remaining convolutional layers. The output of each graph convolutional layer is aggregated utilizing jumping knowledge, which allows for learning on adaptive network depths.[25] The layer aggregation output is fed into a 2 hidden layer MLP with ReLU activation. The resultant hidden embedding of size 2 is passed through a softmax layer to give predictions for each node.

There is additional nuance in how directionality is perceived when defining a node's neighborhood for the aggregation operator (Figure 5). The unidirectional variant constrains a node to only aggregate messages transmitted from incoming edges, i.e., parent nodes, for each of K convolutional layers. The bidirectional variant imposes the same constraint for the first $K/2$ convolutional layers. For the remaining $K/2$ convolutional layers, the flow of message passing is reversed and the aggregation is over messages transmitted from outgoing edges, i.e., child nodes.

The bidirectional variant allows a node to account for information propagated from both parent and child nodes, including more complex flow from other parents of the same child node. It is common for multiple, isolated paths to exist that link the starting molecule to the target metabolite(s) (Figure S8 in the Supporting Information). When using the unidirectional variant, information only flows into a node from its parents and the node's receptive field is limited to information preceding it and within its isolated metabolic pathway. The bidirectional variant allows information to flow back up through a node's child metabolites. If a node's child is connected to another pathway, forming a v-shaped structure, then the node is exposed to and able to account for competing metabolic pathways. In addition, the increase in the node's potential receptive field may enable deeper architectures that can learn higher abstractions of structural features.

MPNN Training. We applied the architecture in a fully supervised manner with no transductive test nodes. For a given split, there are training nodes, which are included in the message passing operations and are used to compute the loss, and inductive test nodes, which are not used to compute the loss and, along with all their edges, are not included in the message passing operations. Thus, inductive test nodes remain unobserved in terms of their ground truth label and their local structure. We trained the model end-to-end using the binary cross entropy loss.

We used two variants of a standard technique (cross-validation), to estimate performance of the metabolite scoring models on external test data for model selection and assessment. The cross-validation strategies involve splits of one or more groups as withheld data for testing. Any start-multitarget XenoNet instances that overlap in terms of their intermediate metabolites are withheld together. Grouping networks in this manner ensures that the learning task is not overly easy and that an intermediate metabolite and its local network structure is not available in both the training set and held-out set. In total, there were 254 groups of related start-multitarget networks.

For the neural network approaches, optimal model performance is dependent on hyperparameter configuration. We employed a group 5-fold nested cross-validation protocol to minimize optimistic bias,[26] where hyperparameter sweeps are applied on tunable model parameters (Table S1 in the Supporting Information). An outer 5-fold cross validation splits the network dataset into one held-out fold for model assessment and four training folds for model selection. Model selection over the training folds is achieved by an inner 5-fold cross validation. The model was then trained using the best performing hyperparameters on the entirety of the networks in the training folds and evaluated on the held-out test fold of the outer loop. This process was repeated for each outer loop iteration and we evaluated performance of multiple neural network approaches by comparing the means of their outer generalization scores for each performance metric.

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

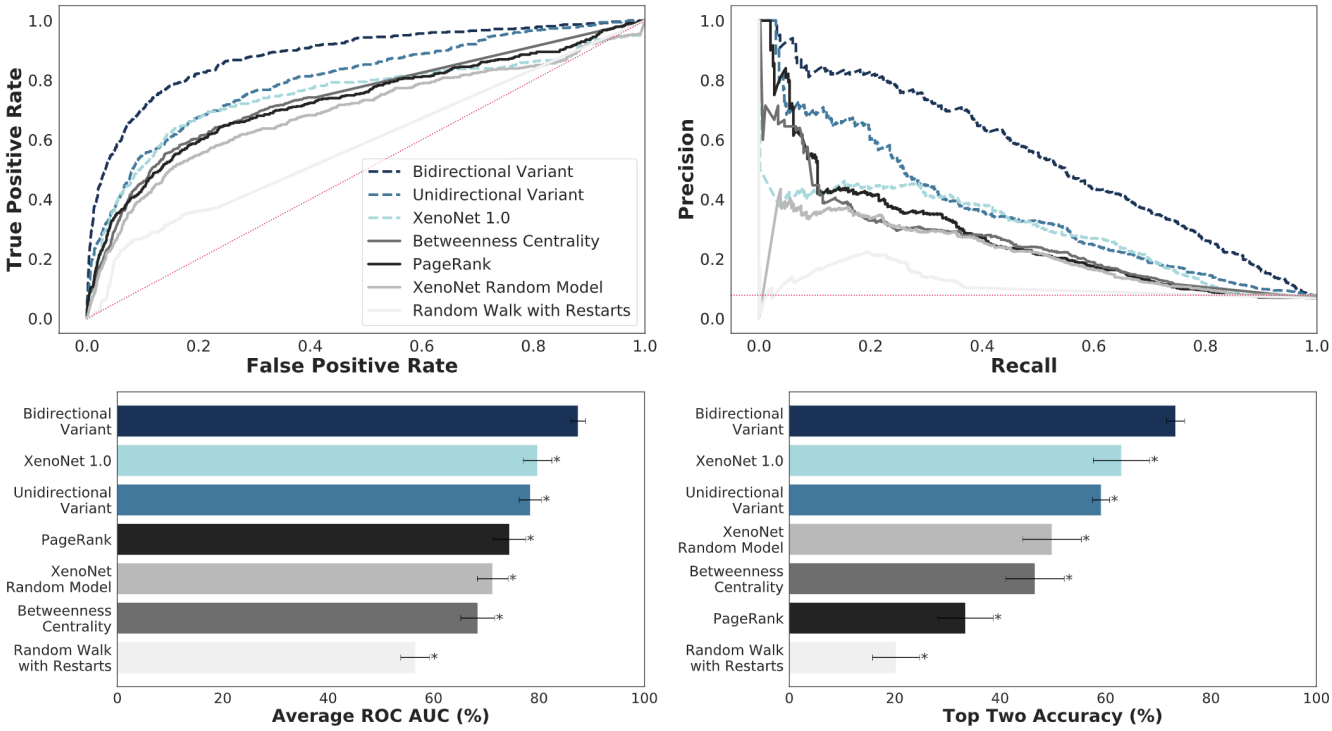


Figure 6. Bidirectional MPNN significantly outperformed the compared methods on ROC AUC (top left) and PR AUC (top right). The dotted crimson line on the ROC curve and PR curve is the no skill curve, the latter at a skew of 1 positive for every 14 negative examples. Bottom left, across the set of 311 networks, the average ROC AUC was measured by calculating how frequently experimentally observed metabolites received higher scores relative to all other metabolites in the network. Bottom right, the top-two accuracy was computed across the same set of networks. For all metrics, we report the performance of the algorithmic approaches as well as the cross-validated scores produced by the MPNN approaches. Asterisks denote performances that were significantly worse than the highest-scoring method, as determined by a paired *t*-test (p-value ≤ 0.05). The paired instances are the networks and the computed values of the average ROC AUC or top-two accuracy when the network is processed by either the highest-scoring model or one of the other models, whose performance is being compared against the highest-scoring model. Error bars represent 95% two-sided confidence intervals.

For inference on training set networks, we utilized a leave-one-group-out cross-validation (LOGOCV) protocol. Networks within the same group are withheld and the rest of the training set networks are used to train a model and make predictions on the withheld networks and all their intermediate metabolites. We repeated this process one-by-one for each group of networks, such that the number of cross-validation folds equaled the number of groups. As a result, the LOGOCV procedure entailed 254 individual, trained models. All models presented here were trained on a CPU (Intel Xeon Processor E5−2630 @ 2.40 GHz and Intel Core i7−8650U CPU @ 1.90 GHz) on a Linux operating system using PyTorch Geometric.

Extraction of Important Connected Subnetworks. XenoNet can produce large, branching networks from which important metabolites may be drawn for closer study. Individual metabolites with high importance that exceeds a classification threshold are straightforward to extract. In some cases, we want to also retain the extracted node's local network structure. To extract important connected subnetworks, we filtered out intermediate-containing paths whose fraction of important intermediates to total intermediates does not exceed an adjustable cutoff. The cutoff ranges from 0 to 1 and becomes more restrictive as it increases. A cutoff of 0 retains all pathways while a cutoff of 1 retains a pathway only when all intermediates are important. We defined a failure case as a network with at least one important intermediate that, upon pruning, results in a disconnected network with no connection between the start and any targets.

RESULTS AND DISCUSSION

We compared the MPNN model against prior work and several algorithmic baselines. First, we quantified how well each model predicted experimentally observed intermediate metabolites. Second, we conducted analysis of sensitivity to input features, robustness to networks constructed with different depth limit and beam width parameters, ability to discern contexts in which the same structure is experimentally observed or not observed, and ability to extract important subnetworks. Third, we used the final model to infer reactive intermediate metabolites as possible facilitators of withdrawn drug toxicity.

Method Comparison. We compared performance of the MPNN and algorithmic approaches on several metrics. All methods acted on the same set of input networks and, if applicable, identical cross validation folds using group 5-fold nested CV.

Intermediate Metabolite Prediction Accuracy. The bidirectional model best separated experimentally observed and unobserved metabolites (Table 1). We quantified this separation by individually computing the area under the receiver operating characteristic curve (ROC AUC) for the intermediate metabolites in each generated network. A ROC curve was produced for each of the 311 networks using each intermediate metabolite's score and respective label, 1 or 0, indicating whether or not that metabolite has been experimentally observed or unobserved with respect to the annotated network. Their ROC AUCs are averaged to

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

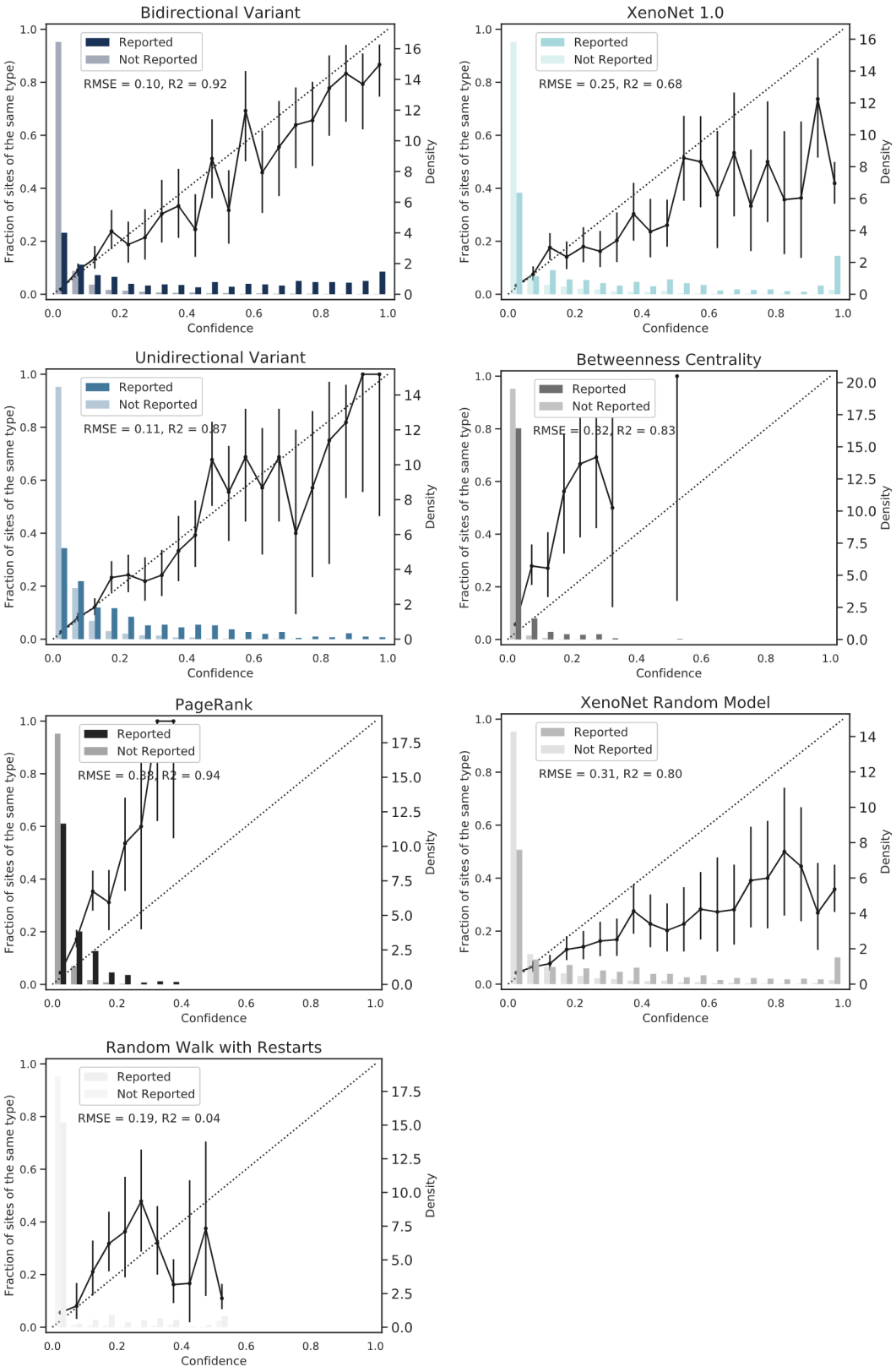


Figure 7. Bidirectional MPNN produced the best-scaled metabolite predictions. The bar graphs plot the distributions for each method across 6606 intermediate metabolites within 311 networks. The solid lines plot the percentage of experimentally observed intermediate metabolites among all intermediates that were assigned the corresponding bin score, marked on the *x*-axis. The diagonal dashed lines indicate the ideal, perfectly scaled prediction.

compute the "average intermediate metabolite ROC AUC" metric. The average ROC AUC compares methods on how well relative rankings of intermediate metabolites within an individual network separate experimentally observed and

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

Table 1. Bidirectional MPNN Architecture Outperformed All Other Methods on Multiple Accuracy and Calibration Metrics[a]

	global ROC AUC	average ROC AUC	global PR AUC	top-two accuracy	reliability, RMSE	reliability, R^2
Bidirectional MPNN Variant	**88.5**	**87.4**	**53.8**	**73.3**	**0.10**	0.92
Unidirectional MPNN Variant	*80.1*	78.4	*36.2*	59.2	0.11	0.87
XenoNet 1.0 Algorithm[10]	75.0	*80.0*	29.0	*63.0*	0.25	0.68
XenoNet Random Model[10]	69.1	71.2	21.0	50.0	0.31	0.80
PageRank[19]	72.9	74.4	27.3	33.4	0.38	**0.94**
Random Walk with Restarts[20,21]	58.5	56.5	11.2	20.3	0.19	0.04
Betweenness Centrality[22,23]	74.2	68.4	25.5	46.6	0.32	0.83

[a]Although PageRank had a higher R^2 value of 0.94, its reliability diagram displays a degenerate case where the model's confidence is unstable. For each metric, the best performing method is shown in bolface font and the second-best method is shown italicized. Method groupings designate new methods (top) and prior methods (bottom).

Table 2. Bidirectional MPNN Architecture That Incorporates Edge Features but Not Node Degree Features Achieves Best Performance

	global ROC AUC (%)	average ROC AUC (%)	global PR AUC (%)	top two accuracy (%)	reliability, RMSE	reliability, R^2
all features	88.2	86.7	53.5	72.4	0.11	0.88
without degree node features	**88.5**	**87.6**	**54.6**	**73.9**	0.08	**0.92**
without reaction type edge feature	83.8	84.2	41.9	64.6	**0.07**	0.89
without weight edge feature	74.4	73.8	34.1	60.9	0.09	**0.92**

unobserved intermediates. In addition, we computed the "global intermediate metabolite ROC AUC", which provides the AUC of a single ROC curve across all generated network's intermediate metabolites and their scores. The global ROC AUC compares methods on their ability to separate intermediate metabolites into experimentally observed and unobserved groups when those intermediate metabolites do not necessarily belong to the same generated network. Similarly, we computed a single precision-recall (PR) curve across all intermediate metabolites and their scores to measure the "global intermediate metabolite PR AUC".

We further assessed separation within individual networks by computing top-two intermediate metabolite performance. For a network, the top two metric assigns a value of 1 if any of its experimentally observed intermediate metabolites received the highest or second-highest score out of all its intermediate metabolites. Otherwise, a value of 0 is assigned to the network. We averaged the top-two metric values across all networks to yield each method's "top-two accuracy".

The bidirectional message passing architecture achieved the best global ROC AUC and PR AUC performances, as well as the best average ROC AUC and top two accuracy (Figure 6). The gap between the bidirectional variant and the XenoNet algorithm is higher for the global-level metrics than the network-level metrics. We theorize that this is due to the XenoNet algorithm's normalization strategy, which does not involve a loss of information at the level of the nodes in an individual network. In contrast, the bidirectional variant exceeds at both global-level and network-level metrics. The bidirectional variant is best able to learn a mapping between a metabolite's local structure and feature information to its metabolic relevance.

Intermediate Metabolite Prediction Calibration. We verified that the bidirectional model's predictions best reflected its confidence. We computed reliability diagrams across all intermediate metabolites (Figure 7). For each method, the predictions of each intermediate metabolite are distributed into 10 equal-width bins between 0 and 1. We computed the percentage of experimentally observed intermediate metabolites in each bin and calculated the root-mean-square error (RMSE) between each bin's midpoint value and its calculated percentage. We also calculated the R^2 of the best fit line, where a method is well-calibrated if its accuracy matches its confidence for each bin. A perfectly scaled prediction will have an RMSE of 0 and a method that produces bins that fit on a perfectly straight line will have an R^2 of 1.

The bidirectional neural network achieved the lowest RMSE of 0.10 and the second highest R^2 of 0.92. Although PageRank had a higher R^2 of 0.94, its reliability diagram displays a degenerate case where the model's confidence is unstable. Hence, the bidirectional neural network was best calibrated, assigning high scores to experimentally observed intermediates and low scores to unobserved intermediates.

LOGOCV Performance of Bidirectional MPNN. We selected the bidirectional MPNN for the final model structure due to its superior performance across all accuracy and calibration metrics. The final model is trained on the full start-multitarget dataset and its LOGOCV predictions are used for inference tasks on training set networks. The optimal hyperparameters used during training are shown in boldface font in Table S1 in the Supporting Information. The optimal aggregation operator summed neighborhood messages, which is unsurprising, since max and mean pooling do not enable networks that are as discriminative as the one-dimensional (1D) Weisfeiler−Leman test and are theoretically less powerful.[27] The optimal layer aggregation operator concatenated each graph convolution layer's output into a vector of size 64.

For the final bidirectional model, we retained the reaction type and weight edge features but not the in-degree and out-degree node features. Moreover, we verified that the bidirectional model is robust when validated with the LOGOCV protocol, as performance for each metric was equivalent or superior to its nested CV performance. We simulated the absence of three feature sets—the pair of in- and out-degree node features, the reaction type edge feature, and the reaction weight edge feature. In each case, the feature is replaced with an uninformative feature value of 1 for all nodes or edges, as applicable. Simulating the absence of in-degree and out-degree node features improved LOGOCV performance (Table 2). One explanation is that useful information

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

contained within the in-degree and out-degree node features is sufficiently learned from the edge weights. The inverse relationship did not hold, as simulating absence of edge weights resulted in decreased LOGOCV performance. Empirically, edge reaction type was also a useful feature.

Generalization of Bidirectional MPNN. We evaluated the model to ensure robustness to input parameter adjustments that influence network construction, correct detection of the presence or absence of metabolites that are annotated as experimentally observed or not observed in different network contexts, and extraction of important intermediates while conserving their local network structure.

Robustness to Network Construction Parameters. The bidirectional MPNN's performance was not strongly influenced by the choice of network construction parameters. This is important because a model that is tightly sensitive to network construction parameters would not generalize. To assess robustness, we applied the bidirectional MPNN, which was trained on networks constructed with a depth limit of 3 and no beam width constraint, on three sets of start-multitarget networks constructed with different depth limits and beam widths (Table S2 in the Supporting Information).

The bidirectional MPNN's LOGOCV global and average ROC AUC performance was robust to substantial decreases in the beam width and depth limit parameters (see Table 3). For global ROC AUC performance, average ROC AUC performance, and reliability R^2, performance was maintained regardless of beam width but did decrease slightly with decreasing depth limit. While RMSE significantly increased with a decrease in depth limit, the maintained correlation is suitable for applications where binarizing the predictions is applicable. Avenues for future improvement divert focus from a deep architecture that chains 1-hop convolutional layers to an alternative strategy that aggregates outputs from multiple shallow networks whose convolutions encode richer, multihop diffusion operators.[28,29]

To sustain performance for different depth limits, jumping knowledge concatenation was necessary. For model variants that do not incorporate layer aggregation, decreased performance for depth limit two networks that may result from oversmoothing, whereby node hidden states converge to an almost uniform distribution and local neighborhood information is lost.[27] The depth limit 2 network has a smaller diameter than the depth limit 3 network, so naively applying a rigid architecture that cannot adjust for network size hurts performance. Oversmoothing is also more likely for approaches that integrate a self-loop update. Notably, we do not witness symptoms of the bottleneck phenomenon: an exponentially growing amount of information from too many neighbors in a widening network must be oversquashed into a fixed length vector, whereby important long-range information may be lost.[30]

Accurate Performance on Subset of Contextual Intermediate Metabolites. The bidirectional MPNN accurately designated authenticity of the same metabolite in different networks. Importantly, the model does not blindly assign authenticity to the same metabolite regardless of its network context. We denote a contextual intermediate metabolite as a metabolite that is annotated as experimentally observed and experimentally unobserved in at least one network each. We evaluated global ROC AUC and PR AUC on a subset of 83 unique, contextual intermediate metabolites. Global performance metrics, e.g., global ROC AUC and PR AUC, are useful for assessing the model's ability to assign good relative predictions for metabolites across multiple networks (as opposed to within a single network). In total, the intermediate metabolites are annotated as experimentally observed in 179 instances and experimentally unobserved in 119 instances.

The bidirectional neural network detected nuanced differences in network contexts that lead the same intermediate metabolite to have been present in one setting and absent in another (Figure 8). The bidirectional model achieved a global ROC AUC of 86.0% and a global PR AUC of 90.0% on the subset of contextual intermediate metabolites. Furthermore, the mean of the experimentally observed intermediate metabolite predictions (0.34) differs significantly from the mean of the experimentally unobserved intermediate metabolite predictions (0.039) based on an unpaired *t*-test (p-value $\leq$0.05) and are separated by the optimal binarization threshold of 0.066.

Extraction of Important Connected Subnetworks. The bidirectional MPNN extracted important connected subnetworks while retaining global ROC AUC performance across multiple cutoffs (Figure S9 in the Supporting Information). Thus, the model can be flexibly used with sustained performance at several different subnetwork extraction cutoffs, depending on application context, sensitivity to possible false positive intermediates, and desire to maximize or minimize retained important or unimportant intermediates, respectively.

Across 311 networks with an average network size of 24 $\pm$ 16 nodes, 1143 of 6606 intermediates had scores that surpassed the metabolite authenticity binarization threshold.

Table 3. Incorporating Jumping Knowledge, The Bidirectional Model Performance Is Invariant to Changes in the Beam Width and Depth Limit Parameters during Network Construction[a]

	global ROC AUC (%)	average ROC AUC (%)	reliability, RMSE	reliability, R^2
With Layer Aggregation				
depth limit 3	88.5	87.6	0.08	0.92
depth limit 3 beam width 5	86.5	87.4	0.08	0.95
depth limit 2	84.1	85.6	0.15	0.84
depth limit 2 beam width 5	83.2	84.5	0.16	0.84
Without Layer Aggregation				
depth limit 3	88.3	86.6	0.10	0.88
depth limit 3 beam width 5	84.3	86.1	0.18	0.78
depth limit 2	71.1	68.8	0.22	0.81
depth limit 2 beam width 5	72.1	70.5	0.24	0.75

[a]Without jumping knowledge, model performance is not maintained when applied to networks whose depth limit parameter differs from the depth limit used to construct the training set networks. Networks constructed with a depth limit of 2 and without layer aggregation resulted in a statistically significant drop in performance based on an unpaired t-test (p-value $\leq$0.05). Results remain unchanged for PR AUC and top two accuracy, but we do not report them as the data sets have different class skew relative to the original start-multitarget network data set. Unlike PR curves, ROC curves are insensitive to changes in class distribution and so the measure will not change if the fundamental classifier performance does not.[31] Furthermore, top two accuracy adopts an optimiztic bias due to the global decrease in class imbalance and the local decrease in the average number of nodes per network.

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

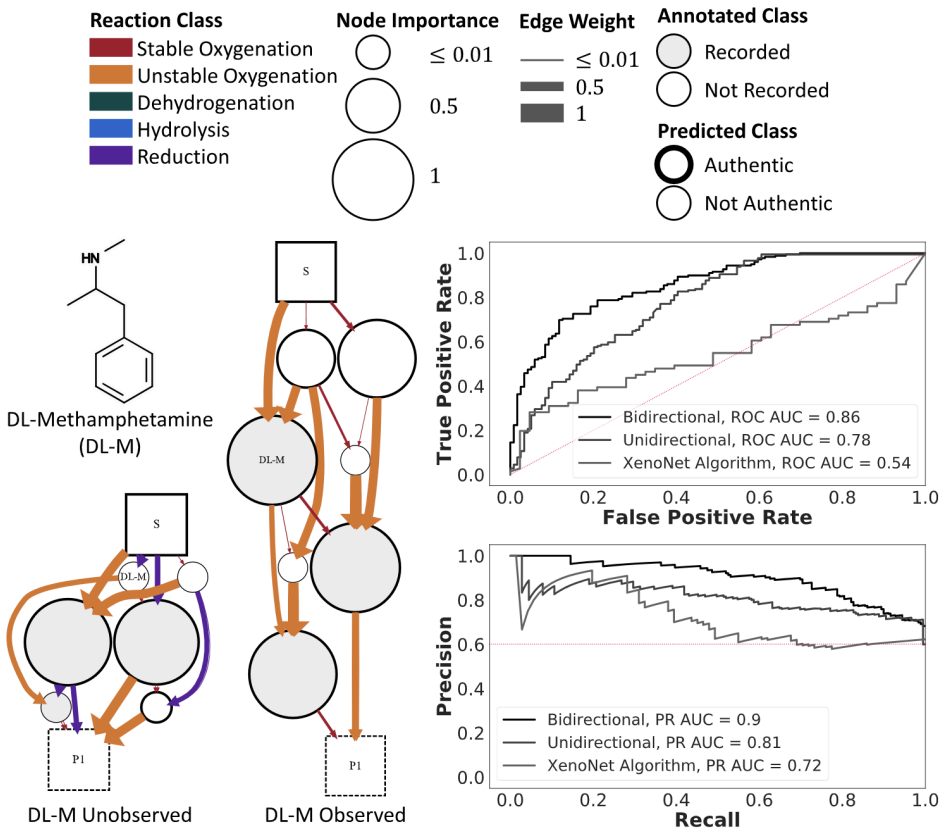


Figure 8. Bidirectional MPNN correctly adjusts predictions for metabolites that are present or absent in different networks. (Left) The intermediate metabolite D,L-methamphetamine is experimentally observed in the right predicted network and is not experimentally observed in the left predicted network.[32] Gray bubbles designate intermediate metabolites whose predictions exceed the optimal binarization threshold. The model assigns scores of 0.91 and 5.2×10^{-2} to D,L-methamphetamine in the contexts in which it is and is not experimentally observed, respectively. D,L-Methamphetamine occurs as an intermediate metabolite in three other networks, not shown. In two of the networks, it is experimentally observed and assigned scores of 0.87 and 0.95. In the remaining network, it is not experimentally observed and assigned a score of 6.1×10^{-4}. (Right) The bidirectional model achieved a global ROC AUC of 86.0% and a global PR AUC of 90.0% on the subset of contextual intermediate metabolites.

A cutoff of 1 resulted in 722 and 0 important and unimportant intermediates, respectively, and reduced the average network size to 4 ± 2 nodes. However, this excluded 421 important intermediates and resulted in 71 failure cases. Alternatively, a cutoff of 0.5 was the maximum cutoff that kept all 1143 important intermediates and did not cause a single failure case. Though a 0.5 cutoff results in 0 failure cases, the average network size was reduced to 15 ± 11 nodes and 2,664 unimportant intermediates are retained. For later case studies, we used a cutoff of 1. If the cutoff failed to extract a subnetwork, we decreased it by increments of 0.05 until success was achieved.

Generalization on External Data. We further assessed the bidirectional MPNN on the GLORY test set of 29 substrates and 81 products[33] that have been unseen by the model. We generated 29 networks in a manner that was consistent with the method used in XenoNet.[10] To evaluate the bidirectional MPNN's ability to generalize, we filtered out spurious metabolites with the binarization threshold previously computed on our metabolic network data set. We did not use the GLORY reference set to recalibrate the optimal binarization threshold.

The bidirectional MPNN outperformed the GLORY and SyGMa,[11] which are the closest comparable works in the literature (see Table 4). The bidirectional MPNN improves upon the XenoNet algorithm, which maximizes recall at the expense of precision, via identifying and filtering out false positives. The ROC AUCs of the bidirectional MPNN, the unidirectional MPNN, XenoNet, GLORY, and SyGMa were 82.6%, 66.8%, 73.3%, 67.6%, and 50.1%, respectively. The unidirectional MPNN performed worse than the XenoNet algorithm and GLORY, demonstrating that unidirectional flow is not sufficient for generalization to an external test set.

Table 4. Bidirectional MPNN Outperforms the XenoNet Algorithm, GLORY, and SyGMa on the GLORY Test Set (In Contrast, the Unidirectional MPNN Performs Worse than the XenoNet Algorithm and GLORY)[a]

	ROC AUC (%)	precision	recall
bidirectional MPNN variant	**82.6**	**0.34**	0.86
unidirectional MPNN variant	66.8	0.10	0.80
XenoNet 1.0 Algorithm[b]	73.3	0.06	**0.89**
GLORY[c]	67.6	0.08	0.83
SyGMa[d]	50.1	0.15	0.74

[a]For each metric, the best performing method is shown in boldface font. [b]Data taken from ref 10. [c]Data taken from ref 33. [d]Data taken from ref 11.

Inferring Unknown Reactive Metabolites of Withdrawn Drugs. Previously, we validated the bidirectional model's ability to score intermediate metabolites as authentic or spurious. A practical application is to identify unreported intermediate metabolites produced during metabolism of

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

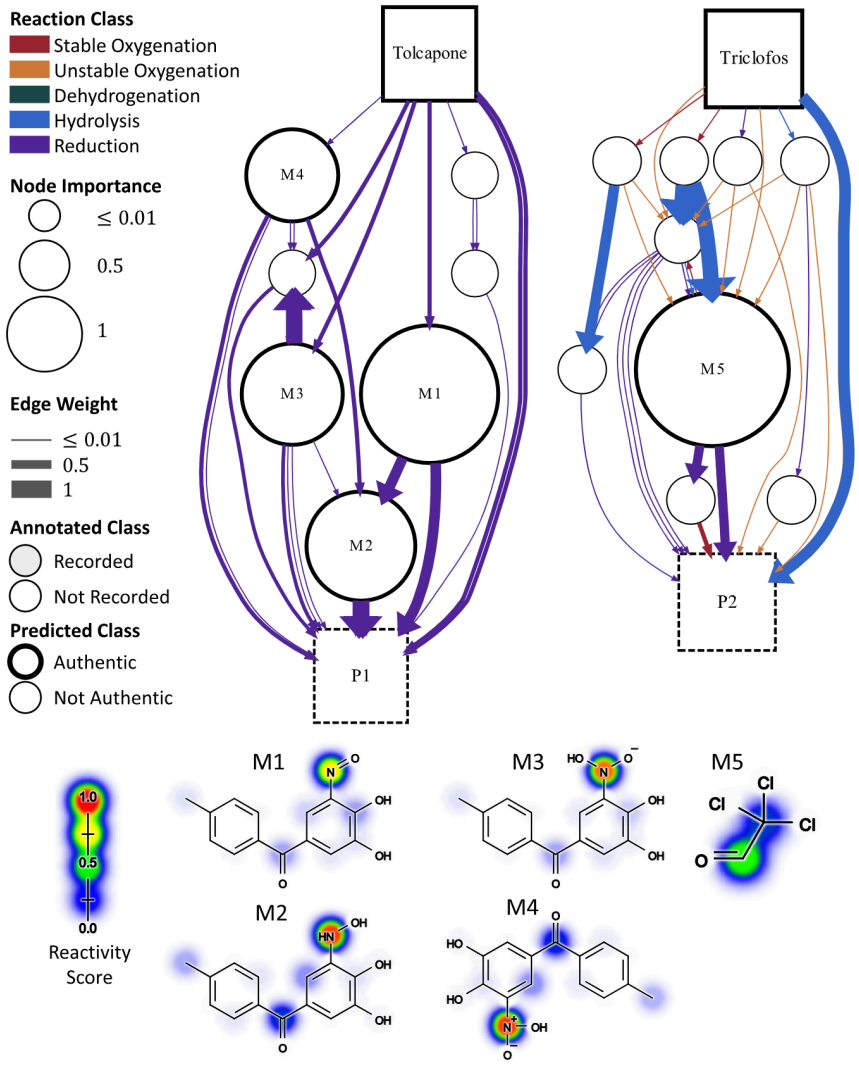


Figure 9. Model identified formation of intermediate(s) that are likely to precede known metabolite end points. The metabolic network for tolcapone posits four unrecorded intermediates with high metabolite scores. The structures of the 4 intermediates, M1−M4, is provided along with an overlay of DNA reactivity predictions. The metabolic network for triclofos identifies a probable intermediate with high protein reactivity, M5 (chloral), which is a known irritant.[37]

withdrawn drugs. Ideally, the model can be confidently used as a fast, inexpensive method for experimentalists to understand step-by-step formation of reactive metabolites, which will help in designing safer drugs. The withdrawn drugs could undergo modification to prevent formation of problematic intermediate metabolites while retaining the intended therapeutic effect.

Identifying Important, Unrecorded Intermediate Metabolites. We applied the fully trained bidirectional model on 70 networks of withdrawn drugs, consisting of 2832 intermediate metabolites. In some cases, the model provides new speculations for mechanisms of a drug's toxicity (Figure 9, left). For instance, the metabolic network for tolcapone, which was introduced for the treatment of Parkinson's disease but later withdrawn due to idiosyncratic hepatotoxicity, posits four unrecorded intermediates with high metabolite scores and high DNA and protein reactivity scores during metabolism to a known reactive metabolite end point.[34−36] In other cases, predicted reactive intermediates are consistent with previous reports (Figure 9, right). The network for metabolism of the withdrawn sedative triclofos to its active metabolite trichloroethanol identifies a probable intermediate with high protein reactivity, chloral, which is a known irritant.[37]

Among the withdrawn drug networks, the bidirectional MPNN classified 550 important intermediate metabolites using an optimal threshold. We computed the optimal threshold using Youden's index and the global ROC curve derived from the bidirectional model's LOGOCV predictions.[38] The optimal threshold for binarizing metabolite scores was 0.066. Moreover, 90 and 105 of the intermediates had reactivity scores greater than 0.5 for protein and DNA, respectively. We provide the full set of predictions in the "DrugBank_withdrawn_drug_intermediate_metabolites_all.csv" file and the subset of 550 intermediate metabolites and associated substrate molecule, target metabolites, metabolite score, and reactivity scores in the "DrugBank_withdrawn_drug_intermediate_metabolites_of_interest.csv" file.

Case Study of Nimesulide Bioactivation. Model predictions were consistent with results reported in the literature and exposed new insights into potential mechanisms of toxicity. For example, nimesulide is a nonsteroidal anti-inflammatory drug used in the treatment of acute pain. Nimesulide is not

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

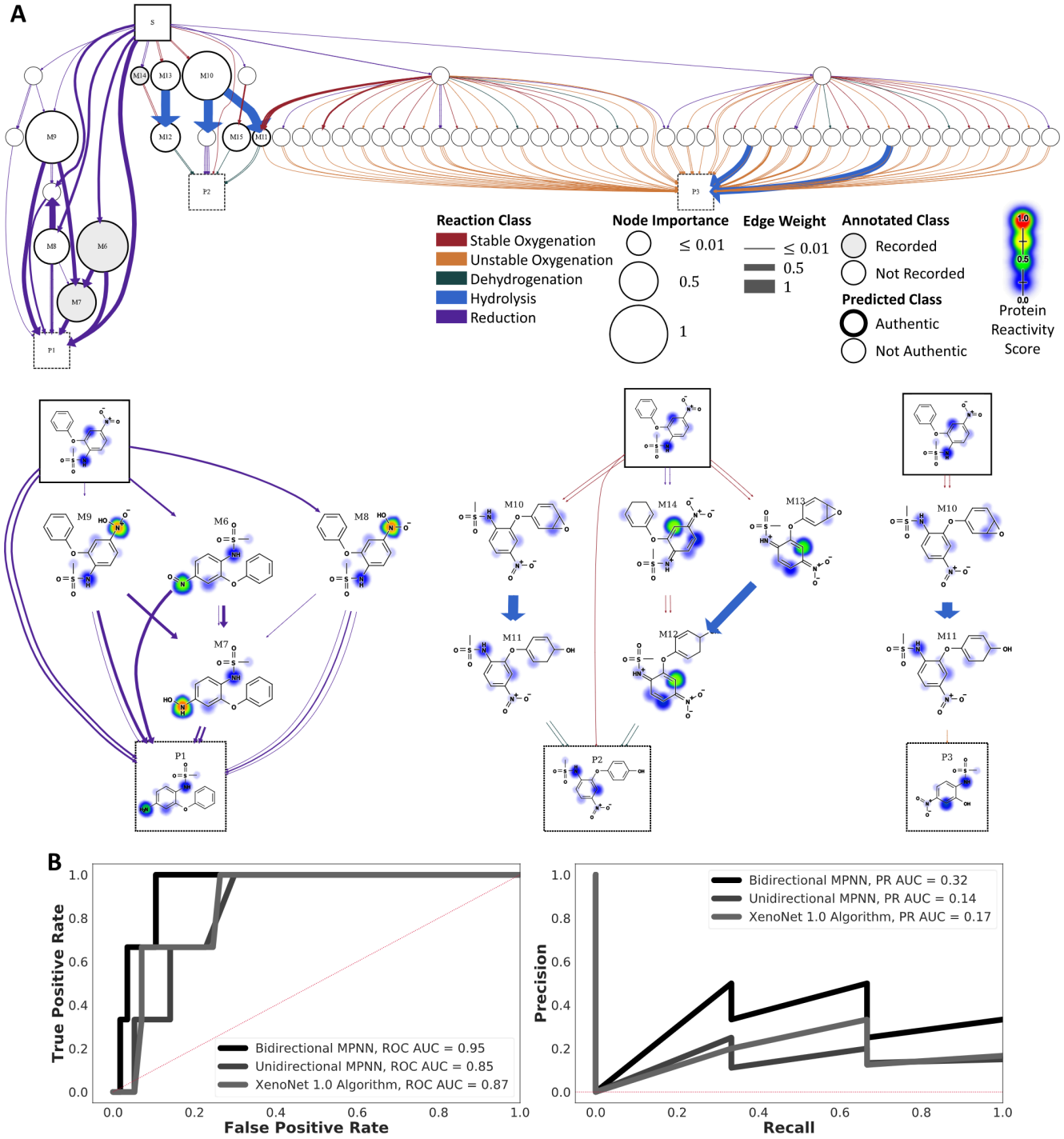


Figure 10. Combination of XenoNet and the bidirectional MPNN pinpointed the dominant bioactivation pathway nimesulide → M6 → M7 → target.[36] (A) XenoNet inferred the presence of 59 intermediate metabolites linking nimesulide and three target metabolites. Gray bubbles designate intermediate metabolites whose predictions exceed the optimal binarization threshold. The bidirectional MPNN extracted three local subnetworks, each one for a different target, using the strictest threshold setting for subnetwork extraction. The left subnetwork contains likely intermediates that are highly predicted to react with protein. The three reactive intermediates are M6, M7, and M8 and there is a fourth, nonreactive and transient intermediate, M9. The remaining two subnetworks do not have intermediates with high site of reactivity predictions. (b) Presence of M6, M7, and M14 is corroborated by th literature. Compared to the unidirectional MPNN and XenoNet algorithm, the bidirectional model assigns higher authenticity to M6, M7, and M14. The bidirectional model's predictions for the remaining important intermediates suggests their presence in a more focused subnetwork context relative to the initial network containing 59 intermediates. The ROC and PR curves were calculated across the 59 intermediate metabolites, 3 of which are reported in the literature.

availabile in the United States and has been established as a cause of acute liver injury.[39,40] Nimesulide's precise mechanism of injury is unknown, although it is thought to be related to production of an intermediate that enables an idiosyncratic reaction during metabolism in the liver.[41] During metabolism of nimesulide to three targets, the model infers the presence of

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

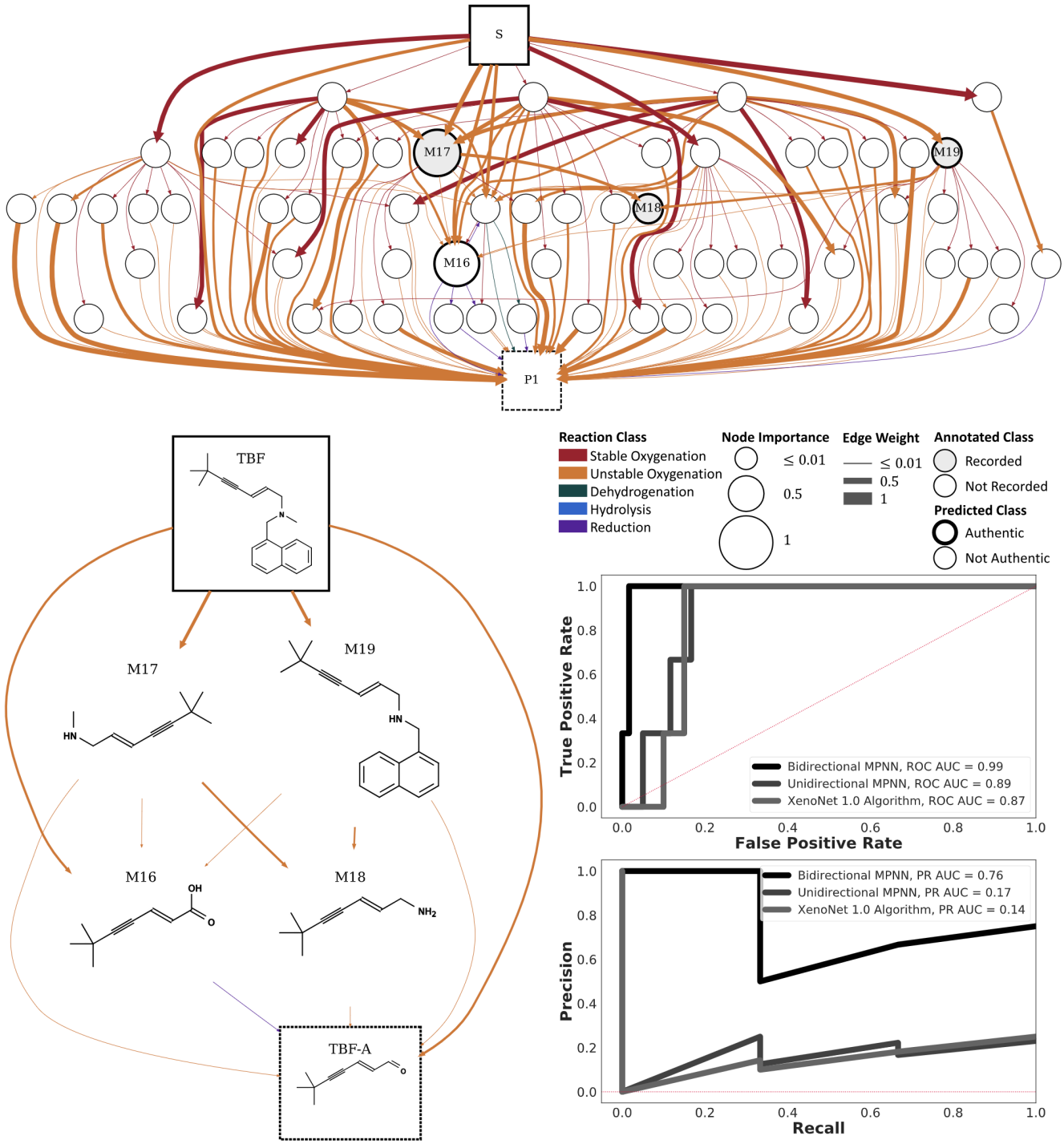


Figure 11. Combination of XenoNet and the bidirectional MPNN inferred M17, M18, and M19 (TBF-D) as bioactivation components of TBF to TBF-A, which aligns with prior experimental validation.[44,45] (Top) XenoNet inferred the presence of 62 intermediate metabolites linking TBF and TBF-A. Gray bubbles designate intermediate metabolites whose predictions exceed the optimal binarization threshold. (Bottom) The model extracted a subnetwork of four intermediates containing TBF-D, M16, M17, and M18 and correctly ranked the pathways traversing these intermediates. Compared to the unidirectional MPNN and XenoNet algorithm, the bidirectional MPNN best assigns authenticity to M17, M18, and TBF-D while minimizing the amount of potential false positive intermediates (bottom right). The ROC and PR curves were calculated across the 62 intermediate metabolites, of which 3 are reported in the literature.

59 intermediate metabolites, of which 10 surpass the optimal metabolite score threshold and 3 (M6, M7, and M8) are predicted to be highly reactive (see Figure 10). Figure 10, middle left, shows the subnetwork containing the three reactive intermediates and M9, another intermediate above the optimal threshold. The presence of M6, M7, and the target is corroborated by experimental evidence and, importantly, the target is thought to be a latent reactive metabolite associated with highly reactive metabolites, specifically M7.[42,43] The model also prioritizes authenticity of nitrogen reduction of nimesulide into M6, which has a basis in bioactivation mechanisms of nitroaromatics. Specifically, intermediates M6

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

and M7 in the metabolism of nimesulide → M6 → M7 → target are both hypothesized to covalently bind to proteins.[36] This bioactivation pathway is also detected by XenoNet. While we did not find literature evidence for M8 or M9, the model views their presence as a possible intermediate state leading to M7.

Case Study of Terbinafine Bioactivation. While not a withdrawn drug, the model similarly extracts a local subnetwork of importance and recapitulates experimental observations regarding bioactivation of terbinafine (TBF) into *E*-6,6-dimethyl-2-hepten-4-ynal (TBF-A) .[44,45] Out of 62 intermediates (Figure 11A), the model scored 4 as relevant to TBF's bioactivation into TBF-A and the local network structure of TBF, TBF-A, and the 4 intermediates is shown in Figure 11B. Desmethyl terbinafine (TBF-D), *N*-methyl-6,6-dimethyl-2-hepten-4-yn-1-amine (M17), and 6,6-dimethylhept-2-EN-4-YN-1-amine (M18) have literature evidence supporting mediation of TBF to TBF-A bioactivation. The presence of 6,6-dimethylhept-2-en-4-ynoic acid (M16) is not recorded in the literature, possibly because of the improbability of M16 undergoing metabolism to TBF-A. In the context of the network structure, the model correctly identifies the most likely pathway as being the direct pathway from TBF to TBF-A, which is an improvement from our prior computational work.[46] The model also correctly ranks the unreported pathway TBF → M16 → TBF-A as last and the pathway TBF → M17 → M18 → TBF-A as second to last. However, the model equates the likelihood of pathways TBF → TBF-D → TBF-A and TBF → M17 → TBF-A, for which experimental evidence favors the former pathway. Nevertheless, the model is capable of constructing a large network of possible metabolites linking TBF to TBF-A, use the metabolite scores to filter out irrelevant metabolites and retain all three intermediates associated with TBF's hypothesized bioactivation into TBF-A, and adequately rank the remaining pathways.

Estimating Important Metabolites Missed by Screening Assays. Our computational approach identified problematic, reactive metabolites that have high metabolic relevance and yet are liable to evade standard screening assays. Trapping studies are designed to address the difficulty in detecting reactive metabolites due to their ephemeral nature. Typically, a trapping agent, e.g., GSH or cyanide, that has a high likelihood of conjugating to reactive metabolites is selected.[3] Formation of a GSH conjugate can be detected via mass spectrometry and indicates the presence of a reactive metabolite. Cyanide and GSH can be applied as trapping agents for hard and soft electrophilic reactive molecules, respectively.[47,48] Due to the possession of only a single type of nucleophilic site, cyanide and GSH may not reflect all possible reactions observed within biologically relevant macromolecules, which often contain both hard and soft nucleophiles oriented across a variety of chemical structures. Hence, nucleophilic trapping assays may overlook the presence of potentially harmful electrophiles.

Across the set of 550 experimentally relevant intermediate metabolites, the estimate produced totals of 51 (9.3%) and 48 (8.7%) metabolites predicted to be exclusively reactive toward DNA and protein, respectively, but not reactive with traditional nucleophilic traps. For each metabolite, we estimated its probability of forming either DNA or protein adducts, but neither cyanide nor GSH adducts, by multiplying the reactivity score for DNA or protein by 1 minus the cyanide reactivity score times 1 minus the GSH reactivity score. The resulting probability is termed the adjusted DNA molecule-level reactivity score (MRS) and the adjusted protein MRS. To estimate the amount of metabolites that selectively react with either DNA or protein, we summed the adjusted DNA MRS and adjusted protein MRS, respectively, for all metabolites. Further experimental validation is necessary to confirm the presence and reactivity of specific missed metabolites.

Limitations and Future Directions. Network construction assumes that only pathways that lead to the target metabolite(s) are meaningful. There could be alternative competing pathways that are not recorded in the network because the pathway does not lead to one of the target metabolites. Since the model only accounts for pathways reported in the XenoNet network, it cannot adjust its predictions on potentially important, but unrecorded, competing pathways. As a solution, network construction could be modified to track and record all predicted reactions and inferred structures in the background to keep two different network states—one with only pathways that terminate at the target metabolite(s) and a second with all pathways. To minimize memory requirements, the second network state could be restrained to only retain pathways that meet certain likelihood criteria. The second network state could then be utilized by the metabolite scoring model.

Although our bidirectional MPNN achieved state-of-the-art performance, it was unable to excise all false positives in the nimesulide (6 false positives) and terbinafine (1 false positive) case studies. While absence of evidence regarding existence of metabolites is not evidence of absence, we acknowledge that the withdrawn drug networks may still contain spurious intermediate metabolites. Future work may improve the model by recognizing the limitations intrinsic to the MPNN paradigm. There are potential issues of bottlenecks or oversmoothing that were referenced earlier, but also issues regarding representational capacity. Without modifications that can lead to less practical architectures,[49,50] MPNNs remain, at most, as powerful as the 1-Wiesfeilr−Lehman test and are unable to discriminate certain graph structures, including simple, yet important, triangles.[51,52]

There are cases where reactive metabolites may be missed simply because they are not intermediates in the metabolism of the starting molecule to one of its targets. In such cases, it is useful to only define the starting molecule and not constrain termination of paths to any target metabolites. We did not directly validate the model on substrate-only networks (Figure 3, top), but based on robustness of the model to variations in beam width and depth limit, we expect maintained performance for inference on substrate-only networks.

Lastly, there is a lack of publicly accessible datasets concerning the task of predicting metabolite authenticity with respect to withdrawn drugs, which may be valuable in retrospective validation for benchmarking future methods. Future work may extend the literature review that we applied in the evaluation of Tolcapone, Triclofos, Nimesulide, and Terbinafine case studies to the rest of our withdrawn drug networks.

■ CONCLUSIONS

This study established and validated a novel metabolite formation model based on a bidirectional MPNN incorporating edge conditioned convolutions and jumping knowledge. The bidirectional MPNN overcomes degenerate cases exhibited by prior work and can aggregate a greater diversity of features, including categorical edge features and local

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

network structure. Incorporating metabolite reactivity further informs specific, testable hypotheses for use by experimentalists in understanding reactive metabolite formation. The bidirectional MPNN can accurately predict experimentally observed and unobserved metabolites, outperforming all compared methods on multiple accuracy and calibration metrics on a dataset of 311 networks and 6606 intermediate metabolites. Moreover, it is robust to networks of varying depth and breadth, detects when a metabolite may be formed or not, depending on different network contexts, and allows for extract of metabolic subnetworks. Metabolite predictions can be used to determine sequential metabolic transformations or relevant metabolic subnetworks that are mediated by previously unknown, potentially reactive, intermediate metabolites that are worth further experimental study of their role in driving toxicity. To demonstrate, we used the metabolite formation model to produce hypotheses for bioactivation mechanisms of drugs associated with idiosyncratic reactions but inconclusive etiology. On a set of networks generated for 70 withdrawn drugs, the model also provided valuable insight on the 9.3% and 8.7% of metabolites with high formation scores that selectively react with DNA or protein, respectively, but are liable to eluding standard screening assays. We anticipate that analysis of formation and reactivity of intermediate metabolites and their local metabolic subnetworks will become central to future experimental investigations.

■ ASSOCIATED CONTENT

Data Availability Statement

The "Start_Multitarget_Training_Dataset.json" file contains the 311 metabolic networks generated by XenoNet and annotated using the AMD. The "Start_Multitarget_DrugBank_Withdrawn_Dataset.json" file contains metabolic networks constructed for 70 DrugBank withdrawn drugs. The metabolic networks are stored in JSON format and each network is most easily parsed via the NetworkX library in Python. The "Drug_withdrawn_drug_intermediate_metabolites_all.csv" file contains 2832 intermediate metabolites from the DrugBank withdrawn drug networks and their associated substrate molecule, target metabolites, metabolite score, and reactivity scores. The "DrugBank_withdrawn_drug_intermediate_metabolites_of_interest.csv" file contains a subset of 550 intermediate metabolites that the model classifies as experimentally relevant. This information is available at http://pubs.acs.org/.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c01383.

Example networks that demonstrate XenoNet's updated capabilities, the Phase I SoM's calibration and class thresholds with respect to each reaction class, relation of note degree to class labels, distribution of information silos that limit the unidirectional MPNNs learning capacity, hyperparameters for model selection, a summary of the start-multitarget data set variants used to evaluate robustness of the bidirectional MPNN to network construction parameters, influence of the optimal binarization threshold on extracted subnetworks, and additional details on the unidirectional and bidirectional MPNN training architectures (PDF)

(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

Author

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

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c01383

Funding

Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health (NIH), under Award Nos. R01LM012222 and R01LM012482 and by the National Institute of General Medical Sciences of the NIH, under Award No. R01GM140635. Computations were performed using the facilities of the Washington University Center for High Performance Computing, which were partially funded by the NIH (under Grant Nos. 1S10RR022984-01A1 and 1S10OD018091-01). The content is completely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We 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. We also thank the developers of the open-source chemoinformatics tools Open Babel and RDKit, of which we made extensive use.

■ ABBREVIATIONS

AMD, Accelrys Metabolite Database
AUC, area under the receiver operating characteristic curve
IADR, idiosyncratic adverse drug reaction
MPNN, Message Passing Neural Network
ROC, receiver operating characteristic
RMSE, root-mean-square error
SoM, site of metabolism

■ REFERENCES

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

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

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

(4) Arrowsmith, J.; Miller, P. Trial watch: phase II and phase III attrition rates 2011−2012. *Nat. Rev. Drug Discovery* **2013**, *12*, 569.

(5) Stepan, A. F.; Walker, D. P.; Bauman, J.; Price, D. A.; Baillie, T. A.; Kalgutkar, A. S.; Aleo, M. D. Structural Alert/Reactive Metabolite

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

Concept as Applied in Medicinal Chemistry to Mitigate the Risk of Idiosyncratic Drug Toxicity: A Perspective Based on the Critical Examination of Trends in the Top 200 Drugs Marketed in the United States. *Chem. Res. Toxicol.* **2011**, *24*, 1345−1410.
(6) Mckim, J. Building a Tiered Approach to In Vitro Predictive Toxicity Screening: A Focus on Assays with In Vivo Relevance. *Comb. Chem. High Throughput Screening* **2010**, *13*, 188−206.
(7) Bylund, J.; Macsari, I.; Besidski, Y.; Olofsson, S.; Petersson, C.; Arvidsson, P. I.; Bueters, T. Novel Bioactivation Mechanism of Reactive Metabolite Formation from Phenyl Methyl-Isoxazoles. *Drug Metab. Dispos.* **2012**, *40*, 2185−2191.
(8) Isin, E. M.; Elmore, C. S.; Nilsson, G. N.; Thompson, R. A.; Weidolf, L. Use of Radiolabeled Compounds in Drug Metabolism and Pharmacokinetic Studies. *Chem. Res. Toxicol.* **2012**, *25*, 532−542.
(9) Zhang, D.; Krishna, R.; Wang, L.; Zeng, J.; Mitroka, J.; Dai, R.; Narasimhan, N.; Reeves, R. A.; Srinivas, N. R.; Klunk, L. J. Metabolism, Pharmacokinetics, and Protein Covalent Binding of Radiolabeled Maxipost (BMS-204352) in Humans. *Drug Metab. Dispos.* **2005**, *33*, 83−93.
(10) Flynn, N. R.; Dang, N. L.; Ward, M. D.; Swamidass, S. J. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation. *J. Chem. Inf. Model.* **2020**, *60*, 3431−3449.
(11) Ridder, L.; Wagener, M. SyGMa: Combining Expert Knowledge and Empirical Scoring in the Prediction of Metabolites. *ChemMedChem.* **2008**, *3*, 821−832.
(12) 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. Cheminform.* **2019**, *11*. DOI: 10.1186/s13321-018-0324-5
(13) 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.
(14) Hughes, T. B.; Dang, N. L.; Kumar, A.; Flynn, N. R.; Swamidass, S. J. Metabolic Forest: Predicting the Diverse Structures of Drug Metabolites. *J. Chem. Inf. Model.* **2020**, *60*, 4702−4716.
(15) 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.
(16) Yan, Z.; Li, J.; Huebert, N.; Caldwell, G. W.; Du, Y.; Zhong, H. Detection of a Novel Reactive Metabolite of Diclofenac: Evidence for CYP2C9-mediated Bioactivation via Arene Oxides. *Drug Metab. Dispos.* **2005**, *33*, 706−713.
(17) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*; International Convention Centre, Sydney, Australia, 2017; pp 1263−1272.
(18) Newman, M. *Networks: An Introduction*; Oxford University Press, 2010; Chapter 7.
(19) Page, L.; Brin, S.; Motwani, R.; Winograd, T. *The PageRank Citation Ranking: Bringing Order to the Web*. 1998.
(20) Tong, H.; Faloutsos, C.; Pan, J. *Fast Random Walk with Restart and Its Applications*. Sixth International Conference on Data Mining (ICDM'06). 2006; pp 613−622.
(21) Jin, W.; Jung, J.; Kang, U. Supervised and extended restart in random walks for ranking and link prediction in networks. *PLoS One* **2019**, *14*, e0213857.
(22) Brandes, U. A faster algorithm for betweenness centrality. *J. Math. Sociol.* **2001**, *25*, 163−177.
(23) Brandes, U. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks* **2008**, *30*, 136−145.
(24) Simonovsky, M.; Komodakis, N. Dynamic Edge-Conditioned Filters in Convolutional Neural Networks on Graphs. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017; pp 29−38.
(25) Xu, K.; Li, C.; Tian, Y.; Sonobe, T.; Kawarabayashi, K.-i.; Jegelka, S. Representation Learning on Graphs with Jumping Knowledge Networks. *Proceedings of the 35th International Conference on Machine Learning*, Stockholm, Sweden, 2018; pp 5453−5462.
(26) Cawley, G. C.; Talbot, N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. *J. Mach. Learn. Res.* **2010**, *11*, 2079−2107.
(27) Xu, K.; Hu, W.; Leskovec, J.; Jegelka, S. How Powerful are Graph Neural Networks? *CoRR* **2018**, abs/1810.00826.
(28) Wu, F.; Souza, A.; Zhang, T.; Fifty, C.; Yu, T.; Weinberger, K. Simplifying Graph Convolutional Networks. *Proceedings of the 36th International Conference on Machine Learning*, 2019; pp 6861−6871.
(29) Klicpera, J.; Weiß enberger, S.; Günnemann, S. Diffusion Improves Graph Learning. *Advances in Neural Information Processing Systems*, 2019.
(30) Alon, U.; Yahav, E. *On the Bottleneck of Graph Neural Networks and its Practical Implications*, 2021.
(31) Fawcett, T. An introduction to ROC analysis. *Pattern Recogn. Lett.* **2006**, *27*, 861−874. ROC Analysis in Pattern Recognition.
(32) Sato, M.; Hida, M.; Nagase, H. Analysis of dimethylamphetamine and its metabolites in human urine by liquid chromatography−electrospray ionization−mass spectrometry with direct sample injection. *Forensic Sci. Int.* **2002**, *128*, 146−154.
(33) 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*, 402.
(34) Limban, C.; Nuţă, D. C.; Chiriţă, C.; Negres, S.; Arsene, A. L.; Goumenou, M.; Karakitsios, S. P.; Tsatsakis, A. M.; Sarigiannis, D. A. The use of structural alerts to avoid the toxicity of pharmaceuticals. *Toxicol. Rep.* **2018**, *5*, 943−953.
(35) Jorga, K.; Fotteler, B.; Heizmann, P.; Gasser, R. Metabolism and excretion of tolcapone, a novel inhibitor of catechol-O-methyltransferase. *Br. J. Clin. Pharmacol.* **1999**, *48*, 513−520.
(36) Boelsterli, U.; Ho, H.; Zhou, S.; Leow, K. Y. Bioactivation and hepatotoxicity of nitroaromatic drugs. *Curr. Drug Metab.* **2006**, *7*, 715−727.
(37) *PubChem. Compound LCSS for CID 6407, Chloral*. National Center for Biotechnology Information, 2021; https://pubchem.ncbi.nlm.nih.gov/compound/Chloral.
(38) Youden, W. J. Index for rating diagnostic tests. *Cancer* **1950**, *3*, 32−35.
(39) Kwon, J.; Kim, S.; Yoo, H.; Lee, E. Nimesulide-induced hepatotoxicity: A systematic review and meta-analysis. *PLoS One* **2019**, *14*, e0209264.
(40) Lewis, J.; Stine, J. *Drug-Induced Liver Disease*; Elsevier, 2013; pp 369−401.
(41) *LiverTox: Clinical and Research Information on Drug-Induced Liver Injury*; National Institute of Diabetes and Digestive and Kidney Diseases, Bethesda, MD, 2016; Chapter Nimesulide.
(42) Bernareggi, A. Clinical Pharmacokinetics of Nimesulide. *Clin. Pharmacokinet.* **1998**, *35*, 247−274.
(43) Yang, M.; Chordia, M. D.; Li, F.; Huang, T.; Linden, J.; Macdonald, T. L. Neutrophil- and Myeloperoxidase-Mediated Metabolism of Reduced Nimesulide: Evidence for Bioactivation. *Chem. Res. Toxicol.* **2010**, *23*, 1691−1700.
(44) 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.
(45) Barnette, D. A.; Davis, M. A.; Flynn, N.; Pidugu, A. S.; Swamidass, S. J.; Miller, G. P. Comprehensive kinetic and modeling analyses revealed CYP2C9 and 3A4 determine terbinafine metabolic clearance and bioactivation. *Biochem. Pharmacol.* **2019**, *170*, 113661.
(46) 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.
(47) Ma, S.; Subramanian, R. Detecting and characterizing reactive metabolites by liquid chromatography/tandem mass spectrometry. *J. Mass Spectrom.* **2006**, *41*, 1121−1139.
(48) Meneses-Lorente, G.; Sakatis, M. Z.; Schulz-Utermoehl, T.; Nardi, C. D.; Watt, A. P. A quantitative high-throughput trapping

assay as a measurement of potential for bioactivation. *Anal. Biochem.* **2006**, *351*, 266−272.
(49) Maron, H.; Ben-Hamu, H.; Serviansky, H.; Lipman, Y. Provably Powerful Graph Networks. *Advances in Neural Information Processing Systems*, 2019.
(50) Morris, C.; Ritzert, M.; Fey, M.; Hamilton, W. L.; Lenssen, J. E.; Rattan, G.; Grohe, M. Weisfeiler and Leman Go Neural: Higher-Order Graph Neural Networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, 2019; pp 4602−4609, .
(51) Arvind, V.; Fuhlbrück, F.; Köbler, J.; Verbitsky, O. On Weisfeiler-Leman invariance: Subgraph counts and related graph properties. *J. Comput. Syst. Sci.* **2020**, *113*, 42−59.
(52) Chen, Z.; Chen, L.; Villar, S.; Bruna, J. Can Graph Neural Networks Count Substructures?. In *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems*, December 2020; Association of Computing Machinery, 2020; Article No. 871, pp 10383−10395.

Downloaded via WASHINGTON UNIV on October 08, 2024 at 17:00:41 (UTC). See https://pubs.acs.org/sharingguidelines for options on how to legitimately share published articles.

https://doi.org/10.1021/acs.jcim.2c01383
J. Chem. Inf. Model. 2023, 63, 1675−1694