Computational Approaches for Screening Drugs for Bioactivation, Reactive Metabolite Formation, and Toxicity

Noah Flynn
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds

Part of the Artificial Intelligence and Robotics Commons, and the Chemistry Commons

Recommended Citation
https://openscholarship.wustl.edu/art_sci_etds/2491

This Dissertation is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.
Computational Approaches for Screening Drugs for Bioactivation, Reactive Metabolite Formation, and Toxicity
by
Noah Flynn

A dissertation presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2021
St. Louis, Missouri
# Contents

List of Figures vii  
List of Tables xi  
List of Abbreviations xiii  
Acknowledgements xiv  
Abstract xvi  

## 1 Introduction  
1.1 Idiosyncratic Adverse Drug Reactions ..................................... 1  
1.2 Drug Metabolism ............................................................................. 2  
1.3 Current Methods for Reactive Metabolite Screening ......................... 4  
1.3.1 Experimental Approaches ................................................................. 4  
1.3.2 Computational Approaches ............................................................... 6  
1.4 Integration of Metabolism and Reactivity Models to Study Bioactivation ... 8  

## 2 XenoNet: Inference and Likelihood of Intermediate Metabolite Formation  
2.1 Introduction ................................................................................... 10  
2.2 Data and Methods ........................................................................... 13  
2.2.1 The Metabolic Network Data Set ...................................................... 13  
2.2.2 The Metabolic Network Generator XenoNet .................................. 16  
2.2.3 The Naive XenoNet Variant .............................................................. 17  
2.2.4 Heuristic XenoNet Variants ............................................................. 19  
2.2.5 Metabolite Scoring Algorithm ......................................................... 23  
2.3 Results & Discussion ..................................................................... 25  
2.3.1 Comparison Between Metabolic Network Variants ....................... 25  
   Time-cost ............................................................................................ 26  
   Path Recall .......................................................................................... 28  
   Intermediate Metabolite Recall ......................................................... 32  
   Intermediate Metabolite Recovery .................................................... 34  
2.3.2 Performance on the Full Metabolic Network Data Set .................. 35
Network Coverage, Path Recall, Metabolite Recall, and Metabolite Recovery .......................... 37
Path and Metabolite Rankings ................................................. 39
2.3.3 Comparison to Prior Work ........................................... 43
2.4 Model Limitations & Future Directions ................................ 48
2.5 Conclusion ............................................................. 51

3 Message Passing Neural Networks Improve Prediction of Metabolite Importance 52
3.1 Introduction ............................................................. 52
3.2 Results and Discussion .................................................. 56
  3.2.1 Method Comparison .................................................. 56
    Intermediate Metabolite Prediction Accuracy ....................... 56
    Intermediate Metabolite Prediction Calibration .................... 59
    LOGOCV Performance of Bidirectional MPNN ....................... 59
  3.2.2 Generalization of Bidirectional MPNN ........................... 62
    Robustness to Network Construction Parameters .................. 63
    Accurate Performance on Subset of Contextual Intermediate Metabolites 64
    Extraction of Important Connected Subnetworks ................... 67
  3.2.3 Inferring Unknown Reactive Metabolites of Withdrawn Drugs .... 68
    Identifying Important, Unrecorded Intermediate Metabolites .... 68
    Case Study of Nimesulide Bioactivation .............................. 69
    Case Study of Terbinafine Bioactivation ............................ 71
    Estimating Important Metabolites Missed by Screening Assays .... 71
  3.2.4 Limitations and Future Directions ................................ 74
3.3 Conclusion ............................................................. 75
3.4 Materials and Methods ................................................ 76
  3.4.1 XenoNet ............................................................ 76
  3.4.2 Relevant Extensions to XenoNet ................................ 77
    Beam Search ............................................................ 77
    Multitarget Network Construction .................................... 78
    Reactivity Model Integration .......................................... 80
  3.4.3 Single Start-Multitarget Metabolism Data Set ................. 80
  3.4.4 DrugBank Data Set ................................................. 82
  3.4.5 XenoNet Start-Multitarget Network Structures & Features .... 83
  3.4.6 Algorithmic Approaches ......................................... 87
    XenoNet Metabolite Scoring Algorithm ................................ 87
    Graph Analysis Algorithms .......................................... 88
  3.4.7 Message Passing Neural Network Architectures ................. 90
    MPNN Training ......................................................... 92


4 ReactNet: Prediction of Phase II Metabolite Formation

4.1 Introduction

4.2 Data & Methods

4.2.1 Phase II Metabolism Data Set

4.2.2 ReactNet: Phase II Metabolism Network Generator

4.3 Results & Discussion

4.3.1 Model Performance & Method Comparisons

4.4 Conclusion

5 Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors

5.1 Introduction

5.2 Methods

5.2.1 Materials

5.2.2 Data to assess quinone model performance toward isoxazole-containing molecules

5.2.3 Modeling bioactivation pathways for isoxazole-containing molecules

5.2.4 Experimental bioactivation of isoxazole-containing BET inhibitors as test cases

5.2.5 Characterization of BET inhibitor reactive metabolites trapped as glutathione adducts

5.3 Results

5.3.1 Mining the Accelrys Metabolite Database for possible isoxazole bioactivations

5.3.2 Assessment of model accuracy to predict electrophilic conjugated systems involving isoxazoles

5.3.3 Application of coupled modeling approach to predict bioactivation of BET inhibitors

5.3.4 OXFBD02 underwent extensive metabolism and bioactivation down competing pathways

5.3.5 Metabolism and bioactivation of OXFBD04 was very limited

5.3.6 There was no evidence of I-BET151 bioactivation despite limited metabolism

5.4 Discussion

5.4.1 Quinone model predictions for reactive non-quinones involving 4-amino-5-methyl-isoxazole

5.4.2 Modeled bioactivations of the 3,5-dimethyl isoxazole present in BET inhibitors
<table>
<thead>
<tr>
<th>Page</th>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.4.3</td>
<td>Metabolism and bioactivation of OXFBD02 and 04, but not I-BET151, into quinones</td>
</tr>
<tr>
<td>141</td>
<td>5.4.4 Novelty and relevance of extended quinone-methides</td>
</tr>
<tr>
<td>142</td>
<td>5.4.5 Advances in isoxazole bioactivation studies revealed current limitations of those efforts</td>
</tr>
<tr>
<td>143</td>
<td>5.5 Conclusion</td>
</tr>
<tr>
<td>144</td>
<td>6 Modeling the Bioactivation and Subsequent Reactivity of Drugs</td>
</tr>
<tr>
<td>145</td>
<td>6.1 Introduction</td>
</tr>
<tr>
<td>149</td>
<td>6.2 Methods</td>
</tr>
<tr>
<td>149</td>
<td>6.2.1 Bioactivation Training Data</td>
</tr>
<tr>
<td>151</td>
<td>6.2.2 DrugBank</td>
</tr>
<tr>
<td>152</td>
<td>6.2.3 Bioactivation Descriptors</td>
</tr>
<tr>
<td>154</td>
<td>6.2.4 Combined Path- and Molecule-level Bioactivation Model</td>
</tr>
<tr>
<td>159</td>
<td>6.3 Results and Discussion</td>
</tr>
<tr>
<td>159</td>
<td>6.3.1 Bioactivation Pathway Prediction Accuracy</td>
</tr>
<tr>
<td>170</td>
<td>6.3.2 Accuracy at Predicting Molecule Bioactivation</td>
</tr>
<tr>
<td>175</td>
<td>6.3.3 Inferring the Toxicity Mechanism of Withdrawn Drugs</td>
</tr>
<tr>
<td>180</td>
<td>6.4 Model Limitations</td>
</tr>
<tr>
<td>183</td>
<td>6.5 Conclusion</td>
</tr>
<tr>
<td>184</td>
<td>7 Machine learning liver-injuring drug interactions with non-steroidal anti-inflammatory drugs (NSAIDs) from retrospective cohort</td>
</tr>
<tr>
<td>184</td>
<td>7.1 Introduction</td>
</tr>
<tr>
<td>188</td>
<td>7.2 Materials and methods</td>
</tr>
<tr>
<td>188</td>
<td>7.2.1 Study population and study design</td>
</tr>
<tr>
<td>189</td>
<td>7.2.2 Polypharmacy data: Twosides database</td>
</tr>
<tr>
<td>190</td>
<td>7.2.3 DILI definition</td>
</tr>
<tr>
<td>191</td>
<td>7.2.4 Estimating percent relative effect</td>
</tr>
<tr>
<td>192</td>
<td>7.2.5 Drug interaction network (DIN)</td>
</tr>
<tr>
<td>193</td>
<td>7.3 Results &amp; discussion</td>
</tr>
<tr>
<td>195</td>
<td>7.3.1 Diclofenac dependent risk and DILI</td>
</tr>
<tr>
<td>200</td>
<td>Comparison to data mining algorithms: diclofenac dependent DILI risk</td>
</tr>
<tr>
<td>204</td>
<td>7.3.2 Metabolic context of a potential and rare hepatotoxic interaction</td>
</tr>
<tr>
<td>205</td>
<td>Comparison to data mining algorithms: NSAID dependent risk to DILI outcomes</td>
</tr>
<tr>
<td>207</td>
<td>Comparison to data mining algorithms: NSAID dependent DILI risk</td>
</tr>
<tr>
<td>210</td>
<td>7.4 Model limitations &amp; future directions</td>
</tr>
<tr>
<td>212</td>
<td>7.5 Conclusion</td>
</tr>
<tr>
<td>214</td>
<td>8 Conclusion and Future Directions</td>
</tr>
</tbody>
</table>
Contents

Bibliography 215
List of Figures

1.1 Structural alerts do not adequately account for molecule context ............ 7
2.1 Metabolic network data set construction and overview of XenoNet ............ 14
2.2 XenoNet infers metabolic pathways connecting a substrate and a product from their structures in three steps ........................................ 18
2.3 Top-N and substructure matching heuristics ........................................ 20
2.4 XenoNet Metabolite Scoring Algorithm .............................................. 24
2.5 Time-cost varies greatly between the seven XenoNet variants .................... 27
2.6 Path recall calculation for metabolic networks ...................................... 30
2.7 Path-recall performance is dependent on path length .............................. 31
2.8 Intermediate metabolite recall calculation for metabolic networks .......... 32
2.9 The combined substructure matching and top-N specific heuristic model has the highest intermediate metabolite recall across all depths .................... 33
2.10 Recovery rate calculation for metabolic networks ................................ 34
2.11 XenoNet accurately predicts metabolic networks .................................. 36
2.12 Example paths XenoNet succeeded or failed in finding for paths of length 1, 2, and 3 ........................................................................ 38
2.13 Path ranking calculation ................................................................. 40
2.14 XenoNet is superior to published metabolite prediction models .............. 46
2.15 XenoNet is superior to GLORY MaxCoverage in terms of recovery rate ... 47
2.16 XenoNet, GLORY, SyGMA, and BioTransformer are not able to infer a pathway that links the substrate and metabolite for the 4 cases shown .......... 49
3.1 Our proposed model overcomes XenoNet’s deficiencies, designates intermediate metabolites of greatest importance, and extracts important connected subnetworks for focused review ........................................ 55
3.2 The bidirectional MPNN significantly outperformed the compared methods on ROC AUC and PR AUC ................................. 58
3.3 The bidirectional MPNN produced the best-scaled metabolite predictions .. 60
3.4 The bidirectional MPNN correctly adjusts predictions for metabolites that are present or absent in different networks ................................. 66
List of Figures

3.5 Optimal subnetwork extraction cut-off value depends on application context, sensitivity to possible false positives, and whether the cut-off results in a failure case .......................... 67
3.6 The model identified formation of intermediate(s) that are likely to precede known metabolite end points ................................................................. 70
3.7 The combination of XenoNet and the bidirectional MPNN pinpointed the dominant bioactivation pathway nimesulide → M6 → M7 → P1 ............................. 72
3.8 The 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 .............................. 73
3.9 Example process of beam search using a beam width of 2 and depth limit of 3 . 79
3.10 XenoNet extensions for multiple target metabolites and reactivity predictions 81
3.11 The model has differing levels of confidence in its reaction type specific predictions and incorporation of reaction type information may assist learning relative comparisons between predictions across different reaction types .............. 84
3.12 The optimal thresholds for binarizing the Rainbow model’s site-level predictions are different for each reaction type .................................................. 85
3.13 In the absence of node features, node in- and out-degree may be exploited as an informative feature with no additional computational overhead to compute ........................... 86
3.14 The XenoNet algorithm calculates metabolite scores using a three step process . 89
3.15 Message passing neural network architecture ........................................... 93
3.16 The bidirectional MPNN takes advantage of information flow in both directions, from parent to child and vice versa, to extend each node’s receptive field and improve learning ................................................................. 94
3.17 The first variant of the proposed MPNN architecture only aggregates messages from a node’s parents and results in an inability for many competing pathways to be considered .................................................. 95

4.1 Phase II-mediated detoxification pathways compete with bioactivation pathways, which are a major contributor of adverse drug reactions .................. 100
4.2 We cover five Phase II reaction types: UGT, GST, SULT, NAT, and TPMT ... 101
4.3 General neural network structure of the SoM models for Phase II metabolism . 104
4.4 Example ReactNet output visualized for hydroxy-methyl-vinyl-ketone ......... 107
4.5 ReactNet significantly outperformed GLORYx and SyGMa when evaluating performance across the full data set, but also when evaluating performance subset by reaction type .................................................. 109

5.1 Alternate pathways for isoxazole bioactivation into electrophilic conjugated systems ................................................................. 120
5.2 Two examples of AMD pathways involving isoxazole cleavage .................. 121
List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.3</td>
<td>Quinone model predictions for bioactivation of isoxazole-containing molecules</td>
<td>122</td>
</tr>
<tr>
<td>5.4</td>
<td>Model predicted bioactivation of 4-amino-5-methyl-isoxazole group into enimines</td>
<td>124</td>
</tr>
<tr>
<td>5.5</td>
<td>Analysis of model bioactivation predictions for isoxazoles with either a phenyl or pyridine substituent</td>
<td>125</td>
</tr>
<tr>
<td>5.6</td>
<td>Analysis of model bioactivation predictions for isoxazoles with either a methyl or ethyl substituent</td>
<td>126</td>
</tr>
<tr>
<td>5.8</td>
<td>Model accuracy analysis for molecule- and pair-level scores for predicted conjugated electrophiles</td>
<td>128</td>
</tr>
<tr>
<td>5.9</td>
<td>Quinone model atom pair and atom scores are different</td>
<td>130</td>
</tr>
<tr>
<td>5.10</td>
<td>Predicted bioactivation pathways and subsequent adducts for OXFBD molecules</td>
<td>131</td>
</tr>
<tr>
<td>5.11</td>
<td>Modeled I-BET151 bioactivation pathways</td>
<td>132</td>
</tr>
<tr>
<td>5.12</td>
<td>Competing bioactivation pathways for OXFBD series of BET inhibitors</td>
<td>133</td>
</tr>
<tr>
<td>5.13</td>
<td>Fluorescence and MS chromatograms of putative quinone adducts for OXFBD02 and OXFBD04 from human liver microsomal reactions</td>
<td>135</td>
</tr>
<tr>
<td>5.14</td>
<td>Mass spectroscopic product ion spectra for predicted quinone metabolite adducts for OXFBD02</td>
<td>137</td>
</tr>
<tr>
<td>5.15</td>
<td>Mass spectroscopic product ion spectra for predicted quinone metabolite adducts for OXFBD04</td>
<td>138</td>
</tr>
<tr>
<td>6.1</td>
<td>This study models four common bioactivation pathways: quinone formation, nitroaromatic reduction, thiophene sulfur-oxidation, and epoxidation</td>
<td>147</td>
</tr>
<tr>
<td>6.2</td>
<td>Bioactivation descriptors were computed using site-level metabolism predictions, metabolite structure predictions, and reactivity predictions</td>
<td>153</td>
</tr>
<tr>
<td>6.3</td>
<td>Structure of the bioactivation model</td>
<td>157</td>
</tr>
<tr>
<td>6.4</td>
<td>Visualization of pathway bioactivation is elusive because it requires consolidating multiple pathway bioactivation scores that can overlap at the same site</td>
<td>158</td>
</tr>
<tr>
<td>6.5</td>
<td>The model produced accurate cross-validated bioactivation pathway predictions</td>
<td>161</td>
</tr>
<tr>
<td>6.6</td>
<td>Binary-class global pathway AUC performance</td>
<td>163</td>
</tr>
<tr>
<td>6.7</td>
<td>The global pathway AUC was computed for the top 5 performing pathway-level models and bioactivation descriptors, in addition to several of the top performing individual topological descriptors</td>
<td>166</td>
</tr>
<tr>
<td>6.8</td>
<td>The neural network computed the most well-scaled predictions of bioactivation pathways, with predictions that were highly probabilistic</td>
<td>168</td>
</tr>
<tr>
<td>6.9</td>
<td>Case studies on ethacrynic acid, mianserin, tolmetin, and imipramine</td>
<td>169</td>
</tr>
<tr>
<td>6.10</td>
<td>Molecule bioactivation scores accurately identified bioactivated molecules</td>
<td>172</td>
</tr>
<tr>
<td>6.11</td>
<td>The neural network produces the best-scaled, highly probabilistic molecule predictions on the molecule-level scoring task</td>
<td>173</td>
</tr>
<tr>
<td>6.12</td>
<td>Comparison of cross-validated molecule predictions against the fully trained model predictions</td>
<td>174</td>
</tr>
</tbody>
</table>
6.13 Using the ROC curve of the cross-validated molecule bioactivation scores, we compute a threshold score for molecule bioactivation .................................................. 176
6.14 To contextualize the bioactivation predictions on the discussed withdrawn drugs, we evaluated their Tanimoto similarity against the training data set on ECFP6-derived fingerprints .................................................................................. 179
6.15 Case studies of model predictions on alclofenac, benzbromarone, safrazine, zimelidine, and astemizole .......................................................... 181

7.1 Model architecture and proposed modeling framework of independent and dependent relative effects ................................................................. 194
7.2 The drug interaction network (DIN) outperforms the compared methods on ROC AUC and, along with MGPS, outperforms the remaining compared methods with regards to PR AUC ................................................. 201
7.3 We examined whether the selection criteria for positive controls has a significant influence on performance by evaluating ROC AUC, PR AUC, and BC for each method over PRR cutoffs of 2 through 20, inclusive ........................................ 203
7.4 The drug interaction network results in comparable performance with MGPS, RR and ROR on the task of binarizing NSAIDs by DILIrank severity scores 208
7.5 The drug interaction network results in comparable performance with RR and ROR on the task of binarizing NSAIDs by the percentage of NSAID liver injury cases ................................................................. 209
## List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Comparison of results on metrics of path recall, intermediate recall, and time cost between substructure matching heuristic variants with radii of 1, 2, or 3</td>
<td>22</td>
</tr>
<tr>
<td>2.2</td>
<td>Intermediate Metabolite Recovery Rate of XenoNet Variants</td>
<td>35</td>
</tr>
<tr>
<td>2.3</td>
<td>Comparison between XenoNet and a random model</td>
<td>43</td>
</tr>
<tr>
<td>2.4</td>
<td>XenoNet outperforms published methods on the GLORY test set</td>
<td>45</td>
</tr>
<tr>
<td>3.1</td>
<td>The bidirectional MPNN architecture outperformed all other methods on multiple accuracy and calibration metrics</td>
<td>61</td>
</tr>
<tr>
<td>3.2</td>
<td>The bidirectional MPNN architecture that incorporates edge features but not node degree features achieves best performance</td>
<td>62</td>
</tr>
<tr>
<td>3.3</td>
<td>We evaluated the bidirectional variant for inference on networks constructed using different parameter values</td>
<td>63</td>
</tr>
<tr>
<td>3.4</td>
<td>Incorporating jumping knowledge, the bidirectional model performance is invariant to changes in the beam width and depth limit parameters during network construction</td>
<td>65</td>
</tr>
<tr>
<td>3.5</td>
<td>Hyperparameters used for model selection during nested cross validation inner loop</td>
<td>97</td>
</tr>
<tr>
<td>4.1</td>
<td>Reaction SMARTS used by the metabolite prediction algorithm.</td>
<td>105</td>
</tr>
<tr>
<td>4.2</td>
<td>ReactNet outperforms published methods in terms of ROC AUC on the AMD-derived Phase II metabolism data set</td>
<td>108</td>
</tr>
<tr>
<td>4.3</td>
<td>ReactNet outperforms published methods in terms of ROC AUC on the GLO- RYx Phase II data set</td>
<td>110</td>
</tr>
<tr>
<td>6.1</td>
<td>Bioactivation descriptors used to predict bioactivation pathways</td>
<td>153</td>
</tr>
<tr>
<td>6.2</td>
<td>Molecule descriptors used to predict bioactivation pathways</td>
<td>154</td>
</tr>
<tr>
<td>6.3</td>
<td>Each pathway-level model’s separate AUC performance for each of the four metabolic transformations and their aggregate multi-class AUC performance</td>
<td>165</td>
</tr>
<tr>
<td>7.1</td>
<td>Characteristics of hospitalizations in cohort</td>
<td>189</td>
</tr>
<tr>
<td>7.2</td>
<td>Assessment of positive and negative controls for diclofenac model validation</td>
<td>198</td>
</tr>
<tr>
<td>7.3</td>
<td>The top 12 diclofenac interactions, as predicted by the model</td>
<td>199</td>
</tr>
<tr>
<td>7.4</td>
<td>Performance metrics comparing Drug Interaction Network to baselines</td>
<td>200</td>
</tr>
</tbody>
</table>
List of Tables

7.5 Predicted interactions between meloxicam and several CYP 3A4 inhibitors . . 205
7.6 Ranking the 8 studied NSAIDs by mean percent relative effect . . . . . . . . 206
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADRs</td>
<td>adverse drug reactions</td>
</tr>
<tr>
<td>AMD</td>
<td>accelrys metabolite database</td>
</tr>
<tr>
<td>AUC</td>
<td>area under the curve</td>
</tr>
<tr>
<td>DILI</td>
<td>drug-induced liver injury</td>
</tr>
<tr>
<td>EHR</td>
<td>electronic health record</td>
</tr>
<tr>
<td>GSH</td>
<td>glutathione</td>
</tr>
<tr>
<td>IADRs</td>
<td>idiosyncratic adverse drug reactions</td>
</tr>
<tr>
<td>MBS</td>
<td>molecule bioactivation score</td>
</tr>
<tr>
<td>PBS</td>
<td>pathway bioactivation score</td>
</tr>
<tr>
<td>QSAR</td>
<td>quantitative structural- activity relationship</td>
</tr>
<tr>
<td>ROC</td>
<td>receiver operating characteristic</td>
</tr>
<tr>
<td>SoM</td>
<td>site of metabolism</td>
</tr>
<tr>
<td>SoR</td>
<td>site of reactivity</td>
</tr>
</tbody>
</table>
Acknowledgements

I am exceptionally grateful to so many people who supported me through my Ph.D. training. First and foremost, I thank my advisor, Josh Swamidass, for giving me the opportunity to pursue research under his supervision. I appreciate the freedom he’s entrusted me with to pursue a variety of research directions, while being a constant source of knowledge, rich ideas, and thoughtful critiques that have developed my skills as a communicator and scientist. I am also indebted to my thesis committee: Drs. Gregory Bowman, Jeremy Buhler, Malachi Griffith, and Fuhai Li. Thank you for your valuable insight and guidance over the years, encouraging me to question and reevaluate not only my scientific assumptions, but also my professional and personal goals.

I am also happy to have collaborated with fantastic, talented researchers within and adjacent to the Swamidass Lab. Arghya Datta and Mick Ward, thank you for your warmth and friendship. I cannot imagine my graduate school experience without your presence and inspiring work ethics as we initiated our graduate school journeys in tandem. Lena Dang, Tyler Hughes, and Matt Matlock, thank you for providing a solid foundation of prior work and an amazing software suite to build off. To other current and past members: Jon Marsh, Rohit Farmer, Kathryn Sarullo, Lev Morgan, and Jack Elliot-Higgins, thank you for your scientific assistance and wonderful conversations. Dr. Grover Miller, Dustyn Barnette, and Allie Schmidt, thank you for your deep metabolism expertise that facilitated my exploration of novel computational methods in concert with experimental techniques. It was a pleasure and I learned a lot; my training would not have been the same without you all.

Finally, thank you to my wife, Astha Saxena. Thank you for pushing me to take risks and proactively seek projects and opportunities outside my comfort zone. I could not have reaped
today’s rewards without your guidance, drive, and nourishment.

Noah Flynn

Washington University in St. Louis

August 2021
For my loving wife, Astha, my generous mother and father, and the many family and friends integral to my life.
Cytochrome P450 enzymes aid in the elimination of a preponderance of small molecule drugs, but can generate reactive metabolites that may adversely conjugate to protein and DNA, in a process known as bioactivation, and prompt adverse reaction, drug candidate attrition, or market withdrawal. Experimental assays are low-throughput and expensive to perform, so they are often reserved until later stages of the drug development pipeline when the drug candidate pools are already significantly narrowed. Reactive metabolites also elude in vivo detection, as they are transitory and generally do not circulate. In contrast, computational methods are high-throughput and cheap to screen millions of potentially toxic molecules during early stages of the drug development pipeline. This work computationally models sequences of metabolic transformations, i.e., pathways, between an input molecule and a corresponding, optional reactive metabolite(s). Additionally, an accurate graph neural network model was developed to assess importance of intermediate metabolites and extract connected subnetworks of relevance to bioactivation. Connecting multiple site of metabolism and structure inference models, we developed an integrated model of metabolism and reactivity to evaluate bioactivation risk driven by epoxidation, quinone formation, thiophene sulfur-oxidation, and nitroaromatic reduction. We applied this framework to an understudied substructure, the isoxazole ring, that is gaining traction in a class of drugs known as bromodomain inhibitors that may potentially drive quinone formation. Finally, we attend to toxicity associated with drug-drug interactions, particularly with NSAID usage reported in electronic health records.
Chapter 1

Introduction

1.1 Idiosyncratic Adverse Drug Reactions

Adverse drug reactions (ADRs) are a core limiting factor in the development of drugs and their distribution. ADRs account for 4.9–7.7% of hospital admissions [1, 2] and are a leading cause of death in the U.S. [3]. Moreover, vulnerable populations face high risks – 39% of ADRs in pediatric patients end up being life-threatening or fatal and 11.4–35.5% of geriatric hospital admissions are due to drug-related incidences [4, 5]. Incidence of ADRs in the USA are estimated to cost up to 30.1 billion dollars annually due to increased hospitalization, prolongation of hospital stay, and visits to emergency departments [6].

Approximately 10–15% of ADRs are dose-independent and idiosyncratic (IADRS), which makes them notoriously difficult to plan for and unable to be evaded by simple dose adjustments [7]. IADRs occur at very low frequencies of 1 in 10,000 to 1 in 100,000 [7], allowing them to evade detection until the drug has gained increased exposure at large population levels that are not tractable in clinical trials. IADRs commonly affect the skin, blood cells, and especially the liver. At least 17% of liver transplant cases and 50% of acute liver failure cases can be traced to IADRs [8, 9]. Consequently, IADRs are a major reason for drug clinical trial termination and drug-induced liver injury, which is the leading cause of drug withdrawal from the market [10, 11]. Partly due to a poor ability to screen early on for IADRs, one in five drugs that has been on the market for at least two years ends up being withdrawn or tagged with a black box warning.
Chapter 1. Introduction

Bioactivation, which is the metabolic activation of an extraneous drug, or xenobiotic, into a reactive metabolite, is a common precursor of ADRs.

1.2 Drug Metabolism

An understanding of basic drug metabolism is necessary for understanding the origin of ADRs. Orally administered drugs must commit a first pass through the hepatic portal system and the liver, where most drug metabolism occurs, before reaching the general circulatory system. To be absorbed effectively, an oral drug must be able to leave the aqueous medium of the digestive tract and cross a non-polar bilayer into the aqueous blood of the hepatic portal system. Thus, oral drugs tend to be lipophilic. Drugs will eventually undergo elimination via a combination of metabolic and excretory processes. Metabolism facilitates the alteration of drugs into more polar forms, which have a higher rate of clearance by the kidneys, so that their effects may be removed from the body.

The process of transforming a drug into a more polar form is described by phase I and phase II reactions. Phase I reactions introduce or unmask polar functional groups, e.g. –OH, –NH₂ or –SH, on the parent molecule. Most phase I products are still not polar enough for rapid elimination and will undergo a phase II reaction. Most phase II reactions are conjugation reactions, whereby a polar molecule is attached to a suitable polar “handle” that is already present on the drug or has been introduced by a phase I reaction. The resulting conjugate has greatly increased polarity and is more likely to be safely excreted. Common polar moieties that are conjugated include glucuronate, acetate, and sulfate [13].

The Cytochrome P450 superfamily of oxidative enzymes aids in the elimination of lipophilic drugs and the metabolism of 90% of all drugs is accounted for by 6 of the human P450s [14, 15]. The P450 enzymes are the most relevant to the bioactivation processes surrounding drugs, as they collectively have more substrates than any other enzymes and several of their reaction products have reaction chemistry due to their strong electrophilic properties [16]. While P450
enzymes typically aid in the detoxification of drugs, sometimes they generate reactive metabolites, or metabolites that can react with protein and DNA, in a process defined as bioactivation.

The interaction between reactive metabolites and proteins can deleteriously alter protein structure and protein folding. In particular, a reactive metabolite can affect the immune system by binding to protein as a hapten and generating antigenic determinants for the adaptive immune system [17]. Halothane, a general anesthetic, is metabolized by CYP2E1 to form trifluoroacetyl chloride. Trifluoroacetyl chloride forms hepatic protein adducts associated with liver injury. For halothane, detection of circulating IgG antibodies that bind to the halothane-protein adduct confirmed that adduct formation results in a drug-specific immune response [18]. Similar studies have implicated diclofenac, tienilic acid, and sulfamethoxazole as vectors for drug-induced immune hepatitis [19, 20, 21].

On the other hand, interactions with nucleic acids can alter DNA structure or gene expression. Consequences of changes in DNA can invoke carcinogenicity and teratogenicity. An example of the former include the ubiquitous alkylaniline–containing compounds and their associations with several types of cancer due to DNA damage mediated by nitrenium ion chemistry, formation of quinone imines, and formation of reactive oxygen species (ROS) [22]. An example of the latter is the anticancer drug cyclophosphamide, which forms reactive metabolites at concentration levels that can overwhelm the relatively low concentration of detoxifying enzymes in the embryo–fetus, such as epoxide hydrolases and glutathione transferases [23]. Formation of adducts on a developmentally important protein or gene has been shown to negatively alter the development of the embryo–fetus.

Unexpected interactions between multiple drugs can also heighten the risk of ADRs. The burden of metabolizing drugs in the liver is primarily incurred by six P450 enzymes and predicting drug behavior assumes proper functionality of these enzymes. In some cases, P450–mediated metabolism of a drug may result in a reactive metabolite that does not release from the enzyme’s active site and blocks access from other substrates, termed mechanism–based inactivation. Mechanism–based inactivation can cause irreversible inhibition of an enzyme and
break the above assumption [24]. A drug which causes irreversible inhibition can increase toxicity risk if the inactivated enzyme is necessary for metabolizing either an endogenous substrate or additional drugs being administered, as in polypharmacy therapies. Prior to its withdrawal, mibefradil was administered as a calcium-channel inhibitor. However, mibefradil was also a potent inhibitor of CYP3A4/5 when given at therapeutically relevant concentrations and suppressed metabolism of other drugs which depended on CYP3A4-mediated pathways, such as simvastatin. Patients who received simvastatin in conjunction with mibefradil reported cases of myopathy and rhabdomyolysis [25]. Due to mibefradil’s interactions with other drugs, it was withdrawn from the market in 1998.

1.3 Current Methods for Reactive Metabolite Screening

1.3.1 Experimental Approaches

Experimental approaches are geared towards detecting reactive metabolites or their conjugated forms. Trapping studies detect if conjugates have formed and can be used to characterize reactive metabolite structure. Covalent binding studies are able to quantify the amount of conjugation. Both types of studies can be performed in vivo and in vitro.

Trapping studies have been 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 [26]. Formation of a GSH conjugate indicates the presence of a reactive metabolite and the GSH conjugate can be detected via mass spectrometry.

Trapping studies have limitations including: bias by trapping agent, low throughput, and that no intermediate structures are tracked. The first consideration in trapping studies is that selection of a trapping agent will bias the possible reactive metabolites that can be trapped. Specifically, GSH has a soft nucleophilic sulfhydryl group which is not likely to react with hard electrophiles. Furthermore, it is necessary to select the appropriate metabolic system depending
on the type of toxicity being studied. For instance, the myeloperoxidase enzyme system is suited to drugs implicated in hematological toxicity [27]. In addition, trapping studies are low throughput and ill-suited to screening a large pool of candidate compounds. Another limitation is that only the formation of an adduct is monitored, but not any prior metabolic transformations that led to the reactive metabolite. Intermediate metabolites between the drug’s initial structure and the detected reactive metabolite are likely to be missed, as most reactive metabolites require more than one metabolic step [28].

Covalent binding studies use a radiolabeled drug to quantify the amount of adducts formed by reactive metabolite binding [29]. A difficulty in covalent binding studies lies in establishing a baseline level of covalent binding which would implicate a drug as toxic. A recent study examined P450–mediated covalent binding of nine hepatotoxins and nine non–hepatotoxins to liver microsomes, S9 fractions, and hepatocytes and found no correlation between the extent of covalent binding and toxicity [7]. Nevertheless, it is generally agreed upon that lower covalent binding is optimal for reduced toxicity. Like trapping studies, covalent binding studies are also low throughput. Covalent binding studies require a radiolabeled drug and synthesizing radioactive substances for a large number of compounds is time consuming, resource intensive, and requires quality assurance that the reactive metabolites are appropriately labeled with radioactive isotopes.

Each of the previously mentioned approaches is still prone to generating false positives if their results are not contextualized. This relates to a discrepancy between bioactivation in standard in vitro systems and alternative clearance pathways in in vivo systems. Raloxifene, a drug used in the treatment of osteoporosis, has displayed in vitro CYP3A4 irreversible inactivation, microsomal covalent binding, and formation of GSH conjugates [30]. These results are due to the potential for raloxifene to form reactive quinonoid species due to alteration of its phenolic groups. However, the principle metabolic pathway in vivo involves scavenging a precursor of the reactive quinonoid via glucuronidation of the aforementioned phenolic groups [31]. Ultimately, raloxifene rarely results in IADRs.
1.3.2 Computational Approaches

In silico methods are appealing for their ease of use, reduction of resources and expenses, and because they can avoid problems engendered by experimental time scales, i.e. the transience of reactive metabolites. Structural alerts conservatively filter out drugs which contain substructures that strongly correlate with toxicity. Quantitative structure–activity relationship (QSAR) models quantify a relationship between the physicochemical properties of a drug and the drug’s biological activity. Site of metabolism (SOM) models attempt to establish the likelihood of each site on the drug undergoing perturbation by an enzymatic system and the likelihood of such a perturbation resulting in a reactive metabolite.

A common approach to early detection of molecules that may cause IADRs is to filter out those molecules containing structural alerts – chemical structures that have been previously implicated in the formation of reactive metabolites. Advantages of structural alerts include their ease of use and low cost. In addition, structural alerts are present in 78–86% of drugs implicated in toxicity studies [32].

A limitation of structural alerts is that they are prone to flagging safe drugs as toxic and toxic drugs as safe [33] (Figure 1.1). This limitation is an artifact of how structural alerts are aggregated. To be an effective filter, an extensive list of structural alerts must be collated from literature. An extensive list is necessary because structural alerts as predictions for future toxicity do not generalize to structures that have not been characterized well or cited as being potentially toxic. However, such a list will capture structures that have important medicinal properties, such as thiophenes, phenols, and isoxazoles. In fact, around 50% of the 200 most frequently prescribed drugs in the US contain at least one structural alert, yet most of these drugs do not yield IADRs [32]. In summation, the key deficiency of structural alerts is their inability to effectively model metabolism and reactivity of a drug. Because structural alerts do not account for the metabolic pathways a compound can go through, they cannot assess whether the structural alert is actually bioactivated, whether bioactivation of the structural alert competes with favored non–bioactivating pathways, nor the rate at which a reactive metabolite undergoes
Chapter 1. Introduction

Figure 1.1: Structural alerts do not adequately account for molecule context, including metabolic processes. Drug metabolism with regards to a structure is ambiguous and structural alerts can be misleading and overestimate toxicity hazard or completely miss compounds that are actually hepatotoxic but do not contain a structural alert. While the thiophene structural alert in methapyrilene is the site of bioactivation, the same alert structure in eprosartan is not. The differences in biotransformation lead to drastic differences in the safety profiles of the two compounds. While methapyrilene is withdrawn from the market due to hepatotoxicity, eprosartan is safe.

QSAR models attempt to identify the relationship between a drug’s structural, physical, and chemical properties to their ability to modulate the drug’s biological activity. While keeping most of a drug’s physicochemical properties constant, one or two of the properties may be modified and related to a corresponding change in the drug’s activity via regression methods. QSAR models have been shown to successfully identify reactive metabolites and inform how structural characteristics of these molecules, such as functional groups, influence biological activity [34, 35]. However, QSAR models are usually only trained on small sets of structurally similar molecules. QSAR predictions are yielded from interpolation within the training domain; QSAR predictions fail to extrapolate outside the range of the training domain. Consequently, the utility of QSAR models is restricted to closely related molecules that belong to the same family, act at the same target, and have the same mechanism of action [36].

SOM models attempt to address the prior method’s issues with generalizing to samples outside of the training set. As the name implies, SOM models predict which sites in a drug are
most likely to be altered by an enzymatic system. The construction of SOM models is diverse. SMARTCyp, an early reactivity–based SOM model for P450-mediated drug metabolism, uses activation energies for P450–mediated reactions involving the substructure of interest. These activation energies are pre–computed using extensive density functional theory calculations for accuracy and saving these calculations for look–up upon execution [37].

SMARTCyp has been validated against CYP3A4 and six other CYP isoforms [38], but still suffers when encountering substructures that do not have reactivity scores accessible via the look–up table. A method which has generalized better is the RS–Predictor model, which incorporates support vector machines and has greater granularity in its observations – able to yield atom-level predictions as opposed to predictions for pre–defined substructures [39]. Additional works seek further improvements through the use of neural networks, which allow for probabilistic predictions and a priority to be placed on those sites of most interest for experimental validation [40].

A predecessor to the work discussed in this dissertation, XenoSite, uses neural network SoM models to predict important vectors of reactive metabolite formation such as quinone formation, epoxidation, reduction, and other phase I reactions. XenoSite also includes reactivity models against DNA, protein, glutathione (GSH), and cyanide. These models have been validated with high predictive accuracy, with area under the receiver operator curve (AUC) scores of 90% or better.

1.4 Integration of Metabolism and Reactivity Models to Study Bioactivation

Chapter 2 details development of an integrated approach, XenoNet, to modeling sequences of metabolic transformations, i.e., pathways, between an input molecule and potential reactive metabolites. XenoNet incorporates site of metabolism models for predicting the probability of each transformation in the pathway along with structure inference models for inferring the
Chapter 1. Introduction

structure of the resultant metabolite. Chapter 3 extends XenoNet, using a graph neural network to assess importance of intermediate metabolites and extract connected subnetworks of relevance to bioactivation. This approach is applied to withdrawn drugs with unknown mechanisms of toxicity to model their Phase I and Phase II metabolic pathways, along with a ranking of which intermediate metabolites are most relevant for further experimental validation. Chapter 4 describes further integration of SoM and structure inference models for Phase II metabolism.

Chapters 5 and 6 are concerned with an approach that directly outputs a molecule’s bioactivation risk. To test the validity of this concept, I first applied an ad hoc framework to an understudied substructure, the isoxazole ring, that is gaining traction in a class of drugs known as bromodomain inhibitors, but may be a potential driver of quinone formation (chapter 5). I then extend this idea to an automated approach via a model for bioactivation risk driven by epoxidation, quinone formation, thiophene sulfur-oxidation, and nitroaromatic reduction (chapter 6). Finally, chapter 7 attends to toxicity resulting from drug-drug interactions, particularly with NSAID usage reported in electronic health records.
Chapter 2

XenoNet: Inference and Likelihood of Intermediate Metabolite Formation


2.1 Introduction

Bioactivation risk can be allayed by early identification of reactive metabolites. Current experimental approaches, such as trapping studies or covalent binding studies, are geared towards detecting reactive metabolites or their conjugated forms [26, 29]. However, such methods are resource-intensive, time-consuming and can be expensive and biased by experimental design choices, such as the type of trapping agent used. In silico methods can operate at a higher throughput and under shorter time scales. A common computational approach is to try to predict the site on a drug molecule where a reactive metabolite might form [23]. Reactive metabolite identification can then be followed up by rational drug modification that avoids potential reactive metabolite formation while retaining the desired pharmaceutical effect of the now altered drug [15, 21, 22].
Chapter 2. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

However, there is a dearth of commonly used techniques to monitor sequential metabolic transformations, and resulting intermediate metabolites, required to form a reactive metabolite. This is an important task since the majority of drug molecules require more than one metabolic transformation, and therefore an intermediate metabolite, to form a reactive metabolite [28]. The fulfillment of such a goal would reveal important intermediate metabolites that are required to form reactive metabolites. Consequently, this would open up additional avenues for preventing reactive metabolite formation and subsequent IADRs, as one could modify drug molecules to avoid these intermediates.

A recently published study highlights the need and the ability to use in silico methods to identify intermediate metabolites that are necessary to form reactive metabolites [41]. The authors identified an important intermediate metabolite by combining a site of metabolism model with a model than can infer metabolite structures. Specifically, they discovered that terbinafine (TBF), an antifungal drug known to cause toxicity [42], forms a previously unidentified intermediate, desmethyl terbinafine (TBF-D), which has since been shown to be an important precursor in the formation of the reactive metabolite TBF-A [43]. Further experimental investigations of TBF and TBF-D metabolism into TBF-A by P450 isoymes was later able to reveal the degree of involvement of CYP2C9 and 3A4 in TBF’s metabolic clearance and bioactivation potential [44, 45]. Knowledge of the intermediate, TBF-D, provides a better mechanistic understanding of reactive metabolite formation and could inform potential modifications that would reduce the bioactivation potential of TBF. However, this approach has not been generalized. Instead, the author’s approach involved manual application of separate models in an ad hoc manner to understand how TBF is bioactivated.

In this work, we explore a generalized approach that combines two types of models: Site of metabolism (SOM) models, which identify atoms, or bonds, that are liable to be metabolized, and metabolite structure inference models, which can infer the structures of potential metabolites formed during metabolism of a given molecule. With regard to SOM prediction, there are
several previously published phase I SOM prediction models that are freely available. SMART-Cyp, RSPredictor, SOMP, MetaPrint2D, FAME 3, Site of Metabolism Estimator (SOME), and He et al. are a few examples of the methods available [38, 46, 47, 49, 50, 51, 52]. We employ our Rainbow XenoSite model, which was developed previously as part of our ongoing effort to develop a collection of free, usable metabolism and reactivity models, referred to collectively as XenoSite [53]. Our motivation for doing so is that Rainbow XenoSite, compared to the aforementioned models, is the only model that produces well-scaled, probabilistic outputs in a reaction-type specific manner that, in conjunction with knowledge of the site of metabolism, allow for unambiguous inference of the metabolite structures that a molecule is likely to form [54]. Furthermore, our prior work has also shown that Rainbow XenoSite covers the largest proportion of known human phase I metabolic reactions and its coverage of reaction types includes many important bioactivation reactions not covered by the other models.

Comparatively, metabolite structure inference models are a less explored space than SOM models. Previously published and freely available methods worth noting include BioTransformer, RD-Metabolizer, SyGMa, and GLORY [55, 56, 57, 58]. The earliest approach of those listed, SyGMa, presents a rule-based method for predicting the potential child metabolites produced from a given parent metabolite. SyGMa’s reaction rules cover both phase 1 and phase 2 metabolism and are augmented by probability scores to allow ranking of the predicted metabolites. SyGMa’s probability scores for reaction rules are derived from statistical analysis on a large data set of experimental metabolic reactions. A second method, BioTransformer, combines a knowledge-based approach with a machine learning approach to predict small molecule metabolism across a wide-range of contexts, including CYP450-mediated metabolism. The former includes use of MetXBioDB, a biotransformation database of annotated, experimentally-derived metabolic reactions that informs a reaction knowledge base for metabolite prediction. The latter involves use of CypReact for CYP specificity prediction [59]. In contrast, GLORY accomplishes metabolite prediction without reliance on metabolic reaction data sets and instead implements rules derived from scientific literature and chemistry knowledge.
applies FAME 2 to enable filtering out potential false-positive predictions and ranking of predicted metabolites by their likelihood of occurrence.

We employ our own metabolite structure inference model, called the Metabolic Forest, which we have previously established and validated as an accurate tool for predicting metabolic structures in comparison against RD-Metabolizer [60]. Similar to the aforementioned methods, Metabolic Forest uses a rule-base approach for metabolite inference. However, an elusive problem common to all of the above methods is the generation of large numbers of false positive metabolite structures. Use of an SOM model in tandem with a metabolite structure inference model can remedy this by allowing early filtration of metabolites that would result from low-value predictions and a basis for ranking the remaining predicted metabolites in accordance to their likelihood of being legitimate.

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

2.2 Data and Methods

2.2.1 The Metabolic Network Data Set

To construct our metabolic network data set, we used a previously described data set that contains 20,736 in vitro and in vivo human phase I reaction records filtered from the Accelrys
Chapter 2. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

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

**Figure 2.1:** Metabolic network data set construction and overview of XenoNet. A) Multiple experimentally observed reactions from the AMD can be linked to an annotated network. Using these annotated networks, we can evaluate how well different metabolic-network generating algorithms can infer observed intermediate metabolites. B) XenoNet is a metabolic network predictor that, given a substrate and a target product as inputs, can infer the metabolic pathways connecting two input molecules and the corresponding likelihood of each pathway. In XenoNet, the Metabolic Forest algorithm is applied iteratively to generate a tree of potential pathways that span multiple metabolic transformations. During construction of this tree, pathways between the starting molecule and a target molecule can be enumerated. The likelihood of each step in a pathway can then be computed from the five-color predictions by Rainbow XenoSite. The five-colors are used to allow for ease of use when visualizing the networks that XenoNet generates. Each of the major groups corresponding to one of the five-colors is further subdivided into more nuanced reaction type classes that are used to annotate the edges in the generated network. These more detailed edge annotations are accessible through the XenoNet network object.
Chapter 2. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

Metabolite Database (AMD), a literature-derived database. Each reaction record was classified into one of the five phase-I metabolism classes: stable oxygenation, unstable oxygenation, dehydrogenation, hydrolysis, and reduction reactions and manually labeled with the site of metabolism (SOM). The final data set contains 10,280, 5,811, 2,794, 3,869, and 1,590 stable oxygenation, unstable oxygenation, dehydrogenation, hydrolysis, and reduction reactions, respectively. Due to their low prevalence, some phase I reaction types such as tautomerization, isomerization, rearrangement, radical formation, hydration, deacylation, denitrogenation, and decarbonylation were not included in our data set.

These reaction records were converted into graph representations where the nodes represent the substrates and the products, and the edges represent metabolic transformations. These graphs were then linked together to construct networks based on shared nodes—identical chemical structures. Figure 2.1A illustrates an example of how these records are collated into an annotated graph. In this example, one record may indicate that a substrate molecule (S) undergoes a reaction that generates a metabolite (M). A second record may indicate that (M) undergoes a reaction to generate another metabolite (P). A third record indicates that (S) is directly metabolized into (P). Therefore, these three records can be linked to show that the substrate (S) is connected to the downstream metabolite (P) through the metabolite (M), serving as an intermediate node. At the end of the network collation process, all networks that are induced subgraphs of another network are removed. The final data set contains 17,054 annotated metabolic networks with at least one direct path connecting each substrate molecule to its recorded product. Each of the networks represents a unique substrate-product pair. Approximately 91% of the paths across all annotated networks require three or fewer metabolic transformations. Although we cannot share the exact chemical structures from the proprietary AMD, we provide the AMD reaction records for each metabolic network in our data set in the “Metabolic_Network_Dataset.json” file.
2.2.2 The Metabolic Network Generator XenoNet

We built XenoNet, a metabolic network predictor that, given a substrate and a target product as inputs, can infer the metabolic pathways connecting two input molecules and the corresponding likelihood of each pathway. Two previously developed, in-house models are combined and augmented to develop XenoNet. The first is Rainbow XenoSite, a deep learning phase I metabolism model that, given a molecule as input, can accurately predict sites of metabolism (SOMs) for each of the 20 phase I reaction types via five reaction classes.[54] In the current work, we utilize the same set of transformation rules that define the reaction classes and reactions types in Rainbow XenoSite. The second is the Metabolic Forest, which takes the substrate molecule as input and infers possible metabolite structures that are one metabolic step away.[60] In addition to inferred metabolite structures, the Metabolic Forest also outputs the specific SOMs and the corresponding metabolic transformations that act on those SOMs to produce the inferred metabolite. Metabolic Forest infers metabolite structures of a given substrate using breadth-first searches and transformation rules encoded through a combination of reaction SMARTS, resonance pair rules, and resonance structure rules.

Briefly, in XenoNet, the Metabolic Forest algorithm is applied iteratively to generate a tree of potential pathways that span multiple metabolic transformations (Figure 2.1B). During construction of this tree, pathways between the starting molecule and a target molecule can be enumerated. Through the application of Rainbow XenoSite, the likelihood of each step in a pathway can then be computed. XenoNet’s weighted metabolic pathways are stored using a graph-based data structure where each molecule is a node, and each metabolic transformation is an edge. Thus, unlike the Metabolic Forest, XenoNet not only infers metabolite structures but also predicts the likelihood of the metabolic transformation that the input molecule undergoes to form the inferred metabolite structures. Whereas the Metabolic Forest outputs a list of the inferred metabolite structures it enumerates, XenoNet encodes its inferred metabolite structures, along with edge-level data such as SOM predictions, in a network object. This network object supports additional functionality, such as the ability to compute the likelihoods of paths.
and metabolites in the network. In addition, the network object readily supports conversion to a NetworkX MultiDiGraph class object for further utility. Lastly, XenoNet incorporates a depth-first search strategy that yields much lower memory requirements compared to the breadth-first search strategy employed by Metabolic Forest. XenoNet was implemented in Python with the 2018.09.03 release of RDKit.[61] Seven different variants of XenoNet were implemented, and the best variant was selected as the final version. The detailed descriptions of each step are in the following sections.

2.2.3 The Naive XenoNet Variant

The first XenoNet variant is a naive model that employs a brute-force approach to network construction, which enumerates all possible paths between the starting molecule and the target molecule. Given a query consisting of a substrate-product pair of molecules, Metabolic Forest is iteratively applied to construct a metabolic tree of successive intermediates in a depth-first manner until a pathway connecting the queried pair is found or the depth limit is reached. Next, all discovered paths from the substrate to the target product are used to construct a new metabolic network. Once the search completes, and the structures in the network are known, the whole network is given as input to Rainbow Xenosite. In this step, each transformation edge is given a prediction score depending on the reaction class that it belongs to. As an example, we show how the process works when the model is given 1,3-butadiene and hydroxymethylvinyl ketone as its substrate and product pair inputs (Figure 2.2).

In addition to detecting and removing cycles, such as repeatedly adding and removing a hydroxyl group, from the network, we prevent wasteful computations by using two parameters: the depth limit and the time limit. The depth limit parameter specifies the maximum length of the searched paths. The time limit parameter specifies the maximum amount of time that the network generation can not exceed. During path enumeration, we construct a separate metabolic tree to keep track of all pathways that connect the queried pair. If the metabolic network exceeds its allotted time, the current state of the partially generated network is saved.
Chapter 2. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

Figure 2.2: XenoNet infers metabolic pathways connecting a substrate and a product from their structures in three steps. First, using Metabolic Forest [60], a depth-first search scans across a space of possible metabolite structures for paths that terminate at the given product metabolite. Second, the discovered paths are used to construct a new metabolic network with the input substrate and product as terminal nodes. Third, Rainbow XenoSite [54] yields predictions on the metabolic transformation edges of the constructed metabolic network. As an example, we show here how the process works when the model is given 1,3-butadiene (in solid-line box) and hydroxymethylvinyl ketone (in dash-line box) as its substrate and product pair inputs. The search starts at 1,3-butadiene and explores as far as possible along a branch of metabolic transformations up to a limited depth, before backtracking to continue the search along other branches (left). Only pathways that go from 1,3-butadiene to hydroxymethylvinyl ketone are retained to construct a new metabolic network (right). Once all metabolites and directed edges linking them are in the network, the network is passed into Rainbow XenoSite to compute the likelihood for each metabolic transformation, shown as the numbers next to the edges.
This brute-force approach is only a baseline to demonstrate the necessity of clever algorithms to make the problem tractable.

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

### 2.2.4 Heuristic XenoNet Variants

In addition to the naive variant, we develop six other XenoNet variants. To limit the branching factor of potential metabolic trees, we use different combinations of a top-$N$ heuristic and a substructure matching heuristic in six other variants of XenoNet (Figure 2.3).

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

The top-$N$ heuristic limits the tree to grow as $x^N$, where $x$ is the number of steps in a pathway, and $N$ is the number of metabolic transformations that can be explored for each molecule. Without this heuristic, $N$ can effectively be on the order of hundreds of metabolic transformations. Furthermore, the top-$N$ heuristic has two forms: reaction-agnostic top-$N$ and reaction-specific top-$N$. In reaction-agnostic top-$N$, the top-$N$ child metabolites are selected to explore in the next step, as previously discussed. However, reaction-agnostic top-$N$ assumes that the probabilities generated by the Rainbow XenoSite for each of the 20 reaction types are comparable, which is an over-simplification. Therefore, we also employ a reaction-specific top-$N$,
Chapter 2. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

Figure 2.3: Top-N and substructure matching heuristics. To limit the branching factor of potential metabolic trees, we use top-N and substructure matching heuristics. As an example, let’s consider a pair of a substrate (in solid-line box) and a target product (in dash-line box), as shown above. For the first generation of metabolites, XenoNet discovers seven structures, one of which is our target product. However, only the target product and structures that meet top-N and/or substructure heuristics would be considered in the next step of the search, and the remaining metabolites would be ignored. The top-N heuristic filters for metabolite structures that receive N highest scores. The numbers next to the arrows are scores assigned by Rainbow XenoSite for the corresponding metabolic transformation. Only two metabolites with the scores of 0.60 and 0.45 pass the Top-2 filter. The substructure matching heuristic filters for metabolite structures produced via transformation at sites expected to lead to the target product. In the substrate structure, the site of structural difference between the substrate and target compound, i.e. the site of metabolism, is circled in red. In each metabolite structure, the site of structural difference between the metabolite and the substrate is circled in red if it contains the site of metabolism and grey if it does not. Only two metabolites with red circles pass the substructure matching filter. Different combinations of top-N and substructure matching heuristics are used to construct six XenoNet variants.
where each of the 20 reaction types is considered separately. In the reaction-specific top-$N$, the child metabolites produced from the top-$N$ transformations within each of the reaction types are eligible for further exploration. For example, when $N$ is set equal to 3, the top-3 subsequent metabolites formed via each reaction type would be considered for the next step, allowing for a maximum of 60 child metabolites to be considered in the next step.

One final variant of the top-$N$ heuristic, hereon referred to as the optimal thresholds heuristic, limits the acceptable child metabolites based on whether the metabolic transformation required to produce them is above a probability threshold. We used a separate threshold for each of the five major reaction classes – stable oxygenation, unstable oxygenation, dehydrogenation, hydrolysis, and reduction. To define the thresholds, we used the optimal point on the cross-validated ROC curves computed from Rainbow XenoSite’s atom-level reaction predictions on its training data for each reaction class.[54] The threshold was computed to optimize both sensitivity and specificity using the Youden index.[62] For example, a metabolic transformation with the epoxidation reaction type must have a score greater than the stable oxygenation reaction class threshold for the produced child metabolite to be retained for further processing at the next step of the search.

In the substructure matching heuristic, XenoNet only generates intermediate metabolite structures that result from applying reaction rules to sites expected to lead to the product metabolite. Concretely, metabolic transformations are only computed for sites where the current intermediate metabolite being assessed differs from the product metabolite, as well as their immediately neighboring sites. For instance, if a substrate-product pair only differ by hydroxylation at a single site, then the reaction rules will only be applied to that differing site and its adjacent sites, rather than every site in the substrate. Substructure matching can be enhanced by allowing a parameter for the radius from sites of differences. As described previously, the radius is one, i.e., we find the sites where differences are present and only run the SOM model on those sites and their adjacent sites. The radius hyperparameter for substructure matching was also evaluated for radii of 2 and 3 (Table 2.1), but these variants did not perform better than substructure
Chapter 2. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

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

Table 2.1: Comparison of results on metrics of path recall, intermediate recall, and time cost between substructure matching heuristic variants with radii of 1, 2, or 3. Increasing radius increases the network generation time cost without increasing performance across the path recall and intermediate recall metrics. In general, the weakness of the substructure matching heuristic is that it misses paths where any single step in the path requires a metabolic transformation to occur at a site that does not pass the substructure matching filter. In theory, increasing the radius hyperparameter would increase the number of viable sites of metabolism for a given molecule being processed and diminish the number of missed paths. In practice, the exchange in greater time cost may be the factor that mitigates the advantage that a greater radius hyperparameter would otherwise bring.

Matching with a radius of 1. Ultimately, substructure matching was evaluated with respect to the other heuristics with a radius of 1 only. Some reaction rules, such as dehydrogenation, can operate on two distinct sites in a molecule. If the reaction rule specifies two sites, then both sites must be in the set of valid sites derived from the substructure matching heuristic.

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

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

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

2.2.5 Metabolite Scoring Algorithm

Several of the metrics used for evaluating our models require a way to rank the relative importance of metabolites in a given network. Thus, except for the substrate that is always assigned a score of 1.0, we calculate metabolite scores for all other compounds in the network through three steps. First, we start with the raw score $w_{M_j \rightarrow M_k}$ that Rainbow XenoSite assigns to the metabolic transformation between a metabolite $M_j$ to one of its children $M_k$ (Figure 2.4a). Second, the raw score $w_{M_j \rightarrow M_k}$ is normalized over all metabolic transformation from $M_j$ to its children (Figure 2.4b):

$$W_{M_j \rightarrow M_k} = \frac{w_{M_j \rightarrow M_k}}{\sum_{M_x \in M_j^{\text{children}}} w_{M_j \rightarrow M_x}} \tag{2.1}$$
Figure 2.4: Metabolite score. To rank the relative importance of metabolites in a given network, we calculate metabolite scores. There are three steps in this process. a) First, each metabolic transformation in the network is assigned with a raw prediction by Rainbow XenoSite. b) Second, the raw prediction is normalized using Equation 2.1. c) Third, the substrate is assigned a score of 1.0, and downstream metabolites are assigned with scores computed using Equation 2.2. Because a node can only be scored after all of its parents have been scored, the third step is carried out in multiple layers. The dashed-line dividing the network into segments corresponds to the computation of the metabolite scores for subsequent layers over the network. On the first layer, for example, scores could be computed for A. The computed score for A, displayed as numbers adjacent to each metabolite, is 0.25. The number of layers required to compute all metabolite scores is equal to the path of maximum length. The maximum length path in this network requires five steps. Computing metabolite scores for all metabolites in the network shown required five layers.
Third, the metabolite score $F_{M_j}$ is a weighted sum of the normalized $W_{M_i \rightarrow M_j}$, where $M_i$ is one of the parents of $M_j$:

$$F_{M_j} = \sum_{M_i \in M_j^{\text{parents}}} F_{M_i} \times W_{M_i \rightarrow M_j} \quad (2.2)$$

Because a node can only be scored after all of its parents have been scored, the third step is carried out in multiple layers (Figure 2.4c).

### 2.3 Results & Discussion

In the following sections, we examine the inner workings of XenoNet, our metabolic network predictor. First, we compared the seven variants of XenoNet on a randomly sampled subset of our metabolic network data set to identify the optimal model. The most efficient XenoNet variant was chosen as the final XenoNet model. Second, we assess the performance of the final XenoNet model on the full metabolic network data set. Third, we compare XenoNet to prior work on the task of predicting metabolite formation with respect to phase I metabolism.

#### 2.3.1 Comparison Between Metabolic Network Variants

The main goal of XenoNet is to enumerate and assign prediction scores to sequences of metabolic transformations (edges) and intermediates (nodes) in pathways that connect a substrate-product pair. We generated seven XenoNet variants and wanted to select the best variant as our final model. However, the time cost of producing networks poses a significant computational challenge when running each method across the full metabolic network data set. Inferring pathways of depth 3 or higher, along with model predictions, can take a long time. As such, it is infeasible to run each XenoNet variant on the full data set. Instead, we randomly sample from our full data set a subset of 1,024 substrate-product pairs to compare the model variants. In addition to the naive XenoNet that uses no heuristics, the other six variants employ either substructure matching, top-$N$ heuristics, or a combination of both. The choice of heuristic influences the
paths that are enumerated from the substrate molecule to the product molecule. Consequently, heuristic choice influences the presence of an edge in the predicted graph but does not influence the scoring of the edges provided by the Rainbow model. As such, metrics for comparing performance between heuristics focused on the time cost of network generation, path recall, intermediate metabolite recall, and intermediate metabolite recovery. Metrics for comparing the relevance of edge predictions during metabolic network construction were not compared. The model variant with the best performance on this subset would be the final XenoNet model for analysis on the full data set.

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

**Time-cost**

An effective algorithm would be able to quickly identify a pathway to known metabolites in all cases and identify known intermediate metabolites. A brute force, depth-first search alone is not tractable since the time complexity of searching the metabolic forest is \( O(n^d) \), where \( n \)
FIGURE 2.5: Time-cost varies greatly between the seven XenoNet variants. Each method was allotted 30 minutes for network generation of each substrate-product pair at a depth limit of 3 steps from the substrate. Top) Across 1,024 substrate-product pairs, only the substructure matching, top-$N$ specific, and top-$N$ agnostic variants produced above 70% fully generated networks within the allotted time. The optimal thresholds heuristic fully generated almost 50% of the networks. While the naive and the combination variants are the least restricted, they only produced at most 20% fully generated networks. Bottom) A similar trend occurred when comparing run time distributions. The substructure matching, top-$N$ specific, and top-$N$ agnostic variants, on average, take less than 10 minutes to generate a network. The time-distribution for the naive and the combination variants is much broader, and many runs hit the 30 minute timeout.
is the number of metabolites for a given molecule and $d$ is the depth of the tree—the number of metabolic steps that is allowed between the starting metabolite and the target metabolite. In some instances, $n$ can be on the order of $10^3$. In this comparison, each method was allotted 30 minutes for network generation of each substrate-product pair at a depth limit of 3 steps from the substrate. In the same allotted time limit, the optimal thresholds heuristic fully generated almost 50% of the networks. Across 1,024 substrate-product pairs, only the substructure matching, top-$N$ specific, and top-$N$ agnostic variants can produce above 70% fully generated networks within the 30 minutes allotted time (Figure 2.5, top panel).

In contrast, while the naive and the combination variants are the least restricted in terms of the metabolite structures generation, they are only able to produce at most 20% fully generated networks within 30 minutes. A similar trend is found when comparing the run time distributions (Figure 2.5, bottom panel). The substructure matching, top-$N$ specific, and top-$N$ agnostic variants, on average, take less than 10 minutes to generate a network. The time distributions for the naive and the combination variants are much broader, and many runs hit the 30 minute time-out before producing a fully generated network. For applications where run time cost needs to be conserved, the substructure matching, top-$N$ specific, and top-$N$ agnostic XenoNet variants may be the most optimal. Nevertheless, a fast method is not useful if it cannot correctly identify known intermediate metabolites. We hypothesize that the combination methods traverse more of the metabolite space for each partially generated network than the naive method. We test this hypothesis in the following sections using path recall, intermediate metabolite recall, and intermediate metabolite recovery metrics.

**Path Recall**

Path recall is a metric designed to measure how well a model captures annotated paths (Figure 2.6). Given an input substrate-product pair, a model would output a graph with paths that lead from the substrate to the product. Ideally, these predicted paths would be the same as
the annotated paths that were derived from literature. However, depending on the various experimental details, annotated paths may miss some intermediates, especially short-lived and reactive metabolites like epoxides or quinones. In contrast, predicted paths consistently include even short-lived metabolites because they are built with fixed and comprehensive reaction-rule sets. Consequently, an annotated path from the substrate to the target product is considered as being captured if 1) all of its nodes are contained in a predicted path, and 2) the order of traversal through these nodes are the same in the annotated and predicted path. The paths vary from a length of one, i.e., a direct path linking the substrate to the product, up to the depth limit specified during the generation of the predicted network. For each pair of substrate and product molecules, the proportions of annotated paths of a certain length that were captured are calculated. For each length classification, the path recall of a test set is the average of the captured proportions at that specific path length across the entire test set.

The length-specific path recalls were evaluated over all metabolic networks in the test set (Figure 2.7). All 1,024 networks contain annotated paths with a length of 1, but only 60 and 24 of these networks contain annotated paths with a length of 2 and 3, respectively. For all methods, the path recall decreases as the path length increases. This is expected since the longer a path gets, the probability that a model generates all of its intermediates and places them in the correct order diminishes.

No single method has the highest length-specific path recall across all three path lengths. While heuristic approaches capture shorter-length annotated paths better, the naive model is superior in capturing longer-length annotated paths. For example, at a path length of 1, both combination approaches perform best with a path recall of 0.89. For a path length of 2, the top-$N$ reaction specific combination approach performs best with a path recall of 0.44. For a path length of 3, the naive method performs best with a path recall of 0.26. The performance trend highlights the trade-off between employing heuristics to speed up network generation at the cost of constraining the set of possible child metabolites. The constraining factor grows exponentially as path length grows. Though the naive method generally takes the longest time
Figure 2.6: Path recall calculation. Path recall is a metric designed to measure how well a model captures annotated paths. Here, we show how path recall is calculated for a hypothetical data set of five substrate-product pairs. For each substrate-product pair, a predicted graph is generated within a depth limit of 3. An annotated path from the substrate to the target product is considered as being captured if 1) all of its nodes are contained in a predicted path, and 2) the order of traversal through these nodes are the same in the annotated and predicted path. For each pair of substrate and product molecules, the proportions of annotated paths of a certain length that were captured are calculated. For each length classification, the path recall of a test set is the average of the captured proportions at that specific path length across the entire test set.
Chapter 2. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

**Figure 2.7:** Path-recall performance is dependent on path length. While heuristic approaches capture shorter-length annotated paths better, the naive model is superior in capturing longer-length annotated paths. The performance trend highlights the trade-off between employing heuristics to speed up network generation at the cost of constraining the set of possible child metabolites. The constraining factor grows exponentially as path length grows.

To iterate over all possible child metabolites branching out from a given parent metabolite, it has the highest expressivity – the greater the variety and quantity of metabolites it can infer. If network generation were allowed to run without a time limit, the naive method would be expected to have the best performance. In contrast, each heuristic makes certain assumptions about the importance of child metabolites to eventuating into the product metabolite. As the path length increases, the constraints imposed by the heuristic compound over each set of child metabolites produced at each step in the path. Eventually, the heuristic’s trade-off worsens performance relative to the naive method.

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

<table>
<thead>
<tr>
<th>Annotated Graphs</th>
<th>Predicted Graphs</th>
<th>Annotated Intermediates</th>
<th>Predicted Intermediates</th>
<th>Captured Intermediates</th>
<th>Depth</th>
<th>Intermediate Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>S → A → P</td>
<td>S → A → P</td>
<td>Depth 2 to Reach A</td>
<td>Depth 2 to Reach A</td>
<td>A</td>
<td>2</td>
<td>( \frac{1}{2} = 1.0 )</td>
</tr>
<tr>
<td>S → A → B</td>
<td>S → A → B</td>
<td>∅</td>
<td>∅</td>
<td>∅</td>
<td>2</td>
<td>( \frac{1}{2} = 0.5 )</td>
</tr>
<tr>
<td>S → A → B → C → D → E → F</td>
<td>S → A → B → C → D → E → F</td>
<td>Depth 2 to Reach A</td>
<td>Depth 3 to Reach B</td>
<td>B</td>
<td>2</td>
<td>( \frac{1}{4} = 0.25 )</td>
</tr>
<tr>
<td>A → B → C → D → E → F</td>
<td>A → B → C → D → E → F</td>
<td>C → B</td>
<td>E</td>
<td>B</td>
<td>2</td>
<td>( \frac{3}{6} = 0.5 )</td>
</tr>
<tr>
<td>Empty Network</td>
<td>∅</td>
<td>∅</td>
<td>∅</td>
<td>∅</td>
<td>2</td>
<td>( \frac{1+2+3+25}{72} = 0.58 )</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>( \frac{2+5}{10} = 0.5 )</td>
</tr>
</tbody>
</table>

FIGURE 2.8: Intermediate metabolite recall calculation. Intermediate recall is a metric designed to measure how well the model can infer experimentally observed intermediates–compounds on the paths from the substrate to the target product. Here, we show how intermediate recall is calculated for a hypothetical data set of five substrate-product pairs. For each substrate-product pair, a predicted graph is generated within a depth limit of 3. For each input substrate-product pair, the proportions of experimentally observed intermediates that can be inferred in the predicted graph of certain minimal depth are calculated. For each minimal depth classification, an intermediate recall of a test set is the average of these depth limit-specific proportions across the entire data set.

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

Intermediate recall is a metric designed to measure how well the model can infer experimentally observed intermediates–compounds on the paths from the substrate to the target product (Figure 2.8). Intermediate recall is depth-limit dependent. To infer a certain annotated intermediate, a predicted network needs a minimal depth limit. As an example, for an intermediate with
Chapter 2. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

a specified minimum depth of 2, any method for predicting a network that potentially contains a path that both infer the intermediate and terminate at the target product would need to have the depth limit set to 2 at minimum. Thus, for each input substrate-product pair, the proportions of experimentally observed intermediates that can be inferred in the predicted graph of certain minimal depth are calculated. For each minimal depth classification, an intermediate recall of a test set is the average of these depth limit-specific proportions across the entire data set.

![Graph](image.png)

**Figure 2.9:** The combined substructure matching and top-\(N\) specific heuristic model has the highest intermediate metabolite recall across all depths.

Among the randomly sampled 1,024 test set substrate-product pairs, there are 60, 82, and 97 annotated networks that contain intermediates that require XenoNet’s depth limit to be set to, at minimum, 2, 3, and 4 or more, respectively. There are 516, 756, and 1164 metabolites in the 60, 82, and 97 annotated networks, respectively. The depth-specific intermediate recalls were computed over these 60, 82, and 97 annotated networks (Figure 2.9).

For all methods, the intermediate metabolite recall decreases as the minimum depth required to reach the intermediate metabolite increases. This phenomenon is consistent with the fact that while the chemical diversity of annotated graphs would vary greatly depending on experimental conditions, predicted graphs strictly follow a set of transformation rules. Because of such
Chapter 2. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

difference in granularity, the discrepancy between the annotated graphs and predicted graphs grows exponentially with the depth of the networks.

The combined substructure matching and top-$N$ specific heuristic model has the highest intermediate metabolite recall across all depths. For this variant, the intermediates metabolite recalls are 0.49, 0.39, and 0.27 for networks of depth 2, 3, and 4 or more, respectively. The naive model, in contrast, only achieves intermediate metabolite recalls of 0.47, 0.36, and 0.25 for networks of depth 2, 3, and 4 or more, respectively. Other models have even lower intermediate metabolite recalls (Figure 2.9). The result is further evidence that the combined substructure matching and top-$N$ specific heuristic mode is the optimal XenoNet variant that can decrease its time cost for inferring metabolites without hampering its ability to infer known metabolites.

Intermediate Metabolite Recovery

![Intermediate Metabolite Recovery Diagram]

**Figure 2.10:** Recovery rate calculation. Here we show how recovery rate is calculated for a set of three hypothetical metabolic networks. Left) Predicted metabolic networks with a depth limit of one of three different substrates, $S_1$, $S_2$, and $S_3$, and their known metabolites are shown. Metabolites $M$ that are both predicted and experimentally observed are in darker grey. Metabolites $M$ that are predicted but not experimentally observed or vice versa are in lighter grey. Right) The proportion of networks that have fractions of known metabolites predicted by the model above a certain threshold is computed. In our hypothetical database of three metabolic networks, 100% of substrates have at least 0 – 20% of their known metabolites predicted, and 66.7% of the substrates ($S_1$ and $S_2$) have at least 60% of their known metabolites predicted. Only 33.3% of the substrates ($S_1$) have at least 70% of their known metabolites predicted.
To measure the ability of a model to capture known metabolites, we use the recovery rate metric. Recovery rate is the proportion of networks that have fractions of known metabolites predicted by the model above a threshold (Figure 2.10).

We assess the intermediate metabolite recovery rate of all seven XenoNet variants on the randomly sampled 1,024 network subset (Table 2.2). The combined substructure matching and top-$N$ reaction specific model is the best model across all thresholds. The naive model is also competitive at thresholds ranging from 0.6 to 1. Both of these methods can infer all intermediate metabolites in 44% of the networks.

<table>
<thead>
<tr>
<th>Method</th>
<th>0.1</th>
<th>0.2</th>
<th>0.3</th>
<th>0.4</th>
<th>0.5</th>
<th>0.6</th>
<th>0.7</th>
<th>0.8</th>
<th>0.9</th>
<th>1.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
<td>0.54</td>
<td>0.54</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
<td>0.44</td>
</tr>
<tr>
<td>Substructure Matching, top-$N$ Reaction Specific</td>
<td>0.62</td>
<td>0.62</td>
<td>0.62</td>
<td>0.59</td>
<td>0.59</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
<td>0.49</td>
<td>0.44</td>
</tr>
<tr>
<td>Substructure Matching, top-$N$ Reaction Agnostic</td>
<td>0.60</td>
<td>0.60</td>
<td>0.60</td>
<td>0.57</td>
<td>0.57</td>
<td>0.46</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td>0.40</td>
</tr>
<tr>
<td>Substructure Matching</td>
<td>0.57</td>
<td>0.55</td>
<td>0.55</td>
<td>0.51</td>
<td>0.51</td>
<td>0.41</td>
<td>0.40</td>
<td>0.40</td>
<td>0.29</td>
<td>0.27</td>
</tr>
<tr>
<td>Top-$N$ Reaction Specific</td>
<td>0.57</td>
<td>0.57</td>
<td>0.57</td>
<td>0.56</td>
<td>0.46</td>
<td>0.41</td>
<td>0.38</td>
<td>0.35</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>Top-$N$ Reaction Agnostic</td>
<td>0.52</td>
<td>0.51</td>
<td>0.49</td>
<td>0.44</td>
<td>0.41</td>
<td>0.35</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>Optimal Thresholds</td>
<td>0.58</td>
<td>0.50</td>
<td>0.44</td>
<td>0.44</td>
<td>0.44</td>
<td>0.40</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
<td>0.37</td>
</tr>
</tbody>
</table>

Table 2.2: Intermediate Metabolite Recovery Rate of XenoNet Variants

The combined substructure matching and top-$N$ reaction specific variant is the best model across thresholds. The naive model is also competitive at thresholds ranging from 0.6 to 1. The highest values in each threshold are bolded.

### 2.3.2 Performance on the Full Metabolic Network Data Set

Comparisons between variants of metabolic network construction were drawn on a small fraction of our metabolic network data set. From these comparisons, we selected the best performing variant, the combined substructure matching and top-$N$ reaction-type specific heuristic model, to run on our entire data set of 17,054 metabolic networks. Networks were generated with a depth limit of 3 and a max time limit of 30 minutes per network.
Chapter 2. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

Figure 2.11: XenoNet accurately predicts metabolic networks. Performance of XenoNet across the full metabolic network data set on the same metrics assessed during the comparison between all heuristics on the subset of 1,024 networks. (A) Path recall stratified across paths of length 1, 2, and 3. (B) Intermediate metabolite recall stratified across minimum network depth limit required to reach the intermediate metabolite of 2, 3, and 4 or more. (C) Intermediate metabolite recovery across recovery rate thresholds of 0.1 to 1. The results for all three plots are close to the initial results for each metric when validated on the subset of 1,024 samples and supports that the comparison between the XenoNet variants generalizes to performance on the full data set.
Network Coverage, Path Recall, Metabolite Recall, and Metabolite Recovery

In terms of coverage, 2,804 of the networks reached completion within the 30 minutes allotted for network generation. The partial network states reached by the remaining 14,250 networks were also preserved for analysis. In total, 14,882 of the networks found at least one path between the substrate-product pair. 1,490 networks timed out at the 30 minute limit before finding any valid paths between the substrate and the target product. The remaining 673 of the networks reached completion within the time limit but did not find any valid paths. In total, the total time required to generate networks for the full metabolic network data set was approximately 468,000 minutes.

We also evaluate the final XenoNet’s ability to infer known pathways and intermediates between a given substrate-product pair of molecules on the 17,054 pairs using path recall, intermediate metabolite recall, and intermediate metabolite recovery metrics (Figure 2.11).

First, we assess path recall. In our data set, all 17,054 networks contain annotated paths with a length of one, but only 817 and 240 of these networks contain annotated paths with a length of two and three, respectively. The path recall for path lengths of one, two, and three steps is 0.88, 0.46, and 0.26 across 17,054, 817, and 240 networks, respectively (Figure 2.11A). Examples of paths that were successfully elucidated are highlighted in Figure 2.12, along with examples for which a valid path was not found. XenoNet’s path recall for a path length of 1 is comparable to the accuracy of Metabolic Forest (88.4-88.8%). [60] This result is expected because, despite aiming at different tasks, the two models use the same ruleset and their data was built using the same set of AMD reactions. XenoNet’s path recall at a path length of 1 is only slightly lower than the path recall assessed for Metabolic Forest, which can be attributed to the heuristics limiting the extent of the chemical transformation space.

Second, we assess intermediate recall. In our data set, there are 817, 1,041, and 1,143 annotated networks that contain intermediates which require XenoNet’s depth limit to be set to, at minimum, 2, 3, and 4 or more. XenoNet’s depth-specific intermediate metabolite recalls at depth
Figure 2.12: Examples of paths that XenoNet succeeded in finding and failed in finding for paths of length 1, 2, and 3. The example paths for each case were randomly sampled from the total set of paths. XenoNet has an easier time of finding paths with distinct, explicit steps.
limits of 2, 3, and 4 or more are 0.46, 0.38, and 0.30, respectively (Figure 2.11B). Unsurprisingly, there is a drastic drop-off in intermediate metabolite recall when going to metabolites that would require a minimum depth limit of greater than 3 to reach. Since XenoNet’s depth-limit parameter was set to 3, we expect low recall for intermediates which, based on their annotated graph, would require a minimal depth limit of 3 to locate.

Recall that the path recall for paths of length one, two, and three steps over the subset of 1,024 networks was 0.89, 0.44, and 0.18 using the same heuristic. Furthermore, the intermediate recall for minimum network depths of two, three, and four or more metabolic steps over the subset of 1,024 networks were 0.49, 0.39, and 0.27. It is reassuring that the recall metrics computed over the subset of 1,024 networks translated to recall evaluation over the full data set. Most of the recall metrics are similar in value, with more notable increases in performance on the full data set for depth 3 path recall and depth 2 intermediate metabolite recall. Regardless, the results on the whole metabolic network data set are broadly consistent with the results across the subset of data used to compare various heuristics in the previous section.

Last but not least, we assess XenoNet’s performance in regard to intermediate metabolite recovery on the full data set. The recovery rate only drops $\sim 13\%$ from $\sim 75\%$ at 0.1 threshold to $\sim 62\%$ at 1.0 threshold (Figure 2.11C). Inversely, 25% of networks (2,163) found 0 of the known intermediate metabolites. About half of these 25% networks are a result of networks that could not find any valid paths.

**Path and Metabolite Rankings**

The ability to enumerate multiple metabolites and metabolic pathways alone is useful, but not enough for triage purposes. Knowing the likelihood of each metabolite or pathway would help differentiate between true positives and false positives. Ideally, an excellent metabolic network generator would always be able to assign experimentally observed pathways or metabolites with higher likelihood than not-observed pathways or metabolites. The more frequently the model assigns the highest probability to pathways/metabolites that have been observed, the
Chapter 2. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

<table>
<thead>
<tr>
<th>Annotated Graphs</th>
<th>Predicted Graphs</th>
<th>Annotated Paths</th>
<th>Predicted Paths &amp; Likelihood</th>
<th>Most Likely Predicted Path</th>
<th>Is Most Likely Predicted Path in Annotated Set?</th>
</tr>
</thead>
<tbody>
<tr>
<td>![Path A]</td>
<td>![Path B]</td>
<td>S → P</td>
<td>S → A → P: -0.602</td>
<td>S → A → P</td>
<td>Yes</td>
</tr>
<tr>
<td>![Path C]</td>
<td>![Path D]</td>
<td>S → P</td>
<td>S → A → P: -2.00</td>
<td>S → A → P</td>
<td>No</td>
</tr>
<tr>
<td>![Path E]</td>
<td>![Path F]</td>
<td>S → P</td>
<td>S → C → B → P: -0.0969</td>
<td>S → P</td>
<td>Yes</td>
</tr>
<tr>
<td>![Path G]</td>
<td>![Path H]</td>
<td>S → P</td>
<td>S → B → P: -2.30</td>
<td>S → B → P</td>
<td>No</td>
</tr>
</tbody>
</table>

Fraction of Highest Likelihood Paths in Annotated Path Set

\[
\frac{1 + 0 + 1 + 0}{4} = 0.5
\]

**Figure 2.13:** Path ranking calculation. This metric is designed to measure how well the predicted path with the highest likelihood of traversal corresponds to a known path in the annotated graph. Here, we show how path ranking is calculated for a hypothetical data set of five substrate-product pairs. The likelihood of each predicted pathway is the logarithmic sum over the scores that Rainbow XenoSite assigns to the metabolic transformations in that pathway. We then rank all predicted pathways in a given network by their likelihoods. If the highest-likelihood path of a given network has an exact match in the annotated path set, then that network is assigned a score of 1. Otherwise, that network is assigned a score of 0.
more confident users would be that, when applied to new data, the highest likelihood pathway and metabolite would be designated to a pathway and metabolite with a high chance of existing. In this section, we explore how well XenoNet model would predict experimentally observed metabolites and pathways with higher likelihood than not-observed metabolites and pathways.

The path ranking metric is designed to measure how well the predicted path with the highest likelihood corresponds to an annotated path (Figure 2.13). Here, the likelihood of each predicted pathway is the logarithmic sum over the scores that Rainbow XenoSite assigns to the metabolic transformations in that pathway. We then rank all predicted pathways in a given network by their likelihoods. If the highest-likelihood path of a given network has an exact match in the annotated path set, then that network is assigned a score of 1. Otherwise, that network is assigned a score of 0. Path ranking of a set is the average score across all networks in the set (Figure 2.13). Note that, unlike in the assessment of the path recall metric, the predicted path with the highest likelihood is assessed for an exact match—containing the same set of metabolites in identical order—in the annotated path set. This is because every annotated network contains a one-step path that directly links the substrate to the product, and every captured path is an approximate match to this one-step path. Allowing approximate matches would inflate the path ranking.

Top-one, -two, and -three metabolite ranking metrics are designed to measure how well the predicted metabolite with the highest metabolite scores (Figure 2.4) corresponds to an observed metabolite. Top-$N$ accuracy of a set is the fraction of predicted networks that have experimentally observed metabolites among the $N$ predicted metabolites with the highest metabolite scores. Note that an obstacle can arise in computing metabolic scores for networks that contain a cycle. Though an individual enumerated path is not allowed to have a cycle, a cycle can still arise in the global network structure. Consider the case of a network with the following two paths: $S \rightarrow M_1 \rightarrow M_2 \rightarrow P$ and $S \rightarrow M_2 \rightarrow M_1 \rightarrow P$. The method for computing metabolite scores, as initially described, would result in an infinite loop due to the one-step cycle between $M_1$ and $M_2$. Since networks with a cycle show up in under 1% of the networks predicted during our experiments, they are ignored when computing metabolite scores for the following results.
Chapter 2. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

over the full metabolic network data set.

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

Among the 817 metabolic networks with at least one intermediate in our data set, XenoNet was able to complete 710 networks. The path ranking, top-\(N\), and average AUC metrics were calculate over these 710 networks (Table 2.3). Overall, XenoNet was able to accurately predict experimentally observed pathways with 93.6\% path-rank accuracy. The model also accurately predicted experimentally observed metabolites with 51.9\%, 66.2\%, 77.5\%, and 78.7\% top-one, -two, -three, and average AUC accuracies, respectively.

To assess the value of scoring pathways and metabolites using predictions from the Rainbow XenoSite model, we compare XenoNet’s performance on path ranking, top-\(N\), and average AUC metrics to a model where these predictions are permuted (Table 2.3). Specifically, for each of the predicted networks, while the nodes and edges are kept in the original order, the scores assigned by Rainbow XenoSite to the edges are randomly permuted. The values computed for the ’randomly permuted weights’ case are computed over 10 trials. Overall, the randomly permuted model predicts experimentally observed pathways with 61.8\% path-rank accuracy. The randomly permuted model also predicts experimentally observed metabolites with 31.9\%, 46.4\%, 60.7\%, and 62.5\% Top-one, -two, -three, and average AUC accuracies, respectively.

We use two different statistical tests to assess the statistical significance of the difference in the performance of XenoNet and the randomly permuted model. For path ranking and top-\(N\) metrics, we use McNemar’s test. McNemar’s test is a paired nonparametric statistical hypothesis test for evaluating the disagreements between two cases [63]. In the context of classification, McNemar’s test can be used to interpret whether both models make different errors and the difference in the relative proportions of those errors. The two cases are the unmodified XenoNet and the randomly permuted XenoNet, and a contingency table is constructed using both cases.
based on the numbers of highest-ranked pathways or metabolites that are experimentally observed versus those that are not. The null hypothesis of marginal homogeneity would mean that there is no effect in regards to where the edge weights are shuffled in the network, i.e., the two cases should have the same error rate [64]. McNemar’s test yielded a \( p \)-value less than 0.001 for both path-ranking and Top-\( N \) metrics, so there is strong evidence to reject the null hypothesis.

For average AUC metric, we use a paired \( t \)-test. Comparing the performance between the two cases via a paired \( t \)-test yielded a \( p \)-value less than 0.001. In summary, XenoNet performs better than a randomly permuted model across all considered metrics.

<table>
<thead>
<tr>
<th>Highest Likelihood Path in Observed Paths</th>
<th>Top-One Intermediate Metabolite</th>
<th>Top-Two Intermediate Metabolite</th>
<th>Top-Three Intermediate Metabolite</th>
<th>Average AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unmodified Weights</td>
<td>93.6%</td>
<td>51.9%</td>
<td>66.2%</td>
<td>78.7%</td>
</tr>
<tr>
<td>Randomly Permuted Weights</td>
<td>61.8%</td>
<td>31.9%</td>
<td>46.4%</td>
<td>62.5%</td>
</tr>
</tbody>
</table>

**Table 2.3: Comparison between XenoNet and a random model**

### 2.3.3 Comparison to Prior Work

No published model does the exact main task that XenoNet is designed for: to take a pair of start-target molecules as inputs and output a network of metabolic pathways between them. The closest comparable works in the literature include GLORY, BioTransformer, and SyGMA.[58, 55, 57] However, the main task of these methods is to receive a molecule as input and output computationally predicted metabolites. Though these prior works do not readily support XenoNet’s main task of interest, XenoNet does support the main task of these prior works and so we can compare XenoNet to them in some capacity.

To enable this comparison, we first ran the naive XenoNet on GLORY’s reference set of 848 parent molecules with 1,588 known metabolites.[58] This reference set was used by GLORY’s MaxEfficiency mode to decide upon a SOM probability cutoff that could be used as a preliminary filter and by GLORY’s MaxCoverage mode to develop a priority score for ranking predicted metabolites. After XenoNet produced networks for each of the 848 parent molecules
in the reference set, precision-recall curves for each of the 20 phase I reaction types were used to define an optimal, reaction-type-specific threshold for filtering predicted metabolites. For each rule, the filtering threshold was set to the lowest threshold that did not reduce the original recall across the reference set’s known metabolites. Similar to how the thresholds were used by GLORY, they are used to filter predicted metabolites for a test substrate after its metabolic network is generated by XenoNet.

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

Next, we tested the naive XenoNet variant on the curated test set of 29 substrates and 81 products that was used in GLORY[58] to compare against BioTransformer[55] and SyGMa[57]. None of the 29 substrates in this test set were contained within the GLORY reference set. In this test, XenoNet was set to generate metabolic networks using only the substrates without any secondary target molecules to enforce paths to terminate at. We allotted a 5-minute time-limit for XenoNet to generate each metabolic network and set the network’s depth limit to 1. The generation of networks included both the inference of metabolite structures as well as assigning edge weights via SOM predictions generated by Rainbow XenoSite. The final list of predicted metabolites is filtered using the thresholds described in the previous paragraph. Production of
the metabolites from the 29 test set substrates by XenoNet was computed on a single Intel Xeon Processor E5-2630 v3 CPU using a Linux operating system. The total run time using a single core was 80.5 minutes, and the average run time per parent molecule was 2.78 minutes.

<table>
<thead>
<tr>
<th></th>
<th>GLORY, MaxCoverage Mode</th>
<th>GLORY, MaxEfficiency Mode</th>
<th>SyGMa</th>
<th>BioTransformer</th>
<th>XenoNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Precision</td>
<td>0.08</td>
<td>0.16</td>
<td>0.15</td>
<td>0.17</td>
<td>0.06</td>
</tr>
<tr>
<td>Recall</td>
<td>0.83</td>
<td>0.64</td>
<td>0.74</td>
<td>0.72</td>
<td>0.89</td>
</tr>
<tr>
<td>Total Number of Predicted Metabolites</td>
<td>793</td>
<td>327</td>
<td>406</td>
<td>344</td>
<td>1179</td>
</tr>
<tr>
<td>Number of successfully predicted reported metabolites</td>
<td>67</td>
<td>52</td>
<td>60</td>
<td>58</td>
<td>72</td>
</tr>
<tr>
<td>AUC</td>
<td>67.6%</td>
<td>–</td>
<td>50.1%</td>
<td>–</td>
<td>73.3%</td>
</tr>
<tr>
<td>Top-1</td>
<td>68.97%</td>
<td>68.97%</td>
<td>0%</td>
<td>–</td>
<td>72.41%</td>
</tr>
<tr>
<td>Top-2</td>
<td>72.41%</td>
<td>72.41%</td>
<td>48.28%</td>
<td>–</td>
<td>75.86%</td>
</tr>
<tr>
<td>Top-3</td>
<td>75.86%</td>
<td>75.86%</td>
<td>68.97%</td>
<td>–</td>
<td>79.31%</td>
</tr>
</tbody>
</table>

Table 2.4: XenoNet outperforms published methods on the GLORY test set. For each metric, the best method’s value is displayed in bold. Values for both GLORY variants, SyGMa, and BioTransformer were extracted from the initial comparisons made in Bruyn Kops, et al. [58]

XenoNet outperformed both GLORY variants, SyGMa, and BioTransformer at multiple metrics (Table 2.4). A metabolite is considered to be a true positive when it is both predicted and experimentally observed. Precision is the proportion of true positives among all predicted metabolites, and recall, also known as sensitivity, is the proportion of true positives among all experimentally observed metabolites. The precision values on the test set for XenoNet, GLORY MaxCoverage, GLORY MaxEfficiency, SyGMa, and BioTransformer are 0.06, 0.08, 0.16, 0.15, and 0.17, respectively. The recall values on the same test set for XenoNet, GLORY MaxCoverage, GLORY MaxEfficiency, SyGMa, and BioTransformer are 0.89, 0.83, 0.64, 0.74, and 0.72, respectively. It should be noted that while XenoNet has lower precision than other models, it has higher recall. The trade-off between precision and recall is dependent on the chosen threshold, so either metric is not reliable to compare between models. The receiver-operating curves (ROC) and their corresponding areas under the curve (AUC) are more reliable metrics than precision and recall. The AUC of XenoNet, GLORY MaxCoverage, and SyGMa

45
are 73.3%, 67.6%, 50.1%, respectively (Table 2.4). We were unable to construct a ROC curve and calculate the AUC for BioTransformer based on their publication, but their recall of 0.72 is lower than other models if all of them were set to have the same precision of 0.17 (Figure 2.14).

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

**Figure 2.14:** XenoNet is superior to published metabolite prediction models. The receiver-operating characteristic curves (ROC) and their corresponding areas under the curve (AUC) are reliable metrics to compare between models. The AUC of XenoNet, GLORY MaxCoverag, and SyGMa are 73.3%, 67.6%, 50.1%, respectively. We were unable to construct a ROC curve and calculate the AUC for BioTransformer based on their publication, but their recall of 0.72 is lower than other models if all of them were set to have the same precision of 0.17.

Moreover, XenoNet predicts many more metabolite structures than other models because it has a broader chemical transformation rule set. XenoNet’s rule set was derived from Rainbow XenoSite’s database, which covered 92.3% of phase I reactions. In contrast, the most extensive
rule sets by previous models, GLORY, was based on FAME 2, which covered only 48.0% of phase I reactions [54]. While this feature lowers XenoNet’s precision, it could also discover metabolites that were missed by experimental methods. Depending on the experimental assay and conditions used, certain metabolites are not easily detected. Computational prediction could serve as a guide for future experiments.[41, 43]

In terms of recovery rate, XenoNet outperformed GLORY MaxCoverage Mode at all thresholds (Figure 2.15). GLORY MaxCoverage Mode was chosen for comparison because it previously out-performed BioTransformer, SyGMA, and GLORY MaxEfficiency Mode with regard to the recovery rate metric[58]. XenoNet can predict at least 50% of the known metabolites for 96.6% of the parent molecules in the test data set, while GLORY MaxCoverage did so for 90% of the parent molecules. The proportion of parent molecules that have all of their known metabolites predicted is 72.4% for XenoNet and 62.0% for GLORY MaxCoverage.

![Figure 2.15: XenoNet is superior to GLORY MaxCoverage in terms of recovery rate. Across all thresholds, XenoNet’s recovery rate is higher than GLORY’s.](image)

In terms of absolute numbers, XenoNet captured 72 of the 81 known metabolites in the test set. The method with the next highest number of captured metabolites, GLORY MaxCoverage, yields 67 of the known metabolites. There were 9 metabolites of 8 parent molecules that
Chapter 2. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

XenoNet was not able to predict when the models’ depth limit was 1. These 9 metabolites were also missed by published models. However, XenoNet is designed to infer metabolic structures from preceding known, potentially reactive, metabolites. We wanted to test whether XenoNet’s ability to specify a given target metabolite in addition to a start metabolite could allow for the detection of pathways between each of the parent molecules and their missed metabolite(s).

XenoNet was run at a depth limit of 3 on 9 substrate-product pairs of molecules representing the parent molecule and one of their missed metabolites. For 5 of the 9 networks, paths were found linking the parent molecule and their previously missed metabolite. The 4 remaining pairs of parent molecules and their missed metabolites that no method found a valid path for are included in Figure 2.16. On manual inspection, most of the missed cases require metabolic transformations that are phase II reaction types or rare phase I reaction types that our Rainbow XenoSite model does not yet account for. We plan to expand our rule set to cover these reactions in future work.

2.4 Model Limitations & Future Directions

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

Besides, the current heuristics could be modified further to resolve weaknesses highlighted by the previous comparisons raised between the XenoNet variants. With regards to the top-$N$
Figure 2.16: XenoNet, GLORY, SyGMA, and BioTransformer are not able to infer a pathway that links the substrate and metabolite for the 4 cases shown.
Chapter 2. XenoNet: Inference and Likelihood of Intermediate Metabolite Formation

heuristics, the relationship between reaction-type specific and reaction-type agnostic variants can be explored further. Specifically, we need to find the best way to tune the value of $N$ on large data sets of substrate-product pairs outside of the AMD data set.

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

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

An additional mode allowing for multiple targets to be specified in the input could also be implemented to improve the functionality of XenoNet. Furthermore, XenoNet infers metabolites and assigns predictions to the metabolic transformations that preclude them, but it doesn’t capture the full bioactivation process. Previously, we have developed a SOM model for predicting reactivity of a molecule with respect to DNA, protein, GSH, and cyanide [65]. A natural next step is to incorporate this reactivity model into XenoNet as a way to predict whether an inferred metabolite is likely to be reactive.

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

2.5 Conclusion

We have established a method, XenoNet, that combines a SOM model with a structure inference model for the enumeration of metabolic pathways between a known parent molecule and target molecule and the intermediate metabolite structures that link them. XenoNet can predict experimentally observed pathways and intermediate metabolites with high accuracies. Our method can also function in a similar capacity to prior methods, such as BioTransformer and GLORY, when only given a parent molecule as input. When given the task, XenoNet outperforms prior methods across multiple metrics. While we have yet to model the full bioactivation pathway potential between two molecules or starting from a single molecule, we anticipate the successful incorporation of reactivity models to XenoNet’s workflow in the near future. Incorporation of reactivity into XenoNet is a natural extension of the current work and will hopefully cement XenoNet as an informative tool that experimentalists can use to generate specific, testable hypotheses for understanding reactive metabolite formation. Importantly, if it helps experimentalists discover otherwise unknown intermediates, they can then use that knowledge to modify drug molecules to prevent the formation of the intermediate(s) that are antecedent to reactive metabolite formation.
Chapter 3

Message Passing Neural Networks Improve Prediction of Metabolite Importance


3.1 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. [28] Metabolism is generally beneficial, making drugs more hydrophilic and easier to excrete. However, metabolism can transform drugs to pernicious reactive metabolites that may conjugate to DNA or off-target proteins and result in adverse events. [9, 26] 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. [66, 32, 67] 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 [68] and covalent binding studies [69, 70], 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 data set of manually labeled metabolic networks.

We previously developed XenoNet, which receives an input substrate molecule and optional target metabolite(s) and generates a metabolic network. [71] 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. [57, 55] 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 3.1A). We previously proposed the XenoNet algorithm [71] to output a “metabolite formation score” for each metabolite as a proxy for its importance. 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. XenoNet
does not add individual cycle-containing paths, but this doesn’t prohibit formation of cycles in
the network (Figure 3.1B), 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 3.1C). 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. [54] 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 perfor-
mane 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 impor-
tance. 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 embed-
dings represent an optimal normalization strategy and allow comparison of metabolites within
an individual network and across multiple networks. Finally, the learned model is robust to
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

Figure 3.1: 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 designates intermediate metabolites of greatest importance and extracts connected subnetworks for focused review.
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 3.1D).

3.2 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.

3.2.1 Method Comparison

An underlying cause of adverse events is production of reactive metabolites. Knowledge of likely intermediate metabolites linking the ingested drug and known structural end points may advise rational drug modifications that preclude formation of undesired, reactive intermediates and divert transformations down benign metabolic pathways. 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. 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 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 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 3.2). 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.
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

Figure 3.2: The 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. 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.
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 3.3). For each method, the predictions of each intermediate metabolite are distributed into ten 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. Though 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 data set and its LOGOCV predictions are used for inference tasks on training set networks. The optimal hyperparameters used during training are bolded in Table 3.5. The optimal aggregation operator summed neighborhood messages, which is unsurprising as max and mean pooling do not enable networks that are as discriminative as the 1-dimensional Weisfeiler-Leman test and are theoretically less powerful. [72] 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- and out-degree node features. Moreover, we verified that the bidirectional model is
FIGURE 3.3: The bidirectional MPNN produced the best-scaled metabolite predictions. The bar graphs plot the distributions for each method across 6,606 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.
TABLE 3.1: The bidirectional MPNN architecture outperformed all other methods on multiple accuracy and calibration metrics. Though PageRank had a higher $R^2$ of 0.94, its reliability diagram displays a degenerate case where the model’s confidence is unstable. For each metric, the best performing is bolded and the second-best method is italicized. Method groupings designate new methods (top) and prior methods (bottom).

<table>
<thead>
<tr>
<th>Method Group</th>
<th>Global ROC AUC</th>
<th>Average ROC AUC</th>
<th>Global PR AUC</th>
<th>Top Two Accuracy</th>
<th>Reliability RMSE</th>
<th>Reliability R2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bidirectional MPNN Variant</td>
<td>88.5%</td>
<td>87.4%</td>
<td>53.8%</td>
<td>73.3%</td>
<td>0.10</td>
<td>0.92</td>
</tr>
<tr>
<td>Unidirectional MPNN Variant</td>
<td>80.1%</td>
<td>78.4%</td>
<td>36.2%</td>
<td>59.2%</td>
<td>0.11</td>
<td>0.87</td>
</tr>
<tr>
<td>XenoNet 1.0 Algorithm</td>
<td>75.0%</td>
<td>80.0%</td>
<td>29.0%</td>
<td>63.0%</td>
<td>0.25</td>
<td>0.68</td>
</tr>
<tr>
<td>Betweenness Centrality</td>
<td>74.2%</td>
<td>68.4%</td>
<td>25.5%</td>
<td>46.6%</td>
<td>0.32</td>
<td>0.83</td>
</tr>
<tr>
<td>PageRank</td>
<td>72.9%</td>
<td>74.4%</td>
<td>27.3%</td>
<td>33.4%</td>
<td>0.38</td>
<td>0.94</td>
</tr>
<tr>
<td>XenoNet Random Model</td>
<td>69.1%</td>
<td>71.2%</td>
<td>21.0%</td>
<td>50.0%</td>
<td>0.31</td>
<td>0.80</td>
</tr>
<tr>
<td>Random Walk with Restarts</td>
<td>58.5%</td>
<td>56.5%</td>
<td>11.2%</td>
<td>20.3%</td>
<td>0.19</td>
<td>0.04</td>
</tr>
</tbody>
</table>
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

<table>
<thead>
<tr>
<th>TABLE 3.2: The bidirectional MPNN architecture that incorporates edge features but not node degree features achieves best performance.</th>
</tr>
</thead>
<tbody>
<tr>
<td>All Features</td>
</tr>
<tr>
<td>Without Degree Node Features</td>
</tr>
<tr>
<td>Without Reaction Type Edge Feature</td>
</tr>
<tr>
<td>Without Weight Edge Feature</td>
</tr>
</tbody>
</table>

robust when validated with the LOGOCV protocol, as performance for each metric was equivalent or superior to its nested CV performance. We simulated 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 absence of in- and out-degree node features improved LOGOCV performance (Table 3.2). One explanation is that useful information contained within the in- 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.

3.2.2 Generalization of Bidirectional MPNN

We evaluated the final model’s robustness. Specifically, we ensured that the model is robust to input parameter adjustments that influence network construction, correctly detects the presence or absence of metabolites that are annotated as experimentally observed or not observed in different network contexts, and extracts important intermediates while conserving their local network structure.
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

Table 3.3: We evaluated the bidirectional variant, which was trained on networks constructed with a depth limit of 3 and no beam width constraint, for inference on networks constructed using different parameter values. The parameters for each variant are a depth limit of 2 with no beam width constraint, a depth limit of 3 and beam width of 5, and a depth limit of 2 and beam width of 5. A depth limit of 2 is selected as this is the lowest depth limit setting than can result in intermediate metabolites. A beam width of 5 is the lowest recommended beam width setting, as any lower does not result in meaningful decreases in time cost while increasing the chance that paths of interest are missed.

<table>
<thead>
<tr>
<th></th>
<th>Depth Limit 3</th>
<th>Depth Limit 3</th>
<th>Depth Limit 2</th>
<th>Depth Limit 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Networks</td>
<td>311</td>
<td>253</td>
<td>155</td>
<td>151</td>
</tr>
<tr>
<td>Number of Experimentally Observed Intermediates</td>
<td>516</td>
<td>302</td>
<td>191</td>
<td>183</td>
</tr>
<tr>
<td>Number of Experimentally Unobserved Intermediates</td>
<td>6090</td>
<td>1077</td>
<td>229</td>
<td>176</td>
</tr>
<tr>
<td>Mean Number of Nodes per Network</td>
<td>24</td>
<td>8</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Max Number of Nodes per Network</td>
<td>102</td>
<td>16</td>
<td>16</td>
<td>14</td>
</tr>
<tr>
<td>Min Number of Nodes per Network</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Robustness to Network Construction Parameters

The bidirectional MPNN’s performance was not strongly influenced by choice of network construction parameters. This is important because a model 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 3.3).

The bidirectional MPNN’s LOGOCV global and average ROC AUC performance was robust to substantial decrease in the beam width and depth limit parameters (Table 3.4). 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, multi-hop diffusion operators. [73, 74]
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

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 2 networks may result from over-smoothing, whereby node hidden states converge to an almost uniform distribution and local neighborhood information is lost. [72] The depth limit 2 networks have a smaller diameter than the depth limit 3 networks, so naively applying a rigid architecture that cannot adjust for network size hurts performance. Over-smoothing 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 has to be over-squashed into a fixed length vector, whereby important long-range information may be lost. [75]

Accurate Performance on Subset of Contextual Intermediate Metabolites

The bidirectional MPNN accurately designated importance of the same metabolite in different networks. Importantly, the model does not blindly assign importance 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 3.4). 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
TABLE 3.4: Incorporating jumping knowledge, the bidirectional model performance is invariant to changes in the beam width and depth limit parameters during network construction. 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. Greyed rows specify network construction parameters with significant drop in performance. 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. [76] Furthermore, top two accuracy adopts an optimistic bias due to the global decrease in class imbalance and the local decrease in the average number of nodes per network.

<table>
<thead>
<tr>
<th>With Layer Aggregation</th>
<th>Global ROC AUC</th>
<th>Average ROC AUC</th>
<th>Reliability RMSE</th>
<th>Reliability R²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth Limit 3</td>
<td>88.5%</td>
<td>87.6%</td>
<td>0.08</td>
<td>0.92</td>
</tr>
<tr>
<td>Depth Limit 3</td>
<td>86.5%</td>
<td>87.4%</td>
<td>0.08</td>
<td>0.95</td>
</tr>
<tr>
<td>Beam Width 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Depth Limit 2</td>
<td>84.1%</td>
<td>85.6%</td>
<td>0.15</td>
<td>0.84</td>
</tr>
<tr>
<td>Depth Limit 2</td>
<td>83.2%</td>
<td>84.5%</td>
<td>0.16</td>
<td>0.84</td>
</tr>
<tr>
<td>Beam Width 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without Layer Aggregation</td>
<td>Depth Limit 3</td>
<td>88.3%</td>
<td>86.6%</td>
<td>0.10</td>
</tr>
<tr>
<td>Depth Limit 3</td>
<td>84.3%</td>
<td>86.1%</td>
<td>0.18</td>
<td>0.78</td>
</tr>
<tr>
<td>Beam Width 5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Depth Limit 2
Beam Width 5
71.1% 68.8% 0.22 0.81
Depth Limit 2
Beam Width 5
72.1% 70.5% 0.24 0.75
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.

\[
\begin{array}{c|c|c}
\text{Annotated Class} & \text{Recorded} & \text{Not Recorded} \\
\hline
\text{Node Importance} & \leq 0.01 & 0.5 & 1 \\
\text{Reaction Class} & \text{Stable Oxygenation} & \text{Unstable Oxygenation} & \text{Dehydrogenation} & \text{Hydrolysis} & \text{Reduction} \\
\text{Edge Weight} & \leq 0.01 & 0.5 & 1 \\
\text{Predicted Class} & \text{Important} & \text{Not Important} \\
\end{array}
\]

**Figure 3.4:** The bidirectional MPNN correctly adjusts predictions for metabolites that are present or absent in different networks. Left) The intermediate metabolite DL-methamphetamine is experimentally observed in the right predicted network and is not experimentally observed in the left predicted network. The model assigns scores of 0.91 and $5.2e-2$ to DL-methamphetamine in the contexts in which it is and is not experimentally observed, respectively. DL-methamphetamine occurs as an intermediate metabolite in 3 other networks, not shown. In 2 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 $6.1e-4$. 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.
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

Extraction of Important Connected Subnetworks

The bidirectional MPNN extracted important connected subnetworks while retaining global ROC AUC performance across multiple cut-offs (Figure 3.5). Thus, the model can be flexibly used with sustained performance at several different subnetwork extraction cut-offs depending on application context, sensitivity to possible false positive intermediates, and desire to maximize or minimize retained important or unimportant intermediates, respectively.

![Figure 3.5: Optimal cut-off value depends on application context, sensitivity to possible false positives, and whether the cut-off results in a failure case. A cut-off of 1 minimized the number of unimportant intermediates to 0 and kept 722 important intermediates. The average network size was reduced from 24 ± 16 nodes to 4 ± 2 nodes. However, 421 important intermediates were excluded and 71 failure cases resulted. Alternatively, a cut-off of 0.5 was the maximum cut-off that kept all 1,143 important intermediates and did not cause a single failure case. The average network size was reduced to 15 ± 11 nodes. The error bars for the average number of nodes represents the standard deviation. Though a 0.5 cut-off results in 0 failure cases, 2,664 unimportant intermediates are retained.](image)

Across 311 networks with an average network size of 24 ± 16 nodes, 1,143 of 6,606 intermediates had scores that surpassed the metabolite importance binarization threshold. A cut-off 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 cut-off of 0.5 was the maximum cut-off that kept all 1,143 important intermediates and did not cause a single failure case. Though a 0.5 cut-off 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 cut-off failed to extract a subnetwork, we decreased it until success.

### 3.2.3 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 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 intended therapeutic effect.

**Identifying Important, Unrecorded Intermediate Metabolites**

We applied the fully-trained bidirectional model on 70 networks of withdrawn drugs, comprising 2,832 intermediate metabolites. In some cases, the model provides new speculations for mechanisms of a drug’s toxicity (Figure 3.6, 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 4 unrecorded intermediates with high metabolite scores and high DNA and protein reactivity scores during metabolism to a known reactive metabolite end point. [78, 79, 80] In other cases, predicted reactive intermediates are consistent with previous reports (Figure 3.6, right). The network for metabolism of the withdrawn sedative triclofos to its active metabolite trichloroethanol identifies a probable intermediate with high protein reactivity, chloral, that is a known irritant. [81]
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. [62] 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 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 available in the US and has been established as a cause of acute liver injury. [82, 83] Nimesulide’s precise mechanism of injury is unknown, though thought to be related to production of an intermediate that enables an idiosyncratic reaction during metabolism in the liver. [84] During metabolism of nimesulide to three targets, the model infers the presence of 59 intermediate metabolites, of which 10 surpass the optimal metabolite score threshold and 3 (M6, M7, and M8) are predicted to be highly reactive (Figure 3.7). Figure 3.7, middle left, shows the subnetwork containing the 3 reactive intermediates and M9, another intermediate above the optimal threshold. Presence of M6, M7, and the target are corroborated by experimental evidence and, importantly, the target is thought to be a latent reactive metabolite associated with highly reactive metabolites, specifically M7. [85, 86] The model also prioritizes importance of nitrogen reduction of nimesulide into M6, which has a basis in bioactivation mechanisms of nitroaromatics. Specifically, intermediates M6 and M7 in the metabolism of nimesulide → M6 → M7 → target are both hypothesized to covalently bind to proteins. [80] This bioactivation pathway is also detected by XenoNet. While we did not find
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

**Reaction Class**
- Stable Oxygenation
- Unstable Oxygenation
- Dehydrogenation
- Hydrolysis
- Reduction

**Node Importance**
- \( \leq 0.01 \)
- 0.5
- 1

**Edge Weight**
- \( \leq 0.01 \)
- 0.5
- 1

**Annotated Class**
- Recorded
- Not Recorded

**Predicted Class**
- Important
- Not Important

**Reactivity Score**

**Figure 3.6:** The model identified formation of intermediate(s) that are likely to precede known metabolite end points. The metabolic network for tolcapone posits 4 unrecorded intermediates with high metabolite scores. The structures of the 4 intermediates, M1 through 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), that is a known irritant. [81]
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). [45, 44] Out of 62 intermediates (Figure 3.8A), 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 3.8B. 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. Presence of 6,6-Dimethylhept-2-en-4-ynoic acid (M16) is not recorded in 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. [87] The model also correctly ranks the unreported pathway TBF → M16 → TBF-A as last and the pathway TBF → M17 → M18 → TBF-A as second 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 3 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 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,
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

Figure 3.7: XenoNet and the bidirectional MPNN pinpointed the dominant bioactivation pathway nimesulide → M6 → M7 → target. [80] A) XenoNet inferred the presence of 59 intermediate metabolites linking nimesulide and three target metabolites. 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 3 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 literature. Compared to the unidirectional MPNN and XenoNet algorithm, the bidirectional model assigns higher importance 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, of which 3 are reported in literature.
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

Figure 3.8: The 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. [45, 44] Top, XenoNet inferred the presence of 62 intermediate metabolites linking TBF and TBF-A. Bottom, the model extracted a subnetwork of 4 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 importance 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 literature.
a trapping agent, e.g. GSH or cyanide, that has a high likelihood of conjugating to reactive metabolites is selected [26]. Formation of a GSH conjugate can be detected via mass spectrometry and indicates presence of a reactive metabolite. Cyanide and GSH can be applied as trapping agents for hard and soft electrophilic reactive molecules, respectively. [88, 89] Due to 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 towards 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.

3.2.4 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 metabo-
lite(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 sec-
ond network state could then be utilized by the metabolite scoring model.

The graph neural network architectures adhered to the MPNN paradigm, which has intrinsic 
limitations. There are potential issues of bottlenecks or over-smoothing that were referenced 
earlier, but also issues regarding their representational capacity. Without modifications that 
can lead to less practical architectures [90, 91], 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. [92, 93]

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 3.10 
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.

3.3 Conclusion

This study established and validated a novel metabolite formation model based on a bidi-
rectional 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 network structure. 
Incorporating of 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 com-
pared methods on multiple accuracy and calibration metrics on a data set of 311 networks and
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

6,606 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.

3.4 Materials and Methods

3.4.1 XenoNet

XenoNet integrates multiple machine-learning approaches to modeling P450 metabolism and reactivity with a rule-based structure inference model. We briefly summarize core properties of XenoNet’s function, but refer to implementation nuances, different heuristic variants for searching the metabolite space, and prior validation from primary literature. [71]

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, multi-edge graph-based data structure. Specifically, we 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). In tandem, our deep learning Phase I SoM model, Rainbow XenoSite, provides a probabilistic prediction of the corresponding metabolic transformation. We then reapply this procedure on the inferred child metabolites to build out a widening tree of potential pathways across multiple metabolic transformations.

A path is retained in the final network output if it terminates at a user-defined target metabolite. 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.

### 3.4.2 Relevant Extensions to XenoNet

Here, we describe relevant improvements to XenoNet. These improvements include modification of the optimal search algorithm, enumeration of paths that lead to multiple target structures, and incorporation of reactivity predictions.

**Beam Search**

XenoNet now finds paths leading from a start molecule to target metabolites using beam search (Figure 3.9). 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 step later and the early choice of the wrong metabolite may cumulatively lead to a poorer overall path probability over time. Ultimately, a
poor choice at any search step commits us to a poor future, but we cannot know at that step that the choice was poor.

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 instead retain only the top \(N\) scoring forks at each search step. We are not just choosing the top \(N\) based on the probability of the latest metabolites. Instead, the choice is 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.

Several modifications allow for the number of retained paths to exceed the beam width. First, paths that terminate at a target no longer contribute to the beam width threshold once they’ve been stored in the constructed network. Second, at the penultimate search step, we enumerate all possible candidate child metabolites for each of the candidate paths and retain all paths (the path up to the penultimate step plus a candidate child metabolite) that fulfill the search criteria, e.g., the 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 potentially overlapping networks (Figure 3.10). For example, our metabolic network data set used to validate XenoNet had approximately 17,000 substrate molecule and target metabolite pairs. [71] Multiple pairs shared the same substrate molecule. By defining a case for one substrate molecule and multiple target metabolites, we decreased the number of networks to 9,686 and reduced redundant computation.
Figure 3.9: 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. 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 $S \to I_{1.1} \to I_{2.3}$ and $S \to I_{1.3} \to I_{2.9}$. Even if the two-step pathway $S \to I_{1.1} \to P$ 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 $S \to I_{1.1} \to I_{2.3} \to P$, $S \to I_{1.3} \to I_{2.9} \to P$, and $S \to I_{1.1} \to P$.
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

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. [65] 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 3.10). Later, we demonstrate how this modification allows for screening of missing or unknown intermediate metabolites and assessing reactivity of inferred intermediates.

3.4.3 Single Start-Multitarget Metabolism Data Set

We extracted 17,054 annotated networks from the metabolic network data set described in prior work. [71] 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). Each annotated network is defined by a substrate molecule, a target metabolite, and zero or more experimentally observed intermediate metabolites. The metabolic network data set was originally designed such that each entry represented the annotated network for one substrate molecule and one target metabolite. Thus, 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 into a single annotated network with multiple targets.

Next, we extracted annotated metabolic 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 retained generated networks that inferred the presence of at least one intermediate metabolite recorded in the corresponding annotated network. For each generated network, each intermediate metabolite was assigned a
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

Figure 3.10: Top) Example network structures for all three cases - no target metabolite, one target metabolite, and multiple target metabolites - are shown. 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 [94].
target label — 1 if the intermediate metabolite is reported in its corresponding annotated network, 0 otherwise. The final XenoNet start-multitarget network data set consisted of 311 pairs of generated networks and their annotated counterparts.

Across the XenoNet start-multitarget network data set, there were 676 annotated intermediate metabolites, of which 467 were unique. However, we considered all 676 metabolites since the same metabolite can be present in different contexts across multiple metabolic networks. We want the model to detect structural conditions under which the same metabolite may be authentic or spurious. XenoNet inferred 516 of the 676 annotated intermediate metabolites, of which 96 annotated intermediates were not inferred because they require a depth limit greater than 3 to discover. Across all generated networks, 516 intermediate metabolites are labeled as observed, or 1, and the remaining 6,090 predicted intermediate metabolites are labeled as unobserved, or 0. We provide the start-multitarget networks in the “Start_Multitarget_Training_Dataset.json” file.

## 3.4.4 DrugBank Data Set

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 data set. We removed withdrawn drugs that were present as substrate molecules in the training data set. The final DrugBank data set comprised 70 annotated networks for each remaining withdrawn drug and 70 generated networks containing 2,832 intermediate metabolites. We provide the start-multitarget networks used for each network in the DrugBank withdrawn drug data set in the “Start_Multitarget_DrugBank_Withdrawn_Dataset.json” file.
3.4.5 XenoNet Start-Multitarget Network Structures & Features

We propose neural network models designed around the message passing framework. [95] The particular form of the message passing neural network (MPNN) is 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 must terminate 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 of 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 3.11), and support different optimal binarization thresholds (Figure 3.12). We do not have a SoM model for epoxide opening, so we 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- and out-degree are easily exploited with minimal computational overhead. Furthermore, we noticed that nodes with high in-degree and low out-degree (stable metabolites) tend to be experimentally observed and nodes with low in-degree and high out-degree (transient metabolites) tend to be experimentally unobserved (Figure 3.13). The in- and out-degrees are calculated using the
FIGURE 3.11: The model has differing levels of confidence in its reaction type specific predictions and incorporation of reaction type information may assist learning relative comparisons between predictions across different reaction types. Individually, site-level predictions specific to unstable oxygenation, reduction, and hydrolysis reaction types are well calibrated. Site-level predictions specific to stable oxygenation and dehydrogenation reaction types are not well calibrated. However, only the hydrolysis site-level predictions are probabilistic, as the unstable oxygenation and reduction site-level predictions do not range from 0 to 1.
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

**Figure 3.12:** The optimal thresholds for binarizing the Rainbow model’s site-level predictions are different for each reaction type. Thus, encoding reaction type information allows the model to adjust influence of predictions based on the Rainbow model’s variability in confidence between different reaction types. To define the binarization thresholds, we computed the Youden index as the optimal point on the cross-validated ROC curves computed from the Rainbow model’s site-level reaction predictions on its training data for each reaction type. [62, 54] The site-level ROC curve is displayed across 10,280, 5,811, 2,794, 1,590, and 3,869 sites of stable oxygenation, unstable oxygenation, dehydrogenation, reduction, and hydrolysis respectively. The light colored arrows for each reaction type point out the corresponding location of the optimal site-level cutoff on that reaction type’s respective ROC curve. The optimal cutoff for binarizing site-level predictions was 0.015, 0.0063, 0.0034, 0.0031, and 0.0083 sites of stable oxygenation, unstable oxygenation, dehydrogenation, reduction, and hydrolysis, respectively.
weighted paths that enter and exit each node. If multiple edges of same directionality exist between a pair of nodes, only the highest weighted edge is considered.

![Figure 3.13](image)

**Figure 3.13:** In the absence of node features, node in- and out-degree may be exploited as an informative feature with no additional computational overhead to compute. Though there is no clear demarcation, nodes with high in-degree and low out-degree tend to be experimentally observed and nodes with low in-degree and high out-degree tend to not be experimentally observed. Intuitively, a node with high in-degree is likely to be visited during metabolism and have a better chance of being observed. In contrast, a node with a high out-degree is likely to be transient or not as stable and have a lesser chance of being observed. This instability may cause the intermediate metabolite to be missed during experimental study, but it is possible the intermediate metabolite is short-lived enough to not be pernicious. The in- and out-degrees are calculated for each intermediate node using the weighted paths that enter and exit from the node. If multiple edges exist of same directionality between a pair of nodes, only the highest weighted edge is considered. Start and target nodes are left out as they are not being predicted on and significantly skew the distribution to high out-degrees and in-degrees, respectively.

We later use reactivity scores in conjunction with learned metabolite scores to assess 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.

### 3.4.6 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 node importance (Figure 3.14). Each edge represents a metabolic transformation and is weighted by a raw score predicted by our Phase I SoM model. Formally, a metabolic transformation between metabolite \( M_j \) and one of its children, \( M_k \), is weighted by raw score \( w_{M_j \rightarrow M_k} \). We normalize each raw score, \( w_{M_j \rightarrow M_k} \), by the sum of raw scores over all metabolic transformations from \( M_j \) to its children. The normalization step is computed using Equation 3.3 and results in a normalized score, \( W_{M_j \rightarrow M_k} \), between a metabolite, \( M_j \), and one of its children, \( M_k \).

\[
W_{M_j \rightarrow M_k} = \frac{w_{M_j \rightarrow M_k}}{\sum_{M_x \in M_j \text{children}} w_{M_j \rightarrow M_x}} \tag{3.1}
\]

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 Equation 3.4. The resultant metabolite score, \( F_{M_j} \), is a weighted sum of the normalized scores, \( W_{M_i \rightarrow M_j} \), 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 \( M_i \).
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

\[ F_{M_j} = \sum_{M_i \in M_j^\text{parents}} F_{M_i} \times W_{M_i \rightarrow M_j} \]  \hfill (3.2)

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.

However, 73 (12\%) of the start-multitarget XenoNets violate the acyclic assumption. To enable the XenoNet algorithm to handle cyclic networks, a pre-processing step is applied to eliminate cycles. Cycles are identified and eliminated by removing the minimum number of edges necessary, in order of edges with the smallest prediction values. This pre-processing step is only applied with 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

There is copious research related to centrality and measuring node importance. [96] For comparison, we utilized PageRank, a common centrality measure used to compute importance of web pages (nodes) in web graphs. PageRank interprets incoming edges as votes towards 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
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

A) Unnormalized Predictions

\[ S \rightarrow I_1 \rightarrow I_3 \rightarrow I_2 \]

B) Normalized Predictions

\[ S \rightarrow I_1 \rightarrow I_3 \rightarrow I_2 \]

C) Metabolite Scoring

\[ S \rightarrow I_1 \rightarrow I_3 \rightarrow I_2 \]

**Figure 3.14:** The 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 Equation 3.3. C) Third, the substrate is assigned a score of 1.0 and downstream metabolites are scored using Equation 3.4. 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.
other nodes in the network, the PageRank implementation is non-trivial and we refer to primary literature for comprehensive description. [97]

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. [98, 99] Betweenness centrality considers the extent to which a node contributes to paths between all other nodes and we used the shortest-path variant. [100, 101]

3.4.7 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, \( h_u^{(0)} \), 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, \( m_{N(u)}^{(k)} \), is defined by aggregating the hidden embeddings received from its neighborhood, \( N(u) \). The node’s hidden embedding for the current iteration is updated based on its hidden embedding at iteration \( k - 1 \), \( h_u^{(k-1)} \), and the aggregated message, \( m_{N(u)}^{(k)} \). After \( K \) iterations of message passing, we define each node \( u \)’s output embedding as \( h_u^{(K)} \).

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 Equations 3.3 and 3.4, are derived from prior work, namely the edge conditioned con-
volutional layer. [95, 102] Each multi-layer perceptron (MLP) is parameterized by learnable network
weights $\Theta$ that are not shared and are specific to each layer $k$. This enables the edge weight in-
terpretations, 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 convo-
lution 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 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 con-
necting 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.

$$m^{(k)}_{N(u)} = \sum_{v \in N(u)} h^{(k-1)}_v \cdot \text{MLP}^{(k)}_{\Theta}(e_{u,v})$$

Equation 3.3

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

$$h^{(k)}_u = \Theta^{(k)}_{\text{self}} h^{(k-1)}_u + m^{(k)}_{N(u)} + b^{(k)}$$

Equation 3.4

We combined multiple edge conditioned convolutional layers to form a deeper architecture
for node prediction (Figure 3.15). 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 which 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. [103] 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 3.16). 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 3.17). 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. I.e., the
FIGURE 3.15: A) Example unidirectional metabolite scoring model with 2 graph convolutional layers. The message passing update for each node is calculated using an edge conditioned convolutional layer. [95, 102] The output hidden embedding vector is processed by subsequent layer normalization and ReLU activation before being submitted to the next graph convolution layer. Graph convolution layers are aggregated and processed by a 2-layer MLP. The final layer output ranges from zero to one and reflects the metabolic importance of the metabolite.
B) Example bidirectional metabolite scoring model with 4 graph convolutional layers, in which the second block of components reverses flow of message passing from child to parent nodes.
Figure 3.16: The bidirectional MPNN takes advantage of information flow in both directions, from parent to child and vice versa, 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 parent 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 parameterized 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 grey 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.
Figure 3.17: The first variant of the proposed MPNN architecture only aggregates messages from a node’s parents and results in an inability for many competing pathways to be considered. Assuming the first variant’s strategy, we compute the number of disconnected components that form information silos per network. In this case, we define an information silo as a group of intermediate nodes for which none of the intermediate nodes can be influenced by the hidden embedding state of any other intermediate nodes in the network, regardless of the \( k \)-hop neighborhood considered. On average, a network contains 5 information silos (minimum 1, maximum 18). Comparatively, the second variant’s strategy results in 0 information silos for all networks.
inductive test nodes remain unobserved both 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 [104], where hyperparameter sweeps are applied on tunable model parameters (Table 3.5). An outer 5-fold cross validation splits the network data set 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.

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.
Chapter 3. Message Passing Neural Networks Improve Prediction of Metabolite Importance

**TABLE 3.5:** Hyperparameters used for model selection during nested cross validation inner loop. For the bidirectional variant used in follow-up case studies, the optimal hyperparameter values are bolded.

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

### 3.4.8 Extraction of Important Connected Subnetworks

XenoNet can produce large, branching networks that important metabolites may be drawn from 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 cut-off. The cut-off ranges from 0 to 1 and becomes more restrictive as it increases. A cut-off of 0 retains all pathways while a cut-off 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.
Chapter 4

ReactNet: Prediction of Phase II Metabolite Formation

This chapter is adapted from a manuscript in preparation: Noah R. Flynn and S. Joshua Swamidass, ReactNet: Prediction of Phase II Metabolite Formation

4.1 Introduction

Xenobiotic metabolism facilitates alteration of drugs into more polar forms, which have a higher rate of clearance, so that their effects may be removed from the body. Functionalization (Phase I metabolism), conjugation (Phase II metabolism), and excretion divide metabolism in distinct processes. Functionalization introduces or exposes polar functional groups on the chemical structure of the parent molecule, which can bioactivate the molecule into a reactive metabolite. Electrophilic, reactive metabolites may adversely conjugate to nucleophilic sites in DNA or proteins. [19, 20, 21, 22, 23] Toxicity resulting from reactive metabolites is a frequent cause of drug withdrawal and drug-induced liver injury. [10, 11]

The liver requires a defense against the continuous break down of drugs into electrophilic, reactive metabolites, otherwise it risks extensively damaging itself (Figure 4.1A). Conjugation with charged species, such as sulfate, glucuronic acid, or glutathione (GSH), is relevant to drug toxicity because it aids in detoxification of electrophilic compounds and production of
Chapter 4. ReactNet: Prediction of Phase II Metabolite Formation

more polar metabolites for transportation and excretion. An example of the utility of GST-mediated detoxification is with the drug paracetamol. Small amounts of paracetamol may be metabolized via P450 enzymes, specifically CYP2E1, into the reactive metabolite N-acetyl-p-benzoquinonimine (NAPQI, Figure 4.1B). GSH attacks NAPQI to form a less toxic, more polar conjugate in the form of mercapturic acid, which is readily excreted from the body. However, at concentration levels beyond therapeutic doses, paracetamol can deplete the GSH stores and result in adverse drug reactions (ADRs). Depletion of GSH allows NAPQI to readily bind to proteins via cysteine residues. Due to the rapid increase of aminotransferase levels, it is thought that hepatocyte necrosis takes place rapidly upon depletion of GSH. [105]

Experimental screening tools for human toxicity drug properties have been established, primarily using S9 fractions and hepatocytes. [109, 110, 111, 112, 113] Computational equivalents provide higher throughput and greater accessibility. In particular, site of metabolism (SoM) models predict liability of each site on a molecule to undergo perturbation by an enzymatic system and structure inference models generate the metabolite resulting from such a perturbation. [114, 115, 116, 117, 118, 119, 120, 121, 54, 122]

Comparatively, SoM and metabolite structure inference models for Phase II metabolism have had less attention and development than models for Phase I metabolism. The five most important contributors, which account for approximately 90% of Phase II metabolism, are the uridine diphosphate glucuronosyl transferase (UGT), glutathione transferase (GST), sulfotransferase (SULT), N-acetyltransferase (NAT), and topurine S-methyltransferase (TPMT) enzyme families (Figure 4.2). [123, 124, 125] Phase II metabolism accounts for approximately 30% of reported metabolites [126] and accurate models of Phase II metabolism are necessary for understanding how defects in Phase II enzymes contribute to drug toxicity. [127, 123, 128, 129, 130, 131, 132]

Currently, the set of freely available tools that support probabilistic SoM prediction and metabolite structure generation for the aforementioned Phase II processes is limited to SyGMA and GLORYx. [57, 133] SyGMA ranks predicted metabolites based on probability scores that
Chapter 4. ReactNet: Prediction of Phase II Metabolite Formation

**Figure 4.1:** ADRs account for 4.9–7.7% of hospital admissions [1, 2] and are a leading cause of death in the U.S. [3]. A) Approximately 10–15% of ADRs are dose-independent and idiosyncratic (IADRS), which makes them notoriously difficult to plan for and unable to be evaded by simple dose adjustments [7]. Reactive metabolites often cause ADRs and IADRs. Carbamazepine, for example, is known to undergo epoxidation into carbamazepine-10,11-epoxide. The electrophilic epoxide may bind to nucleophilic protein sites as a hapten complex and provoke an adverse immune response. [106, 107, 108] B) Acetaminophen metabolism demonstrates how detoxification pathways compete against bioactivation pathways. The majority (85%) of administered acetaminophen is metabolised into readily excreted polar conjugates via UGT or SULT. At therapeutic doses, 15% of administered acetaminophen is metabolised by Phase I oxidation into NAPQI. GSH attacks NAPQI to form a less toxic, more polar conjugate in the form of mercapturic acid, which is readily excreted from the body. At concentration levels beyond therapeutic doses, NAPQI production overwhelms GSH stores and remaining NAPQI readily binds to proteins via cysteine residues and results in hepatotoxicity. [105]
Chapter 4. ReactNet: Prediction of Phase II Metabolite Formation

Figure 4.2: We cover five Phase II reaction types in this study. The uridine diphosphate glucuronosyl transferase (UGT), glutathione transferase (GST), sulfo-transferase (SULT), N-acetyltransferase (NAT), and top-purine S-methyltransferase (TPMT) enzyme families cover approximately 90% Phase II metabolism. [28] We display example reactions for glutathionation, sulfation, glucuronidation, acetylation, and methylation.
are derived from statistical analysis on a large set of experimental reaction records reported in the (discontinued) MDL Metabolite database. GLORYx uses FAME 3 to provide probabilistic SoM predictions based on machine learning, specifically using extremely randomized trees with two-dimensional circular descriptors. SyGMA and GLORYx both incorporate reaction rules for inferring potential child metabolite structures produced from metabolic transformation of a parent molecule. Whereas SyGMA derives its reaction rules from analysis of the MDL Metabolite database, GLORYx does not rely on a specific metabolite data set and instead derives rules from scientific literature and chemistry knowledge.

In this work, we built upon our previously developed tool for network-level metabolism modeling, XenoNet, by developing a tool specifically for Phase II metabolism, termed ReactNet. ReactNet incorporates the XenoSite model suite to provide SoM predictions for Phase II reactions and extends the Metabolic Forest with new reaction rules for inferring resultant metabolite structures. [117, 121, 134, 120] On a Phase II metabolism data set constructed from 18240 reaction records, ReactNet outperforms both GLORYx and SyGMA.

4.2 Data & Methods

4.2.1 Phase II Metabolism Data Set

We assembled a chemically diverse training data set of 18240 Phase II reaction records from the literature-derived Accelrys Metabolite Database (AMD). Each reaction record reports a pair of molecules, where the substrate is the parent molecule and the product is the metabolite. This data set was composed of 6417, 2211, 1059, 918, and 910 parent molecules known to be reactive with UGT, SULT, GSH, NAT, and TPMT, respectively, and their corresponding metabolites. In accordance with the AMD license agreement, we are not allowed to disclose the structures of molecules in the networks. To enable rebuilding the complete Phase II metabolism data set, we provide all reaction and molecule AMD registry numbers in the Supporting Information.
4.2.2 ReactNet: Phase II Metabolism Network Generator

We built ReactNet, which, given an input substrate, provides molecule-level and site-level predictions and generates inferred metabolite structures. ReactNet combines several previously developed, in-house models for accomplishing this task. ReactNet uses XenoSite’s deep learning Phase II metabolism model’s, which can accurately predict sites of metabolism (SoMs) for reactivity to UGT and GSH. \[134, 120\] ReactNet further extends our structure inference model, the Metabolic Forest, with updated or new rule sets for inferring possible metabolite structures that result from Phase II reactions involving GST, UGT, SULT, NAT, and TPMT one metabolic step away. \[121\]

The neural network-based SoM models for the Phase II reaction types all conform to the same general structure. The bioactivation model is a feedforward neural network, with one molecule layer, one input layer, two hidden layers, and two output layers (Figure 4.3). The input layer computes atom-level and molecule-level descriptors from the input molecule’s structure. The first output layer calculates atom-level scores (ALS) for each atom’s interaction with one of the five Phase II enzymes. The second output layer computes a single molecule-level scores (MLS) for each input molecule.

We trained each network in two stages. In the first stage, we trained the atom-level network to compute accurate ALS. For an input molecule, each of its atoms is considered as a possible SoM for the reaction type being modeled. Each atom is represented by a vector containing the atom-level descriptors that define it. Thus, the data set passed into the atom-level network is a matrix with one column per atom-level descriptor and one row per atom. The atom-level network learns an association between each atom’s descriptors and a binary target vector that indicates whether the atom is a known SoM. Model weights are iteratively adjusted using gradient descent on the cross-entropy error, learning to assign high ALS for experimentally observed SoMs and low ALS for all other sites.

In the second training stage, we trained the molecule-level network to compute accurate MLS. The molecule-level networks learns an association between each molecule’s descriptors,
First, the atom-level network is trained using atom-level data, where sites involved in Phase II metabolic processes that lead to conjugation are labeled positive.

Second, the molecule-level network is trained using data where adduct-forming molecules are labeled positive.

**FIGURE 4.3:** General structure of the SoM models for Phase II metabolism. The diagram on the left demonstrates the flow of data through the model. The models consist of one molecule layer, one input layer, hidden layers (two in this example), and two output layers. First, molecule-level and atom-level descriptors are calculated from an input molecule’s structure. Atom-level descriptors are submitted to the hidden layer, which computes atom-level scores (ALS) for the input molecule and each site (atom) with respect to a Phase II reaction, e.g., glutathionation. Each ALS ranges from zero to one, reflecting the probability of conjugation to a conjugate at a specific site within the input molecule. For this study, we incorporated separate neural network-based SoM models for GST, UGT, SULT, NAT, and TPMT. Next, molecule-level descriptors and the top 5 ALS are submitted to a second hidden layer, which computes a molecule-level score (MLS) for the input molecule. MLS also represent the probability of a molecule undergoing conjugation by any of the enzymes considered and, like ALS, ranges from zero to one. A chemical structure is represented by the molecule node. The other circles are probabilistic scores between 0 and 1. Blocks are vectors of real numbers. On the right, site-level data are illustrated on the top (with sites of predicted conjugation circled) and molecule-level data are illustrated on the bottom (with the molecules of predicted conjugation circled).
the top 5 ALS, and the molecule’s classification as experimentally observed to form the Phase II conjugate or not. The data set passed into the molecule-level network is a matrix with one column per molecule descriptor or top 5 ALS and one row per molecule. Network weights are similarly adjusted using gradient descent on the cross-entropy error. The models utilize leave-one-out cross-validated predictions for any molecule inputs that were in their training data sets.

The Metabolic Forest outputs SoMs, the corresponding metabolic transformations that act on those SoMs, and the inferred metabolite structures through use of transformation rules encoded through reaction SMARTS. Reaction SMARTS are derived from SMARTS patterns to express reaction queries initiated on a small set of localized atoms within a molecule. Similar reaction language concepts include SMIRKS [135, 136] and SMILES. [137] The exact Reaction SMARTS used for each rule, developed by manual inspection, adjustments, and reference to prior work, are referenced in Table 4.1. Combined, the SoM neural network models and rule-based Metabolic Forest can output probabilistic predictions and inferred structures mediated by Phase II metabolic transformations (Figure 4.4).

**Table 4.1: Reaction SMARTS used by the metabolite prediction algorithm.**

<table>
<thead>
<tr>
<th>Rule</th>
<th>Reaction SMARTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acetylation</td>
<td>[#7,#8,#16:h:1]»[*:1][#6]=[#8][#6]</td>
</tr>
<tr>
<td></td>
<td>[c:1][NH2:2]=[c:1][N:2]C=O</td>
</tr>
<tr>
<td></td>
<td>[C;N:1][NH2:2]=[*:1][N:2]C=O</td>
</tr>
<tr>
<td></td>
<td>[C;O:1][NH2:2]=[*:1][N:2]C=O</td>
</tr>
<tr>
<td></td>
<td>[CX4:1][NH1;R:2][CX4:3]=[*:1][N:2][C:3]C=O</td>
</tr>
<tr>
<td></td>
<td>[CH3:1][NH1:2][#6:3]=[*:1][N:2][C:3]C=O</td>
</tr>
</tbody>
</table>
### Chapter 4. ReactNet: Prediction of Phase II Metabolite Formation

<table>
<thead>
<tr>
<th>Rule</th>
<th>Reaction SMARTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glucuronidation</td>
<td><img src="#" alt="Rule Reaction SMARTS for Glucuronidation" /></td>
</tr>
<tr>
<td>Glutathionation</td>
<td><img src="#" alt="Rule Reaction SMARTS for Glutathionation" /></td>
</tr>
</tbody>
</table>
Chapter 4. ReactNet: Prediction of Phase II Metabolite Formation

FIGURE 4.4: Example ReactNet output visualized for hydroxy-methyl-vinyl-ketone. ReactNet predicted structures resulting from reactions with GST, UGT, SULT, and NAT.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Reaction SMARTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Methylation</td>
<td>[c:1][OH1:2]=[c:1][O:2]C</td>
</tr>
<tr>
<td>Sulfation</td>
<td>[#6:1][#8:2]=[<em>:1][</em>:2]S(=O)(=O)</td>
</tr>
<tr>
<td></td>
<td>[#6:1][#8:2]=[<em>:1][</em>:2]S(=O)(=O)</td>
</tr>
<tr>
<td></td>
<td>[c:1][OH1:2]=[c:1][O:2]S(=O)(=O)</td>
</tr>
<tr>
<td></td>
<td>[c:1][NH2:2]=[c:1][N:2]S(=O)(=O)</td>
</tr>
<tr>
<td></td>
<td><a href="">C:5</a>(CC[OH1]):1][OH1:2]=[C:1][O:2]S(=O)(=O)O</td>
</tr>
</tbody>
</table>

4.3 Results & Discussion

4.3.1 Model Performance & Method Comparisons

ReactNet takes an input parent molecule and outputs predictions for Phase II metabolic transformations that may act on the parent molecule and the metabolite structures that result. We
Chapter 4. ReactNet: Prediction of Phase II Metabolite Formation

<table>
<thead>
<tr>
<th></th>
<th>ROC AUC</th>
<th>Recall</th>
<th>Precision</th>
<th>Number of Predicted Metabolites</th>
<th>True Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReactNet</td>
<td>90.9%</td>
<td>0.84</td>
<td>0.15</td>
<td>72857</td>
<td>11284</td>
</tr>
<tr>
<td>GLORYx</td>
<td>72.9%</td>
<td>0.84</td>
<td>0.24</td>
<td>46288</td>
<td>11282</td>
</tr>
<tr>
<td>SyGMa</td>
<td>71.7%</td>
<td>0.81</td>
<td>0.27</td>
<td>40381</td>
<td>10956</td>
</tr>
</tbody>
</table>

Table 4.2: ReactNet outperforms published methods in terms of ROC AUC on the AMD-derived Phase II metabolism data set.

ReactNet evaluated ReactNet’s ability to infer the experimentally observed metabolites reported in our AMD-derived Phase II metabolism data set. Further, we assessed how well ReactNet’s predictions separated and ranked inferred, experimentally observed metabolites from inferred, unreported metabolites. We also compared ReactNet’s performance against GLORYx and SyGMa as part of our assessment.

ReactNet outperformed both GLORYx and SyGMa (Table 4.2). We denote a true positive as a metabolite that is both predicted by the model and experimentally observed. Recall is the proportion of true positives among all experimentally observed metabolites and precision is the proportion of true positives among all predicted metabolites. ReactNet, GLORYx, and SyGMa achieved recall values of 0.84, 0.84, and 0.81, respectively, and precision values of 0.15, 0.24, and 0.27, respectively. The recall-precision trade-off is dependent on the chosen decision threshold, making both metrics unreliable for model comparison. We calculate the area under the receiver-operating curves (ROC AUC), resulting in ROC AUCs of 90.9%, 72.9%, and 71.7% for ReactNet, GLORYx, and SyGMa, respectively.

ReactNet outperformed GLORYx and SyGMa when evaluating performance across the full data set, but also when evaluating performance subset by reaction type (Figure 4.5). We calculated a single, binary-class ROC curve for each model and each reaction type subset (all, acetylation, glucuronidation, glutathionation, methylation, and sulfation) using binary labels for each metabolite (whether the metabolite has been experimentally observed or not) and the molecule-level SoM predictions generated by the model. ReactNet’s ROC AUC performance across the full data set and the acetylation, glucuronidation, methylation, and sulfation subsets...
Figure 4.5: ReactNet outperformed GLORYx and SyGMA when evaluating performance across the full data set, but also when evaluating performance subset by reaction type. We calculated a single, binary-class ROC curve for each model and each reaction type subset (all, acetylation, glucuronidation, glutathionation, methylation, and sulfation) using binary labels for each metabolite (whether the metabolite has been experimentally observed or not) and the molecule-level SoM predictions generated by the model. ReactNet’s ROC AUC performance across the full data set and the acetylation, glucuronidation, methylation, and sulfation subsets was significantly greater than GLORYx and SyGMA (paired t-test, p-value ≤ 0.05).

was significantly greater than GLORYx and SyGMA (paired t-test, p-value ≤ 0.05). SyGMA does not support glutathionation and so a comparison was only made against GLORYx, which achieved equivalent performance to ReactNet.

ReactNet also outperformed GLORYx and SyGMA in terms of ROC AUC on GLORYx’s smaller Phase II metabolism data set containing 582 parent molecules and 690 metabolites (Table 4.3). Similar to performance on our Phase II metabolism data set, all methods achieved equivalent recall and SyGMA achieved the best precision. GLORYx’s Phase II metabolism data set was derived from DrugBank and MetXBioDB metabolism data.
Chapter 4. ReactNet: Prediction of Phase II Metabolite Formation

<table>
<thead>
<tr>
<th></th>
<th>ROC AUC</th>
<th>Recall</th>
<th>Precision</th>
<th>Number of Predicted Metabolites</th>
<th>True Positives</th>
</tr>
</thead>
<tbody>
<tr>
<td>ReactNet</td>
<td>92%</td>
<td>0.76</td>
<td>0.12</td>
<td>4165</td>
<td>516</td>
</tr>
<tr>
<td>GLORYx</td>
<td>78%</td>
<td>0.75</td>
<td>0.16</td>
<td>3167</td>
<td>510</td>
</tr>
<tr>
<td>SyGMa</td>
<td>76%</td>
<td>0.77</td>
<td>0.17</td>
<td>3033</td>
<td>519</td>
</tr>
</tbody>
</table>

TABLE 4.3: ReactNet outperforms published methods in terms of ROC AUC on the GLORYx Phase II data set

4.4 Conclusion

We have established a method, ReactNet, that conflates multiple SoM model’s with a structure inference model for enumeration of a parent molecule’s Phase II metabolites. ReactNet predicted experimentally observed Phase II metabolites with high ROC AUC across a large set of chemically diverse parent molecules. ReactNet’s performance extended to all five Phase II metabolite subsets of the data, outperforming contemporary methods such as GLORYx and SyGMa. We can further extend our previously developed Phase I metabolic network generator, XenoNet, by applying ReactNet to infer possible Phase II conjugates one-step away from all the Phase I network’s metabolites. Incorporating Phase II metabolism naturally extends XenoNet as an informative tool for experimentalists to assess risks of reactive metabolite formation and ADRs.
Chapter 5

Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors


5.1 Introduction

Targeted cancer therapeutics boast superior health outcomes relative to traditional chemotherapy drugs; however, hepatotoxicity poses a major clinical concern for patients undergoing treatment with those drugs [138, 139, 140, 141]. The resulting drug-induced liver injury (DILI) is a significant cause of morbidity and mortality. The persistence and pervasiveness of hepatotoxicity with targeted cancer therapeutics reveals a lack of understanding from concept to clinic of relationships between their structures and potential to cause liver injury. In most cases, these
Chapter 5. Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors

Drugs induce idiosyncratic hepatotoxicity, which often lacks unique clinical or histological signatures for predicting outcomes. Moreover, DILI mechanisms are multifactorial and often elusive to discovery. Nevertheless, a well-established initiating event, especially for idiosyncratic DILI, is the metabolic bioactivation of drugs [142, 143, 144]. Consequently, identification of problematic molecular structures prone to bioactivation and factors impacting those outcomes are critical in drug discovery and development.

Assessment of drug bioactivation risk traditionally relies on experimental studies incurring high costs in time, effort, and resources as well as expertise, and thus, we developed metabolism and bioactivation models to increase the efficiency of drug development. First, we designed deep neural models to predict formation of specific reactive metabolites like epoxides [145] and quinone species [146], as well as bioactivation of structural alerts (furans, phenols, nitroaromatics, and thiophenes) [147]. Importantly, the quinone model is capable of predicting formation of other reactive metabolites within the broader class of conjugated electrophiles despite its name. Second, we recently reported a novel approach for predicting the structures of metabolites arising from bioactivations [60]. The information could facilitate their experimental identification as well as the construction of possible metabolic pathways. Third, the toxicological relevance of reactive metabolites often depends on the ability to adduct with and modify biological molecules such as proteins, glutathione, and DNA [148, 149, 150], and so, we developed models to predict adduct formation [147, 151]. These computational analyses are more accessible than experiments and provide an opportunity to readily explore potential relationships between molecular structure and bioactivation that lead to testable hypotheses.

In practice, we couple high throughput computational studies to experimental efforts to validate predicted relationships and reveal model shortcomings for further refinement into practical tools as shown through our work on terbinafine [152], thiazoles [153, 154] and diphenylamine NSAIDs [155]. In this study, we applied our novel computational and experimental strategy to reveal the potential bioactivation liabilities of bromodomain and extra-terminal (BET) inhibitor drugs.
BET inhibitors comprise a growing class of anti-cancer agents [138, 156, 157] including at least 24 drug candidates in Phase I clinical trials [158]. These drugs target proteins that possess bromodomains capable of binding acetyl-lysines present on histones. That action regulates transcription of oncogenes and anti-apoptotic proteins [159, 160]. Consequently, drugs mimicking the modified amino acid are effective at disrupting bromodomain functions in cancer cells. In fact, the widely studied inhibitor I-BET151 incorporates 3,5-dimethyl-isoxazole as an isostere of acetyl-lysine for its mode of action. This feature generated a strongly efficacious inhibition against leukemic cell line proliferation [161, 162] and fostered its progression to clinical trials [163]. Like other early BET inhibitors, I-BET151 induces unwanted toxic side effects often attributed to low selectivity amongst different bromodomains as well as other mechanisms [164]. New generations of BET inhibitors seek to capitalize on the 3,5-dimethylisoxazole [165, 166, 167, 168, 169, 170, 171, 172, 173, 174, 175, 176, 177, 178] motif and incorporate more molecular diversity to improve bromodomain selectivity. This strategy could potentially reduce pan-BET inhibitor-associated side effects yet risks for other adverse outcomes may persist. Isoxazole metabolism and bioactivation resulting in toxicity is rare [179] but possible. Isoxazole substituents may potentiate bioactivation risks in which metabolism generates reactive conjugated electrophilic metabolites such as cyanoacroleins [180, 181] or enimines [182, 183] that react with glutathione. To our knowledge, there are no investigations of these types of bioactivations or others for isoxazole-containing BET inhibitors. We hypothesize that BET inhibitor substituents modulate risks for isoxazoles to undergo bioactivation into reactive metabolites as potential initiators of toxicity.

Herein, we carried out a novel iterative modeling strategy to predict bioactivation pathways involving isoxazoles and validated them experimentally for selected BET inhibitors currently in development. Identification of possible bioactivation pathways for isoxazoles involved coupling deep neural models for quinone species formation, metabolite structures, and then biomolecule reactivity as described previously [155]. Next, we assessed the capacity of our quinone model to identify potential electrophilic metabolites generated for isoxazole-containing molecules in
the Accelrys Metabolite Database (AMD). For select molecules, bioactivation outputs were used to predict metabolite structures from reactions and then reactivity with glutathione. As a practical application of this approach, we predicted possible bioactivation pathways for 32 isoxazole-containing BET inhibitors varying in molecular structure and selectivity toward bromodomain 4 (BRD4(1)) [184]. Subsequently, we selected two promising drug leads: OXBFD02 and OXBFD04. These molecules differ in the substitution of a phenyl group with pyridine leading to significantly enhanced half-life and affinity for BRD4(1) for OXFBD04 relative to OXFBD02 [184]. In this analysis, we also included I-BET151, which is representative of a promising first generation BET inhibitor [161, 185] that may undergo bioactivation. We carried out metabolism studies with those molecules using human liver microsomes. Reactions included dansylated glutathione to trap reactive metabolites and track adducts using the fluorescent dansyl tag [155, 186, 187]. Subsequent mass spectroscopic characterization of the glutathione adducts provided a way to infer structures of the observed reactive metabolites. When combined with observed reaction metabolites, we were able to construct putative bioactivation pathways for the BET inhibitors and then compare the findings to model predictions.

5.2 Methods

5.2.1 Materials

Chemical solvents were purchased from Thermo Fisher Scientific (Waltham, MA). Substrate I-BET151 was purchased from Cayman Chemical (Ann Arbor, MI) and substrates OXFBD02 and OXFBD04 were graciously provided by Dr. Stuart Conway and his laboratory at the University of Oxford (Oxford, United Kingdom). Trapping agent dansyl glutathione trifluoroacetic acid salt was purchased from Toronto Research Chemicals (Toronto, ON, Canada) while internal standard dansylamide was obtained from Millipore Sigma. Reducing agent Tris(2-carboxyethyl)phosphine hydrochloride (TCEP) was purchased from Millipore Sigma.
Chapter 5. Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors

Human liver microsomes pooled from 150 donors (HLM150) were purchased from Corning Gentest (Woburn, MA).

5.2.2 Data to assess quinone model performance toward isoxazole-containing molecules

We identified isoxazole-containing molecules and those undergoing bioactivation from the literature-derived Accelrys Metabolite Database (AMD) to evaluate model performance. We collected reaction data for all isoxazole-containing small molecules catalyzed within the human liver using SMARTS patterns for the isoxazole ring. We built a graph database with Neo4j to connect molecules with observed direct and downstream metabolites up to three metabolic steps away. We queried the graph database to find conjugated electrophile species involving an isoxazole ring by applying our previously developed structure inference model, the Metabolic Forest [60], to identify metabolite pairs formed between the isoxazole-containing molecule and direct or downstream metabolites. For identified molecule pairs, the upstream isoxazole-containing molecule was annotated as forming a conjugated electrophile. Then, we queried our graph to determine whether the isoxazole molecules that form conjugated electrophiles had been experimentally observed to react with glutathione. We identified 344 isoxazole-containing molecules in AMD. Of those, 12 molecules formed enimines, and all but one was reactive to glutathione. Some model evaluation molecules (validation set) were also in model training sets, which may bias predictions, and so we used a hold-out prediction to ensure that we are evaluating model ability to learn general principles of the task, rather than memorize training data.
5.2.3 Modeling bioactivation pathways for isoxazole-containing molecules

We predicted the complete bioactivation pathways for isoxazole-containing molecules by coupling models for quinone (conjugated electrophile) formation, metabolite prediction, and molecular reactivity. First, we revealed bioactivation sites by modeling formation of conjugated electrophiles, namely quinones and enimines [119]. This deep neural network model predicts one- and two-step conjugated electrophile formation by identifying atom pairs at which metabolic oxidation may occur to form conjugated electrophile metabolites with an accuracy of 88.2% based on a ROC analysis. In this case, we evaluated model ability to discriminate between bioactivated (positives) and non-bioactivated (negatives) isoxazole-containing molecules. The assessments relied on the 12 isoxazole-containing molecules undergoing bioactivation (positives) and 332 non-bioactivated isoxazole-containing molecules (negatives) obtained from AMD. From those results, we calculated model accuracy with a ROC-AUC score based on (1) comparing highest bioactivation scores between positives and negatives and (2) proper assignment of sites for conjugated electrophile formation as recorded in AMD. Second, we used quinone model outputs to predict structures for the reactive metabolites using the XenoNet model and scaled their likelihood based on the quinone model predictions [122]. This model uses the input molecule to enumerate pathways of metabolite structures while computing likelihood scores for each pathway. Third, we modeled reactivity of individual quinone or enimine metabolites toward glutathione as a trap [188, 65]. Our reaction models scale scores differently and thus we relied on the quinone model values as the final arbiter of possible bioactivation potential for isoxazole-containing molecules. We calculated accuracy of the reactivity model based on the ROC-AUC score using the highest atom score for each molecule and proper identification of the atom reactive toward glutathione as recorded in AMD. Lastly, we used the Metabolic Forest model to predict glutathione adduct structures based on the glutathione rule within its conjugation rule set [60].
5.2.4 Experimental bioactivation of isoxazole-containing BET inhibitors as test cases

We validated model predictions for bioactivation of selected BET inhibitors through experimental approaches. All reactions employed metabolism by pooled human liver microsomes as a model for the average adult human liver. Wells contained 1 mg/mL HLM150, 1 mM dansyl glutathione, and 50 µM substrate in 100 mM potassium phosphate buffer pH 7.4 with 0.1% DMSO (co-solvent) and were preincubated for 5 min at 37°C with shaking at 350 rpm using a BMG Labtech THERMOstar incubator (Ortenberg, Germany). Reactions were then initiated upon addition of a NADPH regenerating system (0.4 µU µl-1 glucose-6-phosphate dehydrogenase, 10 mM glucose 6-phosphate, 2 mM MgCl2, 500 µM NADP+). Identical mixtures without addition of NADPH regenerating system were incubated as negative controls. After 30, 60, or 90 minutes, aliquots were quenched by adding an 2-fold volume of ice-cold methanol containing an internal standard and a reducing agent (10 µM dansylamide and 5 mM TCEP, respectively) and incubated on ice for 10 min to optimize precipitation of proteins and phosphate buffer [189]. After 2800 g centrifugation at 4°C for 15 min using a Thermo Scientific Sorvall ST 16R Centrifuge (Waltham, MA), the supernatant was transferred to a 96 well full-volume microplate and evaporated to dryness using an Organomation Associates Microvap Nitrogen Evaporator System (Berlin, MA). Dried wells were then resuspended in mobile phase (20:80 H2O:CH3CN + 0.1% formic acid) for HPLC UV/Vis/Fluorescence and mass spectroscopic analysis as described in the following section. Each reaction set was performed in triplicate and replicated two to five times.

5.2.5 Characterization of BET inhibitor reactive metabolites trapped as glutathione adducts

Reactions were analyzed by LC-fluorescence to infer yields and LC-MS to characterize structures as reported previously [155]. Specifically, reaction metabolites were separated by a 4.6x150
mm Waters XSelect HSS C18 3.5 µm column heated to 40°C with a Shimadzu LC-20AB Prominence liquid chromatograph and detected by a Shimadzu RF-10AXL fluorescence detector. Mobile phase consisted of Solvents A (0.1% formic acid/90:10 H2O:CH3CN) and B (0.1% formic acid/CH3CN). A gradient method started with 89% A and held 3 min, decreased to 55% A over 20 min, decreased to 11% A over 10 min. Solvent A was then increased back to 89% over 2 min and held for remainder of run. Flow rate was 1 mL/min and run time 38 min. The fluorescence detector was set to excitation at 340 nm and emission at 525 nm [186, 187, 190] to detect the dansyl tag. At 15 min, both gain and sensitivity were increased to improve visibility of potentially low-yield metabolites. Analyte responses were normalized to the internal standard dansylamide. We then analyzed metabolites and adducts by mass spectrometry to determine parent masses and fragmentation patterns to validate purported structures. Samples were injected onto an Agilent Technologic 1290 Infinity HPLC using the same chromatographic method and column as described above. Analytes were scanned with the Agilent Technologic 6490 Triple Quad LC/MS. The ESI source was operated in negative and positive ion mode, and ion spectra were acquired in full scan mode monitoring the m/z range of 100-1200 amu. Subsequently, product ion spectra were generated from precursor ions with monitoring for fragmentation by collision-induced dissociation with a collision energy of 30 eV and a range of 45-1000 amu in negative ion mode.

5.3 Results

5.3.1 Mining the Accerlys Metabolite Database for possible isoxazole bioactivations

Given limited literature examples of isoxazole bioactivations, we mined examples of molecules and reactions present in the Accerlys Metabolite Database (AMD) for a broader assessment
of possible bioactivations. Among the 20,736 parent molecules, we found 344 isoxazole-containing molecules. Of the 344 isoxazole-containing molecules, 12 formed electrophilic conjugated systems via two different reaction mechanisms involving the isoxazole (Figure 5.1). Using ECFP6-derived fingerprints, the set of 344 isoxazole-containing molecules and the subset of 12 bioactivated molecules each had high internal diversities [191] of 0.85 and 0.70, respectively. The subset of 332 non-bioactivated molecules had a low similarity to a nearest neighbor [191] from the bioactivated subset of 0.212. Two molecules involved oxidation of the isoxazole fused to another ring that ultimately led to cleavage of the isoxazole (Figure 5.1A). The resulting diketo metabolite may be reactive due to conjugation with an adjacent double bond. The inferred metabolite for one molecule is known to be reactive with glutathione, while the metabolites for the other molecule have no evidence of reactivity towards glutathione (Figure 5.2). The ten other electrophilic conjugated systems involved a 4-amino-5-methyl-isoxazole bioactivated into an enimine (Figure 5.1B) and all of them were reactive toward glutathione at the methyl group emphasizing a pattern of bioactivation and reactivity.

5.3.2 Assessment of model accuracy to predict electrophilic conjugated systems involving isoxazoles

Identification of bioactivated 4-amino-5-methyl-isoxazoles provided a test set for assessing the ability of coupled models to predict complete bioactivation pathways for enimine formation, metabolite prediction, and biomolecular reactivity. We then assessed model accuracy toward predicting the occurrence and sites of bioactivations for isoxazole-containing molecules within AMD (Figure 5.3A). First, model predictions identified molecules known to undergo bioactivation with an AUC of 0.90 (0 to 1.0 scale) (Figure 5.3B) with two exceptions in which bioactivation led to cleavage of the isoxazole (Figure 5.2). These findings contrasted with very low predictions for isoxazoles that did not undergo bioactivation, indicating the ability of the model to distinguish between the two possibilities. Second, the model accurately predicted which atoms are the sites of metabolism leading to the enimine for all ten molecules possessing the
Chapter 5. Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors

**Figure 5.1:** Alternate pathways for isoxazole bioactivation into electrophilic conjugated systems. A) When fused to a ring, isoxazole metabolism may lead to ring cleavage to yield a diketo metabolite. B) Methyl substitution of the isoxazole makes possible the formation of an electrophilic conjugated system.
FIGURE 5.2: Two examples of AMD pathways involving isoxazole cleavage. A) The inferred metabolite for one molecule is known to be reactive with glutathione, while B) the metabolites for the other molecule have no evidence of reactivity towards glutathione. For both A) and B), the model predicted lower formation scores relative to the other isoxazole-containing molecules known to form conjugated electrophiles. In practice, the model may be prone to missing potential bioactivation pathways that require cleavage of the isoxazole.
Chapter 5. Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors

**Figure 5.3:** Quinone model predictions for bioactivation of isoxazole-containing molecules. A) Compounds with an isoxazole ring, shaded in grey, can be metabolized into reactive electrophiles like the illustrated enimine. B) The quinone model identifies which isoxazole-containing molecules will undergo bioactivation into a conjugated electrophile involving the isoxazole ring with 90% AUC accuracy. The histogram uses normalized frequencies, so the scale is balanced for the 332 negative (not bioactivated) molecule predictions and the 12 positive (bioactivated) molecule predictions. C) The quinone structure inference model can generate possible quinone and enimine metabolite structures for a queried compound. D) For all 10 enimine-forming molecules, the quinone structure inference model correctly predicts the reported metabolite structure when using the highest atom-pair probability. E) Likelihood of conjugated electrophile formation among drug leads against bromodomains using the quinone model with two types of assessment (see text).
4-amino-5-methyl-isoxazole. Given those positive results, we inferred molecular structures for the corresponding reactive metabolites (Figure 5.3C). The algorithm predicts many possible metabolite structures, so we scaled their likelihood based on the associated highest atom-pair score from the quinone formation model and reported the most probable one as shown in Figure 5.3D. This approach led to the correct identification of the experimentally identified reactive metabolites for all ten molecules, whose 4-amino-5-methyl-isoxazole group underwent conversion into an enimine (Figure 5.4). Third, we predicted the reactivity of those isoxazole enimines with glutathione. As expected, all ten metabolites were highly reactive with glutathione at the correct site of adduction. Further substituent analysis also revealed that model predictions align with our knowledge of metabolism and bioactivation, namely the deactivating effects of a phenyl to pyridine and methyl to ethyl substitutions [192] (Figures 5.5, 5.6). Taken together, the coupling of models provided accurate predictions of 4-amino-5-methyl-isoxazole bioactivations, resulting metabolites, and subsequent reactivity toward glutathione.

5.3.3 Application of coupled modeling approach to predict bioactivation of BET inhibitors

As a practical test case, we used our coupled modeling approach to predict reactive metabolites and potential for glutathione adduction for 32 bromodomain inhibitors containing 3,5-dimethyl isoxazoles (Figures 5.7, 5.8). The quinone model predicted the inhibitors had an elevated relative risk of 2.7 (p-value < 0.001) for undergoing bioactivation relative to isoxazole-containing molecules from our original validation set that do not undergo those reactions (Figure 5.3E). However, the inhibitors also had a lower relative risk of 0.45 compared to positive controls (p-value < 0.01). Given quinone model scores scale well with actual probabilities, we interpret this trend to mean these molecules are of elevated bioactivation risk although that requires experimental validation. Quinone model predictions also reflect general trends for substituent effects begetting quinone formation (Figure 5.7). Specifically, modification of the distal phenyl group with methoxy groups decreased bioactivation, while fluorine and chlorine atoms had no effect
Chapter 5. Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors

Figure 5.4: Model predicted bioactivation of 4-amino-5-methyl-isoxazole group into enimines. The left side displays the ten molecules with an overlay of the quinone formation model predictions and numbered arrows that represent max pair-level predictions. The right-hand side displays the inferred metabolite structures with an overlay of the reactivity model predictions for GSH and max site-level reactivity prediction.
Chapter 5. Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors

Figure 5.5: Model bioactivation predictions for isoxazoles with either a phenyl or pyridine substituent. The model predictions follow a general trend of decreasing magnitude when molecules with the phenyl isoxazole are altered to a pyridine isoxazole. There is a distinguishable gap between the two high-scoring pairs, which are the only two pairs observed to form a conjugated electrophile, from the remaining pairs which have lower scores. We collected pairs of molecules based on a common shared structure other than a specific change in substituents between the molecules in a pair. The quinone formation model was applied to the molecule pairs and the maximum atom pair-level predictions involving the isoxazole ring were plotted to check whether the model can differentiate the impact of the substituent as a function of the remaining, shared structure of the pair of molecules. Two molecule pairs are shown to highlight the impact on the model’s preferred sites. Pair (I) marks the pair with the most extreme decrease and pair (II) marks the only case where the difference in magnitude increased. For pair (I), the site scores at the 4 and 5 positions of the isoxazole substructure decreased from high to moderate and all sites with any formation potential on the phenyl ring decreased to have no formation potential on the pyridine ring. For pair (II), while the max atom-pair did increase, we still see the expected decrease in magnitude of site-specific predictions in the area localized to the phenyl (now pyridine) ring.
Chapter 5. Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors

<table>
<thead>
<tr>
<th>Observed Site of Metabolism</th>
<th>Probability of Formation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(I)</td>
<td>(I)</td>
</tr>
<tr>
<td>(II)</td>
<td>(II)</td>
</tr>
</tbody>
</table>

Figure 5.6: Model bioactivation predictions for isoxazoles with either a methyl or ethyl substituent. The model predictions follow a general trend of decreasing magnitude when molecules with the 5-methyl isoxazole are altered to a 5-ethyl isoxazole. There is a smooth dynamic range over the magnitudes of the maximum atom-pair scores and the observed trend is definitive. We collected pairs of molecules based on a common shared structure other than a specific change in substituents between the molecules in a pair. The quinone model was applied to the molecule pairs and the maximum atom pair-level predictions involving the isoxazole ring were plotted to check whether the model can differentiate the impact of the substituent as a function of the remaining, shared structure of the pair of molecules. Two molecule pairs are shown to highlight the impact on the model’s preferred sites. Pair (I) is representative of the site-specific changes in model predictions, whereby the altered 5-ethyl causes a decrease in model predictions at the 4 and 5 positions of the isoxazole substructure relative to the original 5-methyl. Pair (II) showcases the most extreme change in the model predictions across all the plotted pairs, with the substituent modification not modifying the specificity of possible sites of reaction but lowering the formation score for all the sites.
on bioactivation potential regardless of position on the ring. Lastly, substitution of that phenyl group with pyridine led to a general decrease in overall bioactivation predictions.

For more in-depth predictions, we chose three inhibitors to experimentally assess their bioactivation and subsequent adduction into glutathione adducts. Specifically, an earlier generation BET inhibitor, I-BET 151 [161, 185], and in-development drug leads [184], OXFBD02 and OXFBD04, which differ only in the presence of a phenyl substituent versus pyridine, respectively, were selected. The quinone model predicted molecule-level bioactivation scores of 0.59 (OXFBD02), 0.51 (OXFBD04) and 0.54 (I-BET151). The substitution between OXFBD02 and OXFBD04 then led to a slight decrease in bioactivation. Application of the structure inference
Chapter 5. Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors

![ROC Curve](image)

**Figure 5.8:** Model accuracy analysis for molecule- and pair-level scores for predicted conjugated electrophiles. The model was validated against 718 molecules with 359 of the molecules observed for form conjugated electrophiles. The molecule-level receiver operating characteristic (ROC) curve is displayed across the 718 molecules. The diagonal dashed line indicates performance that achieves 50% area under the ROC curve (AUC). The arrow points to the location on the ROC curve that offers the optimal tradeoff between sensitivity and specificity. The optimal cutoff for binarizing molecule-level predictions is 0.547. The light green arrow points out the corresponding location of the optimal molecule-level score cutoff on the ROC curve. The optimal cutoff for binarizing pair-level predictions is 0.013. The light orange arrow points out the corresponding location of the optimal pair-level score cutoff on the ROC curve.
model indicated all three inhibitors underwent bioactivation into quinone species including a surprising subset involving the isoxazole ring. This outcome contrasts with the enimines predicted with the AMD set of molecules; however, those reactive metabolites are not possible for these compounds. To infer and rank the most probable quinone metabolite structures, we used the more reliable atom-pair score for metabolites (Figure 5.3B) instead of the atom scores despite equivalent AUC performance (Figure 5.9). This performance difference may reflect unlikely outcomes, such as diene formation with both methyl groups when using only atom scores. Furthermore, atom-pair predictions that do not correspond to an inferred structure produced by the structure inference model are discarded and the retained predictions, which map to an inferred quinone structure, are ranked. We predicted glutathione adduct structures for the corresponding quinone species with the highest atom-level reactivity score and inferred possible bioactivation pathways for OXFBD02, OXFBD04, and I-BET151 as shown in Figures 5.10, 5.11, respectively. The number of reported pathways of most probable quinone metabolites and their most probable adducts was expanded to include all experimentally observed metabolites from the parent molecules, with a minimum of the three top pathways if the parent molecule had no experimentally observed metabolites. The highest predictions for bioactivations were relatively low and did not vary significantly when yielding different reactive metabolites. Moreover, the resulting metabolites with conjugated systems possessing a methylene were more reactive with glutathione than quinones in all cases.

5.3.4 OXFBD02 underwent extensive metabolism and bioactivation down competing pathways

In general, OXFBD02 bioactivations would involve sequential oxidations leading to formation of quinone species trapped by dansyl glutathione. As shown in Figure 5.12, initial hydroxylolation of OXFBD02 creates three pathways leading to mono-hydroxylated metabolites (M1, M2 and M3) that can undergo further oxidation into quinone species (Q1, Q2, and Q3) and in one case, an extended quinone-methide (EQ). Pathways 1 and 2 reflect hydroxylation yielding
an ortho-hydroquinone. Pathway 1 leads to an ortho-quinone and subsequently a pair of glutathione adduct isomers (A1a and A1b). In contrast, Pathway 2 bifurcates leading to either an ortho-quinone and a pair of glutathione adduct isomers (A2a and A2b), or a single glutathione adduct arising from an unusual extended quinone-methide (A4). Pathway 3 involves an initial formation of a para-hydroquinone followed by a para-quinone trapped into a pair of glutathione adduct isomers (A3a and A3b). Taken together, the occurrence of all pathways would lead to a 2:2:2:1 pattern of chromatographically resolved adduct peaks.

Experimental studies revealed extensive metabolism of OXFBD02 based on chromatographic resolution and combined detection of analytes by fluorescence and mass spectrometry (MS). The absence of an NADPH-regenerating system (negative control) indicated many background fluorescence signals from the dansyl glutathione reagent or other reaction constituents (Figure 5.13A). The complete reaction led to multiple new peaks in a 2:3:1 pattern that increased over time as expected for metabolites. All peaks were presumably dansylated glutathione adducts given that the parent molecule was spectrally silent under detection conditions. In
Chapter 5. Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors

![Diagram of predicted bioactivation pathways and subsequent adducts for OXFB molecules.](image)

**Figure 5.10:** Predicted bioactivation pathways and subsequent adducts for OXFB molecules. Top) Top six model-derived bioactivation pathways for OXFB02. The top six pathways split evenly between traditional (Q) and extended quinone-methides (EQ) involving the isoxazole and are ranked based on the indicated quinone model score. EQs were highly reactive with glutathione but traditional quinone reactivity was only moderate to low as reflected in the reactivity prediction values. Dashed arrows denote transformations whose predictions are below the model’s binarization threshold. Stars show predicted glutathione adducted metabolites observed experimentally. Bottom) The top six pathways split evenly between traditional (Q) and extended quinone-methides (EQ) and ranked based on indicated quinone model score. EQs were highly reactive to glutathione but traditional quinone reactivity was only moderate to low as reflected in the reactivity prediction values. Dashed arrows denote transformations whose predictions are below the model’s binarization threshold. Stars show predicted glutathione adducted metabolites observed experimentally.
Chapter 5. Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors

Figure 5.11: Modeled I-BET151 bioactivation pathways. Top three pathways led only to extended quinone-methides that are shown ranked based on the quinone model score. The dashed arrow denotes a transformation whose prediction was below the model’s binarization threshold. All three extended quinone-methides are highly reactive to glutathione leading to the indicated adducts.
Chapter 5. Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors

**Figure 5.12**: Competing bioactivation pathways for OXFBD series of BET inhibitors. OXFBD bioactivation pathways require an initial hydroxylation of the internal phenyl group. The three potential sites give rise to three competing pathways as indicated by circled numbers. Pathways 1 and 2 reflect hydroxylation yielding an ortho-hydroquinone. Pathway 1 leads to an ortho-quinone and a subsequent pair of glutathione adduct isomers (A1a and A1b). By contrast, Pathway 2 bifurcates leading to either an ortho-quinone and a pair of glutathione adduct isomers (A2a and A2b) or a single glutathione adduct arising from an extended quinone-methide involving the isoxazole (A4). Pathway 3 involves an initial formation of a para-hydroquinone followed by a para-quinone trapped into a pair of glutathione adduct isomers (A3a and A3b). Analysis of adducts then yields characteristic patterns for pairs of quinones except for the extended quinone in which there is one possibility.
support of that possibility is the elution of those peaks after the dansyl glutathione trap as reported for adducts in other studies [155, 186, 187]. MS analysis of analytes provided further evidence for OXFBD02 bioactivation into adducted quinone species. Initial total ion scans indicated OXFBD02 (295 m/z) oxidation into five mono-hydroxylated metabolites (311 m/z) and three di-hydroxylated metabolites (327 m/z). These hydroxylated metabolites underwent further oxidation leading to reactive quinone species and the extended-quinone methide trapped by dansyl glutathione.

Product ion mass spectrometry analyses confirmed expected parent masses and identified characteristic fragments for expected adducts. As observed by fluorescence (Figure 5.13A), scans for the adduct total ion mass of 849 m/z revealed chromatographic resolution of three analyte clusters (Figure 5.13C). There was a doublet of peaks followed by a triplet and then a single, late-eluting peak suggesting formation of adduct isomers. All analytes possessed the 742 m/z product ion after loss of the characteristic phenoxymethyl group for OXFBD02 indicating none of the adducts involved oxidation of that group. Moreover, putative adducts yielded signature product ions from fragmentation of dansyl glutathione: 234, 252, 361, 378, 487, 505, and 539 m/z (Figure 5.14) [186, 187]. Nevertheless, none of the product ions were distinguishing features among ortho- and para-quinone isomeric adducts or even that for the extended quinone adduct. We then leveraged the chromatographic 2:3:1 pattern for eluted adducts to infer their identification (Figure 5.13A and B). We inferred that the triplet of analytes reflected overlapping elution of the two doublets of ortho-quinone adducts (A1a/b and A2a/b) due to significant structural similarities. The larger central peak corresponded to additivity of fluorescence and mass responses for the co-eluted ortho-quinone adducts. The preceding peak doublet reflected the pair of para-quinone adducts (A3a/b). Finally, the lone, smaller, late-eluting peak was presumably the single extended quinone-methide adduct. Its elongated structure would interact more strongly with the C18 column than the other adducts resulting in the observed longest retention time. Definitive evidence for adduct identities would require nuclear magnetic resonance spectroscopy that is beyond the scope of this study. Nevertheless, the combination of fluorescence
Chapter 5. Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors

**Figure 5.13:** Fluorescence and MS chromatograms of putative quinone adducts for OXFBD02 and OXFBD04 from human liver microsomal reactions. A) OXFBD02 was bioactivated into six fluorescently and mass spectroscopically detectable dansyl-glutathione adducted quinone-species metabolites. B) OXFBD02 was bioactivated into three fluorescently and mass spectroscopically detectable dansyl-glutathione adducted quinone-species metabolites. In both panels, data displayed on the plot are as follows: no metabolic reaction (negative control, pink dotted line), 30 min reaction with fluorescence detection (blue solid line), 60 min reaction with fluorescence detection (black solid line), and 60 min reaction with mass detection (green solid line). The left Y-axis corresponds to data collected from fluorescence detection, whereas the right Y-axis corresponds to data collected from mass detection. Adduct identities are labeled according to the naming strategy displayed in Figure 5.12 and based on patterns of eluted adducts as discussed in Results.
and mass responses with the chromatographic pattern of elution indicated OXBFD02 underwent bioactivation mainly into reactive ortho- and para-quinones and to a lesser extent, an atypical extended quinone-methide.

### 5.3.5 Metabolism and bioactivation of OXFBD04 was very limited

The possible metabolic pathways for OXFBD04 are the same as those for OXFBD02 due to structural similarities (Figure 5.12). Nevertheless, substitution of a phenyl group with pyridine led to significantly different extents of metabolism and bioactivation. The fluorescence chromatogram from the human liver microsomal reactions indicated no distinguishing peaks from background (Figure 5.13B), yet labeled adducts could be beyond the limits of detection and/or co-eluting with background peaks. Complementary total ion scans showed that OXFBD04 (296 m/z) underwent metabolism into only two mono-hydroxylated metabolites (312 m/z). Initial scans of proposed quinone species and extended quinone-methide adduct total ion masses at 850 m/z yielded a 2:1 pattern with two peaks of similar intensity followed by a smaller third one (Figure 5.13D). Product ion mass spectrometry resulted in several characteristic dansyl glutathione fragments (324, 487, 505, and 539 m/z) for putative trapped quinone species adducts (Figure 5.15), albeit with less abundant responses than those observed for OXFBD02 adducts. As with OXFBD02, no product ions were distinguishing features among possible adducts. Nevertheless, the elution pattern was consistent with a pair of quinone adduct isomers followed by the extended quinone-methide adduct. The identity of the isomeric adducts are likely A2a/b due to their origination from the same mono-hydroxylated metabolite (M2). In following, OXFBD04 seems to preferentially undergo metabolism and subsequent bioactivation down a single pathway among three possibilities.
FIGURE 5.14: Mass spectroscopic product ion spectra for predicted quinone metabolite adducts for OXFBD02. The results reflect the integrated spectra and product ion fragments for peaks eluting in a 2:3:1 ratio based on intensity in Figure 5.13C. Characteristic dansyl glutathione fragments m/z 234, 252, 361, 378, 487, 505, and 539 have been reported previously in negative ion mode by Gan and colleagues [186, 187]. Site of hydroxylation did not impact observed fragment patterns. Assignments for the respective fragments are indicated in the accompanying molecular structures for the corresponding adducts. Peak assignments for adducts (Figure 5.12) are discussed in detail under Results. Panels A and B shows spectra for the probable pair of adduct isomers (A3a/b) from a traditional quinone in Pathway 3 (Q3). Panels C, D and E depict the unresolved pairs of adduct isomers from Pathway 1 (A1a/b) and Pathway 2 (A2a/b). Finally, spectra for the last eluting peak corresponds to the adduct (A4) from the extended quinone methide (EQ) in Pathway 2.
FIGURE 5.15: Mass spectroscopic product ion spectra for predicted quinone metabolite adducts for OXFBD04. The results reflect the integrated spectra and product ion fragments for peaks eluting in a 2:1 ratio based on intensity in Figure 5.13D. Characteristic dansyl glutathione fragments m/z 234, 252, 361, 378, 487, 505, and 539 have been reported previously in negative ion mode by Gan and colleagues [186, 187]. Site of hydroxylation did not impact observed fragment patterns. Assignments for the respective fragments are indicated in the accompanying molecular structures for the corresponding adducts. Peak assignments for adducts (Figure 5.12) are discussed in detail under Results. Panels A and B show spectra for the probable pair of adduct isomers (A2a/b) from a traditional quinone (Q2), while Panel C shows that for the lone, lower intensity peak for A4 arising from the extended quinone methide (EQ).
5.3.6 There was no evidence of I-BET151 bioactivation despite limited metabolism

Analysis of human liver microsomal reactions for I-BET151 indicated limited metabolism. As observed for OXFBD04, there were no apparent unique fluorescent peaks generated during the metabolism of I-BET151 as a function of time. Subsequent total ion scans showed elution of the parent drug (417 m/z) and two mono-hydroxylated metabolites (433 m/z), but no observable di-hydroxylated or tri-hydroxylated metabolites. Moreover, total ion mass spectrometry revealed no peaks corresponding to predicted parent masses for 952 or 984 m/z corresponding to possible glutathione-reactive quinone species metabolites from I-BET151 metabolism. Taken together, MS and fluorescence analyses failed to reveal any evidence for quinone species adduct suggesting I-BET151 does not undergo bioactivation.

5.4 Discussion

5.4.1 Quinone model predictions for reactive non-quinones involving 4-amino-5-methyl-isoxazole

Despite its name, our quinone model was able to predict bioactivation of molecules containing 4-amino-5-methyl-isoxazole into rare, and yet observed enimines [182, 183]. This new finding could reflect model bias from the training set; however, the high diversity in structure within and between test and training sets made that possibility not likely. In fact, diversity in the training data has taught the model to adjust the magnitude of predictions due to alkyl substituent effects. Replacement of the methyl with an ethyl group on the isoxazole decreased model scores. This effect reflects the hindrance of the orbitals of the ethyl group to adopt the ethene structure versus methylene for the methyl group as observed in the reported isoxazole enimines. Moreover, the model accurately predicted the decrease in bioactivation potential between phenyl and pyridine to form quinones when attached to the isoxazole. The trend may reflect the impact
of electron withdrawing effects of the pyridine nitrogen that suppresses oxidation typically by cytochromes P450 [192, 193, 194]. After bioactivation of the 4-amino-5-methyl isoxazole, the exposed methylene was very reactive and targeted for adduction by glutathione as reported in the literature [182, 183]. This outcome demonstrated potential for the quinone model to predict reactive metabolites other than quinones within the broad class of conjugated electrophiles and thus could have more applications in predicting bioactivations.

5.4.2 Modeled bioactivations of the 3,5-dimethyl isoxazole present in BET inhibitors

As a practical application, we predicted the bioactivation of a series of 32 established and under development BET inhibitors to determine whether 3,5-dimethyl isoxazole was able to form a reactive metabolite. The isoxazole lacks the 4-amino group, which played a critical role in the previously reported isoxazole bioactivations [164, 140]. For this test set, all inhibitors displayed decreased bioactivation risk relative to molecules known to form enimines. This finding suggests that the 4-amino group on the isoxazole favored bioactivation according to the model. While 3,5-dimethyl isoxazole led to lower bioactivation potential, the modeled likelihoods were greater than that for isoxazole-containing molecules contained within AMD that do not undergo bioactivation. Model scores varied with substituents on the distal phenyl ring indicating the ability of the model to scale their impact on bioactivation. Specifically, methoxy groups decreased bioactivation, while fluorine and chlorine atoms had no effect regardless of position on the ring. In addition, substitution of that phenyl group with pyridine led to a general decrease in overall bioactivation predictions. These differences may reflect effects or a lack of effects on chemical reactions as well as enzyme specificities but require further study to validate the trends and determine underlying causes. Consequently, we selected three of these BET inhibitors, namely, OXFBD02 and OXFBD04 (in-development drug leads) [184] and I-BET151 (first generation BET inhibitor) [161, 162] for more in-depth modeling analyses and importantly, confirming predicted bioactivations with experimental studies.
5.4.3 Metabolism and bioactivation of OXFBD02 and 04, but not I-BET151, into quinones

Molecular structures among the selected inhibitors impacted bioactivation possibilities and subsequent interpretation of the findings. OXFBD02 and 04 share a common isoxazole-phenyl scaffold and differ only in the presence of an additional distal phenyl or pyridine substituent, respectively. For the OXFBD molecules, the structural difference led to a 10-fold increase in metabolite stability for OXFBD04 over OXFBD02 in a half-life assay [184]. Our metabolic studies corroborated the observed effect on metabolism and demonstrated the likely dominance of cytochromes P450 in metabolism based on the requirement for NADPH [195]. Surprisingly, the suppressed bioactivation was not due to deactivating effects of nitrogen in the ring [192, 193, 194], because the distal pyridine and phenyl substituents did not undergo metabolic activation into a reactive quinones (Figure 5.12). The observed effect has to be indirect such as modulations of enzyme specificity. For the pair of OXFBD molecules, the overall model scores reflected a decrease in metabolic bioactivation of OXFBD04 over OXFBD02. This finding aligned with the phenyl to pyridine substituent effects on quinone bioactivation observed with molecules present in AMD. In other words, the quinone model successfully applied lessons learned from molecules very different from these BET inhibitors even if the mechanism for this effect remains unclear. I-BET-151 also possesses a distal pyridine substituent but the scaffold is very different with the isoxazole linked to a tricyclic moiety making comparisons not possible with OXFBD molecules.

Metabolism yielded two general possible outcomes for bioactivations; molecules underwent oxidation into traditional ortho- or para-quinones and novel extended quinone-methides involving the isoxazole. For the top six predictions, modeled OXFBD bioactivations reflected multi-step reactions resulting in an even split between both types of quinones and their glutathione adducts. Among them, model predictions highly favored formation of extended quinone-methides. Overall, four of the six predicted adducts were verified by experimental studies. Nevertheless,
experimental studies showed that the dominant OXFBD bioactivations were formation of traditional quinones with extended quinone-methides being minor metabolites. Moreover, the efforts provided more insights and evidence for reaction steps and specific pathways. The extensively metabolized OXFBD02 yielded mono- and di-hydroxylated metabolites that ultimately fed into three competing bioactivation pathways (Figure 5.12). These possibilities arose from hydroxylation of the phenyl scaffold at one of three sites. Further oxidation yielded a reactive quinone species with two sites of attack by glutathione, so that the subsequent glutathione trapping led to pairs of adduct isomers. Importantly, hydroxylation of the phenyl ring para to the isoxazole created an alternate bioactivation pathway leading to an extended quinone-methide involving both isoxazole and phenyl rings. By contrast, suppressed metabolism of OXFBD04 yielded just two mono-hydroxylated metabolites and corresponding reactive quinones in Pathway 2 (Figure 5.12). The metabolites of other pathways for OXFBD04 were beyond the limit of detection for our highly sensitive assay or simply not present. For I-BET151, modeled bioactivations were similar in magnitude as those for OXFBD molecules, yet pathways led only to extended quinone-methides with a highly reactive methylene group. Unlike those predictions, experimental studies showed minimal I-BET151 metabolism leading to some hydroxylated metabolites and no evidence of any trappable reactive metabolites. Taken together, modeling was capable of predicting a variety of quinones including novel extended quinone-methides but fell short of accurately scaling the likelihood of their formations. Nevertheless, we report the first experimental evidence for novel extended quinone-methides involving the 3,5-dimethyl isoxazole scaffold.

5.4.4 Novelty and relevance of extended quinone-methides

Energetically, traditional quinones would be more accessible to oxidative transformations. Those reactions involve breakage of the aromaticity of the phenyl ring only as opposed to the phenyl and isoxazole rings for the extended quinone-methide. In fact, experimental evidence for extended quinone species is rare [196]. The most well-known examples include the selective
Chapter 5. Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors

estrogen receptor modulators (SERMs) raloxifene and arzoxifene, in which bioactivation of the benzothiophene scaffold results in extended quinone species with very short half-lives and possibly minimal toxicity [197, 198, 199, 200, 201]. By contrast, bioactivation of the estrogen diethylstilbestrol into an extended quinone species may contribute to teratogenic effects of the drug leading to its withdrawal from the market in 1971 [202, 203]. It is possible that extended quinone species are more prevalent but escape detection due to short half-lives and/or insufficient experimental designs to capture their formation. Additionally, there is no uniform structure for an extended quinone species making detection more challenging. In our case, the quinone model predicted that all three BET inhibitors underwent bioactivation into extended quinone species and we experimentally validated their occurrence for the two OXFBD molecules. Those findings suggest that formation of the extended quinone-methide may be inherent with the drug scaffold albeit a very minor pathway. Nevertheless, if model scores correlate with toxicity, then these results may have implications for these drug leads. The modeled bioactivation scores (0.51 and 0.59) were similar in magnitude to that for estradiol (0.60) from a previous study by our group [119]. Estradiol is well-known to undergo bioactivation and induce toxic risks suggesting those possibilities may exist for OXFBD molecules [204].

5.4.5 Advances in isoxazole bioactivation studies revealed current limitations of those efforts

These modeling and experimental studies provided an uncommon, concerted effort to study isoxazole bioactivation, but indicated gaps remain. While very rapid and accessible, modeling did not accurately scale the relative probabilities of traditional quinones and extended quinone-methides based on model scores. The molecule-level model output is well-calibrated, but atom-pair-level scores can lead to underconfident predictions and are not probabilistic. We partially resolved this problem by computing an optimal threshold for binarizing scores. In future applications, it may be advantageous to interpret scores as probabilistic for assessing confidence in predictions. Improving calibration of pair-level model predictions may also improve rankings
Chapter 5. Novel bioactivation of isoxazole-containing bromodomain and extra terminal domain (BET) inhibitors

of inferred metabolites to better reflect experimental results. Another non-intuitive feature of the quinone model is that atom and atom-pair scores can be very different. Sometimes high predictions do not correspond to valid reactive metabolite structures, with atom scores having less reliability than atom-pair scores. Validity of predictions was improved by using the structure inference model to filter out highly predicted sites that do not produce an inferred structure. For the experimental studies, analysis of reactive metabolites relied on set reaction conditions and no steady-state kinetics that could be used to extrapolate the potential in vivo relevance of bioactivation pathways [44, 205]. Moreover, there are no reported studies investigating and confirming that bioactivation of isoxazole-containing molecules contributes to in vitro or in vivo toxicity. Nevertheless, knowledge of the relative significance of bioactivation pathways could aid in developing future studies to resolve those issues.

5.5 Conclusion

Taken together, our computational and experimental studies yielded the first bioactivation evidence for 3,5-dimethyl isoxazole-containing BET inhibitors. In the process, we demonstrated that the deep neural quinone model prediction of enimines indicates the ability to model multiple types of conjugated electrophiles and thus, have broader applications. Improvements in model scaling events relative to one another would further expand on those possibilities. For the BET inhibitors, the occurrence of bioactivations shows a need for subsequent studies to establish relevance in toxicological mechanisms and these findings provide guidance on the potential molecular initiating events. Importantly, our combined investigative approach provides a method to not only identify these risks but also facilitate the development of BET inhibitor leads less prone to bioactivations.
Chapter 6

Modeling the Bioactivation and Subsequent Reactivity of Drugs


*Equal contribution as first author

6.1 Introduction

Adverse Drug Reactions (ADRs) are a major challenge for global public health. Independent investigations of diverse populations implicated ADRs in 6.5%[206], 8.1%,[207], 8.7%[208], and 12.8%[209] of hospital admissions. Similarly, a meta-analysis of U.S. hospital prospective studies found that 6.7% of patients had severe ADRs, with a fatality rate of 0.32%.[210] Extrapolating from those results, the FDA estimated that annual ADRs in the U.S. cause over 2,216,000 hospitalizations and more than 106,000 deaths.[211] Furthermore, predicated on the meta-analysis’s accuracy, the FDA proposed that ADRs are the fourth leading cause of death in the U.S., exceeding automobile deaths, diabetes, AIDS, pulmonary disease, and pneumonia.[211] Some ADRs are traceable to the pharmacological effects of certain drugs, and this
mechanistic understanding can inform efforts to reduce risk.[212] However, a subset of ADRs—idiosyncratic adverse drug reactions (IADRs)—have elusive etiologies.

These IADRs strike seemingly at random, with unpredictable and often severe symptoms. Most commonly, IADRs cause liver disorders but can also induce dangerous skin diseases, including Stevens-Johnson syndrome and toxic epidermal necrolysis, as well as dangerous blood disorders such as agranulocytosis or aplastic anemia.[213, 214, 215, 216, 217, 218, 219, 220, 221, 222, 223] In the U.S., IADRs are responsible for about half of all acute liver failure cases and 15% of liver transplants.[224] Nevertheless, IADRs are rare overall, only occurring in about 1 in 10,000 to 1 in 100,000 patients.[225] As a result, many IADR-causing drugs can slip through all stages of preclinical trials, which even in their largest phase generally only have about 3,000 patients.[213] After approval and market release, however, exposure to much larger patient populations can reveal a drug’s hidden risk. Indeed, already-approved drugs are most commonly withdrawn from the market due to intolerable numbers of IADR cases.[226, 227, 228, 229, 230] Even if a drug is not withdrawn, it may be labeled with a “black box” warning by the FDA, significantly curtailing its profitability.[231]

Devising early-detection methods for IADRs would reduce patient morbidity and mortality. Furthermore, pharmaceutical developers would avoid heavily investing in drugs that ultimately are too risky to use. Unfortunately, IADRs are generally intractable to study and difficult to replicate in humans or animal models. It is not clear why IADRs only affect some individuals, or why the same drug can cause different IADRs in different patients. However, a growing body of evidence suggests that many IADRs are induced by a specific mechanism: bioactivation, the focus of this study (Figure 6.1).

In bioactivation, enzymes convert drugs into electrophilically-reactive metabolites that covalently bind to nucleophilic sites within biological macromolecules, including DNA and (off-target) proteins. Metabolite-DNA adducts may be mutagenic or even carcinogenic[224, 242,
Chapter 6. Modeling the Bioactivation and Subsequent Reactivity of Drugs

Figure 6.1: This study models four common bioactivation pathways: quinone formation, nitroaromatic reduction, thiophene sulfur-oxidation, and epoxidation. Top row, lumiracoxib, a cyclooxygenase-2 selective inhibitor, was withdrawn from several countries after several cases of severe liver damage.[232, 233] This toxicity was traced to the formation of a reactive quinone-imine metabolite that conjugates to off-target proteins, inducing deleterious immune responses.[232, 233] Second row, nitrofurantoin, an antibiotic, carries a risk of acute liver failure [234], which is thought to be caused by reduction of nitrofurantoin’s nitroaromatic group to a reactive nitroso.[235] Third row, zileuton, a 5-lipoxygenase inhibitor used to treat asthma, has been restricted in its use due to rare cases of severe hepatotoxicity, which has been traced to oxidation of the sulfur in its thiophene motif, producing a highly reactive sulfur-oxide.[236] Bottom row, furosemide, a diuretic, confers a risk of idiosyncratic hepatitis due to production of a reactive epoxide metabolite.[237, 238, 239, 240, 241]
and metabolite-protein adducts can disrupt protein function or trigger toxic immune responses. [244, 20, 21] In fact, many IADRs have been linked to overzealous autoimmune attacks set off by the production and covalent binding of reactive metabolites. Due to the threat conferred by bioactivation, drug developers strive to avoid advancing candidates that produce reactive metabolites, in order to reduce the risk of investing in IADR-causing drugs.

Screening assays for reactive metabolites are often used, but have some limitations. Reactive metabolite trapping studies [183] and covalent binding studies [69, 245] are commonly used techniques to detect and understand bioactivation. Trapping studies detect if conjugates have formed and can be used to characterize reactive metabolite structure. Typically, a trapping agent, e.g., the use of glutathione (GSH) or cyanide as proxies for protein and DNA, that has a high likelihood of conjugating to reactive metabolites is selected. [246] Formation of a GSH conjugate indicates the presence of a reactive metabolite and the GSH conjugate can be detected via mass spectrometry. [183] On the other hand, covalent binding studies are able to quantify the extent of conjugation through the use of a radiolabeled drug. [247] In general, screening assays may not accurately reflect endogenous metabolism, consume time and resources, and require physical synthesis of each compound under consideration.

Instead, computational models have the potential to rapidly screen possible structures for bioactivation risk, thereby flagging problematic molecules or providing a short list of molecules for experimental validation. A widely used and simple method for identifying problematic structures is to cross-reference a drug candidate to a list of structural alerts. [237] Usually, structural alerts contain substructures that have known bioactivation mechanisms that contribute to documented toxicity risk. However, structural alerts have limited utility - they are not sufficient to declare a molecule as toxic and they can misclassify toxic molecules as safe. Since structural alerts are determined retrospectively, they have no predictive power for new and understudied motifs. [33] However, there is a lack of more complex models that can generalize to understudied cases and account for contextual nuances of drug metabolism.
In this study, we build a model that jointly models metabolism and reactivity, thereby producing bioactivation predictions. We model four types of metabolism that often produce reactive metabolites: quinone formation, nitroaromatic reduction, thiophene sulfur-oxidation, and epoxidation (Figure 6.1). These pathways are chosen because we have well-developed models for these metabolic routes, including an accurate model of quinone formation[119], epoxidation[248], and a phase I metabolism model that includes predictions for nitroaromatic reduction and thiophene sulfur-oxidation. [54] Since the reaction type is accounted for during prediction, we can infer the structure of the resultant product using a structure inference model. [60] Next, to each inferred structure we apply a previously developed model for predicting reactivity to protein or GSH. Once we have predictions regarding metabolism of the input molecule and reactivity of its inferred metabolites, we can use these predictions to train a deep neural network that will predict bioactivation at both the molecule- and pathway-level. Molecule-level bioactivation aims to predict whether the input molecule will undergo bioactivation, while pathway-level bioactivation aims to predict which of the metabolic transformations and inferred structures will lead to bioactivation.

6.2 Methods

6.2.1 Bioactivation Training Data

We assemble a heterogeneous data set of bioactivation reactions from the literature-derived Accelrys Metabolite Database (AMD). Each reaction takes place in humans, human cells, or human liver microsomes. Four types of metabolic transformations are extracted: quinone formation, epoxidation, nitroaromatic reduction, and thiophene sulfur-oxidation. Each of these metabolic transformations are well-known bioactivation mechanisms and, in previous work, we built models that predict whether molecules will be subject to each type of metabolism. Quinone formation [119] and epoxidation [248] were modeled independently, and nitroaromatic reduction and thiophene sulfur-oxidation were included in a model of diverse phase I reactions. [54]
In total, we extract from the experimental data 210 quinone formations, 174 epoxidations, 4 nitroaromatic reductions, and 15 thiophene sulfur-oxidations. For each parent molecule, we enumerate all possible metabolites for each of these four pathways, producing 7580 quinone formations, 6100 epoxidations, 48 nitroaromatic reductions, and 15 thiophene sulfur-oxidations. We then select for bioactivated parent molecules by taking the set of reactions for each of the four types of metabolic transformations and retaining those parent molecules whose possible metabolites are known to directly conjugate to protein or glutathione (GSH). The bioactivated parent molecule’s metabolic transformation and resultant, conjugating metabolite are labeled as an experimentally observed bioactivation pathway. However, the same parent molecule can be present multiple times if it is indicated with multiple possible, experimentally observed bioactivation pathways. To handle this case, we merge all duplicate parent molecules into a single representation per unique parent molecule, with all of its experimentally observed bioactivation pathways labeled.

In contrast with an experimentally observed bioactivation pathway, a parent molecule might undergo one of the 4 studied metabolic transformations, but the resultant metabolite is not known to conjugate to protein or GSH. Cases where a bioactivated parent molecule only contains bioactivation pathways would not be a good test of the model’s ability to predict the correct pathway within the bioactivated parent molecules. Instead, we want the model to clearly distinguish between pathways that are experimentally observed to bioactivate and pathways that are experimentally unobserved to bioactivate. As a final step, we filter out any molecules that did not have at least one experimentally unobserved bioactivation pathway. This procedure produces a total of 332 bioactivated training molecules.

We also select molecules not experimentally observed to be bioactivated. In this case, we are referring to all molecules that could potentially undergo at least one of the 4 metabolic transformations being studied that precede reactive metabolite formation, but whose resultant metabolites are not experimentally observed to conjugate to protein or GSH. To select these molecules, we start with the same set of reactions from the AMD. Next, we filter out all molecules that
are intrinsically reactive: those that are known to directly conjugate to protein or GSH. Additionally, we remove all molecules that belong to the set of bioactivated molecules. Finally, we select all molecules with at least one possible bioactivation pathway among the 4 metabolic transformations considered in this study. This procedure produces a large pool of 30690 negative molecules.

From the pool of negative molecules, we randomly select 332 molecules—equivalent to the total number of bioactivated molecules—to form a final data set of 664 molecules. The remaining negative molecules are randomly batched into external test sets, each consisting of 332 molecules, with the same proportion of molecules representing each pathway as the negative molecules from the training set. None of these molecules are considered during training. Instead, they represent additional tests for the final trained model to evaluate how well the model generalizes to unseen cases.

Under the AMD licensing agreement, we are not able to share the exact chemical structures of the product metabolites and their corresponding reactions used in the data set. However, we provide all reaction and molecule AMD registry numbers, as well as the parent molecule structures, in the Supporting Information, which is enough information to reconstruct the data set and replicate our results.

### 6.2.2 DrugBank

DrugBank is used as another source of external data and was downloaded from www.drugbank.ca (accessed July 3rd, 2020). We later use the DrugBank external data set during evaluation of the final bioactivation model’s ability to generate hypotheses for the toxicity drivers of withdrawn drugs. To identify withdrawn, small-molecule drugs, we filter entries designated by both the "small-molecule" annotation and the "withdrawn" group. This release of DrugBank includes the structures of 221 drugs. Prior to submission to the model, we remove any drugs also present in our training data, resulting in a total of 201 molecules. We provide all withdrawn drugs assessed by the model and their structural information in the Supporting Information.
### 6.2.3 Bioactivation Descriptors

For this study, we synthesize several previous models of metabolism and reactivity to design specific bioactivation descriptors that are inputs to a neural network (Figure 6.2). Using previously designed models for quinone formation [119], epoxidation [248], nitroaromatic reduction, and thiophene sulfur-oxidation [54], formation scores are computed for each possible transformation for a given input molecule. The formation score for each reaction type represents the likelihood of the parent molecule undergoing the respective metabolic transformation. Next, the actual structures of each of these possible metabolites is generated using an in-house metabolite structure predictor. [60] After enumeration of the metabolite structures, the atom-level reactivity scores with respect to both glutathione (GSH) and protein are computed for both the substrate and the product molecule. [120] Next, the atom-level reactivity deltas between the metabolite and substrate are calculated for both GSH and protein by subtracting the atom’s reactivity prediction in the product from the atom’s reactivity prediction in the substrate. Finally, a bioactivation score is computed by multiplying the maximum atom-level reactivity delta by the corresponding formation score.

In some cases, molecules being evaluated were present in the training sets for the models used to generate the formation scores or atom-level reactivity scores. To ensure unbiased evaluation of each molecule, we use versions of the phase I metabolism models and reactivity models that were retrained with all of the molecules in their original training sets except for the molecule being evaluated. The predictions from the newly trained models are then used to derive the bioactivation descriptors. We also compute simple molecule descriptors, such as molecular weight and the total number of atoms. Overall, we generate 20 descriptors for each possible metabolite, including 5 bioactivation descriptors (Table 6.1) and 15 molecule descriptors (Table 6.2).
FIGURE 6.2: Bioactivation descriptors were computed using site-level metabolism predictions, metabolite structure predictions, and reactivity predictions. Top, the predictions generated by a previously developed epoxidation model[248] are visualized on styrene. The colored shading indicates site-level epoxidation scores, which reflect the probability that an epoxide will form at each possible location within styrene. Bottom, using a previously developed metabolite structure generator,[60] the exact epoxide structure corresponding to the highest site-level epoxidation score was generated. Next, we applied our previously published reactivity model [120], which has proved useful in other studies as well [249, 250], to predict the atom-level reactivity of both the substrate and the metabolite. Finally, by tracking each atom’s reactivity score in the metabolite from the corresponding atom in the substrate, we calculated atom-level reactivity deltas. Metabolism renders the two carbon atoms more reactive, with predictions increasing from 0.19 to 0.41 and from 0.56 to 0.73.

TABLE 6.1: Bioactivation descriptors used to predict bioactivation pathways.

<table>
<thead>
<tr>
<th>Bioactivation Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formation Score</td>
</tr>
<tr>
<td>GSH Max Atom Reactivity Delta</td>
</tr>
<tr>
<td>GSH Bioactivation Score</td>
</tr>
<tr>
<td>Protein Max Atom Reactivity Delta</td>
</tr>
<tr>
<td>Protein Bioactivation Score</td>
</tr>
</tbody>
</table>
Chapter 6. Modeling the Bioactivation and Subsequent Reactivity of Drugs

<table>
<thead>
<tr>
<th>Molecule Descriptors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molar Refractivity</td>
</tr>
<tr>
<td>Molecular Weight</td>
</tr>
<tr>
<td>Number of Aromatic Bonds</td>
</tr>
<tr>
<td>Number of Bonds</td>
</tr>
<tr>
<td>Number of Double Bonds</td>
</tr>
<tr>
<td>Number of Heavy Atoms</td>
</tr>
<tr>
<td>Number of Hydrogen Bond Acceptors Definition 1</td>
</tr>
<tr>
<td>Number of Hydrogen Bond Acceptors Definition 2</td>
</tr>
<tr>
<td>Number of Hydrogen Bond Donors</td>
</tr>
<tr>
<td>Number of Hydrogens</td>
</tr>
<tr>
<td>Number of Rings</td>
</tr>
<tr>
<td>Number of Single Bonds</td>
</tr>
<tr>
<td>Number of Triple Bonds</td>
</tr>
<tr>
<td>Octanol/Water Partition Coefficient</td>
</tr>
<tr>
<td>Topological Polar Surface Area</td>
</tr>
</tbody>
</table>

### 6.2.4 Combined Path- and Molecule-level Bioactivation Model

Descriptors are computed to create inputs for machine learning algorithms, which find mappings between vectors of numbers, known as features, and labeled examples, known as targets. In this context, the descriptors that compose our features and our targets are a binary column indicating whether a pathway was experimentally observed as a bioactivation pathway in the AMD.

The bioactivation model’s architecture is representative of a feedforward neural network, in which an input is passed into a layer of fully-connected neurons. Each neuron has a weight parameter associated with each of its connections to the input or neurons in the previous hidden layer. An affine combination is computed whereby the weight parameter is multiplied by the input feature to a neuron and adjusted by a learnable bias parameter. The affine combination is passed into a nonlinear function, such as a sigmoid function or a rectified linear unit (ReLU), and the neuron output is the input to each neuron in the next hidden layer. The final layer aggregates the outputs of the previous layer and results in a final model prediction. In comparison, a logistic regressor computes an affine combination between the input features and the model parameters, which is then fed into a sigmoid function that squashes it in the range of 0 to 1. For
both methods, a loss function is used to compute the error between the model output and the true label. The computed error is used to update the model’s parameters to minimize the loss function.

The bioactivation model is a feedforward neural network, with one molecule layer, one input layer, two hidden layers, and two output layers (Figure 6.3). The first output layer calculates pathway bioactivation scores (PBS or pathway score) and the second output layer computes a single molecule bioactivation score (MBS or molecule score) for each input molecule.

We trained this network in two stages. In the first stage, we trained the pathway-level network to compute accurate PBS values. For a given input molecule, each of its experimental or model-inferred pathways is considered as a possible bioactivation pathway. Each pathway is represented by a vector containing the bioactivation descriptors that define it. Thus, the data set passed into the pathway-level network is a matrix with one column per bioactivation descriptor and one row per pathway. Using the first hidden layer, the pathway-level network learns an association between each pathway and its bioactivation descriptors against a binary target vector where the experimentally-observed bioactivation pathways are labeled with a 1.

To train the model, we use gradient descent on the cross-entropy error. In this procedure, the model’s weights are gradually adjusted to assign high pathway scores for experimentally observed bioactivation pathways, and low pathway scores for all other pathways. Each pathway score ranges from 0 to 1 and reflects the probability of a specific bioactivation event at specific sites within a molecule, where each pathway is represented by its 5 bioactivation descriptors. Effectively, the pathway score predicts which pathway(s) will lead to bioactivation, where a pathway consists of the input molecule, one of the 4 metabolic transformations, and the structure and conjugation potential of the resultant metabolite.

The described procedure can generate several different PBS, which may share some atoms in common if the different pathway’s involve multiple metabolic transformations occurring at the same sites. For example, the atoms that make up sites of aromatic epoxidation are often themselves possible sites of quinone formation. This phenomenon can make visualization of
multiple PBS, each of which designates a prediction regarding a different bioactivation pathway, a difficult task. To aid interpretability, we devised a method for visualizing PBS for a given input molecule (Figure 6.4). We first map all PBS to the atom-level. This is straightforward: for a quinone pair prediction, each atom in the pair would be assigned the PBS. Similarly, both atoms of a bond prediction (nitroaromatic reduction and epoxidation) would be assigned the same PBS. This results in a vector of PBS for each atom in the molecule, where each vector entry represents the PBS for each of the atom’s possible bioactivation pathways. Next, we compute the final scores on each atom by using the probabilistic OR function across the vector of predictions assigned to that atom.

In the second training stage, we trained the molecule-level output layer to compute MBS values. Using the second hidden layer, the molecule-level network learns an association between the 15 molecule descriptors, the top 5 pathway scores, and the input molecule’s classification as being experimentally observed to bioactivate or not. Each row of the input data matrix represents a molecule and each column represents either a molecule descriptor or one of the top 5 pathway scores. Similar to the first stage, the weights of the network are trained using gradient descent on the cross-entropy error to assign bioactivating molecules with higher molecule scores than those for non-bioactivating molecules. Each molecule score reflects the probability of the input molecule undergoing bioactivation by any of the pathways considered and, like pathway scores, ranges from zero to one.

To produce pathway and molecule scores for the entire data set, we use a standard practice in machine learning for simulating performance on external data: cross-validation. In this procedure, any metabolically-related molecules are withheld together, and the model is trained on the remaining data. Molecules were separated into metabolically related clusters based on connections through metabolic reactions in the AMD – each cluster was comprised of a molecule and all of its detected parent and sibling molecules. Next, the trained model predicts the pathway scores of the withheld molecules. In total, there are 569 groups of related molecules, so the cross-validation procedure entails training 569 individual models. This process guarantees that
First, the pathway-level network is trained using bioactivation pathway data, where sites involved in metabolic processes that lead to bioactivation are labeled positive.

Second, the molecule-level network is trained using data where bioactivated molecules are labeled positive.

Figure 6.3: Structure of the bioactivation model. The diagram on the left demonstrates the flow of data through the model. The model consists of one molecule layer, one input layer, two hidden layers, and two output layers. First, several descriptors are calculated from an input molecule’s structure. Each pathway has their own set of bioactivation descriptors, and these descriptors are submitted to the hidden layer, which computes pathway bioactivation scores (PBS or pathway score) for the input molecule and each of its pathways. Each pathway score ranges from zero to one, reflecting the probability of a specific bioactivation mechanism at a specific site within the input molecule. For this study, we enumerated four types of metabolic transformations that precede bioactivation, including all possible quinone formations, epoxidations, nitroaromatic reductions, and thio- phenes sulfur-oxidations. Next, molecule descriptors and the top 5 pathway scores are submitted to a second hidden layer, which computes a molecule bioactivation score (MBS or molecule score) for the input molecule. Molecule scores also represent the probability of a molecule undergoing bioactivation by any of the pathways considered and, like pathway scores, ranges from zero to one. A chemical structure is represented by the molecule node. The other circles are probabilistic scores between 0 and 1. Blocks are vectors of real numbers. The stack of plates marked “Pathways” represents that the pathway-level module is replicated across all possible metabolites and their corresponding pathways. On the right, site-level data are illustrated on the top (with sites of predicted bioactivation circled) and molecule-level data are illustrated on the bottom (with the molecules of predicted bioactivation circled). Model output for two pairs of highly similar molecules are illustrated.
FIGURE 6.4: Visualization of pathway bioactivation is elusive because it requires consolidating multiple PBS that can overlap at the same site. Epoxidations and nitroaromatic reductions entail bond cleavage. However, quinones take place on atom pairs, and thiophene sulfur-oxidation occurs on a single sulfur atom. Consequently, while each of these predictions types can be visualized on separate structures (Figure 6.1), this paradigm does not make for easily-understood predictions. Instead, we first map all PBS to the atom level. This is straightforward: for a quinone pair prediction, each atom in the pair would be assigned the initial score. Similarly, both atoms of a bond prediction (nitroaromatic reduction and epoxidation) would be assigned the same initial score. While this mapping moves closer to an interpretable result, the same atom can still be subject to multiple possible bioactivation pathways. For instance, the atoms making up sites of aromatic epoxidation are often themselves possible sites of quinone formation. Therefore, in the last step prior to visualization, we compute the scores on each atom by using the probabilistic OR function across all predictions that included the current atom under consideration. Top left, an example of the resulting, final visualization of pathway bioactivation. Top right and bottom, the underlying metabolism and reactivity processes that constitute some of the possible bioactivation pathways extending from the visualized substrate.
Chapter 6. Modeling the Bioactivation and Subsequent Reactivity of Drugs

each molecule’s predictions are computed by a model that does not contain information about that molecule or closely related molecules.

6.3 Results and Discussion

The bioactivation model’s performance and capabilities are investigated in the following sections. First, we focus on the model’s pathway bioactivation scores (PBS or pathway score). Using several metrics, we quantify how well pathway scores predict the correct pathway(s) within bioactivated molecules. Next, we turn our attention to the model’s second output layer, which produces molecule bioactivation scores (MBS or molecule score). Using similar methodologies as the pathway-level analysis, we measure the performance of molecule scores by several standards. Finally, we use the final bioactivation model to enumerate hypotheses for the toxicity drivers of drugs with currently unknown or poorly understood toxicity mechanisms.

6.3.1 Bioactivation Pathway Prediction Accuracy

For bioactivated molecules, knowledge of their specific bioactivation pathways yields potentially fundamental insights about their possible toxicity mechanisms. Bioactivation takes place at specific sites within molecules, and forms specific reactive metabolite structures. Knowledge of these sites and subsequent reactive structures can potentially guide rational modifications to prevent bioactivation while hopefully retaining a drug’s pharmacological effect.

In the following experiments, we compare the neural network’s performance on various metrics to that of a simpler model—a logistic regressor—trained with identical inputs and cross-validation folds. We also calculate the performance of each descriptor by treating it as a model and simply using its raw values as predictions. These comparisons indicate which descriptors are informative, and whether anything is gained by agglomerating several properties using the neural network or logistic regressor. We use several metrics to assess each model’s performance at predicting bioactivation pathways.
First, we compute the “top-two” performance, a commonly used metric in site of metabolism prediction studies (Figure 6.5). [248, 252, 253, 254, 255] This metric counts a molecule as accurately predicted if any of its experimentally observed bioactivated pathways receive the highest or second-highest PBS for the entire molecule. The total number of correct predictions are divided by the total number of bioactivated molecules and multiplied by 100 to produce the percentage of correct predictions.

Second, we compute the “average path AUC” for calculating the area under the receiver operating characteristic curve (ROC AUC) for each bioactivated molecule, followed by averaging these AUCs. Recall that each bioactivated molecule has multiple metabolites and pathways that were enumerated in the construction of the bioactivation training data, and at least one of the pathways is experimentally observed to bioactivate. Note that a given bioactivated molecule can have multiple pathways specific to one of the four metabolic transformations or zero pathways specific to another type of metabolic transformation. Due to the discrepancy in pathway type representation within each bioactivated molecule, an appropriate multi-class metric is elusive because computing four ROC curves for each bioactivated molecule (one for each pathway type) results in undefined behavior where there is not a sufficient number of either bioactivated or non-bioactivated classes for the computation to make sense. Furthermore, we are more concerned with predicting whether a molecule will be bioactivated and, if so, which pathway leads to bioactivation – no matter the exact mechanism. For this reason, the pathway-level target vector is binary and not multi-class. Thus, we compute a single, binary-class ROC curve for each bioactivated molecule using binary labels for each of the molecule’s pathways (whether the pathway has been experimentally observed to bioactivate or not) and the pathway-level predictions generated by the model. Finally, we compute the AUC for each of these ROC curves and aggregate them via an ordinary arithmetic mean to yield the average path AUC performance. The average path AUC is more sensitive than the top-two metric, because it considers the relative ranking of all pathways within each molecule, and we have often used variations of it in past work. [119, 188, 248, 120]
Figure 6.5: The model produced accurate cross-validated bioactivation pathway predictions. Right, the top-two metric was computed across 332 bioactivated molecules. The top-two metric reflects the percentage of correctly predicted molecules, where a correct prediction is defined as a molecule for which any of its bioactivated pathways received the highest or second-highest score for all possible pathways within that molecule. Left, across the same set of bioactivated molecules, the average path AUC was measured by calculating how often bioactivated pathways received higher scores than all other possible pathways. For both metrics, the performances of the five bioactivation descriptors were reported, as well as the cross-validated scores produced by training with either a neural network or a logistic regressor. Asterisks denote performances that were significantly worse than the highest-scoring method, as determined by a paired t-test using Bonferroni’s correction.[256] In this context, the paired instances are the bioactivated parent molecules and the computed value of the average path AUC or top-two metric when the bioactivated molecule is processed by either the highest-scoring model or one of the other individual models, whose performance is being compared against the highest-scoring model. The error bars represent 95% two-sided confidence intervals. [257]
Chapter 6. Modeling the Bioactivation and Subsequent Reactivity of Drugs

However, five methods have average path AUC performances that are statistically equivalent by paired $t$-tests corrected for multiple comparison via Bonferroni’s correction. [256] The neural network, the protein bioactivation score, the formation score, the GSH bioactivation score, and a logistic regressor had equivalent average path AUCs of 89.98%, 89.62%, 89.38%, 88.50%, and 88.10%, respectively. The top-two performances are slightly more informative, but nevertheless four of the five previously tied methods are also equivalent, with the formation score, the neural network, the logistic regressor, and the protein bioactivation score, having statistically comparable top-two performances of 78.01%, 78.01%, 77.11%, and 76.20%, respectively. Only the GSH bioactivation score can be judged inferior, with its 68.07% top-two performance having a $p$-value of 0.002 when compared to the predictions of the top-scoring method using paired $t$-tests corrected for multiple comparison via Bonferroni’s correction.[256]

To further evaluate the four methods with equivalent average path AUCs and top-two performances, we compute a third metric, the global pathway AUC (Figure 6.6). Unlike the previous metrics, this measure does not consider molecule identity, and instead computes the AUC of a single, binary-class ROC curve across all possible bioactivation pathways within bioactivated molecules. However, the same five methods that had equivalent average path AUCs also had equivalent global pathway AUCs, with the formation score, the neural network, the protein bioactivation score, the GSH bioactivation score, and the logistic regressor performing at 91.41%, 90.92%, 90.51%, 90.10%, and 89.41%, respectively.

Unlike the average pathway AUC, we can compute a multi-class variant of the global pathway AUC and evaluate the pathway-level models for each metabolic transformation being studied. For this measure, we compute the AUC of four individual ROC curves – one for each metabolic transformation. The target vector is still binary – 1 if bioactivated and 0 if not – and we use the term multi-class in this setting to refer to a global pathway AUC analysis on separate subsets of the pathway data that have been stratified across the four pathway types. Each ROC curve is created based on its metabolic transformation’s possible bioactivation pathways within the set of bioactivated molecules, which are annotated based on whether the pathway has
FIGURE 6.6: The binary-class global pathway AUC was computed for several methods across all bioactivated molecules. Asterisks denote performances that were statistically significantly worse than the highest-scoring method, using a false-positive-rate paired t-test.[256] In this context, the paired instances are the experimental or model-inferred pathways and their computed PBS when the pathway is processed by either the highest-scoring model or one of the other individual models, whose global pathway AUC performance is being compared against the highest-scoring model. The error bars represent 95% two-sided confidence intervals computed using the method specified in Cortes and Mohri, which requires the error rate and the numbers of positive and negative samples. [257] With respect to each model’s ROC curve, the threshold for optimal binarization is calculated and applied to classify each sample, from which the error rate can be computed.
been experimentally observed to bioactivate or not, and the model-predicted pathway scores. Each molecule does not contribute just its maximum scored pathway to the global AUC calculation for a given metabolic transformation, but instead contributes all the scored pathways and their possible metabolites. Each AUC, which is measured for one of the four pathway types, is weighted by the prevalence of its corresponding pathway type in the pathway-level training set. For example, the weight of the AUC for the ROC curve specific to quinone forming pathways is \( \frac{7,580}{13,743} \). The multi-class AUC is then given by the sum of the four weighted AUCs. [76]

The pathway-level multi-class global pathway AUC for each model is reported in Table 1, along with each pathway type’s individual global pathway AUC. The same five methods that had equivalent binary-class global pathway AUCs also had equivalent multi-class global pathway AUCs, with the formation score, the neural network, the protein bioactivation score, the GSH bioactivation score, and the logistic regressor performing at 91.2%, 90.8%, 90.3%, 90.3%, and 89.2%, respectively. For each model, pathway performance was highest for both the epoxidation and quinone formation pathway types. In contrast, the pathway types for nitroaromatic reduction and thiophene S-oxidation had a lower, more varied performance. Both the nitroaromatic reduction and thiophene S-oxidation pathway types are much less frequently observed than the epoxidation and quinone formation pathway types in the bioactivation training set, and the disparity in pathway type-specific results may be driven by the disparity in available data.

Ultimately, we place more emphasis on average pathway AUC because it more closely approximates the expected use of the pathway-level model. When inputting a test molecule to see which potential pathways and metabolites may result in bioactivation, a user is likely interested in whether the model correctly identifies the correct pathway(s) of bioactivation, relative to all other pathways inferred by the model. Average pathway AUC reflects this consideration.
Chapter 6. Modeling the Bioactivation and Subsequent Reactivity of Drugs

### TABLE 6.3: Each pathway-level model’s separate AUC performance for each of the four metabolic transformations and their aggregate multi-class AUC performance

<table>
<thead>
<tr>
<th>Model</th>
<th>Epoxidation</th>
<th>Nitroaromatic Reduction</th>
<th>Quinone Formation</th>
<th>Thiophene S-Oxidation</th>
<th>Multi-class Global Pathway AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formation Score</td>
<td>92.8%</td>
<td>65.2%</td>
<td>90.1%</td>
<td>71.4%</td>
<td>91.2%</td>
</tr>
<tr>
<td>Neural Network</td>
<td>92.1%</td>
<td>65.9%</td>
<td>89.9%</td>
<td>64.3%</td>
<td>90.8%</td>
</tr>
<tr>
<td>Protein Bioactivation Score</td>
<td>91.6%</td>
<td>73.2%</td>
<td>89.5%</td>
<td>42.3%</td>
<td>90.3%</td>
</tr>
<tr>
<td>GSH Bioactivation Score</td>
<td>91.5%</td>
<td>88.9%</td>
<td>89.3%</td>
<td>85.7%</td>
<td>90.3%</td>
</tr>
<tr>
<td>Logistic Regressor Protein Atom Reactivity Delta</td>
<td>90.5%</td>
<td>64.4%</td>
<td>88.4%</td>
<td>59.0%</td>
<td>89.2%</td>
</tr>
<tr>
<td>GSH Atom Reactivity Delta</td>
<td>70.4%</td>
<td>70.4%</td>
<td>56.2%</td>
<td>0.07%</td>
<td>62.5%</td>
</tr>
<tr>
<td></td>
<td>63.4%</td>
<td>89.2%</td>
<td>46.6%</td>
<td>42.6%</td>
<td>54.2%</td>
</tr>
</tbody>
</table>

It is intriguing that three of the bioactivation descriptors perform so well compared to training the logistic regressor or neural network. These performances are calculated simply by treating a descriptor as a model prediction, without any training. In previous studies, we have frequently compared our results to individual descriptors, and we have always found that machine learning outperforms the naive descriptor approach.[119, 188, 248, 120] This is as we would expect, due to the greater flexibility of the machine learning approaches, which by design simultaneously consider many different chemical attributes.

One explanation for the above phenomenon is that, in the previous studies, the individual descriptors that were outperformed were either topological or quantum descriptors describing a basic feature of the molecule or its atoms. In this study, the three individual descriptors that perform nearly as well as the model are derived from the outputs of previously trained site of metabolism or reactivity models. These bioactivation descriptors represent information learned by the previously trained models, and may encode more information regarding bioactivation, and metabolic processes in general, compared to the more naive descriptors referenced in prior
Figure 6.7: The global pathway AUC was computed for the top 5 performing pathway-level models and bioactivation descriptors, in addition to several of the top performing individual topological descriptors. Asterisks denote performances that were statistically significant worse than the highest-scoring method, using a false-positive-rate paired t-test. As theorized, the topological descriptors input as features into the neural network, when used individually, perform significantly worse. This is consistent with past work in developing similar models.

A broader comparison of global pathway AUC is made between several individual topological descriptors against the neural network, logistic regressor, and the top 3 bioactivation descriptors (Figure 6.7). In this comparison, the individual topological descriptors all perform significantly worse, as expected and in keeping consistent with prior work.

Additionally, it is noteworthy that the two bioactivation descriptors with much less predictive value are the two reactivity deltas. This suggests that, for this study’s data set, merely calculating the max reactivity delta of a potential product is not very informative for predicting bioactivation, because that metabolite may be very unlikely to occur. All three descriptors that match the model performances include information about the likelihood of the metabolic transformation: the formation score (which is just a prediction of metabolism and does not consider reactivity), and both “bioactivation scores,” computed by multiplying the maximum atom reactivity increase by the formation score. Based on these observations, we infer that, for our data
set, metabolism seems to be more important to evaluate than reactivity for predicting bioactivation. In future work, by expanding the model to include additional bioactivation pathways that produce metabolites with more diverse reactivity, we might observe an increased relevance of the reactivity computations.

As a final evaluation of the various best-performing methods, we compute reliability plots across all of the enumerated pathways within the bioactivated molecules (Figure 6.8). In this measure, the predictions of each method are distributed into ten bins between 0 and 1, each of width 0.1. Next, we compute the percentage of bioactivations in each bin, and calculate the root mean square error (RMSE) between each of these percentages and the midpoint value of that bin. Lower RMSE values indicate prediction methods that correlate better to probabilities, with a perfectly scaled prediction having an RMSE of 0. To assess performance, we also compute the $R^2$ of the best fit line, which has a maximum possible result of 1 for a model that produces score bins that fit on a perfectly straight line.

Note that the RMSE and $R^2$ metrics are dependent on the binning strategy and can be sensitive to the concentration of data points in each bin. To resolve this issue, we also compare each method’s calibration performance using the Brier score, which calculates the mean squared error between the model’s predictions and the observed outcome. A lower Brier score designates improved model calibration and is preferred.

With an RMSE of only 9.5\% and a Brier score of 0.0197, the neural network outperforms all other methods by both measures. In comparison, the Brier scores for the logistic regressor, the protein bioactivation score, and the formation score are 0.022, 0.0230, and 0.0258 respectively. The logistic regressor and the protein bioactivation score have much lower $R^2$ performances of 0.77 and 0.21, respectively. In contrast, the formation score’s $R^2$ of 0.95 almost equals the neural network’s $R^2$ of 0.95. Nevertheless, the formation score’s RMSE of 15\% indicates that it does not produce a probability as well-scaled as the neural network, with its much lower 9.5\% RMSE.

The neural network is the only method that produces both a high $R^2$ and a low RMSE and
Figure 6.8: Of all methods, the neural network computed the most well-scaled predictions of bioactivation pathways, with predictions that were highly probabilistic. The bar graphs plot the normalized distributions for each method across 388 pathways within 332 bioactivated molecules. The solid lines plot in each bin the percentage of bioactivated pathways among all pathways in the corresponding score, denoted on the X-axis. The diagonal dashed lines indicate the ideal perfectly scale prediction. The neural network produced the best scaled prediction of all methods, indicated by the highest correlation to the best fit-line, and the lowest RMSE compared to a perfectly scaled prediction.
Brier score, so we chose to use it for the rest of the study. However, we certainly acknowledge that valid arguments could also be made for choosing one of the simpler methods instead. In addition, models with good correlation but bad RMSE can still be useful in scenarios where binarizing the prediction is applicable and where there is flexibility in shifting the decision threshold for optimal binarization. Ultimately, we chose the neural network’s scores because we felt that their probabilistic nature might be helpful for the next objective: predicting molecule bioactivation. To give an explicit example of the output given by the pathway-level neural network model, four drugs are visualized with their cross-validated PBS produced by the neural network (Figure 6.9).
6.3.2 Accuracy at Predicting Molecule Bioactivation

Given a list of drugs or drug candidates, a useful bioactivation model should accurately predict which structures will be bioactivated. Flagging these potentially toxic molecules and separating them from lower-risk compounds could enable problematic compounds to be triaged for more rigorous testing, set aside if there are acceptable alternatives, or rationally modified to prevent bioactivation while retaining efficacy. To enable the bioactivation model to make these molecule-level bioactivation predictions, we include in the training data molecules that are not bioactivated, despite having structures capable of forming quinones, epoxides, thiophene sulfur-oxides, or aromatic nitrosos.

After the first-stage of training, which produces accurate pathway-level predictions, we perform a second training stage to tune the model to distinguish between bioactivated and non-bioactivated molecules. To assess performance at this objective, we compute the “molecule AUC”: the AUC over all molecules in the training data set. While all the previous AUC values are based on the predictions of the pathway-level model, the molecule-level AUC is based on the predictions of the molecule-level model and only considers the parent molecule and not the whole set of possible metabolites. Specifically, we compute the AUC of a single ROC curve that is generated using the labels for each molecule in the training set (whether that molecule has been experimentally observed to bioactivate or not) and the molecule-level predictions generated by the model. We evaluate several models using this metric.

First, simply assigning each molecule its maximum cross-validated pathway score is the most intuitive approach. This highest pathway score method produces a molecule AUC of 80.23% (Figure 6.10). Alternatively, we perform a second training step to tune the weights of an additional model layer to distinguish bioactivated and non-bioactivated molecules. As descriptors, we use each molecule’s top five PBS as well as all 15 molecule topological descriptors, for a total of 20 descriptors. With these descriptors as input, we compare a logistic regressor and a neural network with ten hidden nodes. The same cross-validation folds are used as for the pathway-level training: each group of similar molecules is withheld in turn, and a
model trained on the remaining data, ensuring that training and testing data is never mixed. The cross-validated scores produced by the logistic regressor and the neural network have molecule AUCs of 80.19% and 81.06%, respectively.

While molecule AUC fails to distinguish the highest pathway score, the logistic regressor, and the neural network, reliability plots indicate the best method: the neural network has both the best RMSE (6.9%) and the best $R^2$ (0.94) (Figure 6.11). In contrast, the reliability performances of the highest pathway score and the logistic regressor are, respectively, RMSE’s of 8.4% and 24%, and $R^2$ values of 0.90 and 0.90. Furthermore, the neural network achieves the lowest Brier score of 0.168, compared to Brier scores of 0.184 and 0.179 for the highest pathway score and logistic regressor, respectively. The reliability plots also show that the neural network does the best job of assigning high scores to bioactivated molecules and low scores to non-bioactivated molecules. Consequently, similar to our decision on the path-level, we select the neural network for the final model structure, while acknowledging that this is a somewhat subjective choice. The final model is trained on the full bioactivation training data set.

As external tests, we use ten of the previously discussed batches of non-bioactivated molecules. Each batch includes 332 unique molecules not present in the training data. Using the final trained model, we predict the MBS of each molecule within each external test set. Next, for each test set, we measure the separation between the MBS of the test molecules and the cross-validated MBS of the bioactivated molecules.

Separation between the bioactivated molecules from the training set and the non-bioactivated molecules from the external sets could simply be due to a lack of calibration between the cross-validated model predictions and the final trained model predictions. Without additional positives being introduced via the external test sets, further evidence is required to demonstrate that the final trained model predictions do not need to be adjusted to compensate for potential differences relative to the cross-validated model predictions. On the bioactivated molecule training subset, we compare the cross-validated molecule predictions against the fully trained model predictions (Figure 6.12). As expected, there is an optimistic bias to the fully trained model predictions,
Figure 6.10: Molecule bioactivation scores (MBS or molecule score) accurately identified bioactivated molecules. While all the previous AUC values are based on the predictions of the pathway-level model, the molecule-level AUC is based on the predictions of the molecule-level model and only considers the parent molecule and not the whole set of possible metabolites. Several methods are compared by their capacity to separate bioactivated and non-bioactivated molecules. These methods include training a neural network or logistic regressor on the top five pathway-level scores and all molecule descriptors, using each molecule’s top pathway-level scores as molecule predictions, and using molecule descriptors as molecule predictions. Based on false positive rate paired t-tests,[256] results are denoted with an asterisk if their performance is significantly lower than the best-performing method. In this context, the paired instances are the bioactivated or non-bioactivated molecules and their computed MBS when processed by either the highest-scoring model or one of the other individual models, whose molecule AUC performance is being compared against the highest-scoring model. As a baseline comparison, an approach that uses structural alerts to filter molecules with at least one alert achieved a molecule AUC of 59.8%. The 35 structural alerts used for screening were derived from literature on minimizing idiosyncratic toxicity mediated by reactive metabolites [262]. The error bars represent 95% two-sided confidence intervals. [257] Right, examples of bioactivated and non-bioactivated molecules are visualized. From left to right, top to bottom: nefazodone (MBS: 0.74), lumiracoxib (MBS: 0.78), buspirone (MBS: 0.18), sertindole (MBS: 0.43). Each experimentally observed site of bioactivation is circled. For each molecule, the colored shading represents PBS, which range from 0 to 1.
Chapter 6. Modeling the Bioactivation and Subsequent Reactivity of Drugs

Figure 6.11: Of all methods, the neural network produces the best-scaled, highly probabilistic molecule predictions. The bar graphs plot the normalized distributions for each method across 388 pathways within 332 bioactivated molecules. The solid lines plot the percentage of bioactivated pathways among all pathways that were assigned the corresponding score bin, denoted on the X-axis. The diagonal dashed lines indicate the ideal perfectly scaled prediction. Of all three methods, the neural network has both the highest correlation to a best fit-line through respective binned percentages, indicated by the $R^2$ values, and the lowest RMSE compared to a perfectly scaled prediction.

which can be seen when comparing the line of best fit to the identity line. The optimistic bias results from the fully-trained model seeing each instance during training. Several notable outliers are present, most of which undergo observed quinone formation. These outliers may have manifested due to having moderate pathway-level scores that lead to greater sensitivity in bias when alternating between the cross-validated model and the fully-trained model. The squared Pearson correlation coefficient of 0.87 ($p$-value $\leq 0.001$) denotes a high correlation between the cross-validated model predictions and the fully trained model’s predictions, which is evidence that the fully trained model’s predictions are comparable to the cross-validated model predictions.

Over the ten external test sets, the molecule AUC is 83.29(96)% equivalent to the accuracy of the MBS on the full cross validated training set of 81.06%. The model successfully generalized to new data, assigning non-bioactivated molecules it had never seen before with much lower scores than bioactivated molecules.
FIGURE 6.12: On the bioactivated molecule training subset, we compare the cross-validated molecule predictions against the fully trained model predictions. As expected, there is an optimistic bias to the fully trained model predictions, which can be seen when comparing the line of best fit to the identity line. The squared Pearson correlation coefficient of 0.87 (p-value less than 0.001) denotes a high correlation between the cross-validated model predictions and the fully trained model’s predictions, which is evidence that the fully trained model’s predictions are comparable to the cross-validated model predictions.
6.3.3 Inferring the Toxicity Mechanism of Withdrawn Drugs

To investigate the model’s utility, we explored its applicability to withdrawn, small-molecule drugs. Ideally, the model will be able to generate hypotheses regarding the underlying toxicity of the withdrawn drugs. After downloading a list of withdrawn drugs from DrugBank and removing any molecules found in our training data, we used the bioactivation model to enumerate hypotheses for the toxicity drivers of the withdrawn drugs. Each of these 201 molecules were submitted to our final trained model, thereby producing a molecule score for each molecule.

The model outputs probabilistic molecule scores, but for the purposes of estimating the total number of bioactivated molecules within the list of withdrawn drugs, it was useful to define an exact score to binarize predictions. To define this cutoff, we analyzed the full ROC curve used to calculate the molecule AUC across all training molecules (Figure 6.13). While several methods for defining this cutoff are possible, in this study we computed the cutoff that optimizes both sensitivity and specificity.[263] This analysis resulted in an MBS threshold of 0.57. Using this threshold, we predicted that among the dataset of withdrawn drugs, 33 are bioactivated.

Upon investigation, we find that some of these predictions are consistent with previously reported experimental results that were not present in our training data (Figure 6.15, top). For example, alclofenac, a non-steroidal anti-inflammatory drug, was introduced to the U.K. in 1972, but withdrawn in 1979[264] due to several incidences of hepatotoxicity and skin rashes.[265] Later, it was determined that an epoxidation reaction takes place at alclofenac’s terminal olefin.[266] Furthermore, this epoxide metabolite directly conjugates to sulphydryl nucleophiles, [267, 268] and consequently is a likely driver of alclofenac’s adverse effects. Indeed, the bioactivation model predicts this epoxidation and assigns alclofenac a very high probability of bioactivation with an MBS of 0.75.

The bioactivation model also agrees with the experimental findings for the gout drug benzbro-marone, withdrawn from Europe in 2004[269] due to cases of severe hepatotoxicity,[270, 271, 272]. At the time, the mechanism of this toxicity was unclear. Later, experiments with human
Figure 6.13: Using the ROC curve of the cross-validated molecule bioactivation scores (MBS), we compute a threshold score for molecule bioactivation. For a given test molecule, applying this threshold converts the continuous MBS to a discrete prediction of bioactivation or non-bioactivation. For some decisions, such as choosing molecules to experimentally test, these simple predictions might be more useful than probabilistic scores. Depending on the relative importance of sensitivity and specificity for a given application, one might define various different thresholds. Sensitivity is the true positive rate, while specificity is the complement of the false positive rate. If sensitivity and specificity are equally important, then the optimal score corresponds to the point on the ROC curve closest to 100% true positive rate and 0% false positive rate. We used this point, corresponding to an MBS of 0.57, to threshold predictions for the rest of the study.
liver microsomes produced glutathione-conjugated metabolites, believed to be adducts of a reactive ortho-quinone.\[273\] This reactive metabolite could be the driver of the benzbromarone’s idiosyncratic reactions. Surprisingly, the bioactivation model similarly predicts quinone formation at this ring, although it scores formation of a para-quinone as more likely than formation of the ortho-quinone. It is not clear whether the human liver microsomes had the resolution to distinguish ortho- and para-quinones, so it seems that the bioactivation model’s hypothesis essentially matches the experimentally-observed reactive metabolite.

In the cases of alclofenac and benzbromarone, the bioactivation model reiterates toxicity mechanisms that have already been reported. These results are examples of the model successfully generalizing beyond its training data since these structures were not included at any training phase. Additionally, the agreement between experiment results and the computational modeling bolsters confidence that predictions of novel toxicity mechanisms for less well-understood drugs be experimentally validated in the future.

In other cases, the model’s predictions represent new hypotheses for the mechanism’s of a drug’s toxicity (Figure 6.15, bottom). For example, the antidepressant safrazine, a monoamine oxidase inhibitor, was withdrawn due to hepatotoxicity \[274\] and has also been associated with neuropathy. \[275\] The model assigned safrazine an MBS of 0.65, above the threshold defined by the optimal point on the ROC curve of cross-validated scores. To the best of our knowledge, a toxicity mechanism has not been proposed for this abandoned drug. From the structural alerts perspective, one motif of concern might be safrazine’s terminal hydrazine, a well-known structural alert. \[237\] However, the model does not predict a bioactivation pathway extending from the terminal hydrazine, which may be because activation of this structural alert was not one of the bioactivation pathways modeled in this study. Instead, the model predicts a bioactivation pathway on a different structural alert of safrazine: a 1,3-benzdioxole motif leading to a reactive quinone. Without experiments, it is difficult to say which bioactivation pathway is more likely, but understanding the toxicity of this drug could inspire a rational modification to improve its safety profile.
Chapter 6. Modeling the Bioactivation and Subsequent Reactivity of Drugs

...
Figure 6.14: To contextualize the bioactivation predictions on the discussed withdrawn drugs, we evaluated their Tanimoto similarity against the training data set on ECFP6-derived fingerprints. Neither alclofenac, benzbromarone, safrazine, zimelidine, nor astemizole had a Tanimoto similarity greater than 0.5 with any of the training set molecules. The vertical dashed line marks the maximum Tanimoto similarity for the distribution.
next step in future work, but is outside of the scope of this study.

Lastly, it is worthwhile to point out cases in which the model yields potential false negatives with respect to literature. 12 of the withdrawn drugs had an MBS below the bioactivation threshold but were annotated as being observed to undergo bioactivation according to the AMD. Of these 12, we found corresponding literature for 6 that referenced potential bioactivation via the 4 pathways studied — zomepirac [284] (MBS: 0.42), nimesulide [285] (MBS: 0.40), thioridazine [286] (MBS: 0.36), troglitazone [287] (MBS: 0.22), phenacetin [288] (MBS: 0.15), and thalidomide [289] (MBS: 0.07). For the remaining 6 drugs, their withdrawal is associated with other causes, such as drug-drug interactions or formation of alternate reactive metabolites than those within the scope of this study. Identifying and addressing the model’s limitations may aid in resolving these false negatives or better understanding why they occur.

### 6.4 Model Limitations

One limitation of the current model is that there are other bioactivation pathways beyond those included in this study. However, the approach demonstrated here is easily extendable to additional pathways by providing training data for those pathways and models for specific reactions of interest. For instance, the phase I metabolism model used for prediction of nitroaromatic reduction and sulfur oxidation also supports epoxidation and provides coverage of 92.3% of AMD phase I reactions [54], and the structure inference model supports the phase I metabolism model’s rule set coverage as well. [60] Another shortcoming is that we only considered one-step bioactivations, where a metabolic reaction produced a metabolite that conjugated to macromolecules. Going forward, by expanding the metabolism scores to included multiple steps, we plan to extend the ability of the model to pick up multi-step bioactivations.

Additionally, the current modeling approach has some inherent limitations due to its reliance on a database of literature-derived reactions. For example, we label molecules as not bioactivated based on a lack of evidence in the database. However, the aphorism “absence of
Chapter 6. Modeling the Bioactivation and Subsequent Reactivity of Drugs

FIGURE 6.15: Top, examples of withdrawn drugs with high bioactivation predictions. We submitted the structures of 201 withdrawn drugs to the bioactivation model, the majority of which were downloaded directly from DrugBank, and supplemented by literature review. In order to generate toxicity mechanism hypotheses for these drugs, we looked at molecules that received high molecule bioactivation scores, yet have not been reported to form reactive metabolites (to the best of our knowledge). From left to right: alclofenac[265, 266] (MBS: 0.75), and benzbromarone (MBS: 0.66). Bottom, hypothesized toxicity drivers of withdrawn drugs. The bioactivation model predicts novel bioactivation mechanisms for several drugs. Three examples are visualized with their metabolism predictions indicated by the colored shading on each parent drug structure, and the reactivity predictions similarly visualized on the structure of the hypothesized metabolite. From left to right, top to bottom: safrazine (MBS: 0.65), zimelidine (MBS: 0.66), and astemizole (MBS: 0.69). For each pairing, the color shading gradient indicates PBS on the substrate and protein reactivity on the metabolite.
evidence is not evidence of absence” is certainly worth remembering, and we view our labeling paradigm more as a necessary assumption than an ideal data set, due to the paucity of alternatives. This caveat is reflected in the lower performance of the molecule-level scores compared to the pathway-level scores. Another data limitation is with respect to understudied or less common pathways, as seen in the model’s worse performance when evaluating potential bioactivation pathways involving nitroaromatic reduction or thiophene sulfur-oxidation compared to those involving quinone formation or epoxidation. Hopefully, future experimental work will yield higher-quality data sets that include both bioactivated and non-bioactivated molecules, with both categories tested by the same assays. Furthermore, the current approach is reliant on several underlying models of metabolism and reactivity. Updates to any of the underlying models will influence the values of the calculated bioactivation descriptors and, as a result, require retraining the bioactivation model.

In the current work, the max atom-level reactivity delta features had less relevance than the formation scores in predicting PBS and MBS. This could be a result of the types of reactivity present in the data set, but it could also be a result of how the reactivity is used. Future work could expand upon more complex feature engineering of reactivity. For example, a product metabolite can result in a large max atom-level reactivity delta without having a high reactivity. Potential avenues of experimentation include altering the type of aggregation function used, e.g. an average function instead of a max function, and the incorporation of reactivity delta features with a feature for the absolute value of the product’s reactivity.

Finally, not all reactive metabolites are toxic. Sometimes, detoxification pathways such as glutathionation are able to effectively clear reactive metabolites before they cause deleterious effects. Other complex phenomenon that influence the likelihood of an adverse event include route of administration, coadministered medications, comorbidities, and genetic variants. We ultimately envision a broader toxicity model, where bioactivation predictions are combined with other important factors, like daily dose and rate of reaction, in order to build a model that explicitly predicts the toxicity risk of a certain molecule.
Bioactivation entails two distinct concepts: metabolism and reactivity. Metabolism gives rise to electrophilic structures—reactive metabolites—that then conjugate to nucleophilic sites within biological macromolecules, such as proteins. Protein-metabolite adducts can incite toxic immune responses, the driving force for many idiosyncratic adverse drug reactions. This study constructs a novel, sophisticated bioactivation model by synthesizing several previous studies of both metabolism [248, 119, 60, 54] and reactivity [188, 120]. Four types of metabolism are included: quinone formation, epoxidation, thiophene sulfur-oxidation, and nitroaromatic reduction, some of the most common reactions that form reactive products. To predict bioactivation, the model first predicts the chances of any of these metabolic events occurring at all possible sites within a given input molecule. Next, the model enumerates the exact structures of those metabolites. These structures are assigned scores reflecting their chances of conjugating to protein or GSH. Finally, using these metabolism and reactivity scores as input, a feedforward neural network predicts bioactivation on both the pathway- and molecule-level.

Across 332 bioactivated molecules extracted from the literature, the model predicts the exact bioactivation pathway with 89.98% AUC. These pathway-level bioactivation predictions make a specific hypothesis about the mechanism of a molecule’s toxicity. Furthermore, with 81.06% AUC the model separates bioactivated and non-bioactivated molecules. Molecule-level bioactivation predictions can be used to rapidly screen a large number of molecules for the key toxicity risk of bioactivation. To demonstrate, we use the bioactivation model to generate hypotheses for the toxicity drivers of several withdrawn drugs associated with idiosyncratic reactions with currently unknown etiology. By adding metabolism models for additional bioactivation pathways, future work will easily expand the utility of the bioactivation model. Nevertheless, this study makes a major step forward towards a fully comprehensive model of metabolism and reactivity.
Chapter 7

Machine learning liver-injuring drug interactions with non-steroidal anti-inflammatory drugs (NSAIDs) from retrospective cohort


*Equal contribution as first author

7.1 Introduction

Synergistic drug combinations, which consist of at least two active pharmaceutical ingredients, form a crucial therapeutic option for the treatment of complex diseases that may manifest multiple conditions, such as cancer and AIDS [290]. The concomitant application of multiple drugs can enhance therapeutic effect and selectivity, delay drug resistance, allow lower dose of each individual drug and combat multiple related targets to address redundancies in disease mechanisms [291, 292, 293, 294]. However, just as multiple drugs can interact in a salubrious manner, they can also interact to cause unintended consequences. Combined drug therapies can
result in an antagonistic effect that is smaller than the additive effect of each individual drug or, worse, can result in synergistic toxicity [293]. In some cases, these drug-drug or polypharmic interactions can result in an adverse drug reaction of clinical significance.

Understanding the potentially adverse consequences resulting from drug-drug interactions is a significant problem with regards to patient safety and clinical outcomes. These adverse effects are reflected by the additive risk of each drug the patient is exposed to, as well as how each drug may alter the pharmacokinetic and pharmacodynamic properties of the other co-prescribed drugs [295]. Certain patient groups, such as the elderly, may also be more susceptible due to decreased mobility, increased body mass and impaired renal and hepatic functions [296]. Prevalence of multimorbidity, the co-existence of two or more chronic health conditions, can range from 27.2% of patients to 67% [297, 298]. In the absence of multimorbidity, certain individual disorders, e.g., cancer, can still require a cocktail of drugs to be treated effectively [299]. One recent longitudinal study reported that 35.8% of U.S. adults take at least five drugs concomitantly [300]. Heightened cases of polypharmacy, almost doubling from 8.2% of cases in 1999 to 15% of cases in 2012, have exacted an estimated toll of 177.4 billion USD to treat the resultant adverse polypharmic interactions [301].

In clinical trials, adverse events that can be observed and distinctly mapped to a specific combination of drugs occur at a level of frequency that would require an intractably large patient sample size to detect. In vitro and in vivo experimental approaches are useful for detecting drug-drug interactions [302, 303, 304, 305], but at an increased expense in terms of resources, monetary cost, labor and time relative to computational approaches. A set of $N$ drugs would require evaluation of $N(N − 1)/2$ pairwise drug combinations. As the number of co-administered drugs increases, there is a combinatorial explosion of possible pairwise drug combinations. In contrast, computational approaches are appealing for rapid, high-throughput screening and early detection of adverse drug-drug interactions. Furthermore, computational approaches can incorporate multiple data sources that increase availability to a wider range of population subgroups and to long-term, post-approval therapeutic contexts not explored in short-term clinical
Chapter 7. Machine learning liver-injuring drug interactions with non-steroidal
anti-inflammatory drugs (NSAIDs) from retrospective cohort

trials [306].

Previous research studies have focused on ranking drug–drug event associations using public databases and spontaneous reports [307]. There exist several data mining algorithms that generate and rank adverse drug associations, or signals, based on projections of the data to two-dimensional contingency tables. Such methods include relative risk (RR), proportional reporting ratio (PRR) and reporting odds ratio (ROR) [308, 309]. More complex dis-proportionality methods build on top of the aforementioned statistical measures of association. Namely, Multi-item Gamma Poisson Shrinker (MGPS) is widely used and is the U.S. Food Drug Administration’s main signal detection algorithm for pharmacovigilance [310].

MGPS is conceptually similar to PRR, but incorporates Bayesian shrinkage to produce dis-proportionality scores that alleviate variability issues with limited data and small case numbers [311, 310]. MGPS assumes that the number of observed counts of a drug combination and adverse event pair is drawn from a Poisson distribution with an unknown mean that can be computed as a function of $\lambda$. The goal is to estimate the $\lambda$s. Each $\lambda$ is assumed to be drawn from a common, 5-parameter prior distribution, which is further assumed to be a mixture of two gamma distributions. Using an empirical Bayes approach, the 5 parameters are estimated such that they maximize the marginal likelihood and empirical Bayesian geometric mean (EGBM) scores are output for each $\lambda$ [312, 311, 310].

Bayesian confidence propagation neural networks (BCPNN) also take a Bayesian approach to signal generation [313]. BCPNNs are similar to feed-forward neural networks, but Bayesian principles are used during learning and inference. Other popular techniques, such as Bayesian logistic regression, have also been used to analyze the effects of drugs in pharmacovigilance studies [314]. Beyond the scope of this study, there also exist methods that operate outside of EHRs and incorporate additional data sources to model polypharmacy at a network level. Recently, Burkhardt et al. [315] have proposed the use of neural embeddings to predict adverse
drug-drug interactions. Another method, Decagon, achieves strong performance on polypharmacy effects with a strong molecular basis via data sets on protein-protein interactions, drug-protein target interactions and known polypharmacy side effects [316]. Approaches such as DeepDDI [317] have been developed to study drug interactions from structures of chemical compounds but have not used EHR datasets and often lack interpretability.

The primary disadvantage of these approaches is that they rely on incomplete datasets, studying specific cases where severe reactions were identified and reported. Though these datasets are large, they are often biased in ways that limit the interpretability, certainty and robustness of results derived from them. For example, an increase in reports of adverse events associated with a drug could be caused by an increase in prescriptions for that drug.

In this study, we propose a logistic regression-based machine learning algorithm that infers drug-drug associations from EHR datasets. The EHR datasets include several avenues of information that have not yet been fully exploited. It is also not biased towards adverse events, since it includes all hospitalizations with and without adverse events. To clarify, the previously mentioned methods are most widely used within the context of spontaneous reporting systems, which primarily collect reports of adverse events made by clinicians or patients to a regulator or product manufacturer[312]. Furthermore, since our model takes into account outcomes from all hospitalizations, it does not suffer from potential under-reporting of unexpected adverse interactions, which is otherwise a common source of signal loss [318].

We hypothesize that statistical modeling on EHR data can identify drug-drug interactions. Our proposed model simultaneously reveals the risk contribution of individual drug and pairs of interacting drugs with respect to a therapeutic outcome, such as an adverse event. Empirically, we have shown that our model can extract meaningful drug-drug associations between a candidate drug, whose potential drug-drug interactions are of interest, and all of its co-prescribed drugs in EHR datasets consisting of less than 400,000 hospitalization records.

As a case study, we have identified drug-dependent risk of nonsteroidal anti-inflammatory drugs (NSAIDs) with respect to drug-induced liver injury (DILI). NSAIDs are one of the most
commonly and widely used class of drugs, yet many of them have been implicated in causing adverse drug reactions [319]. Since NSAIDs are often used, concomitantly, with a variety of co-prescribed drugs across a wide range of therapeutic contexts, the resultant polypharmic interactions may drive some of these adverse drug reactions. Furthermore, NSAIDs are an ideal class of drugs for such a case study, because they are prescribed in a wide variety of contexts and it is anticipated that their widespread use may allow the detection of statistically significant interactions.

7.2 Materials and methods

7.2.1 Study population and study design

The electronic healthcare records (EHR) dataset contains data of 397,064 hospitalizations reported by the BJC HealthCare system in St. Louis, Missouri, USA (Table 7.1) [320]. The 397,064 hospitalizations involve 223,883 unique patients. The earliest inpatient admit date was September 2012 and last discharge date was October 2016. The number of hospitalization cases in the St. Louis area during the data collection period determined the sample size. The hospitalization cohort (aged ≥ 18 years) contains 176,443 (44.44%) male hospitalizations, 189,723 (47.78%) female hospitalizations and 30,878 (7.77%) hospitalizations with no specified gender. The cohort’s median age is 63.2 years (max: 110.4; min: 17.9) and the median hospital stay is 3 days (max: 214; min: 0). Each hospitalization is associated with demographics, diagnoses (23366 ICD9,10 codes), drugs (1083 unique active ingredients) and procedures (13097 ICD 9-CM,10-PCS codes). In this study, we included drugs that were administered orally or via intravenous route.

As a case study of our proposed modeling framework, our study design compared hospitalization records involving the presence or absence of DILI and evaluated the model’s ability to, using these comparisons, derive drug dependent DILI risk that corresponds with knowledge from literature or public databases. To train our proposed modeling framework, each datapoint
was a hospitalization with specific admit and discharge dates. Hence, it is quite plausible that one patient with multiple hospitalizations over time will contribute multiple datapoints to the training set. In order to capture drug interactions during a specific timeline, we performed hospitalization-based analyses rather than a patient-based analyses. A major drawback with patient-based analyses is that there can be significant time differences between two successive hospitalizations and drugs administered during the first hospitalization will, in no plausible way, interact with drugs administered during the second hospitalization. A hospitalization-based analyses addresses this issue, since we can now capture meaningful drug interactions within a specific hospitalization and not across different hospitalization timelines.

\begin{table}[h]
\centering
\begin{tabular}{|c|c|c|c|}
\hline
Characteristic & Quartile & Median (Min, Max) & %DILI positives \\
\hline
Age (years) & Q1 & 82.2 (74.5, 110.4) & 8.6 (1038) \\
& Q2 & 68.5 (63.2, 74.4) & 9.9 (1193) \\
& Q3 & 57.7 (51, 63.2) & 9.7 (1169) \\
& Q4 & 39.2 (17.9, 50.9) & 9.9 (1192) \\
\hline
Length of stay (days) & Q1 & 8 (5, 214) & 48.8 (5866) \\
& Q2 & 4 (3, 5) & 24.7 (2966) \\
& Q3 & 2 (2, 3) & 15.5 (1857) \\
& Q4 & 1 (0, 2) & 11 (1324) \\
\hline
No. of drugs & Q1 & 22 (17, 101) & 42.4 (5092) \\
& Q2 & 15 (13, 17) & 23.5 (2824) \\
& Q3 & 11 (9, 13) & 19.1 (2291) \\
& Q4 & 6 (1, 9) & 14.6 (1750) \\
\hline
No. of diagnoses & Q1 & 24 (19, 88) & 48.8 (5861) \\
& Q2 & 16 (13, 19) & 26.3 (3157) \\
& Q3 & 11 (8, 13) & 17.2 (2063) \\
& Q4 & 6 (1, 8) & 7.8 (933) \\
\hline
\end{tabular}
\caption{Characteristics of hospitalizations in cohort.}
\end{table}

% DILI positives are based on the total DILI positives in the data set. % DILI positives may not sum to 100% due to missing values.

\subsection{7.2.2 Polypharmacy data: Twosides database}

We downloaded the v0.1 release of the Twosides database, which contained data on drug-drug interaction side effects reported up to, and including, the year 2014 [321]. Twosides is based
on analysis of drug-drug interactions mined from the FDA Adverse Event Reporting System (FAERS). In this study, we primarily utilized Twosides to understand the validity of the model’s predictions in the context of known polypharmic toxicity. During analysis of a specific NSAID, we extracted only those Twosides interactions that involved the NSAID with conditions related to hepatotoxicity: DILI, liver injury, hepatocellular injury, mixed liver injury and cholestatic liver injury. To extract positive and negative controls for comparison with our model’s results, we used the proportional reporting ratio (PRR) recorded for each Twosides interaction. The PRR is used as a signal of the drug pairs side-effect association. A PRR of 2 suggests that the adverse event is reported twice as frequently as for individuals receiving co-administration of the drug pair relative to taking the drug alone. For positive controls, we only considered interactions with a PRR equal to or greater than 5. For negative controls, we only considered interactions with a PRR less than 1.

7.2.3 DILI definition

The DILI outcome was computed using a combination of diagnoses and procedure codes, available for each hospitalization. The codes are defined in accordance with the International Classification of Diseases (ICD), which has near-universal availability in EHR systems [322]. DILI can be present with a wide range of severity, from mild and reversible elevation of liver enzymes to permanent liver failure. Mild DILI is more common, not usually reported to the FDA, but well represented in the study population and has a large impact on healthcare costs by increasing length of stay at the hospital. Moreover, most DILI cases result from dose-independent, idiosyncratic injury [323, 324] and similar underlying mechanisms may be present in both mild and severe DILI. As an example, metabolite reactivity commonly causes rash, but can also cause rare, severe hepatotoxicity by the same bioactivation mechanism [325]. Thus, adverse reactions which cause mild DILI may also be associated with severe DILI. For these reasons, we used a definition of DILI that also included low severity cases.
Hospitalizations were deemed “DILI positive” under fulfillment of the following three criteria: (A) having diagnosis codes that indicate the presence of DILI, such as (1) elevation of levels of transaminase, lactic acid dehydrogenase and serum enzymes, (2) poisoning by aromatic or non-opioid analgesics, antipyretics and antirheumatics causing adverse effects in therapeutic use, (3) toxic liver diseases such as cholestasis, hepatitis and hepatic necrosis; (B) not having diagnosis codes that include (1) poisoning by, adverse effect of and under-dosing of systemic antibiotics, (2) alcholic liver diseases, internal injury to liver and inflammatory liver diseases, (3) malignant neoplasm of gallbladder, hepatic bile ducts and small intestine, (4) pancreatic diseases; (C) not having procedure codes involving (1) surgeries on liver such as marsupialization of liver lesion, hepatectomy, lobectomy, laceration, etc., (2) surgeries on gallbladder and biliary tract including cholecystotomy, cholecystostomy, anastomosis, etc. and (3) surgeries on pancreas such as panreatotomy, marsupialization of pancreatic cyst, transplantation of pancreas, etc. Applying the aforementioned definition, we identified 12,014 hospitalizations associated with DILI.

7.2.4 Estimating percent relative effect

In this study, we have reported the effects of drug-drug interactions on DILI outcomes in terms of percent relative effect. We used odds ratio from our models to approximate the relative risk of the independent and candidate drug dependent interactions. In epidemiology, relative risk, or the risk ratio, is defined as the ratio of probabilities of an event in the exposed group to that in the non-exposed group. Odds ratio (OR) is defined as the ratio of the odds of an event in the exposed group to the odds of that event in the non-exposed group. In our dataset, the number of DILI negatives greatly outweighs the number of DILI positives. Hence, we estimated the relative risk as

\[
RR = \frac{\text{probability of DILI in exposed group}}{\text{probability of DILI in non-exposed group}} = \frac{\frac{a}{a+b}}{\frac{c}{c+d}} \approx \frac{\frac{a}{b}}{\frac{c}{d}} = OR
\]

(7.1)
where \( a \) and \( b \) are the respective number of events (DILI positives) and non-events (DILI negatives) in the exposed group and \( c \) and \( d \) are the respective number of events and non-events in the non-exposed group. A risk ratio greater than one suggests an increased risk of DILI in the exposed group, whereas a risk ratio less than one suggests a reduced risk of DILI in the exposed group. Finally, we have computed the percent relative effect (the percent change in the exposed group). In essence, we have considered the non-exposed group as having 100% of the risk and express the exposed group relative to that.

\[
\% \text{ increase or } \% \text{ decrease in relative effect} = \pm (RR - 1) \times 100
\]  

(7.2)

where the \((+)\) sign indicates an increase in percent relative effect and a \((-)\) sign indicates a decrease in percent relative effect in the exposed group.

### 7.2.5 Drug interaction network (DIN)

We have used a logistic regression model to estimate the independent and dependent risk of drugs relative to an outcome variable. Rather than estimating the full pairwise matrix of interactions, the model learns the risk dependent on a single candidate drug, whose potential interactions with other drugs are of interest. Equivalent to learning a single column of a pairwise interaction matrix, this approach dramatically reduces the number of weights to be learned, focusing all modeling effort on a more focused question - what is the independent risk of each drug and what is the additional risk when co-prescribed with the candidate drug?

The logistic regression model has two branches: an independent risk branch and a dependent risk branch (Fig 1, A). The input to the independent risk branch is a binary vector that records whether or not a drug was administered during the hospitalization. The input to the dependent risk branch is the same vector when the candidate drug is prescribed in the hospitalization, otherwise it is a vector of zeros. Conceptually, the presence or absence of the candidate drug acts as a switch that controls the input to the dependent risk branch. Mathematically, the input to the
dependent risk branch is computed as an element-wise multiplication between the binary vector representation of a hospitalization and a binary scalar variable denoting the presence (binary scalar variable is 1) or absence (binary scalar variable is 0) of the candidate drug in that hospitalization. The logistic regression model uses the inputs from both of these branches to estimate the probability of the outcome variable, e.g. DILI in this study, using the maximum likelihood estimation framework. The coefficients, learnt by the model, are then used to compute the percent relative effects of drugs when prescribed independently and co-prescribed alongside the candidate drug of interest, respectively.

Though not considered in this study, we expected that improvements are possible. We point out that continuous variables, such as age, were not used as an input feature in our modeling framework and we only used the binary encoding of presence (represented by 1) or absence (represented by 0) of drugs during a hospitalization timeline as input to our models. For example, encoding the severity of DILI as distinct outcomes would give the model additional information that may yield better estimates. Likewise, encoding the dose for each drug would also reduce noise. We also expected that using a dependent risk input vector for drugs, that are administered on the same days during a hospitalization, would produce better estimates, as drugs without overlapping exposures do not usually interact. However, it appears that these improvements were not necessary to produce clinically relevant results.

7.3 Results & discussion

We have evaluated the proposed framework’s capabilities on three tasks as a demonstration of its utility. We studied the role of diclofenac in hepatotoxicity across the full range of drugs co-prescribed with it in our clinical dataset. We also demonstrated that the model can elucidate a specific hypothesis concerning meloxicam and CYP 3A4 inhibitors. Finally, we ranked the overall hepatotoxic risk of eight commonly prescribed NSAIDs. Where applicable, we also compared the model against several common methods for EHR signal detection.
Chapter 7. Machine learning liver-injuring drug interactions with non-steroidal anti-inflammatory drugs (NSAIDs) from retrospective cohort

Figure 7.1: (A) Model architecture for our proposed modeling framework using logistic regression. (B) Variations between independent and dependent relative effect of drugs. Red and blue respectively correspond to positive and negative controls used during the evaluation of diclofenac dependent risk and DILI. Grey corresponds to all other drugs in the hospitalization cohort that were co-prescribed with diclofenac. (C) Distribution of the Twosides-derived positive and negative controls, with respect to model output for diclofenac. The peak around 0 is suspected to be due to a lack of co-occurrence data for those drugs. (D) Variations between independent and dependent relative effect for diclofenac, after elimination of drugs that did not surpass a diclofenac co-occurrence threshold of 10. (E) Distribution of the Twosides-derived positive and negative controls, after elimination of drugs that did not surpass a diclofenac co-occurrence threshold of 10.
7.3.1 Diclofenac dependent risk and DILI

The risk of liver injury with NSAIDs is normally not substantive. Clinical incidence of severe liver injury, resulting from NSAIDs, is 1–10 cases per 100,000 prescriptions [326], with NSAIDs being widely used and clinically ubiquitous. Less severe DILI with mildly elevated liver enzymes is much more common. Moreover, association of NSAIDs with other hepatotoxic drugs is marked with elevated hepatotoxic risk [327, 328]. Potentially, hepatotoxic medications taken simultaneously with NSAIDs may result in a six to nine times increase in frequency of liver injury [329]. In particular, diclofenac is the most common NSAID associated with hepatotoxicity. In fact, 34.1% of hepatotoxic cases associated with NSAIDs involved the use of diclofenac [330].

To analyze diclofenac’s involvement in DILI risk, we trained a model to estimate both independent risk (IR) and diclofenac dependent risk (DDR) of a given drug. The model’s 10-fold cross-validation AUC is $0.68 \pm 0.009$, with a low standard deviation indicating that the model is not overfit. After the training phase, we evaluated the model on the hospitalization cohort and computed the IR and DDR for the remaining unique active ingredients. Fig 1B visualizes the distribution of IR and DDR associations learned by the model for all drugs present in the hospitalization cohort.

Diclofenac is known to independently cause hepatotoxicity. Hence, most drugs co-administered with diclofenac, in cases that result in DILI, are themselves not likely to be the culprits in causing a DILI outcome via interactions with diclofenac. As expected, Fig 1B shows that the majority of the drugs do not have a positive DDR with respect to DILI risk, regardless of their IR. Nevertheless, two drugs that independently cause hepatotoxicity could combine synergistically to have a stronger hepatotoxic effect. The model identifies a few such drugs that have both a positive IR and a positive DDR that is greater than the drug’s IR. Unsurprisingly, there are also few interactions that have a positive IR and negative DDR, which signifies that, individually, hepatotoxic drugs do not become safer in the presence of diclofenac. Going forward, the drugs of most interest will be those that possess low IR but high DDR.
To evaluate the model, we used diclofenac interactions from Twosides as a reference to extract 71 positive controls and 20 negative controls that are also reported in our EHR data. The distribution of model scores, binned by control type, is shown in Fig 1C. On initial inspection, the model not only indicates potential high-priority diclofenac interactions, but also a relatively high density of drugs with DDR as zero. Since output of DDR as zero may be influenced by a lack of co-occurrence between diclofenac and a given drug, we also filtered out drugs below a co-occurrence threshold and replot the scatterplot and histogram in Fig 1D and Fig 1E, respectively. Based on rationale from prior literature, we set the co-occurrence threshold to 10 [331]. As expected, filtering drugs by a co-occurrence threshold lowers the peak. It is to be noted that the peak for positive controls is lowered more than the peak for negative controls. Thus, there is a greater proportion of positive controls than negative controls that are assigned to DDR values as zero, based on an absence of co-occurrence in the data. Likely, the negative controls are not assigned DDR of 0 because of a lack of co-occurrence but because the reported co-occurrence often results in a negative DILI outcome.

To understand how well the model’s top predictions align with Twosides, we focussed on the top 20 diclofenac interactions from Twosides, sorted by PRR. Of the 20 co-prescribed drugs, 4 were not present in our EHR data. Of the remaining 16 co-prescribed drugs, 14 of the interactions had a positive dependent relative effect (Table 7.2). The remaining 2 interactions might have been missed due to a limitation in data availability. In our EHR data, bisoprolol and rivaroxaban each had 0 hospitalizations that involved a DILI positive case with diclofenac co-prescription. In contrast, the Twosides data set contains 3 DILI positive hospitalizations that involved co-administration of rivaroxaban and diclofenac and 6 DILI positive hospitalizations that involved co-administration of bisoprolol and diclofenac.

In addition, we extracted the bottom 10 diclofenac interactions from Twosides; 8 of which were present in our EHR data. 6 of the 8 interactions had a negative dependent relative effect. One explanation for the 2 missed negative controls is that, depending on the available data in our EHR datasets, it is possible for the model to learn differing associations between drug-drug interactions.
interactions, compared to those associations that can be extracted from Twosides. Whereas the EHR data set and Twosides each report single-digit DILI positive hospitalizations involving diclofenac with simvastatin and omeprazole, Twosides DILI negative hospitalizations are a magnitude greater compared to the EHR data set’s DILI negative hospitalizations ($10^3$ for Twosides compared to $10^2$ for our EHR data set). Regardless, the model’s results are statistically significant for both positive and negative controls via a two-sided Fishers exact test ($p$-value < 0.01).

Afterwards, we also examined the top 20 DILI interactions predicted by the model, as sorted by the dependent relative effect. Of these 20 interactions, 12 had a clinical basis reported in the Twosides dataset (Table 7.3). The other 8 prescribed drugs were not in the Twosides dataset for the interactions we used to filter DILI outcomes.

We further cross-referenced each interaction with results from the literature. Several of the co-prescribed drugs in Tables 7.2 and 7.3 have varying degrees of known hepatotoxic associations. The co-prescribed drugs with reported DILI association in literature include acetaminophen [332, 333], amoxicillin [334, 335, 336, 337], aspirin [338], atorvastatin [339, 335], carbamazepine [334], cefazolin [340], cetirizine [341], ciprofloxacin [342, 339, 343], famotidine [344], fluoxetine [345], metformin [346, 347, 348], pioglitazone [349, 350] and topiramate [351, 352]. Not all the reported DILI associations included concomitant consumption of diclofenac, rather combined use of multiple hepatotoxic drugs, such as diclofenac and the aforementioned drugs, is likely to drive Twosides’ reporting of DILI. As an example, the independent relative effect of amoxicillin is 18%, but it becomes more potent in presence of diclofenac and produces a diclofenac dependent relative effect of 35%. Thus, the model can reflect risk for co-prescribed drugs both in presence or absence of the candidate drug.

It is also possible that, in the predicted interactions of positive dependent relative effect, the co-prescribed drug does not promote increased DILI risk. Generally, the co-prescribed drug may not drive the recorded hepatotoxic outcome, but instead can be used during treatments that involve either NSAID administration or the alleviation of hepatotoxic conditions. As an
Table 7.2: *Assessment of positive and negative controls for diclofenac model validation.*

<table>
<thead>
<tr>
<th>Co-prescribed Drugs</th>
<th>Dependent Relative Effect (%)</th>
<th>Twosides PRR</th>
<th>O+ Rx+</th>
<th>O- Rx+</th>
<th>O+ Rx-</th>
<th>O- Rx-</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lidocaine</td>
<td>47.1</td>
<td>30</td>
<td>4</td>
<td>27</td>
<td>1311</td>
<td>17705</td>
<td>+</td>
</tr>
<tr>
<td>Amoxicillin</td>
<td>35.0</td>
<td>40</td>
<td>2</td>
<td>0</td>
<td>103</td>
<td>1884</td>
<td>+</td>
</tr>
<tr>
<td>Acetaminophen</td>
<td>34.6</td>
<td>40</td>
<td>12</td>
<td>240</td>
<td>3260</td>
<td>93351</td>
<td>+</td>
</tr>
<tr>
<td>Olmesartan</td>
<td>33.8</td>
<td>20</td>
<td>2</td>
<td>16</td>
<td>79</td>
<td>2574</td>
<td>+</td>
</tr>
<tr>
<td>Aspirin</td>
<td>27.1</td>
<td>20</td>
<td>11</td>
<td>295</td>
<td>3840</td>
<td>136147</td>
<td>+</td>
</tr>
<tr>
<td>Omeprazole</td>
<td>26.8</td>
<td>0.91</td>
<td>3</td>
<td>53</td>
<td>325</td>
<td>11248</td>
<td>-</td>
</tr>
<tr>
<td>Pioglitazone</td>
<td>26.8</td>
<td>20</td>
<td>1</td>
<td>0</td>
<td>23</td>
<td>739</td>
<td>+</td>
</tr>
<tr>
<td>Famotidine</td>
<td>16.6</td>
<td>20</td>
<td>6</td>
<td>149</td>
<td>2248</td>
<td>73609</td>
<td>+</td>
</tr>
<tr>
<td>Carbamazepine</td>
<td>11.2</td>
<td>16.7</td>
<td>1</td>
<td>9</td>
<td>56</td>
<td>2205</td>
<td>+</td>
</tr>
<tr>
<td>Cefazolin</td>
<td>9.14</td>
<td>30</td>
<td>4</td>
<td>95</td>
<td>1688</td>
<td>53881</td>
<td>+</td>
</tr>
<tr>
<td>Simvastatin</td>
<td>8.13</td>
<td>0.63</td>
<td>5</td>
<td>130</td>
<td>917</td>
<td>31079</td>
<td>-</td>
</tr>
<tr>
<td>Esomeprazole</td>
<td>7.14</td>
<td>23.3</td>
<td>5</td>
<td>70</td>
<td>3023</td>
<td>51870</td>
<td>+</td>
</tr>
<tr>
<td>Escitalopram</td>
<td>6.51</td>
<td>20</td>
<td>2</td>
<td>39</td>
<td>369</td>
<td>11764</td>
<td>+</td>
</tr>
<tr>
<td>Tamsulosin</td>
<td>0.12</td>
<td>30</td>
<td>1</td>
<td>27</td>
<td>135</td>
<td>7805</td>
<td>+</td>
</tr>
<tr>
<td>Enoxaparin</td>
<td>0.017</td>
<td>22.5</td>
<td>11</td>
<td>304</td>
<td>2667</td>
<td>87483</td>
<td>+</td>
</tr>
<tr>
<td>Oxazepam</td>
<td>0.07</td>
<td>30</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>18</td>
<td>+</td>
</tr>
<tr>
<td>Folic Acid</td>
<td>-0.003</td>
<td>0.67</td>
<td>1</td>
<td>37</td>
<td>1249</td>
<td>33131</td>
<td>-</td>
</tr>
<tr>
<td>Rivaroxaban</td>
<td>-0.007</td>
<td>20</td>
<td>0</td>
<td>10</td>
<td>53</td>
<td>2976</td>
<td>+</td>
</tr>
<tr>
<td>Celecoxib</td>
<td>-0.03</td>
<td>0.91</td>
<td>0</td>
<td>12</td>
<td>177</td>
<td>9248</td>
<td>-</td>
</tr>
<tr>
<td>Adalimumab</td>
<td>-0.13</td>
<td>0.91</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>Bisoprolol</td>
<td>-0.32</td>
<td>20</td>
<td>0</td>
<td>6</td>
<td>17</td>
<td>573</td>
<td>+</td>
</tr>
<tr>
<td>Amlodipine</td>
<td>-3.19</td>
<td>0.63</td>
<td>2</td>
<td>67</td>
<td>1293</td>
<td>35437</td>
<td>-</td>
</tr>
<tr>
<td>Dexamethasone</td>
<td>-24.9</td>
<td>0.83</td>
<td>0</td>
<td>24</td>
<td>1053</td>
<td>19860</td>
<td>-</td>
</tr>
<tr>
<td>Morphine</td>
<td>-37.1</td>
<td>0.71</td>
<td>6</td>
<td>212</td>
<td>3241</td>
<td>84935</td>
<td>-</td>
</tr>
</tbody>
</table>

The top 20 interactions, by PRR, were extracted from our filtered Twosides data set and are used as positive controls. Of the 16 positive controls, 14 were successfully captured by the model. The 2 uncaught positive controls reflect a limit in the data availability, as neither positive control had any cases of DILI and diclofenac co-administration from which the model could learn an association. We also extracted 8 interactions with PRR < 1 to be use as negative controls. Of the 8 negative controls, 6 where successfully captured by the model. O+ and O- designates the DILI outcome’s presence and absence, respectively. Rx+ and Rx- designates whether diclofenac is prescribed or not. Grayed out rows indicate diclofenac-drug interactions that may be undersampled based on a co-occurrence threshold of 10.

Example of the former, co-administration of a proton pump inhibitor, such as esomeprazole, can help to prevent NSAID-associated lesions and damage of the upper gastrointestinal tract [353, 198].
354]. With regards to the latter, lidocaine (Table 7.2) is a local anesthetic used widely for minor surgeries or invasive procedures. In the absence of supporting literature, lidocaine’s predicted association with diclofenac may instead be due to a polypharmic approach to pain treatment.

**Table 7.3: The top 12 diclofenac interactions, as predicted by the model.**

<table>
<thead>
<tr>
<th>Co-prescribed Drugs</th>
<th>Percent Dependent Relative Effect</th>
<th>Twosides PRR</th>
<th>O+ Rx+</th>
<th>O- Rx+</th>
<th>O+ Rx-</th>
<th>O- Rx-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ciprofloxacin</td>
<td>136</td>
<td>10</td>
<td>6</td>
<td>43</td>
<td>921</td>
<td>22878</td>
</tr>
<tr>
<td>Fluoxetine</td>
<td>96.3</td>
<td>3.33</td>
<td>3</td>
<td>36</td>
<td>258</td>
<td>9924</td>
</tr>
<tr>
<td>Cetirizine</td>
<td>95.0</td>
<td>2.5</td>
<td>4</td>
<td>58</td>
<td>351</td>
<td>12576</td>
</tr>
<tr>
<td>Atorvastatin</td>
<td>94.8</td>
<td>10</td>
<td>5</td>
<td>75</td>
<td>1390</td>
<td>48234</td>
</tr>
<tr>
<td>Ondansetron</td>
<td>50.6</td>
<td>5</td>
<td>11</td>
<td>236</td>
<td>2550</td>
<td>76289</td>
</tr>
<tr>
<td>Meloxicam</td>
<td>48.3</td>
<td>3.33</td>
<td>2</td>
<td>18</td>
<td>93</td>
<td>4457</td>
</tr>
<tr>
<td>Lidocaine</td>
<td>47.1</td>
<td>30</td>
<td>4</td>
<td>27</td>
<td>1311</td>
<td>17705</td>
</tr>
<tr>
<td>Metformin</td>
<td>42.0</td>
<td>10</td>
<td>3</td>
<td>62</td>
<td>253</td>
<td>11260</td>
</tr>
<tr>
<td>Topiramate</td>
<td>41.2</td>
<td>2</td>
<td>2</td>
<td>25</td>
<td>44</td>
<td>2423</td>
</tr>
<tr>
<td>Amoxicillin</td>
<td>35.0</td>
<td>40</td>
<td>2</td>
<td>0</td>
<td>103</td>
<td>1884</td>
</tr>
<tr>
<td>Acetaminophen</td>
<td>34.6</td>
<td>40</td>
<td>12</td>
<td>240</td>
<td>3260</td>
<td>93351</td>
</tr>
<tr>
<td>Olmesartan</td>
<td>33.8</td>
<td>20</td>
<td>2</td>
<td>16</td>
<td>79</td>
<td>2574</td>
</tr>
</tbody>
</table>

O+ and O- designates the DILI outcome’s presence and absence, respectively. Rx+ and Rx-designates whether diclofenac is prescribed or not. Grayed out rows indicate diclofenac-drug interactions that may be undersampled based on a co-occurrence threshold of 10.

Of most interest are those co-prescribed drugs with less independent hepatotoxic association reported in the literature, but with a high dependent risk predicted by the model – such as olmesartan and meloxicam. The model assigns olmesartan, an antihypertensive, with a high dependent relative effect of 33.8% and Twosides also records olmesartan with a high PRR of 20. As a result, future cohort studies regarding DILI may find it valuable to examine the potentially hepatotoxic contexts of olmesartan.

Meloxicam, an NSAID, only has a PRR of 3.33, yet the model predicted a high dependent relative effect of 48.3% for the interaction. Based on reports in the literature, multi-NSAID therapies may provoke increased risk of hepatic injury, in addition to GI bleeding and acute renal failure [355]. It is also possible that, once patients show DILI from diclofenac, they are switched to meloxicam and this change in prescription causes a spurious association. We expect
that an improved model, which ensures drugs are co-prescribed at the same time and not just present in the same hospitalization, would resolve this question.

**Comparison to data mining algorithms: diclofenac dependent DILI risk**

We compared the drug interaction network against several data mining algorithms for signal detection - relative risk (RR), reporting odds ratio (ROR), multi-item Gamma Poisson shrinker (MGPS), and a one-layer Bayesian confidence propagation neural network (BCPNN). We used the EBGM and the 2.5% quantile of the posterior distribution of the information component as statistics to rank signals for MGPS and BCPNN, respectively. For MGPS, we use DuMouchel’s priors as a default [311]. First, we evaluated the drug interaction network (DIN), along with the RR, ROR, MGPS and BCPNN methods, on the 71 positive controls and 20 negative controls used in the case study on diclofenac dependent DILI risk. As an interaction-less baseline, we also assess performance of a logistic regressor (LR) whose input feature vector contains diclofenac and all coprescribed drugs. For this comparison, we computed the area under the receiver-operating characteristic curve (ROC AUC), the area under the precision-recall curve (PR AUC), and the biserial correlation (BC). BC is a variant of point biserial correlation adjusted for an artificially dichotomized variable with some underlying continuity. Table 7.4 summarizes performance for each method across each metric with 95% two-sided confidence intervals [356, 357].

<table>
<thead>
<tr>
<th>Method</th>
<th>ROC AUC</th>
<th>PR AUC</th>
<th>BC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug Interaction Network</td>
<td>80.3% ± 2.5%</td>
<td>93.7% ± 1.2%</td>
<td>0.63 ± 0.050</td>
</tr>
<tr>
<td>Relative Risk</td>
<td>57.9% ± 3.7%</td>
<td>83.0% ± 2.3%</td>
<td>0.081 ± 0.033</td>
</tr>
<tr>
<td>Reporting Odds Ratio</td>
<td>58.0% ± 3.7%</td>
<td>83.5% ± 2.3%</td>
<td>0.12 ± 0.033</td>
</tr>
<tr>
<td>Multi-item Gamma Poisson Shrinker</td>
<td>78.3% ± 2.7%</td>
<td>90.5% ± 1.6%</td>
<td><strong>0.67 ± 0.046</strong></td>
</tr>
<tr>
<td>Bayesian Confidence Propagation Neural Network</td>
<td>65.9% ± 3.4%</td>
<td>80.9% ± 2.5%</td>
<td>0.49 ± 0.046</td>
</tr>
<tr>
<td>Interaction-less Logistic Regressor</td>
<td>60.9% ± 3.6%</td>
<td>87.5% ± 1.9%</td>
<td>0.24 ± 0.033</td>
</tr>
</tbody>
</table>
Chapter 7. Machine learning liver-injuring drug interactions with non-steroidal anti-inflammatory drugs (NSAIDs) from retrospective cohort

The drug interaction network, with a ROC AUC of 80.3% and a PR AUC of 93.7%, outperformed all methods in the comparison (Fig 2). In decreasing order, MGPS, BCPNN, LR, ROR and RR each had a ROC AUC of 78.3%, 65.9%, 60.9%, 58.0% and 57.9%, respectively, and a PR AUC of 90.5%, 80.9%, 87.5%, 83.5% and 83.0%, respectively. Consistent with the ROC AUC and PR AUC performance, MGPS and the drug interaction network also outperformed the remaining methods with respective BCs of 0.67 and 0.63. Though the drug interaction and MGPS were equivalent in terms of ROC AUC and BC, the drug interaction network had a significantly higher PR AUC than MGPS.

**FIGURE 7.2:** The drug interaction network (DIN) outperforms the compared methods on ROC AUC (left) and, along with MGPS, outperforms the remaining compared methods with regards to PR AUC (right). The receiver-operating characteristic (ROC) and precision-recall (PR) curves and their corresponding areas under the curve (AUC) are reliable metrics to compare the performance of the models.

Compared to the other methods, the drug interaction network and MGPS did better at extracting relevant signals with respect to adverse events reported in Twosides. This is unsurprising, since both methods are intended to build on top of ROR and RR in a way that mitigates variability issues. BCPNN’s performance on this task should be viewed in light of its intended use cases. The motivation behind BCPNN was to extract drug-adverse event signals on increasing large volumes of spontaneously reported adverse drug reactions [313]. Though BCPNNs
may be suitable for handling large data sets, it appears that they are more limited on smaller EHR data sets as analyzed in this case study.

In terms of specific metrics, the drug interaction network and MGPS presented some performance trade offs. The drug interaction network had superior ROC AUC and PR AUC performance compared to MGPS, but MGPS had a better BC. Given routine usage of MGPS as a method of choice for EHR signal detection by organizations such as the US FDA, it is favorable that the drug interaction network outperformed MGPS on ROC AUC and PR AUC and remained competitive on BC [310].

A sensitivity analysis identified consistency of the performance comparison trends for different values of the positive control PRR cutoff (Fig 3). The negative control PRR cutoff follows from prior work using Twosides that filtered out DDIs of no interest using PRR cutoffs of 1 [358], but the positive control PRR cutoff of 5 is more arbitrary. We examined whether the selection criteria for positive controls have a significant influence on performance by evaluating ROC AUC, PR AUC, and BC for each method over PRR cutoffs of 2 through 20, inclusive. The error bars represent 95% two-sided confidence intervals [356, 357]. Up to a PRR of 10, results stay consistent with minor deviations but an overall trend of the drug interaction network and MGPS outperforming the other methods. Above a very high PRR of 10, trends become less defined with greater deviations in performance rankings between each method across different cutoffs, narrower separation of the point estimates, and larger confidence intervals.

Though the drug interaction network and MGPS perform competitively on the assessed tasks, the drug interaction network requires less analytical overhead to use for signal detection. Adequate use of MGPS may require estimating priors for the underlying 5-parameter distribution, requiring additional reasoning and work. In addition to priors, decisions must be made for selection of a decision metric, the decision threshold for the decision metric, and the ranking statistic. We achieve strong performance using default settings recommended in literature, but other problem contexts may require further tuning [311, 310]. MGPS also assumes that the number of reports follows a Poisson distribution, which may be at odds with adverse event
Chapter 7. Machine learning liver-injuring drug interactions with non-steroidal anti-inflammatory drugs (NSAIDs) from retrospective cohort

Figure 7.3: We examined whether the selection criteria for positive controls has a significant influence on performance by evaluating ROC AUC, PR AUC, and BC for each method over PRR cutoffs of 2 through 20, inclusive. The performance comparison remains consistent up to a PRR cutoff of 10, which is very high in the context of DDI reporting and results in less confident performance estimates and rankings.
data sets that can contain many zero count cells. However, this limitation may be temporary as extensions to MGPS continue to be developed [359].

### 7.3.2 Metabolic context of a potential and rare hepatotoxic interaction

Previously, we demonstrated the application of the model to diclofenac, one of the NSAIDs most commonly involved in hepatotoxic treatment outcomes. However, other NSAIDs that result in liver injury at much lower frequencies may require a more targeted application of the model. As an example, meloxicam has been associated with hepatocellular damage, but at a frequency of less than 0.1% of severe hepatotoxic NSAID events [360].

Previous studies have shown that meloxicam detoxification pathways are mediated in part by CYPs 2C9 and 3A4 [205, 361]. Therefore, we expected that inhibitors of CYPs 2C9 or 3A4, when co-prescribed with meloxicam, may result in increased incidence of DILI. Consequently, we trained a model to examine meloxicam’s involvement in drug dependent risk with respect to DILI (10-fold CV AUC of 0.68 ± 0.005).

We posit that CYP 3A4 inhibitors may limit meloxicam detoxification. Conversely, CYP 3A4 inducers may expedite meloxicam detoxification. As a result, we first looked at the model’s ability to separate CYP 3A4 inhibitors and inducers based on drug dependent DILI risk. Across 30 CYP 3A4 inhibitors and 17 CYP 3A4 inducers in the data set, the model achieves a ROC AUC of 84.6% and hints at a relation between CYP 3A4 modulators, meloxicam, and DILI risk.

We then inspected the model’s predictions for interactions with co-prescribed drugs that are known CYP 3A4 inhibitors and when used alongside meloxicam, were represented by at least 100 hospitalization records. We cross-referenced the model’s results against known interactions reported by Twosides to see whether the model can garner novel insights (Table 7.5).

Of the 6 CYP 3A4 inhibitors analyzed, 5 of them have some clinical basis in Twosides that links them to DILI outcomes when co-prescribed with meloxicam. The model predicted a percent dependent relative effect of 41.1% (p-value < 0.05) for the interaction involving meloxicam
Table 7.5: Predicted interactions between meloxicam and several CYP 3A4 inhibitors.

<table>
<thead>
<tr>
<th>Co-prescribed Drugs</th>
<th>Percent Dependent Relative Effect</th>
<th>Twosides PRR</th>
<th>O+ Rx+</th>
<th>O- Rx+</th>
<th>O+ Rx-</th>
<th>O- Rx-</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diltiazem</td>
<td>54.8</td>
<td>2.5</td>
<td>9</td>
<td>222</td>
<td>806</td>
<td>21661</td>
</tr>
<tr>
<td>Esomeprazole</td>
<td>41.1</td>
<td></td>
<td>10</td>
<td>168</td>
<td>3018</td>
<td>51772</td>
</tr>
<tr>
<td>Omeprazole</td>
<td>34.4</td>
<td>2.9</td>
<td>17</td>
<td>493</td>
<td>311</td>
<td>10808</td>
</tr>
<tr>
<td>Amiodarone</td>
<td>22.3</td>
<td>10</td>
<td>4</td>
<td>101</td>
<td>921</td>
<td>21396</td>
</tr>
<tr>
<td>Ciprofloxacin</td>
<td>8.02</td>
<td>5</td>
<td>6</td>
<td>153</td>
<td>921</td>
<td>22768</td>
</tr>
<tr>
<td>Pantoprazole</td>
<td>5.74</td>
<td>1.7</td>
<td>29</td>
<td>1004</td>
<td>3391</td>
<td>97914</td>
</tr>
</tbody>
</table>

O+ and O- designates the DILI outcome’s presence and absence, respectively. Rx+ and Rx- designates whether meloxicam is prescribed or not. Notably, the model predicted a percent relative effect of 41.1% (p-value < 0.05) for the interaction involving meloxicam and esomeprazole, which is a known CYP 3A4 inhibitor and not recorded in Twosides. Furthermore, combination use of proton pump inhibitors (esomeprazole) with NSAIDs (meloxicam) to allay potential GI bleeding is common practice [353] and so the clinical relevance of this interaction is high.

and esomeprazole, which is a known CYP 3A4 inhibitor and not recorded in Twosides. Furthermore, combined usage of proton pump inhibitors (esomeprazole) with NSAIDs (meloxicam) to allay potential GI bleeding is a common practice [353] and so the clinical relevance of this interaction is high. Still, validity of this complex interaction would require further clinical investigation. Nevertheless, our model offers a high-throughput, less resource intensive alternative for enumerating hypotheses concerning deleterious drug-drug interactions.

### 7.3.3 Comparison of NSAID dependent risk to DILI outcomes

In certain treatment contexts, it is not possible to avoid NSAID use. In general, it would be useful if the model could surmise risk and rank the NSAIDs. Here, we demonstrated how well the model estimates overall DILI percent relative effect for eight NSAIDs. For each NSAID, we trained a separate model to examine that NSAID’s DILI associations. Next, for each NSAID and co-prescribed drug, we constructed a contingency table across two variables: DILI outcome (+ or -) and concomitant NSAID use (+ or -). We only retained significant NSAID and co-prescribed drug interactions, as calculated by Fisher’s exact test. Finally, for each NSAID, we
computed the average dependent relative effect (Table 7.6).

The model separates the 8 drugs into two groups based on the mean percent relative effect ($p$-value < 0.1, one-way ANOVA). To validate model rankings, we referenced DILIrank [362] and NSAID-associated DILI outcome frequencies, as reported in the literature [360]. With respect to liver injury cases, diclofenac, ibuprofen and naproxen show high frequencies of 34.1%, 14.6% and 11.1%, respectively. Diclofenac and naproxen belong to the group of NSAIDs with greater predicted DILI association, whereas ibuprofen belongs to the group of lower DILI association. With respect to DILIrank, where a higher severity denotes greater DILI risk, all 3 NSAIDs with high DILI concern and 4 NSAIDs with low DILI concern were correctly grouped. In this case, naproxen stands out as having low DILI concern, yet being grouped with the NSAIDs with greater predicted DILI association.

**Table 7.6: Ranking the 8 studied NSAIDs by mean percent relative effect.**

<table>
<thead>
<tr>
<th>NSAID</th>
<th>Mean Percent Relative Effect</th>
<th>95% CI</th>
<th>DILIrank Severity Class</th>
<th>Percent NSAID Liver Injury Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indomethacin</td>
<td>56.4%</td>
<td>[32.6%, 80.2%]</td>
<td>8</td>
<td>&lt; 0.1%</td>
</tr>
<tr>
<td>Naproxen</td>
<td>48.2%</td>
<td>[23.1%, 73.3%]</td>
<td>3</td>
<td>11.1%</td>
</tr>
<tr>
<td>Etodolac</td>
<td>42.9%</td>
<td>[20.7%, 65.1%]</td>
<td>8</td>
<td>&lt; 0.1%</td>
</tr>
<tr>
<td>Diclofenac</td>
<td>40.5%</td>
<td>[23.8%, 57.1%]</td>
<td>8</td>
<td>34.1%</td>
</tr>
<tr>
<td>Meloxicam</td>
<td>25.3%</td>
<td>[2.18%, 48.5%]</td>
<td>3</td>
<td>&lt; 0.1%</td>
</tr>
<tr>
<td>Celecoxib</td>
<td>25.2%</td>
<td>[13.7%, 36.6%]</td>
<td>3</td>
<td>&lt; 0.1%</td>
</tr>
<tr>
<td>Ibuprofen</td>
<td>22.4%</td>
<td>[15.8%, 28.9%]</td>
<td>3</td>
<td>14.6%</td>
</tr>
<tr>
<td>Ketorolac</td>
<td>21.3%</td>
<td>[14.2%, 28.3%]</td>
<td>3</td>
<td>&lt; 0.1%</td>
</tr>
</tbody>
</table>

Frequencies are based on a prior study derived from 6,023 hospitalizations [360].

There is ambiguity on the basis chosen for reference due to each NSAID’s prescription patterns and patient exposure – commonly prescribed NSAIDs will contribute to greater cases of liver injury due to greater exposure. As a result, there is known heterogeneity in studies on liver injury case frequency of NSAIDs [335, 363]. For example, model groupings for indomethacin, etodolac and ibuprofen do not conform to the grouping that results from using the frequency of liver injury cases across NSAIDs. However, of the 8 NSAIDs, ibuprofen is the most commonly prescribed across the EHRs and indomethacin and etodolac are the 2 least prescribed.
When grouping the NSAIDs for DILI risk using the DILIrank severity class, model rankings for indomethacin, etodolac and ibuprofen become more clear.

**Comparison to data mining algorithms: NSAID dependent DILI risk**

In addition, we also evaluated the drug interaction network and data mining algorithms on the task of ranking the 8 NSAIDs according to DILI risk. For each method, we only retained significant NSAID and co-prescribed drug interactions as calculated by Fisher’s exact test and we output an aggregate NSAID DILI risk by averaging model DILI risk outputs for each NSAID-drug pair. We normalized the aggregate risks for each method and rendered the heat maps in Fig 4 and Fig 5. Each NSAID is binarized into high DILI risk and low DILI risk based on two separate reference points – the DILIrank severity class and the percentage of NSAID liver injury cases reported in a prior study across 6,023 hospitalizations [360].

With respect to the DILIrank severity class binarization, the drug interaction network, RR, ROR and MGPS methods assign high scores to the three NSAIDs with the most DILI risk – indomethacin, etodolac and diclofenac – and to naproxen, which has low DILI risk according to this reference but a high risk according to the percent NSAID liver injury reference. Interestingly, MGPS also assigns high scores to ibuprofen and ketorolac. Though ibuprofen does have DILI risk according to the second binarization reference scheme, ketorolac is indicated as having low DILI risk for both references. Generally, BCPNN does not perform as favorably compared to any of the other methods on this task.

Due to known heterogeneity in studies on liver injury case frequency of NSAIDs [335, 363] and DILIrank’s status as the largest publicly available annotated DILI dataset [362], we place greater weight on the usage of DILIrank as a reference point for NSAID DILI risk. In a comparison of point biserial correlation (PBC) between the model predictions and DILIrank NSAID risk, the drug interaction network and RR outperform the other three methods. The PBC of the drug interaction network, MGPS, ROR, RR and BCPNN are 0.70, 0.54, 0.56, 0.71 and −0.35. The drug interaction network surpasses MGPS, with the biggest distinction between the
Chapter 7. Machine learning liver-injuring drug interactions with non-steroidal anti-inflammatory drugs (NSAIDs) from retrospective cohort

**Figure 7.4:** The drug interaction network results in comparable performance with MGPS, RR and ROR on the task of binarizing NSAIDs by DILIrank severity scores. BCPNN does not perform as favorably. Interestingly, MGPS also assigns high scores to ibuprofen and ketorolac. Though ibuprofen does have DILI risk according to the second binarization reference scheme, ketorolac is indicated as having low DILI risk for both references.
Figure 7.5: The drug interaction network results in comparable performance with RR and ROR on the task of binarizing NSAIDs by the percentage of NSAID liver injury cases [360]. MGPS is the only method to predict DILI risk for diclofenac, ibuprofen, and naproxen, though, along with BCPNN, it also is the only method to predict DILI risk for ketorolac, which is a false positive for both reference points.
two being that the latter method assigns high risk to ketorolac regardless of the chosen reference point.

7.4 Model limitations & future directions

One limitation of the current study is due to clinical data availability. For certain drugs, the model yielded positive results, but there was ultimately not enough data available to describe such results as significant. Furthermore, results demonstrated are specific to the patient cohort accessible via the available data. Even if the model’s learned associations don’t always reflect reference datasets or literature, such inconsistencies may instead be a reflection of limited data for certain interactions or of a patient cohort that doesn’t reflect those cohorts used to construct the referential data or literature.

The proposed modeling framework was trained using each hospitalization instance as a datapoint. Hence, one patient, having multiple hospital visits will contribute multiple training instances in the training dataset. This was done to capture meaningful drug interactions within each hospitalization timeline. Concatenating multiple hospitalization timelines into a single datapoint for each patient would lead to interactions between drugs not prescribed in the same time window. However, for rare drug interactions, it may so happen that those are from one patient across multiple hospitalizations thereby leading to poor generalization of results.

In this study, our proposed modeling framework was used as a signal detection algorithm capable of estimating the independent and dependent relative risks of drugs on the clinical outcome. We highlighted the potential utility of our modeling framework in estimating risks of drug exposures from relatively small EHR datasets with known denominators rather than from FAERS database where most incidence rates are estimated with unknown denominators. EHR datasets are an under-utilized resource for studying drug interaction discovery and our research study aims to highlight the benefits of using EHR datasets for this purpose.
The results, presented in this study, have been cross-referenced with other published works as well as previously known interactions from the FAERS database. It is quite plausible that factors such as other comorbidities, other drug exposures both within and outside the hospitalization window and length of hospitalization may confound some findings. A key advantage of EHR datasets for drug interaction discovery is that they contain different data streams such as demographics, hospitalization stay and other drug exposures during a hospitalization timeline whereas adverse reports in FAERS database usually do not contain this additional information. However, in EHR datasets, complex underlying causal relationships exist between different variables and the clinical outcome. Adjusting for these confounding factors was not within the scope of this research study. Future studies include using the drug interaction network in conjunction with the proposed framework by Datta et al. \cite{320} to identify and adjust for potential confounding variables. However, for questions in which other pieces of information are necessary, such as drug exposure outside the hospitalization timeline and environmental or behavioral variables, accurate inferences are unlikely to be made solely from EHRs.

Age is often considered an influential confounder in clinical studies involving adverse drug reactions and more than 60% of our hospitalization data did not have any age information associated with them. However, age should not be a confounder for drug interactions which was the key focus of this research study. Also, age was not used as an input variable in our modeling framework in this research study. Furthermore, the findings in this study have been validated using results published in prior studies using FAERS and Twosides databases.

In addition, the manner in which diagnosis, procedure, or other hospitalization codes are used to define possible outcome definitions can lead to ambiguity. Different models can be developed based on the method chosen for applying hospitalization codes or other clinical features, such as the levels of certain aminotransferases or bilirubin, to infer DILI hospitalizations. Ultimately, the method used to define the outcome definition from the available clinical features may depend on the manner in which data was collected for a specific cohort and the target outcome to be studied, e.g., liver, renal, cardiovascular, or other clinical risks.
Lastly, the described approach avoids learning a full pairwise matrix of interactions, which aids in a reduction of learnable parameters and leads to a more focused query. However, multiple models may be required when trying to answer more general queries. Furthermore, a model tasked with predicting many more outputs can lead to a model with better generalization. In future studies, we plan on using interaction detection frameworks \[364\] for interpreting weights in non-linear extensions to the drug interaction network.

7.5 Conclusion

In this work, we propose a modeling framework to study drug-drug interactions that may lead to adverse outcomes using EHR datasets. As a case study, we used our proposed modeling framework to study pairwise drug interactions involving NSAIDs that lead to DILI. We validated our research findings using previous research studies on FAERS and Twosides databases. Empirically, we showed that our modeling framework is successful at inferring known drug-drug interactions from relatively small EHR datasets (less than 400,000 hospitalizations) and our modeling framework’s performance is robust across a wide variety of empirical studies. Our research study highlights the numerous benefits of using EHR datasets over public datasets such as FAERS database for studying drug interactions. In the analysis for diclofenac, the model identified drug interactions associated with DILI, including each co-prescribed drug’s independent risk when administered in absence of the candidate drug, e.g., diclofenac and dependent risk in the presence of the candidate drug. We have explored how prior knowledge of a drug’s metabolism, such as meloxicam’s detoxification pathways, can inform exploratory analysis of how combinations of drugs can result in increased DILI risk. Strikingly, the model indicates a potentially harmful outcome for the interaction between meloxicam and esomeprazole, confirmed by metabolic and clinical knowledge. Though beyond the scope of this computational study, these preliminary results suggest the applicability of a joint approach – models of drug interactions within EHR data streamlined by knowledge of metabolic factors, such as those that
Chapter 7. Machine learning liver-injuring drug interactions with non-steroidal anti-inflammatory drugs (NSAIDs) from retrospective cohort

affect P450 activity in conjunction with hepatotoxic events. We have also studied the ability of the model to rank commonly prescribed NSAIDs with respect to DILI risk. NSAIDs undergo widespread usage and are, therapeutically, valuable agents for relief of pain and inflammation. When use of a class of drugs is unavoidable, it is still valuable to select a specific candidate from that class of drugs that is least likely to incur patients’ harm. These results are important because EHR data is increasingly available and may prove to be a more effective approach in mining drug-drug interactions. We believe that the proposed framework in this study will be widely applicable for understanding drug interactions resulting in diverse adverse outcomes using EHR datasets and pave the way for incorporating future analyses based on dosage responses as well as accounting for comorbidities and confounding.
Chapter 8

Conclusion and Future Directions

We successfully modeled multi-step sequences of metabolic transformations, i.e., pathways, between an input molecule and a corresponding, optional reactive metabolite(s) using a combination of site of metabolism and structure inference models. Additionally, an accurate graph neural network model was developed to assess importance of intermediate metabolites and extract connected subnetworks of relevance to bioactivation. Connecting multiple site of metabolism and structure inference models, we developed an integrated model of metabolism and reactivity to evaluate bioactivation risk driven by epoxidation, quinone formation, thiophene sulfuration, and nitroaromatic reduction. We applied this framework to an understudied substructure, the isoxazole ring, that is gaining traction in a class of drugs known as bromodomain inhibitors that may potentially drive quinone formation. Finally, we attend to toxicity associated with drug-drug interactions, particularly with NSAID usage reported in electronic health records.

However, predicting idiosyncratic adverse drug reactions provoked by bioactivation and reactive metabolite formation is complex. The following fundamental challenges remain. First, our work doesn’t comprehensively cover metabolic transformations such as tautomerization, isomerization, and rearrangement. The structure inference model supports these reactions, but development of corresponding site of metabolism models is necessary. Tangentially, the bioactivation model (chapter 6) would be more effective if generalized to additional Phase I metabolism rules. Second, we can improve our integrated methods for predicting complete
Chapter 8. Conclusion and Future Directions

metabolic networks for a given molecule. XenoNet (chapter 2) is practically functional up to a depth limit of 3, but support for greater depth limits may be possible with better search heuristics and pruning methods. Third, there is a disparity in the amount of data available across the Phase I reaction types. The bidirectional MPNN (chapter 3) may be used to extract metabolic subnetworks that are likely to be authentic, but not recorded in training or test data. Potentially, a cyclic mechanism could be developed where the reactions within these subnetworks are fed back into XenoNet and its foundational models to improve model performance.

Fourth, the models make simplifying assumptions and do not explicitly account for certain aspects of drug administration and metabolism, namely dosage, kinetics, and co-administered medicines. We modeled toxicity due to certain drug-drug interactions (chapter 7), but further work remains for integrating those models and our prior work into a joint approach. Fifth, patient-specific differences in metabolism are not accounted for, such as genetic and environmental factors, but also variations in CYP enzymes across pediatric and adult patients. Sixth, XenoNet struggles to visualize large metabolic networks (greater than 100 molecules) due to its static output. Support for dynamic visualizations would improve user experience due to greater customizability and ease of use.

Accounting for all of these factors appears daunting. The work described in this thesis takes steady, actionable steps towards solving these problems in smaller chunks. Furthermore, our work composes a valuable workhorse, internally and externally, as part of our ongoing effort to develop XenoSite and XenoNet for analysis of small molecule metabolism. We made major steps forward towards a comprehensive model for rapidly screening large numbers of molecules for several key toxicity risks governed by formation of reactive metabolites, bioactivation, and drug-drug interactions. Ultimately, these approaches aid in development of novel medicines with improved safety profiles.
Bibliography


Bibliography


[22] Paul L Skipper et al. “Monocyclic aromatic amines as potential human carcinogens: old is new again”. In: *Carcinogenesis* 31 (2009), pp. 50–58.


Bibliography


220
Bibliography


[75] Uri Alon and Eran Yahav. “On the Bottleneck of Graph Neural Networks and its Practical Implications”. In: (2021).


[97] Larry Page et al. The PageRank Citation Ranking: Bringing Order to the Web. 1998.


Bibliography


Bibliography


[129] GailD. Anderson. “Children Versus Adults: Pharmacokinetic and Adverse-Effect Differences”. In: Epilepsia 43 (June 2002), pp. 53–59. ISSN: 0013-9580, 1528-1167. DOI: 10.1046/j.1528-1157.43.s.3.5.x. URL: http://doi.wiley.com/10.1046/j.1528-1157.43.s.3.5.x (visited on 05/14/2021).


Mary Alexandra Schleiff et al. “Significance of Multiple Bioactivation Pathways for Meclofenamate as Revealed through Modeling and Reaction Kinetics”. en. In: Drug Metabolism and Disposition 49.2 (Feb. 2021), pp. 133–141. ISSN: 0090-9556, 1521-009X. DOI: 10.1124/dmd.120.000254. URL: http://dmd.aspetjournals.org/lookup/doi/10.1124/dmd.120.000254 (visited on 05/03/2021).


233


Bibliography


Bibliography


Bibliography


[231] Cassie Frank et al. “Era of faster FDA drug approval has also seen increased black-box warnings and market withdrawals”. In: Health Affairs 33.8 (2014), pp. 1453–1459.


[240] Qiang Qu et al. “Nrf2 protects against furosemide-induced hepatotoxicity”. In: Toxicology 324 (2014), pp. 35–42.


Bibliography


242


Bibliography


[319] Patricia McGettigan and David Henry. “Use of Non-Steroidal Anti-Inflammatory Drugs That Elevate Cardiovascular Risk: An Examination of Sales and Essential Medicines Lists in Low-, Middle-, and High-Income Countries”. In: PLOS Medicine 10.2 (Feb. 2013), pp. 1–6. DOI: 10.1371/journal.pmed.1001388. URL: https://doi.org/10.1371/journal.pmed.1001388.


Bibliography


248
Bibliography


