GRAPH-BASED FORWARD SYNTHESIS PREDICTION OF BIOCATALYZED REACTIONS

Peter G. Mikhael[*], **Itamar Chinn**[*] **& Regina Barzilay**
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
{pgmikhael,itamarc,regina}@csail.mit.edu

ABSTRACT

The identification of biocatalyzed reaction products plays a critical role in enzyme function prediction, drug discovery, and metabolic engineering. Uncovering the products of biocatalyzed reactions experimentally is both time-consuming and costly, which underscores the urgent need for computational methods. Previous machine learning methods have largely focused on spontaneous, non-biocatalyzed reactions but do not perform well when applied to biocatalyzed reactions specifically. We present a novel approach that harnesses graph-based deep learning to predict the primary products of enzyme-catalyzed reactions, considering both the protein sequence and substrates involved. On the recently published dataset EnzymeMap, we find that our method based on graph-editing outperforms existing transformer-based approaches.

1 INTRODUCTION

A key computational task in biocatalysis is predicting the products of a reaction from an enzyme and its substrates. *In silico* methods for this task enable new opportunities in enzyme discovery, therapeutic development, and metabolic engineering. Current machine learning models have shown initial feasibility at automating this process; however, thus far they rely on information that may not be available for novel chemistry (e.g. Enzyme Commission number). As a result, this limits their practical use as an alternative to experimental methods.

The goal of our work is to improve the generalization capacity of these models to new chemistry. To this end, we assume access to only the molecular structure of the substrates and the enzyme primary sequence, without any additional information. In predicting the products of spontaneous chemical reactions, graph-based methods have outperformed both language model and rule-based approaches. These methods, however, fail to take into consideration the enzyme and therefore experience a significant drop in performance on biocatalyzed reactions. We hypothesized that better generalization can be achieved by a mechanistically-inspired model that captures the biochemical interaction between the enzyme residues and substrate atoms. We demonstrate that these interactions can be learned through a multi-headed cross attention using graph convolutions to encode the substrates as 2D molecular graphs and a protein language model to encode the enzyme's amino acid sequence.

In the context of drug design, the metabolism of small molecule drugs impacts their efficacy, toxicity, and mechanism of action. For example, Fenofibrate, which is used to treat high cholesterol, must first be metabolized into fenofibric acid by liver carboxylesterase 1 in order to become active. Therefore in Appendix A.3, we consider phase II metabolism of small molecule drugs as a potential real-world application and a prime example of generalization to a novel chemical space. Specifically, we use a dataset of drug reactions from DrugBank to predict the products generated by the reaction of a drug with its target enzyme. This is an interesting and challenging generalization scenario since the chemical distribution of therapeutics differs significantly to that of metabolites on which these models are trained.

[*]Equal contribution

We develop our model using the EnzymeMap dataset, consisting of 103,120 pairs of atom-mapped reactions and UniProt-SwissProt proteins. We demonstrate a significant improvement in predicting unseen products on a standard product split. For instance, we obtain 89% accuracy in generating correct products when evaluating the top 10 predictions and outperform current methods that range between 50%-70%. The comparison between our method and previous methods highlights the importance of adequate enzyme encoding. Ignoring the enzyme altogether or utilizing the protein EC numbers leads to significantly worse performance (Probst et al., 2022; Kreutter et al., 2021). Finally we show comparable improvements using a dataset from DrugBank in Appendix A.3.

2 Related Work

See Appendix A.2 for full details.

3 Method

3.1 Biocatalyzed Product Generation

We present here an overview of the method. We first predict whether and how the reactant bonds change conditioned on both the set of reactants and the protein sequence of the associated enzyme (Section 3.1.1). We deterministically perform chemically valid graph edits to obtain all products that can be generated with up to k of the most likely predicted bond changes. We train a second model to retrieve the correct product given the full reaction and the protein sequence (Section 3.1.2).

3.1.1 Reaction Center Prediction

Reactants and products are constructed as 2D graphs $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with node features $v_i \in \mathcal{V}$ and edges $e_{ij} \in \mathcal{E}$. While the bonds we predict correspond to the overall net change between the atom-mapped reactants and products, they are nonetheless dependent on chemical interactions between atoms in the same reactants, atoms in different reactants, and the enzyme amino acid residues. We model each type of interaction and use them together to predict all bond changes.

First, a Graph Attention Network (Brody et al., 2022) f_{local} is used to encode each reactant separately and obtain node embeddings for each atom:

$$A = f_{\text{local}}(\mathcal{V}, \mathcal{E})$$

where $A = \{a_1, a_2, \ldots, a_n\}$ is the set of reactant node features after applying the GNN and $a_i \in \mathbb{R}^d$.

Similarly to Jin et al. (2017), a second model, f_{global}, is then used to encode the interaction between atoms in different molecules by constructing a complete graph from the reactants. Specifically we add an edge between every pair of nodes: $e'_{ij} = [\mathbb{1}_{\text{same}} \parallel \mathbb{1}_{\text{diff}} \parallel e_{ij}]$, where $\mathbb{1}_{\text{same}}$ indicates whether the atoms are in the same molecule, $\mathbb{1}_{\text{diff}}$ indicates whether the atoms are in different molecules, and e_{ij} are the bond features. We set $e_{ij} = \mathbf{0}$ when the atoms are not connected by a chemical bond. We compute a pairwise attention with every atom in the complete graph and obtain the global node embeddings $a'_i \in \mathbb{R}^d$ as a weighted sum:

$$\begin{aligned} \alpha_{ij} &= \sigma(\mathbf{u}^\top \text{ReLU}\left(P_a(a_i + a_j) + P_b e_{ij}\right)) \\ a'_i &= \sum_j \alpha_{ij} a_j \\ A' &= \{a'_1, a'_2, \ldots, a'_n\} \end{aligned}$$

Third, we use ESM-2 (Lin et al., 2023) as the protein encoder f_p to obtain residue-level representations $P = \{r_1, r_2, \ldots, r_m\}$ and perform a multi-headed cross-attention (Vaswani et al., 2017) between the residues and the node embeddings of the reactant graphs a_i:

$$A'' = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

where

$$Q = W^Q A; \qquad K = W^K P; \qquad V = W^V P$$

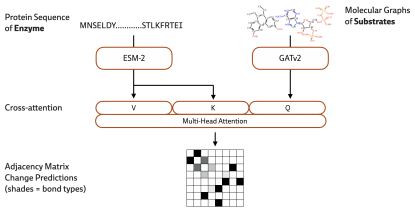

Figure 1: Schematic of the model architecture for predicting the bond changes associated with a given an enzyme and its substrates.

Finally, for each atom pair (i, j), we compute the probability that a particular bond change k occurs between them, which consists of either the loss of a bond or the formation of a single, double, triple, or aromatic bond:

$$\begin{aligned} c_i &= W_a[a_i \parallel a_i' \parallel a_i''] \\ b_{ij} &= W_b e_{ij} \\ s_{ijk} &= W_k \text{ReLU}\left([c_i + c_j \parallel b_{ij}]\right) \end{aligned}$$

To force the model to focus on bond changes associated with substrate, we do no compute the loss over bond changes associated with common co-factors and co-enzymes like ATP, which often comprise most of the bond changes associated with the reaction.

3.1.2 Candidate Product Ranking

Given the predicted bond changes above, we select the top k predictions. We empirically predefine a k' as the maximum number of changes that could occur within a biochemical reaction and construct all sets of size at most k' consisting of chemically valid changes. Each set of bond changes is applied as graph edits on the original reactant graphs to obtain candidate products. We then train a classifier to retrieve the products associated with the ground truth set of changes from the list of all candidate products.

The identity of the correct product depends on the reactants and enzyme, and the most likely products are those whose transition state is stabilized by the enzyme (Martí et al., 2004; Warshel et al., 2006). As a result, we represent a pseudo-transition state using the condensed reaction graph (Hoonakker et al., 2011; Heid & Green, 2021) for each prediction by superimposing the reactants and generated products and concatenating their node and edge features. This aims to incorporate all representations of the predicted reaction and the enzyme together. We then encode the graph structure with a directed message passing neural network f_{rxn} (Yang et al., 2019) to obtain atom-level features a_i and obtain residue-level features r_i of the enzyme using ESM-2.

$$\begin{aligned} a_i &= f_{\text{rxn}}\left(\left[v_i^{(\text{reactants})} \parallel v_i^{(\text{products})}\right], \left[e_{ij}^{(\text{reactants})} \parallel e_{ij}^{(\text{products})}\right]\right) \\ a &= \sum_i a_i; \qquad p = \frac{1}{|P|}\sum_i r_i \\ g &= f_{\text{rank}}([a \parallel p]) \end{aligned}$$

Finally, we aggregate both the reaction graph representations and the protein representations, and pass them together through a small feed-forward network, f_{rank}, to score each proposed reaction.

4 Experimental Setup

See Appendix A.1 for full details.

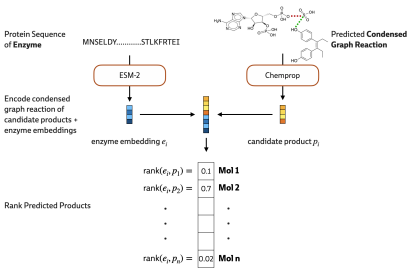

Figure 2: Schematic of the ranking model used to select the correct product from a list of candidates by considering the enzyme and the full predicted reaction. The red and green dashed edges represent bonds that are predicted to be deleted and created, respectively.

5 RESULTS

5.1 ENZYMEMAP

While prior deep learning methods developed specifically for this task use more detailed data on the enzyme identity (either the EC number or the enzyme nomenclature) our method assumes that only the amino-acid sequence of the enzyme and the substrate molecules are known. However, we compare against these methods for completeness. Additionally, we impose a conservation of mass constraint and only generate bond changes whereas existing baselines use free generation to decode the product SMILES.

We test the hypothesis that encoding the protein and molecular structure leads to better generalizability in predicting unseen products. We find that our model is able to generalize better to unseen reaction products and surpasses other models by a considerable margin with a top-1 accuracy of 72.5% relative to 35.3% and 50.5% using enzyme name and EC, respectively (Table 5). Since reactions can have multiple possible products, we expect that not all products can be recovered within the first prediction. Considering the top $k > 1$ predictions, we observe sustained performance gains in recovering all products, approaching 90% accuracy with $k = 10$. We also consider other biologically relevant splits based on protein structure similarity and the reaction classes defined by EC numbers, and observe that our model exhibits comparable on these harder splits, albeit without assuming any additional protein annotations (Appendix A.5).

5.2 IMPACT OF PROTEIN SEQUENCE

Enzymes play an important role in biocatalysis. However, since the molecular structure of the substrates alone provides some information about the potential sites of metabolism (Kirchmair et al., 2015), we sought to evaluate the extent to which these models simply memorize reaction rules versus take into account the impact of the enzyme itself. Here, we show how well each model predicts the products of enzymatic reactions from the reactants alone without enzyme information. We train both the Molecular Transformer architecture (Schwaller et al., 2019) and WLN (Coley et al., 2017) without incorporating any protein information. Since our primary task is to generalize

Table 1: Top-k accuracy of our graph-based method compared to existing approaches for biocatalyzed forward synthesis. Published methods are trained as detailed in their respective GitHub codebases (Appendix A.6). Performance is evaluated on EnzymeMap using a product split.

MODEL	TOP 1	TOP 3	TOP 5	TOP 10
KREUTTER ET AL. (2021)	35.3%	43.6%	46.0%	47.8%
PROBST ET AL. (2022)	50.5%	61.7%	65.4%	68.8%
OURS	72.5%	84.3%	87.3%	89.4%

Table 2: Top-k accuracy of a transformer and graph model that exclude protein information compared to our full model. Performance is evaluated on EnzymeMap using data splits based on a product split.

MODEL	TOP 1	TOP 3	TOP 5	TOP 10
SCHWALLER ET AL. (2019)	35.0%	50.6%	55.5%	58.9%
COLEY ET AL. (2017)	58.3%	75.9%	81.8%	85.2%
OURS	72.5%	84.3%	87.3%	89.4%

to new products, we focus our analysis on the product split. We observe that both models achieve improved performance when the protein sequence is included (Table 2), with the top 1 accuracy of our method obtaining a 14% gain in performance relative to the WLN model (no protein sequence). However, this gap decreases to 4% as more candidates are considered (top k=10). We also find that both graph-based models perform better over sequence-to-sequence models.

While the results of Table 2 suggest that the model is utilizing the protein sequence in improving its final prediction, they provide no indication whether it learns any biologically meaningful properties regarding the protein's catalytic function. Our architecture, however, learns a multi-head cross-attention between the full protein sequence and the latent atom representations of the substrates, yielding attention scores for every residue-atom pair. By summing over the attentions scores across all atoms, we obtain a weighting per residue. We extract active site annotations from the Mechanism and Catalytic Site Atlas (Ribeiro et al., 2018) for both the reference sequences as well their homologs, which are assumed to have identical active sites, and we compare them with the top-scoring residues according to the learned attention scores. We take the residues with the top q-th quantile of attention scores and compute the fraction of annotated active site residues included in that predicted set. We find that our learned attention has a consistently better correspondence with the active site than an equivalent random guess (Figure 3). This suggests that our model is able to learn a functionally meaningful association between the protein sequence and the substrates.

6 CONCLUSION

This paper presents a novel graph-based method for predicting the products of biocatalyzed reactions given a set of substrates and an enzyme sequence. We show that incorporating the enzyme sequence in the input improves performance compared to other methods that include alternative representations of enzymes, namely EC numbers and enzyme names. We report an improvement of 37.2 points in top-1 accuracy against preceding state-of-the-art methods on the EnzymeMap dataset.

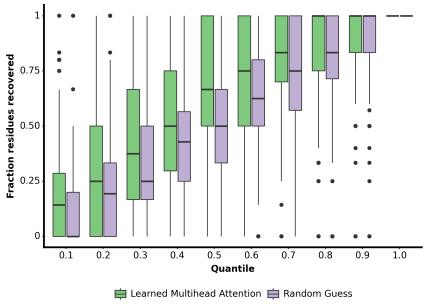

Figure 3: Fraction of true active site residues included in the top q quantile of attention scores extracted from the multi-head cross-attention layer used to predict the bond changes in each reaction. For every quantile, we take a random permutation over all residue indices and select the same number of predictions as in that quantile to obtain a random guess baseline.

Lastly, we note that by relying on enzyme sequence, we widen the utility of our model compared to previous models to encompass unannotated and orphan enzymes.

The results presented also exhibit a number of limitations. While we show that the model has a capacity to generalize to out-of-distribution molecules, like small molecule drugs in Appendix A.3, there still remains room for improvement especially for completely new chemical transformations.

ACKNOWLEDGMENTS

This work is supported by the Jameel Clinic for AI and Health at MIT and Novo Nordisk A/S. We are also grateful for David Sabatini, Doug Wheeler, Mathai Mammen, Esther Heid, and Jeremy Wohlwend for their insightful discussions.

REFERENCES

Samuel E Adams. *Molecular similarity and xenobiotic metabolism*. PhD thesis, University of Cambridge, 2010.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Hangrui Bi, Hengyi Wang, Chence Shi, Connor Coley, Jian Tang, and Hongyu Guo. Non-autoregressive electron redistribution modeling for reaction prediction. In *International Conference on Machine Learning*, pp. 904–913. PMLR, 2021.

John Bradshaw, Matt J Kusner, Brooks Paige, Marwin HS Segler, and José Miguel Hernández-Lobato. A generative model for electron paths. *arXiv preprint arXiv:1805.10970*, 2018.

Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? 2022. URL `https://openreview.net/forum?id=link-to-the-paper-if-available`.

Shuan Chen and Yousung Jung. A generalized-template-based graph neural network for accurate organic reactivity prediction. *Nature Machine Intelligence*, 4(9):772–780, 2022.

Connor W Coley, Regina Barzilay, Tommi S Jaakkola, William H Green, and Klavs F Jensen. Prediction of organic reaction outcomes using machine learning. *ACS central science*, 3(5):434–443, 2017.

Gabriele Cruciani, Emanuele Carosati, Benoit De Boeck, Kantharaj Ethirajulu, Claire Mackie, Trevor Howe, and Riccardo Vianello. Metasite: understanding metabolism in human cytochromes from the perspective of the chemist. *Journal of medicinal chemistry*, 48(22):6970–6979, 2005.

Ferenc Darvas. Metabolexpert: an expert system for predicting metabolism of substances. In *QSAR in environmental toxicology-II*, pp. 71–81. Springer, 1987.

Christina de Bruyn Kops, Martin Šícho, Angelica Mazzolari, and Johannes Kirchmair. Gloryx: prediction of the metabolites resulting from phase 1 and phase 2 biotransformations of xenobiotics. *Chemical research in toxicology*, 34(2):286–299, 2020.

Yannick Djoumbou-Feunang, Jarlei Fiamoncini, Alberto Gil-de-la Fuente, Russell Greiner, Claudine Manach, and David S Wishart. Biotransformer: a comprehensive computational tool for small molecule metabolism prediction and metabolite identification. *Journal of cheminformatics*, 11(1):1–25, 2019.

Janani Durairaj, Alice Di Girolamo, Harro J Bouwmeester, Dick de Ridder, Jules Beekwilder, and Aalt DJ van Dijk. An analysis of characterized plant sesquiterpene synthases. *Phytochemistry*, 158:157–165, 2019.

Arndt R Finkelmann, Daria Goldmann, Gisbert Schneider, and Andreas H Göller. Metscore: site of metabolism prediction beyond cytochrome p450 enzymes. *ChemMedChem*, 13(21):2281–2289, 2018.

Anja Greule, Jeanette E. Stok, James J. De Voss, and Max J. Cryle. Unrivalled diversity: the many roles and reactions of bacterial cytochromes p450 in secondary metabolism. *Nat. Prod. Rep.*, 35:757–791, 2018. doi: 10.1039/C7NP00063D. URL http://dx.doi.org/10.1039/C7NP00063D.

Esther Heid and William H Green. Machine learning of reaction properties via learned representations of the condensed graph of reaction. *Journal of Chemical Information and Modeling*, 62(9): 2101–2110, 2021.

Esther Heid, Daniel Probst, William H Green, and Georg KH Madsen. Enzymemap: Curation, validation and data-driven prediction of enzymatic reactions. 2023.

Matthias Hennemann, Arno Friedl, Mario Lobell, Jörg Keldenich, Alexander Hillisch, Timothy Clark, and Andreas H Göller. Cypscore: Quantitative prediction of reactivity toward cytochromes p450 based on semiempirical molecular orbital theory. *ChemMedChem: Chemistry Enabling Drug Discovery*, 4(4):657–669, 2009.

Frank Hoonakker, Nicolas Lachiche, Alexandre Varnek, and Alain Wagner. Condensed graph of reaction: considering a chemical reaction as one single pseudo molecule. *Int. J. Artif. Intell. Tools*, 20(2):253–270, 2011.

Tyler B Hughes and S Joshua Swamidass. Deep learning to predict the formation of quinone species in drug metabolism. *Chemical research in toxicology*, 30(2):642–656, 2017.

Wengong Jin, Connor Coley, Regina Barzilay, and Tommi Jaakkola. Predicting organic reaction outcomes with weisfeiler-lehman network. *Advances in neural information processing systems*, 30, 2017.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Dhiraj Kalamkar, Dheevatsa Mudigere, Naveen Mellempudi, Dipankar Das, Kunal Banerjee, Sasikanth Avancha, Dharma Teja Vooturi, Nataraj Jammalamadaka, Jianyu Huang, Hector Yuen, et al. A study of bfloat16 for deep learning training. *arXiv preprint arXiv:1905.12322*, 2019.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Johannes Kirchmair, Mark J Williamson, Avid M Afzal, Jonathan D Tyzack, Alison PK Choy, Andrew Howlett, Patrik Rydberg, and Robert C Glen. Fast metabolizer (fame): A rapid and accurate predictor of sites of metabolism in multiple species by endogenous enzymes. *Journal of chemical information and modeling*, 53(11):2896–2907, 2013.

Johannes Kirchmair, Andreas H Göller, Dieter Lang, Jens Kunze, Bernard Testa, Ian D Wilson, Robert C Glen, and Gisbert Schneider. Predicting drug metabolism: experiment and/or computation? *Nature reviews Drug discovery*, 14(6):387–404, 2015.

D Kreutter, P Schwaller, and JL Reymond. Predicting enzymatic reactions with a molecular transformer, chem. *Sci*, 12(25):8648–8659, 2021.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Daniel Lowe. Chemical reactions from us patents (1976-sep2016), 2017. *DOI*, 10:m9, 1976.

Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, University of Cambridge, 2012.

Sergio Martí, Maite Roca, Juan Andrés, Vicent Moliner, Estanislao Silla, Iñaki Tuñón, and Juan Bertrán. Theoretical insights in enzyme catalysis. *Chemical Society Reviews*, 33(2):98–107, 2004.

Homa Mohammadi Peyhani, Anush Chiappino-Pepe, Kiandokht Haddadi, Jasmin Hafner, Noushin Hadadi, and Vassily Hatzimanikatis. Database for drug metabolism and comparisons, nicedrug. ch, aids discovery and design. *bioRxiv*, pp. 2020–05, 2020.

Lars Olsen, Marco Montefiori, Khanhvi Phuc Tran, and Flemming Steen Jørgensen. Smartcyp 3.0: enhanced cytochrome p450 site-of-metabolism prediction server. *Bioinformatics*, 35(17):3174–3175, 2019.

Marina V Omelchenko, Michael Y Galperin, Yuri I Wolf, and Eugene V Koonin. Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biology direct*, 5:1–20, 2010.

Daniel Probst, Matteo Manica, Yves Gaetan Nana Teukam, Alessandro Castrogiovanni, Federico Paratore, and Teodoro Laino. Biocatalysed synthesis planning using data-driven learning. *Nature communications*, 13(1):964, 2022.

António J M Ribeiro, Gemma L Holliday, Nicholas Furnham, Jonathan D Tyzack, Katherine Ferris, and Janet M Thornton. Mechanism and catalytic site atlas (m-csa): a database of enzyme reaction mechanisms and active sites. *Nucleic acids research*, 46(D1):D618–D623, 2018.

Lars Ridder and Markus Wagener. Sygma: combining expert knowledge and empirical scoring in the prediction of metabolites. *ChemMedChem: Chemistry Enabling Drug Discovery*, 3(5): 821–832, 2008.

Anastasia V Rudik, Alexander V Dmitriev, Alexey A Lagunin, Dmitry A Filimonov, and Vladimir V Poroikov. Metatox 2.0: Estimating the biological activity spectra of drug-like compounds taking into account probable biotransformations. *ACS omega*, 2023.

Mikołaj Sacha, Mikołaj Błaz, Piotr Byrski, Paweł Dabrowski-Tumanski, Mikołaj Chrominski, Rafał Loska, Paweł Włodarczyk-Pruszynski, and Stanisław Jastrzebski. Molecule edit graph attention network: modeling chemical reactions as sequences of graph edits. *Journal of Chemical Information and Modeling*, 61(7):3273–3284, 2021.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583, 2019.

Marwin HS Segler and Mark P Waller. Neural-symbolic machine learning for retrosynthesis and reaction prediction. *Chemistry–A European Journal*, 23(25):5966–5971, 2017.

Michel van Kempen, Stephanie S Kim, Charlotte Tumescheit, Milot Mirdita, Jeongjae Lee, Cameron LM Gilchrist, Johannes Söding, and Martin Steinegger. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, pp. 1–4, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Arieh Warshel, Pankaz K Sharma, Mitsunori Kato, Yun Xiang, Hanbin Liu, and Mats HM Olsson. Electrostatic basis for enzyme catalysis. *Chemical reviews*, 106(8):3210–3235, 2006.

David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research*, 46(D1):D1074–D1082, 2018.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

A APPENDIX

A.1 EXPERIMENTAL SETUP

A.1.1 ENZYMEMAP DATASET

We train all models on data derived from EnzymeMap (Heid et al., 2023), which consists of biocatalyzed reactions paired with protein UniProt identifiers and their EC numbers. All reactions are fully atom-mapped, meaning that every atom in the products can be traced back to an atom in the reactants of the reaction. To obtain protein sequences, we consider only reactions associated with UniProt or SwissProt identifiers and pull their sequences from their respective databases. As is standard in the literature, we remove products that occur as reactants in the same reaction, common byproducts, and products with fewer than 4 heavy atoms. We follow Probst et al. (2022) and split reactions with multiple products and exclude reactions with large molecules (> 100 heavy atoms). To control for the size of the proteins, we only consider sequences that are no more than 800-amino acids long. This yields 103,120 enzyme-catalyzed reactions with 20,385 unique chemical reactions, 12,541 enzymes, covering 2743 EC numbers.

We consider several splits of the dataset. In keeping with previous work, our primary test set is constructed using a product split, where no product in the test set is seen in the training set. Additionally, we explore a structure similarity split and an EC split in the appendix. In particular, enzymes are clustered using Foldseek (van Kempen et al., 2023) with a 90% structure overlap and a sequence identity of 0, and enzymes in the same cluster are assigned to the same split. We split the data into train, development, and test datasets with a ratio of $8:1:1$. For the EC split, we held-out reactions in an EC for the test set and split the remaining reactions into ($\sim$89%) train and ($\sim$11%) development. Details on data processing are provided in Appendix A.4.1.

A.1.2 DRUGBANK DATASET

We consider the out-of-distribution chemical domain of drug metabolism and showcase the improved performance of our model as compared to other models on this task. We obtain drug reactions from DrugBank for which a UniProt ID is available. Since our graph-editing procedure requires all reactant molecules present, including co-factors, we obtain from UniProt all reactions annotated for each protein entry and extract the substrates that are common among all reactions of an entry and add them to the corresponding DrugBank data samples. We further focus our analysis on reactions from phase II metabolism and exclude reactions catalyzed by cytochromes. The cytochrome P450 superfamily is known to perform a wide range of chemical transformations and is often non-specific to location or to substrate such as hydroxylation of unactivated C-H bonds, C-C or C-N bond formation, heteroatom oxidation, oxidative C-C bond cleavages, and nitrene transfer Greule et al. (2018). Since these chemical transformations can be stochastic in their location, annotated datasets represent only a small subset of possible products making it hard to evaluate predictions using the same method and so we exclude them. This curated dataset yields 804 reaction-enzyme pairs, with 160 unique proteins and 342 drugs.

A.1.3 BASELINES

We consider the two prior works on biocatalysis prediction as baselines (Kreutter et al., 2021; Probst et al., 2022). We retrain the transformer models on the USPTO and EnzymeMap training sets following the paradigm reported in the publications and detailed in Appendix A.6. Kreutter et al. (2021) uses the enzyme names to encode the protein, therefore we map protein identifiers from EnzymeMap to their annotated name in UniProt. In cases where annotated names are missing we mark the name as "unknown" in order to avoid skipping many samples. Providing the protein name as input to the model suggests that the protein's function has already been studied and its function characterized, thus defining the name of the protein. In cases where the name does not provide any indication of the function, it should not provide useful information to the model (e.g. "unknown"). In these cases the model must rely on the substrates alone. On the other hand, Probst et al. (2022) encodes the first three levels of the EC number of the reaction; similarly, knowing the associated EC number suggests that much of the biochemical reaction is already characterized and provides the model with information that is beyond what we assume to be available at inference time.

A.2 Related Work

Enzyme Modeling Central to correctly predicting the product of a biochemical reaction is learning the function of the enzyme. In fact, depending on the enzyme identity, the same substrates can undergo different chemical transformations (Durairaj et al., 2019). Prior research on biocatalyzed reaction prediction considers two alternative methods for incorporating enzyme information: using enzyme nomenclature (Kreutter et al., 2021) or EC number (Probst et al., 2022). The former method encodes the scientific name of the enzyme using a language model, while the latter relies on expert defined enzyme classes (i.e., their EC numbers). In both cases, enzymes with similar characteristics are likely to exhibit similarity in their encoding. However, both methods only provide limited generalization capability especially for unseen enzymes where categorization information may not be available. Moreover, these methods ignore the rich biological information embedded in protein sequences. In operating on enzyme classes, previous research also disregards the specificity of proteins and treats all enzymes of a particular class as capable of catalyzing the same substrates. In contrast, utilizing sequence information, our method can be applied to unseen enzymes, without relying on functional annotations.

Chemical Reaction Prediction The field of biocatalyzed reaction prediction is still relatively nascent, and prior methods frame the task as a machine translation problem using language models. However, language-based generation does not make use of the fact that the atoms of the reactants are conserved, and small mistakes in generation can lead to widely different molecules. Our work most closely follows graph-based approaches developed for the small molecule, general chemistry, space. These methods leverage this inductive bias and learn the graph edits to apply on the molecular graph encoding of the reactants, and recently demonstrated better generalization than language model based approaches (Jin et al., 2017; Coley et al., 2017; Segler & Waller, 2017; Bradshaw et al., 2018; Bi et al., 2021; Sacha et al., 2021; Chen & Jung, 2022). Our approach builds on the success of these graph based methods and develops it further to include enzyme sequences and exploit the interactions between the sequences and substrates.

Drug Metabolism Prediction In the pharmaceutical industry, drug metabolism screening is typically done through experimental assays. Existing analytical methods largely rely on rule-based approaches (Cruciani et al., 2005; Ridder & Wagener, 2008; Djoumbou-Feunang et al., 2019; Finkelmann et al., 2018; Adams, 2010; Kirchmair et al., 2013; Darvas, 1987; Rudik et al., 2023). For example, Mohammadi Peyhani et al. (2020) used a template-based search to predict that the anticancer drug 5-fluorouracil can be metabolized into competitive inhibitors of native enzymes, which can help explain the observed toxicity of the drug. As an alternative to rule-based approaches, several machine learning methods have emerged. However, these approaches are limited in their reach as they are typically trained on a specific class of enzymes (e.g., cytochrome P450s) (de Bruyn Kops et al., 2020; Hughes & Swamidass, 2017; Olsen et al., 2019; Hennemann et al., 2009). In contrast, our method provides a more general framework that can be applied to any chemical matter while delivering strong performance on a curated dataset from DrugBank (Wishart et al., 2018).

A.3 Predicting Drug Metabolism

Here, we take the metabolism of small molecule drugs as a potential real-world application of our model and showcase the improved performance of our model as compared to others on this task. How a drug is metabolized has important implications for its efficacy, toxicity, and mechanism of action. While some experimental approaches exist to study drug-metabolizing enzymes, there remains a critical need for *in-silico* drug metabolism models to address the cost, time, and human expertise required by *in-vitro* and *in-vivo* methods. We evaluate our model on drug reactions from DrugBank for which a UniProt ID is available and focus on non-cytochrome-catalyzed biotransformations (Wishart et al., 2018). We find that our model is able to predict the correct drug metabolite with a 60.1% top-10 accuracy and outperform other deep learning models (Table 3).

To better understand the errors observed on drug reactions, we manually inspect cases where the model fails to find an exact match to the annotated product within the top ten predictions. In many cases, we find that the model comes close in identifying the reaction type but focuses on incorrect, yet similar, sites of metabolism. For example, the model correctly predicts the reaction in Figure 4(a) to be a hydroxylation but predicts the wrong methyl group to which to add the OH group, though it

is near the true site. We also identify cases where the model predictions are considered wrong as a result of inconsistencies in the databases. Raloxifene (DB00481) is reported to be metabolized by a UDP-glucuronosyltransferase (Q9HAW8) (Figure 4(b)). Since we obtain enzyme co-factors from UniProt, we utilize UDP-α-D-glucuronate as the other substrate in the reaction. Our prediction matches exactly the chemical pattern annotated in UniProt and provided by the Rhea database. However, this appears to be inconsistent with the metabolic reaction of raloxifene in DrugBank and results in our prediction to be considered incorrect. In some cases, the model is not able to fully capture the complexity of the biochemical reaction. The metabolism of morphine (DB00295) consists of the transfer of glucuronic acid and ring breaking (Figure 4(c)). The model is found to be partially correct as it predicts the right glucuronidation site but is unable to identify the bond changes to the ring in any of its top-k predictions.

Table 3: Performance on the DrugBank drug reactions.

Model	Top 1	Top 3	Top 5	Top 10
KREUTTER ET AL. (2021)	28.6%	37.2%	40.8%	43.8%
PROBST ET AL. (2022)	25.7%	33.0%	38.1%	42.8%
OURS	40.7%	56.0%	58.0%	60.1%

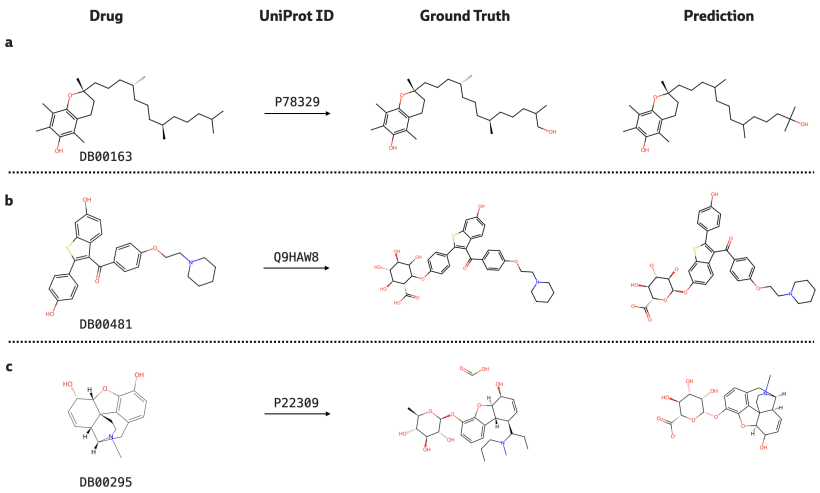

Figure 4: Illustrative examples of errors made by our model, where (a) the predicted reaction type is correct but the reactive site is misclassified; (b) the mistake is possibly due to inconsistencies between databases; and (c) the reaction consists of several changes that the model is unable to fully recover.

A.4 IMPLEMENTATION DETAILS

A.4.1 DATA SPLITS

Product Split We follow prior work and split the data such that there is no overlap of products between the three data splits. We use the `rxn4chemistry` tools (`https://github.com/rxn4chemistry/biocatalysis-model`, Probst et al. (2022)) to pre-process our data and exclude reactions with large molecules ($>$ 100 heavy atoms), those with products with fewer than 4 heavy atom, and those with proteins that are more than 800-amino acids long. This yields 95,318 training samples, 5,037 development samples, and 2,765 test samples.

Structure Split We download predicted protein structures from AlphaFold (Jumper et al., 2021) as `.cif` files. We use Foldseek (van Kempen et al., 2023) to cluster our database of structures using `easy-cluster` with `--min-seq-id` $= 0.0$ and `-c` $= 0.9$. We obtained 13,866 clusters which were split into 80% train, 10% development, and 10% test. Samples were placed in a split according

to the enzyme's cluster identity. This yields 79,443 training samples, 11,208 development samples, and 10,781 test samples.

EC Split For each EC, $ec \in \{1, 2, 3, 4, 5, 6\}$, we held out all reactions with that specific ec number, considering only the top level class. The remaining reactions were split according to product-based split into $\frac{8}{9}$ training and $\frac{1}{9}$ development sets (maintaining the ratio of 8:1:1). The number of samples in each split are provided in Table 4.

Table 4: Number of reactions in each data split when using the top-level EC number to construct the test sets.

HELD-OUT EC	TRAINING SPLIT	DEVELOPMENT SPLIT	TEST SPLIT
1	64,065	8,159	30,896
2	65,652	8,247	29,221
3	68,462	7,335	27,323
4	82,669	10,809	9,642
5	88,352	10,751	4,017
6	89,907	11,192	2,021

A.4.2 MODEL TRAINING

Reaction center prediction We use pre-trained ESM-2 with 35M parameters (`esm2_t12_35M_UR50D`) to encode the enzyme sequences. We use Graph Attention Networks (Brody et al., 2022) for f_{local} with 3 layers, 16 attention heads, and a hidden dimension of 480. We construct a complete graph of the reactants to compute the pairwise attentions across all atom pairs in the f_{global} model. The multi-head cross-attention between the protein residues and reactant atoms is implemented with 4 attention heads. Individual atom representations from each are concatenated and passed through a linear projection layer (3×480 to 480) before predicting pair-wise bond changes.

Candidate product ranking We apply chemprop (Yang et al., 2019), a directed message passing network, on the condensed graph reaction representation of the reactants and candidate products with 5 layers and a hidden dimension of 480. We obtain the mean protein embeddings from ESM-2 (35M parameters) and concatenate them with the graph-level feature representations of the reaction. The final ranking is done with a 2-layer feed-forward network with layer norm (Ba et al., 2016).

Training parameters We use a batch size of 16, learning rate of $1e^{-4}$, learning rate decay of 0.1, and the Adam optimizer (Kingma & Ba, 2014). Training is done with half precision training with bfloat16 (Kalamkar et al., 2019), and we train the reaction center for 20 epochs and the ranker for 5 epochs.

A.4.3 ATTENTION ANALYSIS

Our multi-head cross-attention in the reaction center prediction model is performed between the full protein sequence embedding and the latent atom representations of the substrates, yielding attention scores for every residue-atom pair, $\boldsymbol{A} \in \mathbb{R}^{|\mathcal{V}|\times|P|}$, with $\boldsymbol{A}_{ij} \in [0, 1]$. For every residue, we sum the attentions scores across all reactant atoms and obtain a weight per residue r_i: $\boldsymbol{a}_i = \sum_v \boldsymbol{A}_{vi}$.

Where available, we collect a set of indices, $\mathcal{R}_{AS}$ for each sample in the test set corresponding to the location of annotated active sites from the Mechanism and Catalytic Site Atlas. For each protein p, We take the residues in the top q-th quantile of attention scores and compute the fraction of annotated active site residues included in that predicted set:

$$\hat{\mathcal{R}} = \{i | r_i > k, k = \text{j}^{\text{th}} \text{ quantile of } \boldsymbol{a}\} \tag{1}$$

$$s_p^{(k)} = \frac{|\mathcal{R}_{AS} \cap \hat{\mathcal{R}}|}{|\mathcal{R}_{AS}|} \tag{2}$$

We plot $s_p^{(k)}$ for all test sample proteins with annotations at 10 equally spaced quantile levels. As a control, we compare these scores with those obtained by randomly selecting an equivalent number of indices spanning the length of the protein.

A.5 Performance Along Additional Splits

A.5.1 Performance on Structure Splits

We assign proteins to the training and testing splits based on their Foldseek (van Kempen et al., 2023) cluster identity. We observe that all models achieve similar performance ranging from 60% top-1 accuracy to 80% top-10 (Table 5). While the proteins in the test are expected to assume different 3D folded structures, they may still share catalytic activities with proteins seen during training Omelchenko et al. (2010). For instance, convergent evolution can result in significantly different proteins that catalyze the same reaction. As a result, this can result in data splits where the encoding used in our model does not provide an advantage.

Table 5: Top-k accuracy of our graph-based method compared to existing approaches for biocatalyzed forward synthesis. Published methods are trained as detailed in their respective GitHub codebases (Appendix A.6). Performance is evaluated on EnzymeMap using data splits based on protein structure similarity using FoldSeek (van Kempen et al., 2023).

Model	FoldSeek 90% Split			
	Top 1	Top 3	Top 5	Top 10
Kreutter et al. (2021)	64.5%	77.5%	79.8%	81.2%
Probst et al. (2022)	60.2%	75.0%	77.9%	80.3%
Ours	60.4%	71.7%	75.9%	78%

A.5.2 Performance on EC Splits

The EC system defines seven large classes of biochemical transformations. To measure the generalization across enzyme families and types of chemical transformations, we we trained six models separately, holding out each time all reactions with a specific EC number (only six classes are contained in the EnzymeMap dataset). This constituted the hardest setting among the three splits. Since Probst et al. (2022) utilizes the EC number as an input, we omitted it from this experiment since it would never see the test-set EC during training. We observe that our method is comparable to Kreutter et al. (2021) on ECs 2,3, and 5, better on ECs 1 and 4, and significantly worse on EC 6 (Table 6) . Across ECs, however, both models achieve poor performance in terms of absolute accurate generalization, demonstrating the challenge of truly learning the chemistry underlying enzymatic catalysis.

A.6 Training of transformer-based models

We train existing deep learning model for biocatalysis Kreutter et al. (2021) and Probst et al. (2022) according to the codebases associated with their respective publications: `https://github.com/rxn4chemistry/OpenNMT-py/tree/carbohydrate_transformer` and `https://github.com/rxn4chemistry/biocatalysis-model`. Specifically, we use the same tokenization scheme for the enzyme names with either byte pair encoding or the EC numbers. We pre-process (`onmt_preprocess`) the data with the default parameters of sequence source and target lengths of 3000 and a shared vocabulary, and we train the models (`onmt_train`) simultaneously on the USPTO dataset (Lowe, 2012; 1976) and the same splits of EnzymeMap that we use for our model. We use the default hyper-parameters for training (Table 7).

Table 7: Hyper-parameters used for training the transformer-based models. All are the default provided in the models' respective codebases.

Hyper-parameter	Value

data_weights	(9,1) (for USPTO and EnzymeMap, respecitvely)
seed	42
gpu_ranks	0
world_size	1
train_steps	250,000
param_init	0
param_init_glorot	true
max_generator_batches	32
batch_size	32768
batch_type	tokens
normalization	tokens
max_grad_norm	0
accum_count	1
optim	adam
adam_beta1	0.9
adam_beta2	0.998
decay_method	noam
warmup_steps	8,000
learning_rate	2
label_smoothing	0.1
layers	6
rnn_size	512
word_vec_size	512
encoder_type	transformer
decoder_type	transformer
dropout	0.1
position_encoding	true
share_embeddings	true
global_attention	general
global_attention_function	softmax
self_attn_type	scaled-dot
heads	8
transformer_ff	2048

Table 6: Top-k accuracy of our graph-based method compared to existing approaches for biocatalyzed forward synthesis on different EC-based splits. Each model is trained on all other ECs and tested on the held-out EC.

MODEL	HELD OUT EC	TOP 1	TOP 3	TOP 5	TOP 10
KREUTTER ET AL. (2021)	EC 1 (n=30,896)	8.6%	17.3%	21.9%	26.3%
OURS		9.4%	22.2%	27.2 %	34.0 %
KREUTTER ET AL. (2021)	EC 2 (n=29,221)	9.1%	16.6%	20.9%	26.2%
OURS		8.0%	15.9%	20.7%	25.8%
KREUTTER ET AL. (2021)	EC 3 (n=27,323)	26.8%	47.7%	55.2%	61.7%
OURS		32.1%	45.9%	52.9%	60.7%
KREUTTER ET AL. (2021)	EC 4 (n=9,642)	13.9%	19.8%	23.0%	28.6%
OURS		7.4%	20.1%	29.0%	35.1%
KREUTTER ET AL. (2021)	EC 5 (n=4,017)	2.4%	6.4%	7.1%	11.9%
OURS		1.7%	9.0&	10.9%	12.0%
KREUTTER ET AL. (2021)	EC 6 (n=2,021)	20.6%	41.5%	47.0%	48.3%
OURS		4.1%	15.9%	21.4%	26.1%
KREUTTER ET AL. (2021)	MEAN	13.5%	24.9%	29.2%	33.8%
OURS		10.5%	24.0%	27.0%	32.3%