

Mechanism-Informed Rules Tunably Balance Novelty and Feasibility of Predicted Enzymatic Reactions

Stefan C. Pate, Keith E. J. Tyo, and Linda J. Broadbelt*



Cite This: <https://doi.org/10.1021/acs.jcim.6c01595>



Read Online

ACCESS |



Metrics & More

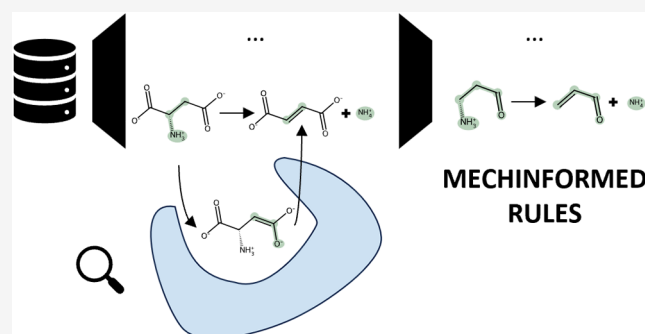


Article Recommendations



Supporting Information

ABSTRACT: Enzymes catalyze reactions with remarkable specificity and can unlock recalcitrant feedstocks that are dilute, complex, and variable in their constituent molecules. While characterized enzymatic reactions cover a wide range of chemistries, there are an undetermined number of cryptic activities for every known one. These cryptic activities can be elicited through rational design, adaptive laboratory evolution, and increasingly, generative models of proteins. However, prior to tuning a catalyst, one must efficiently predict viable novel reactions. In this work, we leverage the growing amount of mechanistic enzyme information, specifically the Mechanism and Catalytic Site Atlas, to construct a set of reaction rules that can meet this demand. By explicitly utilizing mechanistic information, the rule sets developed here more accurately identify molecular structures required for catalysis compared to existing curated and heuristically constructed rules. The 899 Distilled rules are constructed directly from characterized mechanisms and recapitulate 62.5% of atom-mapped reactions from Rhea. The Learned rule set is generated from a classifier trained on structural patterns putatively required for catalytic mechanisms. The Learned rules recapitulate all atom-mapped Rhea reactions and precisely predict mechanism-required atoms (ROC-AUC = 0.98). Additionally, our Learned rules exhibit a more favorable trade-off between novelty and feasibility and provide users with fine-grained control over this trade-off. The rules are compatible with all SMARTS-based reaction network expansion and retrosynthesis software.



INTRODUCTION

Enzymes serve an important role in the synthesis of products ranging from commodity chemicals to pharmaceuticals. They exhibit remarkable regio- and stereo- selectivity, function well at mild temperatures and pressures, and, in the cellular context, can accommodate dilute and complex feedstocks. These characteristics have driven their successful utilization in research and industry.¹ Tens of thousands of reactions and their protein catalysts have been characterized to date,^{2,3} and yet, this may just scratch the surface.

Many undiscovered functions of enzymes are cryptic. These functions, usually referred to as promiscuous, underground, or non-native, tend to be overshadowed by kinetically superior primary activities.^{4,5} However, the slow kinetics of non-native activities are not inevitable; rather they appear to be frozen accidents of evolutionary history. Theories of enzyme evolution suggest that selective pressures drove generalist enzymes toward specificity in many cases, but that enzymes remain labile, which confers evolvability.^{5–8} Leveraging this lability, biotechnologists have modified enzymes to significantly enhance non-native activities, thus engineering systems to synthesize compounds of interest.^{9–11}

The underlying enzymatic reaction network inclusive of non-native activities may be very large, making it impractical to store in its entirety. Instead, the edges of the network can be

represented implicitly by a set of reaction rules, which include only the most important structural molecular features of reactions. In computer-aided synthesis planning (CASP), reaction rules, along with either a target molecule, starting feedstock, or both, are input into a procedure that samples from the underlying reaction network and enumerates linear synthesis paths, trees, or more involved subnetworks.^{12–26} Template-free methods forego reaction rules in favor of a generative model of reactions,^{27–31} but we focus on template-based methods in this study.

The accuracy of reaction rules (or generative model in the case of template-free methods) is critical as incorrect predictions can result in expensive and time-consuming efforts to engineer an enzyme for an activity that was incorrectly predicted, or a desired activity may be abandoned because reaction rules did not correctly predict a viable transformation. An ideal set of reaction rules would, for all possible

Received: May 18, 2026

Revised: June 4, 2026

Accepted: June 8, 2026

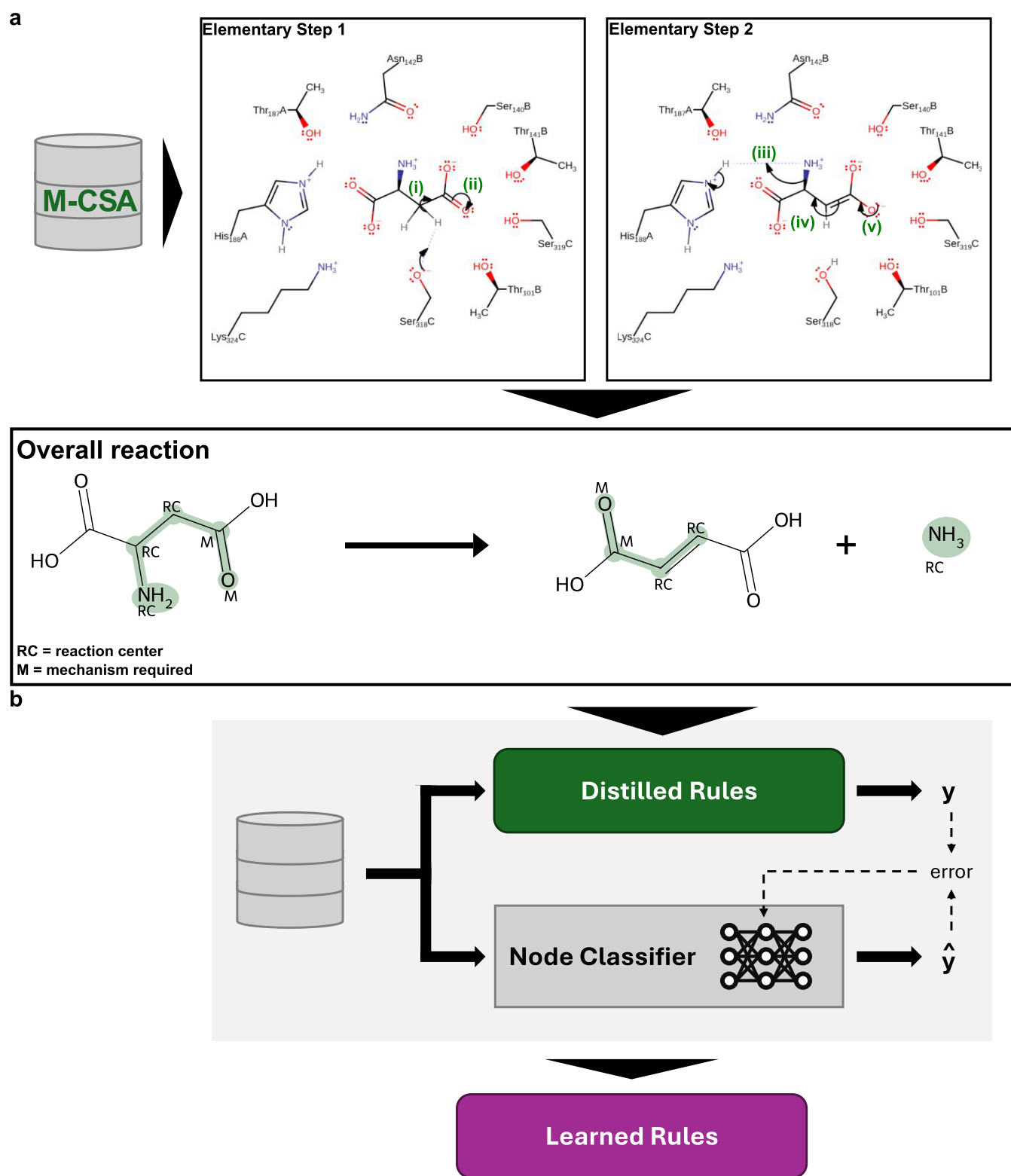


ACS Publications

© XXXX The Authors. Published by
American Chemical Society

A

<https://doi.org/10.1021/acs.jcim.6c01595>
J. Chem. Inf. Model. XXXX, XXX, XXX–XXX



transformations, capture only the reactant structures necessary for catalysis. Aside from the obvious atoms that have bond changes between reactant(s) and product(s), there are often atoms that temporarily donate or hold an unstable charge in an intermediate step but return to the initial bond configuration in the final product. For example, in the aspartate ammonia lyase reaction, a carbonyl group accommodates a negative charge while a double bond shifts and the amine group leaves, ultimately returning to its initial bond conformation (Figure 1a). One can then make a very broad set of viable predictions. A minimal description of a reaction is merely the reaction center (RC).³² However, in the case of enzymatic reactions, RC-only rules miss adjacent structures which are required for the catalysis and make unreasonable predictions as a consequence. Due to the recursive nature of a network search, this issue compounds catastrophically. To address this issue, previous work has added additional molecular structure surrounding the RC within a topological radius, (RC + R).^{33,34} For example, RC + 1 templates would include all atoms in the RC and bonded to it. In the aspartate ammonia lyase example, the template would include the carbonyl carbons in addition to the RC atoms (Figure 1a). RC + 2 also includes atoms that are two bonds from the RC, e.g., the entire carboxylic acid groups in aspartate (Figure 1a). Other approaches use manually curated reaction rules rather than attempt to extract rules from reaction data.^{18,35} Still other data-driven approaches break with the radial mold in various ways. RDChiral includes in its templates any moieties from a curated list appearing adjacent to the RC.^{36–38} EVODEX differentiates between σ and π bonds as it expands the template about the RC.³⁹ The algorithms in EHreact and Ni, 2022 cluster reactions based on common subgraphs, which need not grow radially with respect to the RC.^{40,41}

Here, we develop a new approach to automatically extract enzymatic reaction rules that directly uses mechanistic data rather than rely on various heuristics to approximate it. Our method was made possible by the high-quality Mechanism and Catalytic Site Atlas (M-CSA), which compiles enzyme reaction mechanisms for a thousand overall reactions.⁴² In contrast to other methods which predict elementary steps given an overall enzymatic reaction,^{43–45} we focus on constructing or inferring overall reaction rules based on information from the elementary steps. Our goal is the same as all previous sets of reaction rules, i.e., to capture only reactant structures necessary for catalysis in order to predict viable novel reactions in an efficient manner. Accordingly, we focus on distilling and learning simple, minimal reaction rules from mechanistic data. We show that our approach more accurately identifies or predicts mechanism-required structures and exhibits a more favorable trade-off between novelty and feasibility of predicted reactions.

RESULTS

Curating Mechanistic Data for Enzymatic Reaction Rule Construction and Inference

At the time we accessed it, the M-CSA⁴² database contained 1,166 mechanisms for 1,003 overall reactions. 887 of mechanism entries featured detailed arrow-pushing diagrams represented in Chemical Markup Language (CML). The rest were described in natural language. We translated the elementary steps of these detailed mechanisms from CML to SMILES format, ensuring all reactions were balanced, and

consistently atom-mapped throughout all steps. We were able to recover 673 of the 887 CML-encoded mechanisms into balanced, atom-mapped SMILES. In addition to atom-mapped SMILES, we store indices of all atoms involved in the mechanism of any elementary reaction step. More precisely, we stored the union of all elementary reaction centers. This is a superset of the overall reaction center and comprises a template of a valid reaction rule (Figure 1a). We then reversed all 673 balanced, atom-mapped SMILES leaving 1,346 reactions with all reactant, product, and active site residue structures specified. Finally, we extract just the structures involved in the mechanism to form rule templates and eliminate duplicate rule templates. We refer to the resulting collection of 899 unique rules generated through this procedure as the Distilled rules.

With the Distilled rules in hand, we next tested their coverage of known biochemistry as described in the Rhea database.² Note that we use the terms “cover” and “recapitulate” to indicate that for a certain empirically characterized reaction, a rule successfully generates the correct products given the reactants. The Distilled rules recapitulate 62.5% of atom-mapped enzymatic reactions from the Rhea database² (Table 1). The vast majority of these recapitulated

Table 1. Summary Statistics for Evaluated Rule Sets

Rule set ^{1,2}	Number of templates	% of atom-mapped Rhea reactions recapitulated ³	Median template size	Mean template size
<i>Learned</i> >0.956	3,249	100%	4	5.5
<i>Learned</i> >0.224	5,579	100%	9	9.3
<i>Learned</i> >0.112	7,488	100%	11	12.5
<i>Learned</i> >0.039	9,924	100%	15	17.3
<i>Learned</i> >0.019	12,388	100%	24	26.1
RC + 0	1,032	100%	6	6.7
RC + 1	6,781	100%	8	8.9
RC + 2	9,774	100%	12	13.4
RC + 3	11,664	100%	16	17.2
RC + 4	12,902	100%	21	20.8
Distilled	899	62.5%	7	7.6
RDChiral	4,919	96.6%	10	11.2
Pathway-level clustering	3,597	84.0%	3	4.5
EHreact	10,901	65.9%	45	48.3

¹For the Learned rule sets, the “> X” refers to the threshold used to include atoms with a score > X in the rule. X can take on any value between 0 and 1; the values selected for presentation in Table 1 were chosen such that each Learned rule set had an average template size comparable to one of the RC + R rule sets. ²For RC + R rule sets, the “+ R” value is used to define atoms that are R hops away from the reaction center (RC) atoms. ³This is not a measure of unseen reaction prediction. It is a fraction of known enzymatic reactions covered by each rule set.

reactions (95%) are *not* found in the M-CSA database, making the Distilled rules useful for reaction prediction and retrosynthesis tasks on their own. To cover the remaining 37.5% of reactions not mapped by the Distilled rules, we leveraged the mapping of Distilled rules to the overall enzymatic reactions they correctly recapitulated to construct a labeled data set. We then used this rules-to-reactions mapping to train a parametrized model that could be used to create reaction rules for enzymatic activities not in the M-CSA database (Figure 1b). The model approximates the mapping of the Distilled rules but

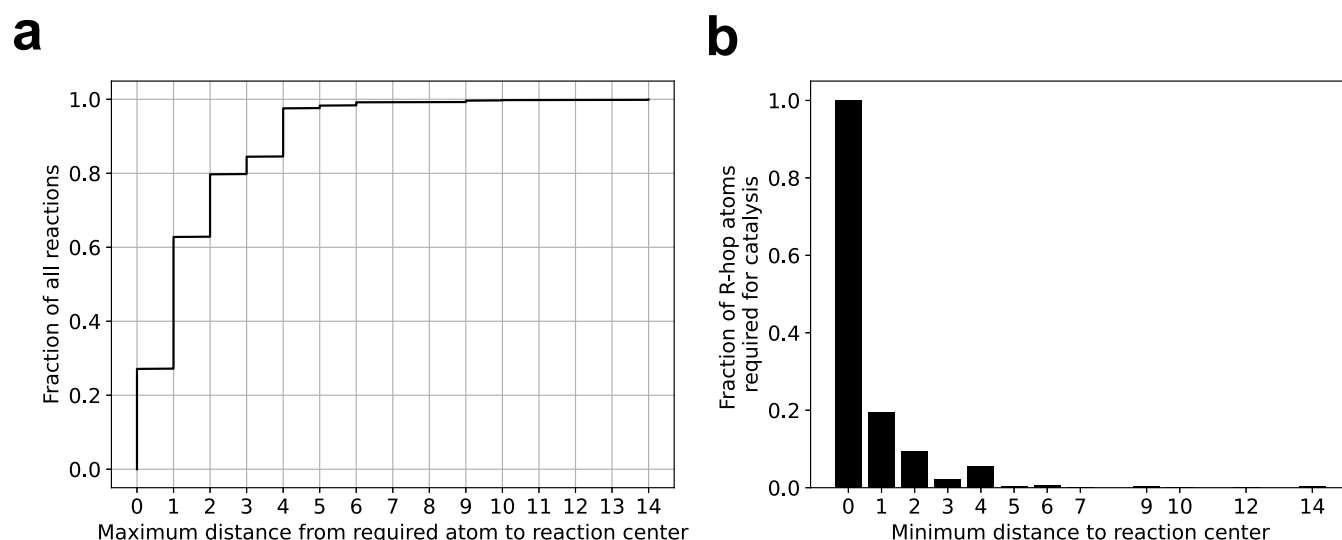


Figure 2. Mechanism-required atoms are sparsely distributed across multiple topological distances from the RC. (a) A CDF of fraction of overall reactions in the M-CSA database against the maximum distance between overall reaction center atoms and mechanism-required atoms. (b) A plot of the fraction of atoms required for catalysis out of all atoms of a given minimum topological distance (# of bonds) from the overall reaction center, R . Fractions at $R > 6$ are all less than 0.003.

is not limited to the 899 templates of this rule set. It accepts any overall enzymatic reaction, and outputs what can be formatted into a reaction rule, achieving full coverage of all atom-mapped overall reactions. We will build up to this rule set, referred to as the Learned rules, in subsequent sections.

Mechanism-Required Atoms are Sparsely Distributed Over Multiple Distances from the Reaction Center

An analysis of enzyme reaction mechanisms reveals the limitations of using an RC + R approach to constructing reaction rules and instead motivates an approach that utilizes mechanistic information. While it is true that mechanism-required atoms tend to be within a few bonds of the reaction center, with nearly 90% located within a single bond (Figure S5a), it is not sufficient for a rule extraction method to include merely the majority of required atoms, aggregately across many reactions. An ideal method would include all atoms required for the mechanism per each reaction while minimizing the amount of unnecessary spectator atoms included in the rule template. This goal is made challenging by the fact that the distance from the reaction center of the most distal of the mechanism-required atoms varies reaction to reaction. The maximum distance varies between zero and four for 98% of all reactions with described mechanisms (Figure 2a). While 90% of mechanism-required atoms are located within a single bond of the overall reaction center, RC + 1 rules would only fully cover the required structural elements for 63% of all reactions (Figure 2a). At the same time, 81% of atoms in RC + 1 are unnecessary for the mechanism (Figure 2b). The trade-off between covering all required atoms and rejecting spectator atoms, i.e., those that are not involved in the mechanism, is harsh for the RC + R strategy. With an RC + 4 rule set, all mechanism-required atoms for 98% of characterized reactions are included, but at the cost of including 3.9 spurious atoms for every required one (Figure S5b). These overly prescriptive reaction rules do not generalize much beyond known enzymatic reactions. However, relative to the Distilled rules derived from M-CSA, which only covered 62.5% of known reactions, RC + 4 rules have the advantage of very high coverage of known reactions. We sought a data-driven solution

to this issue that would both have high coverage of known reactions while minimizing the spurious atoms to allow a higher degree of generalization to suggest novel reactions.

Learning a Parametrized Approximation of Mechanistically Relevant Structural Patterns

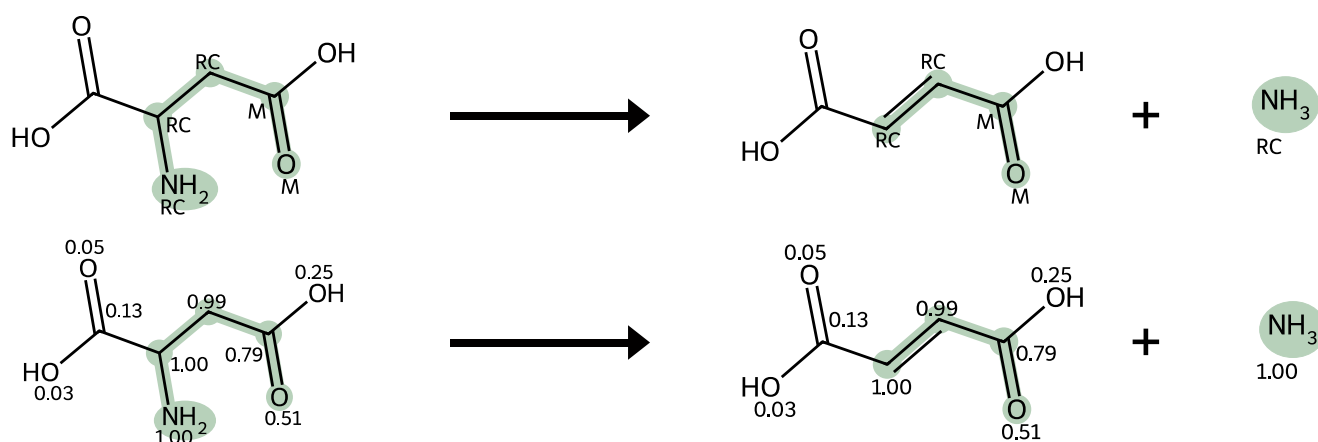
Ultimately, we desired a set of enzymatic reaction rules informed by mechanistic information which generalize to as many known overall enzymatic reactions as possible, and with an added requirement of granting user control over level of rule specificity. We achieved this by first leveraging the Distilled rules to create a labeled data set and then formulating reaction rule writing as a binary classification task that we solved with machine learning.

The number of enzyme reaction mechanisms retrieved from M-CSA was relatively small. Nevertheless, through mapping the derived Distilled rules to Rhea, we assembled a data set of 12,137 reactions with atoms labeled as required for the mechanism or not. When a reaction rule is mapped to a full reaction, in addition to recapitulating the reaction (transforming the reactants into the observed products), one also obtains a correspondence between the rule's template and the reactants' molecular structures. The resulting data labels are binary, and there is one label for each atom in the full overall reaction. For an example, see Figure 3 in which atoms of the positive class—involved in the mechanism—are highlighted green.

Given the structure of this data set, we constructed a model that accepts a featurized reaction graph as an input, and outputs as many binary labels as there are atoms in the reaction (Figure 3).

The classification task associated with this form of data set can be viewed in multiple lights. One can think of a reaction as a data point, making the task a form of multilabel binary classification in which the number of labels, equal to the number of atoms, is variable. Alternatively, one can think of an atom as a data point, and data points are batched together depending on molecular structure. A third interpretation of the task as node classification emphasizes the graphical nature of the data. Motivated by this, we employed GNNs to learn

a



b

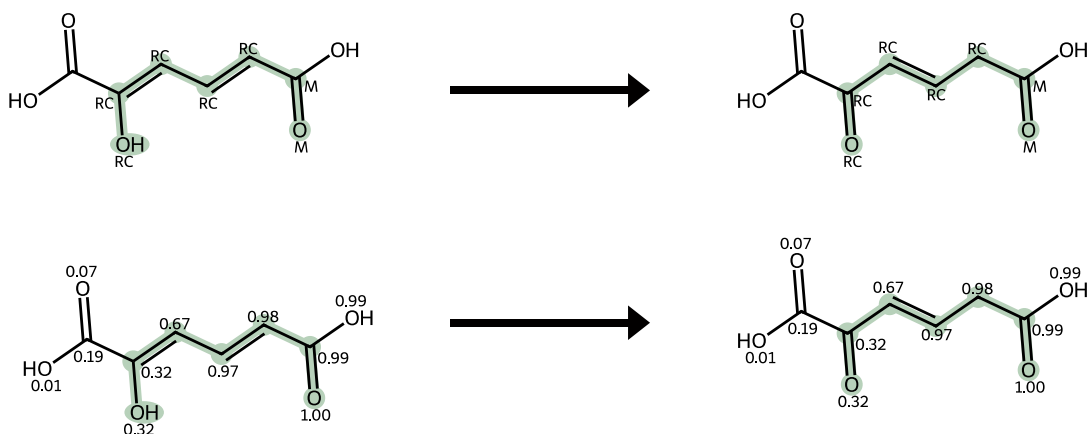


Figure 3. Model-inferred labels closely adhere to Distilled templates in difficult examples. (a) An example of a highly successful labeling of a difficult overall reaction from M-CSA. Top, aspartate ammonia lyase reaction (EC 4.3.1.1) highlighting atoms involved in any elementary step of the catalysis. Bottom, the same reaction visualizing the scores output by a trained model. This model was not trained on any reaction mapped by the Distilled template displayed in green. Note that the scores are the same for a given atom in both the reactants and products, as the score is specific to the atom. (b) An example of a reasonably successful labeling of a difficult overall reaction from M-CSA. Same format as (a) but with 2-hydroxymuconate tautomerase reaction (EC 5.3.2.6). Abbreviations: RC = reaction center, M = mechanism-required.

vector representations of the atom nodes, dependent on their structural surroundings, and fed these vectors into a FFN to output scalar scores that were later made discrete labels by application of a threshold.

We evaluated the model's generalization to mechanisms not present in the M-CSA data set through the use of group 5-fold cross validation in which reactions were first grouped by the mapped Distilled rule prior to being split into the test data set. We employed group 3-fold cross validation nested within the 5 outer splits for hyperparameter optimization. See Methods for details.

We were most interested in decision-threshold-invariant metrics as we designed to grant users control over the value of the decision threshold to tune specificity to their needs. Considering first the interpretation of atoms as data points, the trained node classifier achieves a more favorable trade-off between precision and recall than the RC + R strategy where *R* is the specificity control parameter for this scheme (Figure 4a).

The area under the precision-recall curve (PRC-AUC) for the former was 0.95 while that of the latter was 0.87. On this task, RDChiral^{36,37} performed similarly to RC + 1. The interpretation of reactions as data points suggests a slightly different evaluation in which successful recall of a data point means all mechanism-required atoms for a given reaction are labeled with the positive class. The definition of the precision metric is unchanged. Here again, the trained node classifier outperforms heuristic algorithms with PRC-AUC of 0.85 vs 0.59 for the RC + R method (Figure 4b). Similar results were observed when the node classifier was trained and evaluated exclusively on reactions found in the M-CSA data set. In this case, the node classifier achieved an atom-wise PRC-AUC of 0.93 vs RC + R with 0.89 and a reaction-wise PRC-AUC of 0.76 vs RC + R with 0.64 (Figure S9).

Two examples illustrate how the trained node classifier outperforms RC + R on the binary classification of mechanism-required atoms (Figure 3). Both reactions shown in Figure 3

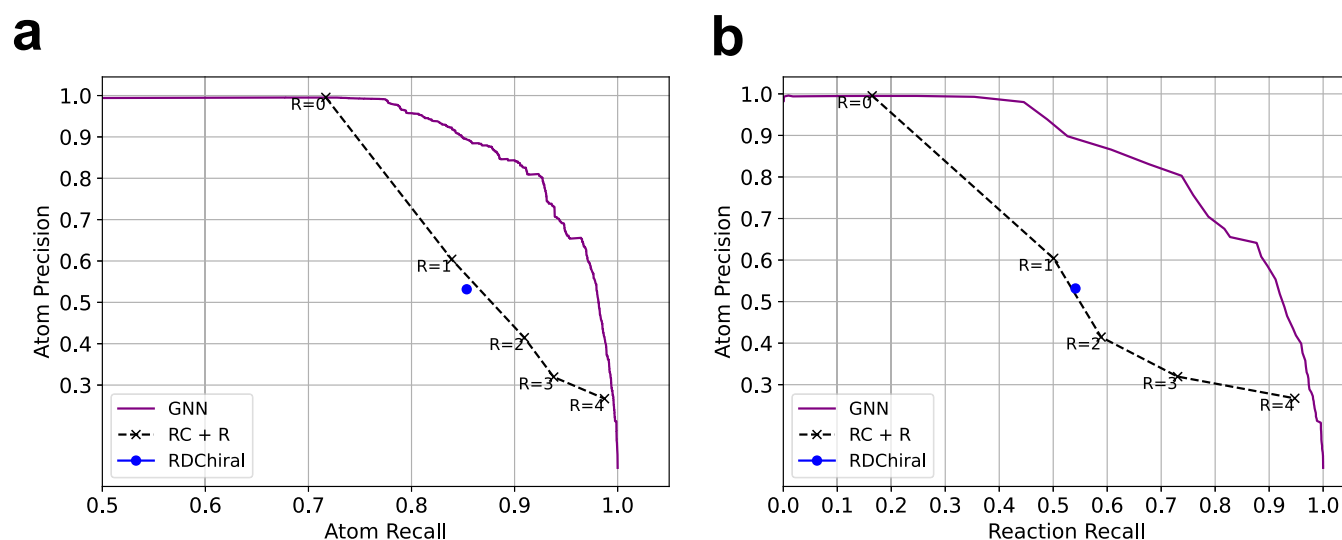


Figure 4. GNN model successfully labels atoms putatively required for the catalysis of enzymatic reactions. (a) Precision-recall curves for the task of labeling atoms as required for the catalysis of an overall reaction. (b) Precision-recall curves for the task of labeling atoms required for the catalysis versus recall evaluated for full overall reactions. For the latter metric, a correctly recalled data point is defined as an overall reaction with all atoms required for the catalysis of that reaction correctly labeled as with the positive class.

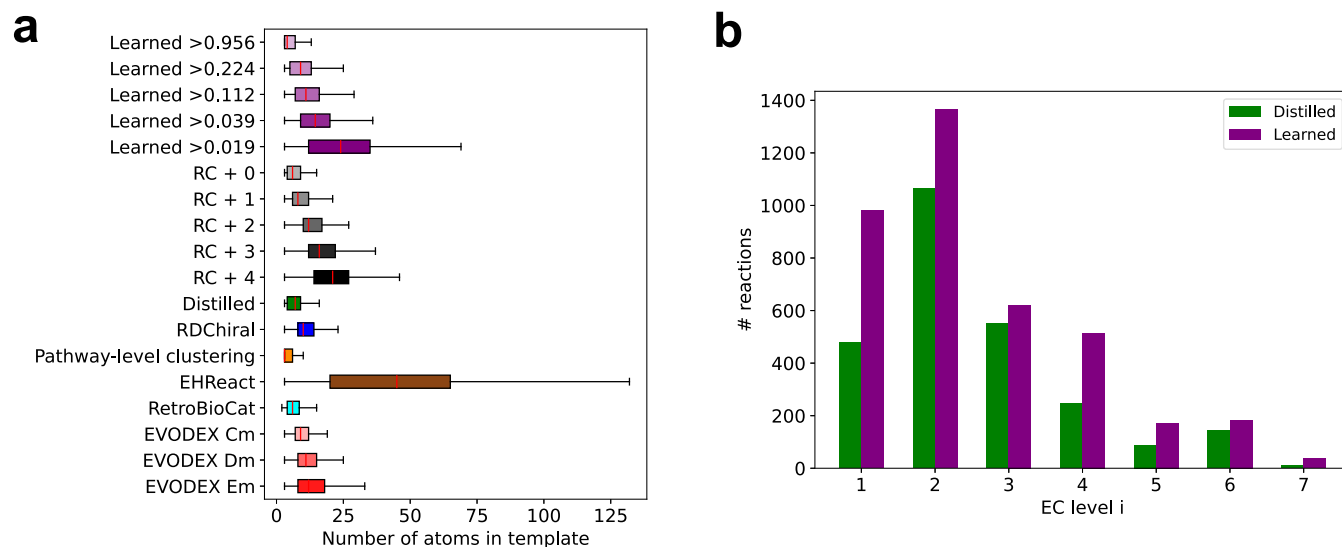


Figure 5. Summary of the specificity and coverage of the Distilled and Learned rule sets. (a) Box and whisker plots depicting the distribution of template sizes, defined as the number of atoms included, from each set of reaction rules that were evaluated in the test reaction network expansions summarized in Figure 6. Median template sizes are represented by red lines and are the same values presented in Table 1. Other ticks represent (from left to right) minimum, first quartile, third quartile, and maximum. (b) Comparison of the coverage of the Enzyme Commission (EC) number classes (i,j,k,l) by the Distilled and Learned rule sets. The y-axis counts the number of unique reactions mapped which are assigned to the EC i digit plotted on the x-axis. Note that all Learned rule sets cover the same known reactions.

are catalyzed via mechanisms that involve atoms outside of the overall reaction center, so reaction center only rules (RC + 0) miss this critical information. Even increasing specificity to RC + 1 leaves out required carbonyl oxygen atoms in both examples and at the same time includes unnecessary backbone carbons in the template. Increasing specificity further to RC + 2 is even more problematic as now all atoms in the reactants are required by the resulting template, thus precluding the prediction of any novel reactions which are not a superset of the observed structure.

In contrast, it is possible to generate more tailored reaction rule templates by tuning the decision threshold of the trained node classifier to assign the atoms to classes. In the best cases, scores given to mechanism-required atoms are much higher

than those given to spectator atoms (Figure 3a). Even in cases where the model makes an error by giving a high score to an unnecessary atom, it fails more gracefully than the RC + R method (Figure 3b).

Generating Learned Rules with a Trained Node Classifier

The successful training and evaluation of the node classifier on Distilled-labeled data lay the groundwork for the Learned rule sets. The procedure consists of labeling atoms from overall enzymatic reactions, extracting structural patterns corresponding to these labeled atoms in SMARTS, and enumerating all unique rule templates. As a demonstration of this procedure, we generated variants of the rule sets for a range of decision thresholds, which is inversely proportional to the resultant

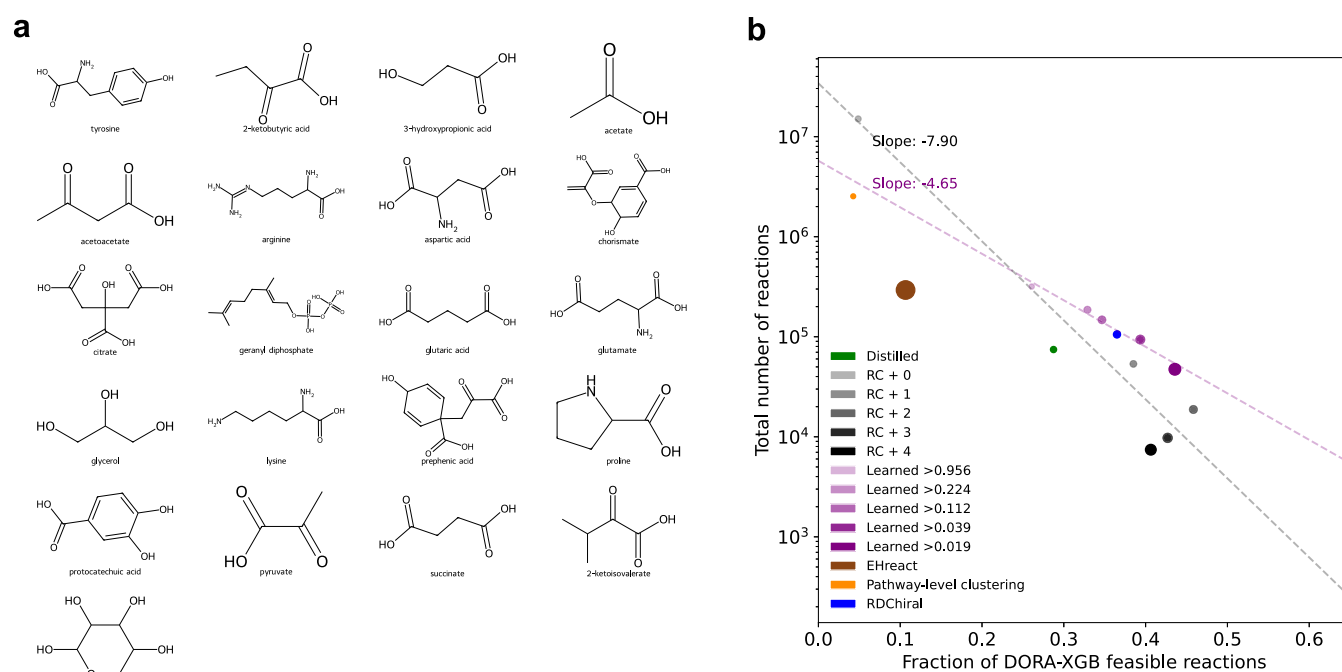


Figure 6. Learned rules exhibit more favorable novelty-feasibility trade-off than RC + R. (a) The molecular structures of the 21 starting molecules which seeded the test network expansion.¹ (b) Total number of unique predicted reactions plotted against the fraction of total reactions labeled as feasible by the DORA-XGB classifier.⁴⁶ Each point corresponds to one reaction network expansion and corresponding reaction rule set used for the expansion. Point size scales linearly with the median template size of the reaction rule set. Dashed lines represent lines of best fit for the RC + R points and Learned points.

rules' specificity. The number of unique templates in the rule set and the average template size increase concomitantly with specificity (Table 1). Beyond spanning a range, we selected decision thresholds to approximately match template size to the five variants of RC + R rules evaluated in this work (Figure 5a). We provide evidence of the superior coverage of known enzymatic reactions by the Learned rules, relative to Distilled, in the form of an aggregate count of reactions recapitulated, which equals that of the reaction center only rules (RC + 0) (Table 1), as well as broken down by Enzyme Commission (EC) number as a proxy for the diversity of chemical transformations covered. We provide this breakdown up to the first EC digit (Figure 5b) and second EC digit (Figure S4) for a more granular view.

Comparison of Mechanism-Informed and Established Rules in Reaction Prediction

Prediction of novel enzymatic reactions is one of the main applications of enzymatic reaction rules. Often this is performed iteratively to compose reactions into multistep synthesis pathways. The ideal set of reaction rules would predict many novel reactions that extend the current reach of enzymatic synthesis while minimizing the number of infeasible reactions predicted which strain processing and memory resources without delivering results that can be realized experimentally. These costs compound with each subsequent reaction network generation step which multiplies the size of the network. To evaluate rule sets' ability to efficiently predict feasible novel reactions, we ran two test network expansions.

In the first network expansion, we seeded the network with a set of starting molecules designated by the Agile BioFoundry as "beachhead molecules" in a graphic adapted from S.Y., et al, 2019 (Figure 6a). We then applied reaction rules for two generations to the starting molecules and the progeny in the

first generation to create a reaction network. For each rule set and corresponding generated reaction network, we counted the number of novel predicted reactions and the fraction of those predictions labeled as feasible by an existing discriminative model, DORA-XGB.⁴⁶ We did not conduct this analysis on RetroBioCat or EVODEX rules since they generate unbalanced reactions and DORA-XGB was trained exclusively on balanced reactions.

An exponential function fit well the novelty-feasibility trade-off for both rule sets with a tunable parameter controlling specificity, Learned with $R^2 = 0.98$ and RC + R with $R^2 = 0.94$ (Figure 6b). The decay rate for the Learned rule sets, -4.65, quantifies a more favorable trade-off relative to the RC + R rules whose decay rate was -7.90. The Distilled rule set achieves feasibility between the most and second most permissive Learned rules but with significantly fewer reactions predicted (Figure 6b). This may be attributed to either the relatively low coverage of enzymatic transformations of the Distilled rule set, particular biases of the discriminative feasibility model, or both. The RDChiral rule set resulted in a reaction network very close to the Learned novelty-feasibility trade-off line, indicating that the manually curated set of important functional groups with which it augments the reaction center is well aligned with mechanistically relevant structural patterns (Figure 6b). Finally, a striking difference between the Learned and RC + R rule sets was the sensitivity to their respective specificity control parameters. For example, increasing from RC + 0 to RC + 1 dramatically reduces the number of reactions 281-fold and increases the fraction which are feasible 8-fold (Figure 6b). By comparison, tuning the decision threshold of the Learned rules yields much finer-grained control over specificity.

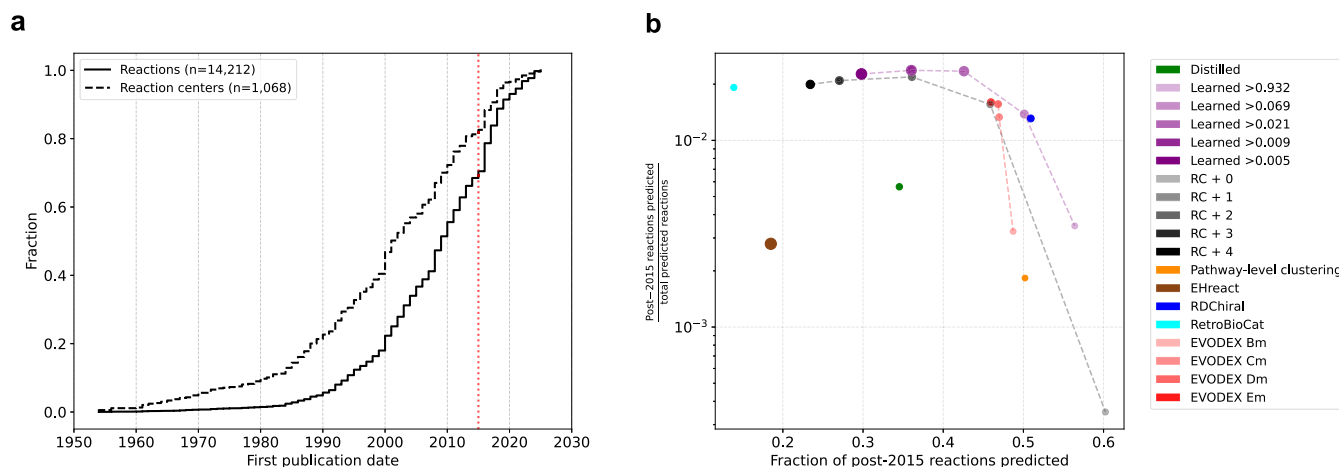


Figure 7. Learned rules effectively balance coverage and efficiency in a time-based split. (a) CDF showing first publication date for unique data set reactions and reaction centers (equivalently RC + 0 rules). The cutoff date chosen for the time split, 2015, is shown in red. (b) Trade-off between fraction of post-2015 characterized reactions predicted by pre-2015 rule sets and the ratio of post-2015 characterized reactions to all reactions predicted by each rule set. Dashed lines connecting the Learned and RC + R rule sets serve as a visual aid of the achievable trade-offs of each set.

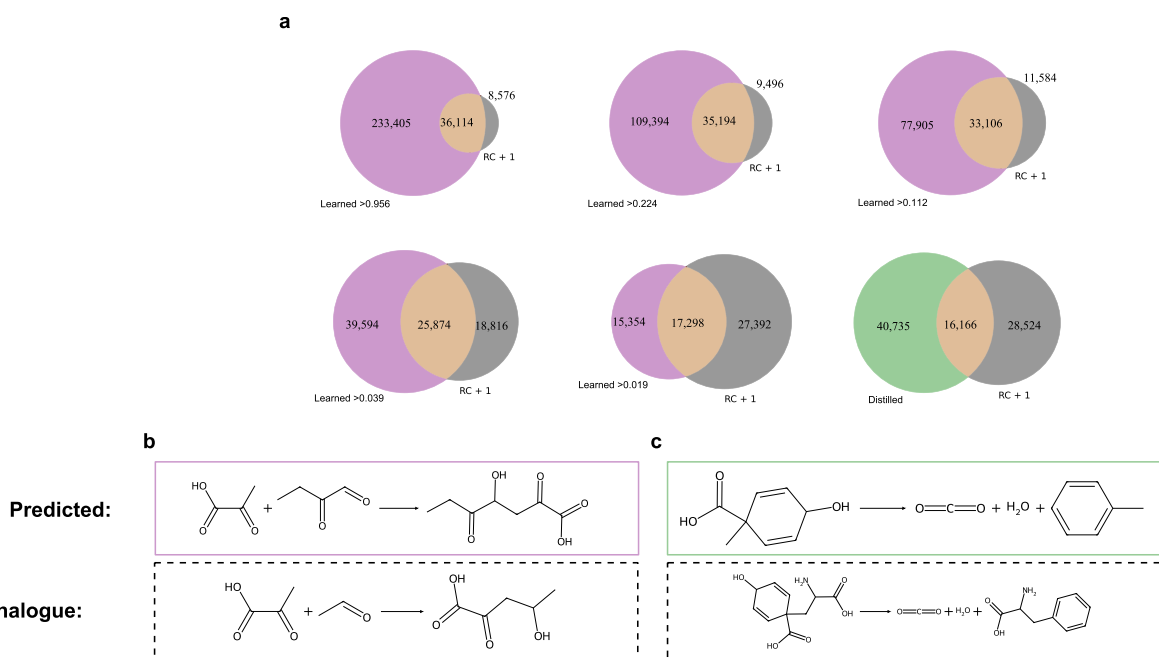


Figure 8. Learned and Distilled rules predict distinct sets of reactions from RC + 1. (a) Venn diagrams contrasting the predicted reactions of the Learned rules or Distilled rules with those predicted by the RC + 1 rules. Purple = Learned, Gray = RC + 1, Tan = Intersection. (b) Top, an example reaction predicted by a rule from the Learned >0.019 set. Bottom, the nearest known analogue, 4-hydroxy-2-oxovalerate aldolase reaction (EC 4.1.3.39), of the predicted reaction. (c) Top, an example reaction predicted by a rule from the Distilled set. Bottom, the nearest known analogue, arogonate dehydratase reaction (EC 4.2.1.91), of the predicted reaction.

DORA-XGB was shown to achieve a high level of performance in predicting feasible reactions;⁴⁶ nevertheless, it may carry biases inherent to its model structure, training data, or optimization procedure. In the second test network expansion, we sought an alternate method to score reaction feasibility that does not rely on a probabilistic model. We achieved this with a time-based split wherein we constructed rule sets based on all reactions first characterized before a cutoff date. Then we performed a one-step expansion, seeded with all compounds involved in reactions characterized after this cutoff date. For each rule set, we quantified the fraction of postcutoff reactions predicted by precutoff rules and the ratio of postcutoff characterized reactions to total reactions

predicted. In total, 4,466 reactions were characterized after a cutoff date of 2015 (Figure 7a). As not all reaction centers (alternatively RC + 0 rules) had been characterized by 2015, even the RC + 0 rules could not predict all post-2015 reactions. These rules predicted 60.2% of post-2015 reactions and did so inefficiently; just 0.035% of all reactions generated were experimentally characterized post-2015 (Figure 7b). The most permissive of the Learned rules (>0.932) predicted 56.4% of post-2015 reactions and were ten times more efficient than the RC + 0 rules (Figure 7b). In general, the Learned rules exhibited a more favorable trade-off between coverage of empirically characterized rules and efficiency than the RC + R rules. The RDChiral rule set performed comparably to the

second most permissive Learned rules (>0.069). Most of the EVOLEX variants performed comparably to the RC + 1 rules. The RetroBioCat and EHreact predicted significantly fewer post-2015 reactions than the other rule sets tested.

The reactions predicted by the RC + 0 rules are a superset of reactions predicted by all the other rule sets, but using this rule set comes at a severe cost due to the sheer volume of reactions its rules predict. Thus, compared to the most expressive of the practical RC + R rules, RC + 1, we highlight that both the Distilled and Learned rules accessed a distinct set of predicted reactions (Figure 8a). We provide two examples of such predicted reactions. In the first, the Learned rules predict a reaction whose nearest known analogue is a 4-hydroxy-2-oxovalerate aldolase reaction (EC 4.1.3.39) (Figure 8b). The RC + 1 rules extract a template based on this known analogue that specifies a primary carbon α to the aldehyde moiety, so it does not generalize to the predicted reaction which has an α ketone group. The second example, predicted by the Distilled rules, shows a dehydration reaction in analogy to an arogenate dehydratase reaction (EC 4.2.1.91) (Figure 8c). Here again, the RC + 1 rule includes a secondary carbon one bond from the reaction center which is unnecessary for the mechanism and fails to generalize to the predicted reaction.

DISCUSSION

In this work, we developed two sets of reaction rules which are novel in their use of mechanistic data. We demonstrated, relative to common heuristics, a trained node classifier more accurately predicts putatively mechanism-required molecular structures, extending the coverage of mechanism-informed rules beyond reactions with characterized mechanisms. Additionally, the Learned rules exhibit a more favorable trade-off between novelty and feasibility of predictions, evidenced by a less negative exponential decay rate. Quantifying the feasibility of predicted reactions is not straightforward, and while DORA-XGB has been shown to be a good proxy, it is still indirect and approximate. However, the time-split experiment, which operationalizes reaction feasibility by alternate means, shows similar trends to those in the experiment with model-based scoring.

Additional benefits of the Learned rules include their tunability and explicit use of mechanistic data. By varying the decision threshold on the node classifier, users can tune rule specificity with more fine-grained control than by varying the radius in an RC + R scheme. Furthermore, our mechanism-informed workflow stands to improve with additional characterization of enzyme reaction mechanisms, unlike prior methods.

A limitation of this work is that stereochemical information was removed prior to rule extraction. Since enzymes are largely stereospecific, this may lead to attrition as users select enzymes to catalyze predicted reactions. We show that given reactions with stereochemical information, our algorithms can correctly extract rules that recapitulate the reactions, inclusive of stereochemistry (Figure S8). This will be useful as adoption of stereoaware reaction generation grows.^{36,39}

In the broader context of computer-aided enzymatic synthesis planning, our rules are designed to be interoperable with most network expansion and retrosynthesis tools as they are encoded in the widely used SMARTS format. Building the correct information into reaction rules (or generative model) will always be more efficient than methods which predict and then filter with a discriminative model or heuristic. Our

method shares many of the limitations of other reaction rules and template-free reaction predictors. It does not account for noncovalent interactions between reactants and amino acid residues in the enzyme active site which will impact activation energies and thus kinetics through processes such as binding and transition state stabilization. These interactions heavily depend on the protein, which one can in various ways engineer to accommodate the required transformation. Nor does our method take thermodynamics into account, which depends on reactant and product concentrations in addition to their structures. However, by increasing the quality of reaction predictions, our method saves computational resources for these important downstream considerations, and moreover minimizes unnecessary experimental effort, thereby making enzymatic synthesis planning more efficient overall.

METHODS

Curating and Processing Data

Overall enzymatic reaction data were downloaded from Rhea.² To standardize the reaction SMILES,⁴⁷ we removed stereochemistry, neutralized charges that vary depending on environmental conditions (e.g., protonating $-OH$ groups), applied standard molecule normalization transformations, and canonicalized molecular SMILES. Standardization was carried out with cheminformatics library RDKit.⁴⁸ All transport reactions (e.g., between cellular components) were removed. We automatically generated RCs for each reaction by applying a set of minimal enzymatic reaction operators³² to reactants, alternately protecting all but one substructure matching the operator's RC template.

We removed stereochemistry to enable the use of the DORA-XGB⁴⁶ tool to compute feasibility scores. However, it is not necessary to do so; our template extraction process supports stereochemistry. As a demonstration of this, we extracted chiral templates from five reactions with stereochemical information and successfully recapitulated the reactions using RDChiral's `rdchiralRunText` function³⁶ (Figure S8).

Mechanistic enzymatic reaction data were downloaded from the Mechanistic & Catalytic Site Atlas (M-CSA).⁴² Elementary reaction steps and arrow-pushing diagrams were converted from Chemical Markup Language (CML) to SMILES using a custom script. We extracted reactants and products from each step, discarding all amino acid residues and cofactors. We then ensured that each elementary step was fully balanced and consistently atom-mapped. Finally, we generated atom-mapped SMILES for the overall reaction consisting of the elementary steps, and stored indices of atoms involved in any elementary step, including primarily those adjacent to bonds which changed order in the step, i.e., the elementary step's reaction center, and atoms involved in coordinate bonds, e.g., with metal ions. These atom indices were used subsequently in the creation of Distilled reaction rules and the construction of the binary classification task ultimately leading to the creation of the Learned reaction rules.

Assembling Reaction Rules

Reaction rules were encoded as SMARTS. The following procedure was followed for RC + R rules, Distilled rules, and Learned rules. For atoms included in reaction rule templates, we specified features: degree, valence, atomic number, formal charge, membership in ring structures, membership in aromatic structures, and number of heteroatom neighbors. Additionally, in the case where reaction rule templates contained disjoint subgraphs within single molecule templates, we connected all subgraphs to the reaction center via anonymous atoms. Anonymous atoms do not have any specified features and thus can match any atom. Bonds in which an anonymous atom was involved were also anonymous; in other words, no features of these bonds were specified. Atom and bond anonymity was achieved with wildcard characters "*" and "~", respectively. We

specified the bond order of bonds occurring between two non-anonymous atoms.

The criteria for inclusion of an atom nonanonymously in a reaction rule template is as follows for different classes of reaction rules. Distilled: all atoms whose indices were saved during the processing of M-CSA data, i.e., part of any elementary step reaction center or coordinate bond. Learned: all atoms labeled with a score above the chosen decision threshold by a trained GNN classifier described further in a section, "Binary classification of atom involvement in catalytic mechanism". RC + R: all atoms located within R bond hops of the reaction center.

RDChiral³⁶ and EVODEX³⁹ reaction rules were extracted from the portion of Rhea overall enzymatic reaction for which we had atom mappings. We applied the template extraction pipelines provided in the publicly available code accompanying each of those publications. Pathway-level clustered reaction rules were constructed by first clustering reactions with similar reactant structures and then writing a reaction rule which includes the maximum common subgraph of all reactions grouped into a cluster. For more details see Ni, 2022.⁴¹ EHreact⁴⁰ rules were generated by applying their pipeline to the Rhea reactions we curated without all-atom mapping. We used the code accompanying the EHreact publication, based on Reaction Decoder Tool,⁴⁹ to get all atom mappings. To generate a final set of rules, we selected the most specific mutual template of each tree as recommended in EHreact.

Finally, we ensured that all reaction rules were unique.

Mapping Rules to Reactions

Using the rdChemReactions module of RDKit, we attempted to apply each reaction rule to each overall reaction from Rhea. If the rule was able to recapitulate the reaction, i.e., generated the observed product(s) from the provided reactant(s), that rule was said to "map" that reaction. In the case of multiple rules mapped to a single reaction, we resolved the conflict with one of two approaches. For the RC + 0 rules, we preferred rules without structural patterns consisting of disjoint fragments of a single whole molecule, and we preferred rules without C–C bond-breaking. After applying these criteria, if there were still multiple mappings, we chose the rule which mapped more overall reactions. For the case of all other reaction rule sets, we selected the rule with the largest template, in terms of number of atoms included in that template.

We modified the rule-reaction-mapping criteria in the case of RetroBioCat and EVODEX rules, which are not always atom-balanced. For these rule sets, we aggregated all generated reactions which mapped subsets of known reaction's reactants to subsets of its products. In other words, we took the union of generated reactions' reactants and the union of generated reactions' products. We then checked that every reactant and product from the known reaction appeared in the left-hand side and right-hand side unions, respectively. If this condition was met, the known reaction was counted as "mapped". Furthermore, we allowed for common currency molecules, such as oxygen, water, or the pair NAD⁺/NADH to be omitted.

Binary Classification of Atom Involvement in Catalytic Mechanism

Data Labeling. We first mapped overall enzymatic reactions from Rhea with the assembled Distilled rules. For this subset of mapped reactions, we labeled each atom included nonanonymously in the Distilled rule with the positive class, "involved in the catalytic mechanism". All other atoms in the reaction were labeled with the negative class.

Model Construction. We constructed GNN-based classifiers using Chemprop⁵⁰ and PyTorch Lightning. We used the default atom and bond features of Chemprop and chose the condensed graph of reaction⁵¹ featurizer.⁵² Input features were transformed into hidden representations via bond-wise message passing layers⁵³ and subsequently fed into a custom prediction head (either linear or feedforward neural network) which output scalar values for each atom in the input reaction.

Training. All model parameters were learned using a binary cross entropy loss function with a positive weight parameter chosen through hyperparameter optimization.

$$l(x, y) = -[mpy \log(\sigma(x)) + (1 - y) \log(1 - \sigma(x))]$$

Where m = negative multiple, p = positive multiplier, y = label, x = model output. We used an Adam optimizer and Noam learning rate schedule with linear warmup.^{50,54}

Hyperparameter Optimization (HPO) and Evaluation. We conducted HPO and subsequently evaluated models using nested cross validation with three inner folds and five outer folds. Data were split using the GroupKFold class from scikit-learn's model_selection module.⁵⁵ Reaction data points were grouped by which Distilled reaction rule mapped them. We performed Bayesian HPO using the package Optuna with which we maximized the area under the precision recall curve.⁵⁶ All optimized hyperparameters and the corresponding values used in the evaluated and production models are provided in Table S1.

Labeling Atoms as Input to Learned Rules. Using the trained production models, we predicted labels for atoms in all overall reactions from Rhea for which we had atom mappings, which is required for featurization as a condensed graph of reactions. Note that this is equivalent to the subset of overall reactions mapped by the RC + 0 rules and is larger than the subset mapped by the Distilled rules. We applied various thresholds to the output atom scores to label them as involved in the catalysis or not. All atoms labeled with the positive class were included in the Learned reaction rule templates.

Evaluating Reaction Predictions of Reaction Rules

To evaluate the reaction rules on their primary task of generating predicted enzymatic reactions, we conducted a test reaction network expansion. The reaction network was seeded with 21 starting compounds designated by the Agile BioFoundry as "beachhead molecules" in a graphic adapted from Lee et al., 2019.¹ We customized the network expansion tool Pickaxe¹² to be able to predict reactions featuring multiple unique reactants, and to return atom-mapped reactions. All applicable reaction rules were applied to this starting set of beachhead molecules and then applied again to the union of beachhead molecules and products of the first generation of predicted reactions.

Calculating Evaluation Metrics from the Test Expansion

In addition to simple counts of reactions and compounds, we utilized two metrics to evaluate rules in the test expansion. Reaction feasibility scores were calculated using the DORA-XGB feasibility classifier using its default settings.⁴⁶ We used only the discrete labels and not the scores output by this classifier. Similarity between predicted reactions and known enzymatic reactions from Rhea was calculated as Jaccard similarity on custom extended-connectivity fingerprints.⁵⁷ In addition to the default atom invariants, we included minimal topological distance to any reaction center atom as an atom feature. Atom features are input into an iterative hashing procedure in the usual way. We restricted ourselves to molecular fragments of radius 3 based on all central atoms within 1 bond of the reaction center. In this way, only fragments contained within 4 bonds of the reaction center were represented in the resulting bit vector of length 1024. This choice was motivated by our finding that the vast majority of atoms involved in the catalytic mechanism are found within 4 bonds of the reaction center (Figure 2).

■ ASSOCIATED CONTENT

Data Availability Statement

All code and data for this work are made available at: <https://github.com/stefanpate/coarse-grain-rxns>

Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jcim.6c01595>.

Additional details and analysis of constructed rules and the trained GNN classifier, including optimal hyperparameters selected through cross-validation (PDF)

AUTHOR INFORMATION

Corresponding Author

Linda J. Broadbelt – Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, United States; Center for Synthetic Biology and Interdisciplinary Biological Sciences Graduate Program, Northwestern University, Evanston, Illinois 60208, United States; orcid.org/0000-0003-4253-592X; Phone: +1 847 467 1751; Email: broadbelt@northwestern.edu

Authors

Stefan C. Pate – Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, United States; Center for Synthetic Biology and Interdisciplinary Biological Sciences Graduate Program, Northwestern University, Evanston, Illinois 60208, United States; orcid.org/0000-0003-0500-6674

Keith E. J. Tyo – Department of Chemical and Biological Engineering, Northwestern University, Evanston, Illinois 60208, United States; Center for Synthetic Biology and Interdisciplinary Biological Sciences Graduate Program, Northwestern University, Evanston, Illinois 60208, United States; orcid.org/0000-0002-2342-0687

Complete contact information is available at: <https://pubs.acs.org/10.1021/acs.jcim.6c01595>

Author Contributions

S.C.P.—conceptualization of study, design and development of methodology, software development, formal analysis, data curation, investigation, writing (original draft), writing (review and editing), visualization. L.J.B. and K.E.J.T. (equal contributions)—conceptualization of study, provision of computing resources, supervision of research activities, funding acquisition, writing (original draft), writing (review and editing).

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was performed as part of the Bio-Optimized Technologies to keep Thermoplastics out of Landfills and the Environment (BOTTLE) Consortium and was supported by AMO and BETO under contract no. DE-AC36-08GO28308 with the National Laboratory of the Rockies (NLR), operated by Alliance for Sustainable Energy, LLC. Stefan Pate was supported in part by the National Institutes of Health Training Grant (T32GM153505 and T32GM008449) through Northwestern University's Biotechnology Training Program. We would like to thank Yash Chainani, Geoffrey Bonnazio, and Kenna Roberts for helpful discussions. This research was supported in part through the computational resources and staff contributions provided for the Quest high-performance computing facility at Northwestern University, which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology.

REFERENCES

- (1) Lee, S. Y.; Kim, H. U.; Chae, T. U.; Cho, J. S.; Kim, J. W.; Shin, J. H.; Kim, D. I.; Ko, Y.-S.; Jang, W. D.; Jang, Y.-S. A Comprehensive Metabolic Map for Production of Bio-Based Chemicals. *Nat. Catal.* **2019**, *2* (1), 18–33.
- (2) Bansal, P.; Morgat, A.; Axelsen, K. B.; Muthukrishnan, V.; Coudert, E.; Aimol, L.; Hyka-Nouspikel, N.; Gasteiger, E.; Kerhornou, A.; Neto, T. B.; Pozzato, M.; Blatter, M.-C.; Ignatchenko, A.; Redaschi, N.; Bridge, A. Rhea, the Reaction Knowledgebase in 2022. *Nucleic Acids Res.* **2022**, *50* (D1), D693–D700.
- (3) Bateman, A.; Martin, M.-J.; Orchard, S.; Magrane, M.; Adesina, A.; Ahmad, S.; Bowler-Barnett, E. H.; Bye-A-Jee, H.; Carpenter, D.; Denny, P.; Fan, J.; Garmiri, P.; Gonzales, L. J. D. C.; Hussein, A.; Ignatchenko, A.; Insana, G.; Ishtiaq, R.; Joshi, V.; Jyothi, D.; Kandasamy, S.; Lock, A.; Luciani, A.; Luo, J.; Lussi, Y.; Marin, J. S. M.; Raposo, P.; Rice, D. L.; Santos, R.; Speretta, E.; Stephenson, J.; Tootoo, P.; Tyagi, N.; Urakova, N.; Vasudev, P.; Warner, K.; Wijerathne, S.; Yu, C. W.-H.; Zaru, R.; Bridge, A. J.; Aimol, L.; Argoud-Puy, G.; Auchincloss, A. H.; Axelsen, K. B.; Bansal, P.; Baratin, D.; Batista Neto, T. M.; Blatter, M.-C.; Bolleman, J. T.; Boutet, E.; Breuza, L.; Gil, B. C.; Casals-Casas, C.; Echioukh, K. C.; Coudert, E.; Cucho, B.; De Castro, E.; Estreicher, A.; Famiglietti, M. L.; Feuerhahn, M.; Gasteiger, E.; Gaudet, P.; Gehant, S.; Gerritsen, V.; Gos, A.; Gruaz, N.; Hulo, C.; Hyka-Nouspikel, N.; Junge, F.; Kerhornou, A.; Mercier, P. L.; Lieberherr, D.; Masson, P.; Morgat, A.; Paesano, S.; Pedruzzi, I.; Pilboud, S.; Pourcel, L.; Poux, S.; Pozzato, M.; Pruess, M.; Redaschi, N.; Rivoire, C.; Sigrist, C. J. A.; Sonesson, K.; Sundaram, S.; Sveshnikova, A.; Wu, C. H.; Arighi, C. N.; Chen, C.; Chen, Y.; Huang, H.; Laiho, K.; Lehtvaslainen, M.; McGarvey, P.; Natale, D. A.; Ross, K.; Vinayaka, C. R.; Wang, Y.; Zhang, J.; The UniProt Consortium. UniProt: The Universal Protein Knowledgebase in 2025. *Nucleic Acids Res.* **2025**, *53* (D1), D609–D617.
- (4) Notebaart, R. A.; Kintsjes, B.; Feist, A. M.; Papp, B. Underground Metabolism: Network-Level Perspective and Biotechnological Potential. *Curr. Opin. Biotechnol.* **2018**, *49*, 108–114.
- (5) Tawfik, O. K. a. D. S.; S, D. Enzyme Promiscuity: A Mechanistic and Evolutionary Perspective. *Annu. Rev. Biochem.* **2010**, *79* (1), 471–505.
- (6) Jensen, R. A. Enzyme Recruitment in Evolution of New Function. *Annu. Rev. Microbiol.* **1976**, *30* (1), 409–425.
- (7) Copley, S. D. Toward a Systems Biology Perspective on Enzyme Evolution. *J. Biol. Chem.* **2012**, *287* (1), 3–10.
- (8) Kim, J.; Kershner, J. P.; Novikov, Y.; Shoemaker, R. K.; Copley, S. D. Three Serendipitous Pathways in *E. Coli* Can Bypass a Block in Pyridoxal-5'-phosphate Synthesis. *Mol. Syst. Biol.* **2010**, *6* (1), No. MSB201088.
- (9) Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem., Int. Ed.* **2018**, *57* (16), 4143–4148.
- (10) Hammer, S. C.; Kubik, G.; Watkins, E.; Huang, S.; Minges, H.; Arnold, F. H. Anti-Markovnikov Alkene Oxidation by Metal-Oxo-Mediated Enzyme Catalysis. *Science* **2017**, *358* (6360), 215–218.
- (11) Oue, S.; Okamoto, A.; Yano, T.; Kagamiyama, H. Redesigning the Substrate Specificity of an Enzyme by Cumulative Effects of the Mutations of Non-Active Site Residues *. *J. Biol. Chem.* **1999**, *274* (4), 2344–2349.
- (12) Shebek, K. M.; Strutz, J.; Broadbelt, L. J.; Tyo, K. E. J. Pickaxe: A Python Library for the Prediction of Novel Metabolic Reactions. *BMC Bioinf.* **2023**, *24* (1), No. 106.
- (13) Zhang, Q.; Sprague, W. W.; Kozarekar, S. S.; Pate, S. C.; Uekert, T.; Broadbelt, L. J. Discovery of Hybrid Chemical Synthesis Pathways with DORAnet. *Digit. Discovery* **2025**, *4* (11), 3109–3125.
- (14) Kumar, A.; Wang, L.; Ng, C. Y.; Maranas, C. D. Pathway Design Using de Novo Steps through Uncharted Biochemical Spaces. *Nat. Commun.* **2018**, *9* (1), No. 184.
- (15) Upadhyay, V.; Anand, M.; Maranas, C. D. novoStoic2.0: An Integrated Framework for Pathway Synthesis, Thermodynamic Evaluation, and Enzyme Selection. *PLOS Comput. Biol.* **2025**, *21* (8), No. e1012516.

- (16) Carbonell, P.; Parutto, P.; Baudier, C.; Junot, C.; Faulon, J.-L. RetroPath: Automated Pipeline for Embedded Metabolic Circuits. *ACS Synth. Biol.* **2014**, *3* (8), 565–577.
- (17) Delépine, B.; Duigou, T.; Carbonell, P.; Faulon, J.-L. RetroPath2.0: A Retrosynthesis Workflow for Metabolic Engineers. *Metab. Eng.* **2018**, *45*, 158–170.
- (18) Finnigan, W.; Hepworth, L. J.; Flitsch, S. L.; Turner, N. J. RetroBioCat as a Computer-Aided Synthesis Planning Tool for Biocatalytic Reactions and Cascades. *Nat. Catal.* **2021**, *4* (2), 98–104.
- (19) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, *365* (6453), No. eaax1566.
- (20) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604–610.
- (21) Koch, M.; Duigou, T.; Faulon, J.-L. Reinforcement Learning for Bioretrosynthesis. *ACS Synth. Biol.* **2020**, *9* (1), 157–168.
- (22) Gao, W.; Mercado, R.; Coley, C. W. *Amortized Tree Generation for Bottom-up Synthesis Planning and Synthesizable Molecular Design* 2022.
- (23) Cretu, M.; Harris, C.; Roy, J.; Bengio, E.; Liò, P. *SynFlowNet: Towards Molecule Design with Guaranteed Synthesis Pathways*. 2024, arXiv:2405.01155. arXiv.org e-Print archive. <https://arxiv.org/abs/2405.01155>.
- (24) Hadadi, N.; Hafner, J.; Shajkofci, A.; Zisaki, A.; Hatzimanikatis, V. ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol.* **2016**, *5* (10), 1155–1166.
- (25) MohammadiPeyhani, H.; Hafner, J.; Sveshnikova, A.; Viterbo, V.; Hatzimanikatis, V. Expanding Biochemical Knowledge and Illuminating Metabolic Dark Matter with ATLASx. *Nat. Commun.* **2022**, *13* (1), No. 1560.
- (26) Chowdhury, A.; Maranas, C. D. Designing Overall Stoichiometric Conversions and Intervening Metabolic Reactions. *Sci. Rep.* **2015**, *5* (1), No. 16009.
- (27) Probst, D.; Manica, M.; Teukam, Y. G. N.; Castrogiovanni, A.; Paratore, F.; Laino, T. Biocatalysed Synthesis Planning Using Data-Driven Learning. *Nat. Commun.* **2022**, *13* (1), No. 964.
- (28) Kreutter, D.; Schwaller, P.; Reymond, J.-L. Predicting Enzymatic Reactions with a Molecular Transformer. *Chem. Sci.* **2021**, *12* (25), 8648–8659.
- (29) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572–1583.
- (30) Tu, Z.; Coley, C. W. Permutation Invariant Graph-to-Sequence Model for Template-Free Retrosynthesis and Reaction Prediction. *J. Chem. Inf. Model.* **2022**, *62* (15), 3503–3513.
- (31) Zheng, S.; Zeng, T.; Li, C.; Chen, B.; Coley, C. W.; Yang, Y.; Wu, R. Deep Learning Driven Biosynthetic Pathways Navigation for Natural Products with BioNavi-NP. *Nat. Commun.* **2022**, *13* (1), No. 3342.
- (32) Ni, Z.; Stine, A. E.; Tyo, K. E. J.; Broadbelt, L. J. Curating a Comprehensive Set of Enzymatic Reaction Rules for Efficient Novel Biosynthetic Pathway Design. *Metab. Eng.* **2021**, *65*, 79–87.
- (33) Duigou, T.; du Lac, M.; Carbonell, P.; Faulon, J.-L. RetroRules: A Database of Reaction Rules for Engineering Biology. *Nucleic Acids Res.* **2019**, *47* (D1), D1229–D1235.
- (34) Kumar, A.; Maranas, C. D. CLCA: Maximum Common Molecular Substructure Queries within the MetRxn Database. *J. Chem. Inf. Model.* **2014**, *54* (12), 3417–3438.
- (35) Henry, C. S.; Broadbelt, L. J.; Hatzimanikatis, V. Discovery and Analysis of Novel Metabolic Pathways for the Biosynthesis of Industrial Chemicals: 3-Hydroxypropanoate. *Biotechnol. Bioeng.* **2010**, *106* (3), 462–473.
- (36) Coley, C. W.; Green, W. H.; Jensen, K. F. RDChiral: An RDKit Wrapper for Handling Stereochemistry in Retrosynthetic Template Extraction and Application. *J. Chem. Inf. Model.* **2019**, *59* (6), 2529–2537.
- (37) Levin, I.; Liu, M.; Voigt, C. A.; Coley, C. W. Merging Enzymatic and Synthetic Chemistry with Computational Synthesis Planning. *Nat. Commun.* **2022**, *13* (1), No. 7747.
- (38) Sankaranarayanan, K.; Heid, E.; Coley, C. W.; Verma, D.; Green, W. H.; Jensen, K. F. Similarity Based Enzymatic Retrosynthesis. *Chem. Sci.* **2022**, *13* (20), 6039–6053.
- (39) Conceicao, L. L.; Lin, H.; Tai, Y.-C.; Xue, G.; Zhang, H.; Zhang, C.; Anderson, J. C. EVODEX: A Cheminformatics Framework for Mechanistic Abstraction and Prediction of Enzymatic Reactions. *J. Chem. Inf. Model.* **2026**, *66*, 4793–4807.
- (40) Heid, E.; Goldman, S.; Sankaranarayanan, K.; Coley, C. W.; Flamm, C.; Green, W. H. EHreact: Extended Hasse Diagrams for the Extraction and Scoring of Enzymatic Reaction Templates. *J. Chem. Inf. Model.* **2021**, *61* (10), 4949–4961.
- (41) Ni, Z. Data-Driven Approaches to Predict and Utilize Enzyme Promiscuity for Novel Biosynthesis Design, Ph.D Thesis; Northwestern University 2022.
- (42) Ribeiro, A. J. M.; Holliday, G. L.; Furnham, N.; Tyzack, J. D.; Ferris, K.; Thornton, J. M. Mechanism and Catalytic Site Atlas (M-CSA): A Database of Enzyme Reaction Mechanisms and Active Sites. *Nucleic Acids Res.* **2018**, *46* (D1), D618–D623.
- (43) Ribeiro, A. J. M.; Riziotis, I. G.; Tyzack, J. D.; Borkakoti, N.; Thornton, J. M. EzMechanism: An Automated Tool to Propose Catalytic Mechanisms of Enzyme Reactions. *Nat. Methods* **2023**, *20* (10), 1516–1522.
- (44) Hartley, A. D.; Upadhyay, V.; Boorla, V. S.; Maranas, C. D. MechFind: A Computational Framework for de Novo Prediction of Enzyme Mechanisms *bioRxiv* 2025.09.25.678649.
- (45) Joung, J. F.; Fong, M. H.; Casetti, N.; Liles, J. P.; Dassanayake, N. S.; Coley, C. W. Electron Flow Matching for Generative Reaction Mechanism Prediction Obeying Conservation Laws. 2025, arXiv:2502.12979. arXiv.org e-Print archive. <https://arxiv.org/abs/2502.12979>.
- (46) Chainani, Y.; Ni, Z.; Shebek, K. M.; Broadbelt, L. J.; Tyo, K. E. J. DORA-XGB: An Improved Enzymatic Reaction Feasibility Classifier Trained Using a Novel Synthetic Data Approach. *Mol. Syst. Des. Eng.* **2025**, *10* (2), 129–142.
- (47) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36.
- (48) RDKit: Open-Source Cheminformatics. <https://www.rdkit.org>.
- (49) Rahman, S. A.; Torrance, G.; Baldacci, L.; Cuesta, S. M.; Fenninger, F.; Gopal, N.; Choudhary, S.; May, J. W.; Holliday, G. L.; Steinbeck, C.; Thornton, J. M. Reaction Decoder Tool (RDT): Extracting Features from Chemical Reactions. *Bioinformatics* **2016**, *32* (13), 2065–2066.
- (50) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2024**, *64* (1), 9–17.
- (51) Hoonakker, F.; Lachiche, N.; Varnek, A.; Wagner, A. A Representation to Apply Usual Data Mining Techniques to Chemical Reactions — Illustration on the Rate Constant of Sn2 Reactions in Water. *Int. J. Artif. Intell. Tools* **2011**, *20* (02), 253–270.
- (52) Heid, E.; Green, W. H. Machine Learning of Reaction Properties via Learned Representations of the Condensed Graph of Reaction. *J. Chem. Inf. Model.* **2022**, *62* (9), 2101–2110.
- (53) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59* (8), 3370–3388.
- (54) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need.

2023, arXiv:1706.03762. arXiv.org e-Print archive. <https://arxiv.org/abs/1706.03762>.

(55) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, É. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12* (85), 2825–2830.

(56) Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M. Optuna: A Next-Generation Hyperparameter Optimization Framework. 2019, arXiv:1907.10902. arXiv.org e-Print archive. <https://arxiv.org/abs/1907.10902>.

(57) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.

The advertisement features a vertical image on the left showing a blue, translucent, spherical object with a yellow, textured, tube-like structure extending from its base, which is surrounded by a green and pink mesh-like structure. The right side of the advertisement has a dark blue background with white and yellow text. The text reads: "CAS BIOFINDER DISCOVERY PLATFORM™", "PRECISION DATA FOR FASTER DRUG DISCOVERY", "CAS BioFinder helps you identify targets, biomarkers, and pathways", and "Unlock insights" in a yellow box. At the bottom right is the CAS logo, which includes the letters "CAS" and a stylized molecular structure, with the text "A division of the American Chemical Society" below it.

CAS BIOFINDER DISCOVERY PLATFORM™

**PRECISION DATA
FOR FASTER
DRUG
DISCOVERY**

CAS BioFinder helps you identify
targets, biomarkers, and pathways

Unlock insights

CAS
A division of the
American Chemical Society