Supporting information for: DORA-XGB: An improved enzymatic reaction feasibility classifier trained using a novel synthetic data approach

Yash Chainani[†1,2], Zhuofu Ni[†1,2], Kevin M. Shebek[1,2], Linda J. Broadbelt[1,2], and Keith E.J. Tyo[1,2],

[†] The authors contributed equally to this work

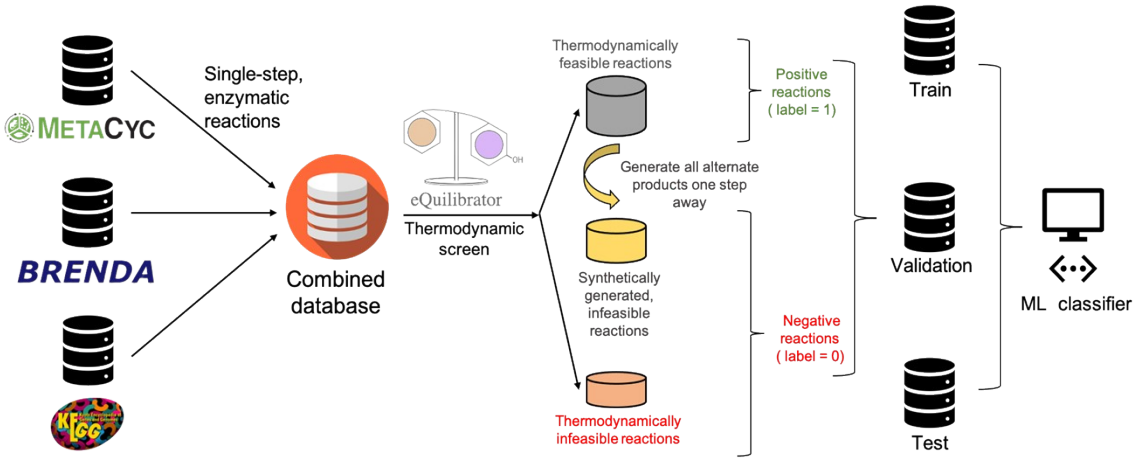
[1]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA

[2]Center for Synthetic Biology, Northwestern University, Evanston, IL, USA

Manuscript in preparation for:
Royal Society of Chemistry Molecular Systems Design and Engineering

1. GENERAL PIPELINE

1.1 Overall workflow for building DORA-XGB.



SI Fig. 1 Overall workflow for developing DORA-XGB

In order to develop our DORA-XGB models, reported reactions from BRENDA, KEGG, and MetaCyc were curated in both directions. A thermodynamic screen was then performed to divide curated reactions into a thermodynamically feasible and infeasible set. From the thermodynamically feasible set of reactions, all products one step away that could hypothetically have been observed under the same biochemical transformation but were never actually observed are generated (our "alternate reaction center" assumption). These synthetically generated infeasible reactions are then combined with the thermodynamically infeasible known reactions found earlier to create a training dataset with both positive and negative reaction data. Stratified train/validation/test splits in an 80/10/10 ratio were then performed to divide positive and negative reaction data into sets for model training, validation, and testing respectively. All model hyperparameters were tuned with a Bayesian hyperparameter optimization procedure as opposed to an exhaustive grid-search or a random-search.

2. DEPLOYMENT AND USAGE

2.1 Using DORA-XGB for the prediction of enzymatic reaction feasibility.

Users can try our consolidated DORA-XGB classifier by providing an enzymatic reaction string as an input to the classifier. The input reaction should be balanced. For a reaction of the form "A + cofactor → B + cofactor", the input string can written as: "A SMILES + cofactor SMILES = B SMILES + cofactor SMILES" or as "A.cofactor>>B.cofactor". The output from DORA-XGB is a feasibility score. The optimum threshold at which an input reaction can be labelled as feasible on the basis of its predicted score has been provided and was determined through analysis of precision, recall, and F1 scores of all models against the test set at 100 linearly spaced thresholds between 0 and 1. The threshold at which a model's F1 score on its corresponding test set is maximized is then reported as its optimum threshold. Users may also choose their own threshold, allowing them to filter fewer or more compounds and reactions in a network expansion based on the threshold used.

3. PREPROCESSING OF DATA

3.1 Complete list of cofactor concentration ratios used in this study.

The following ratios of cofactor concentrations are used in this study when using eQuilibrator to determine the minimum $\Delta_r G'$ value, $\Delta_r G'_{min}$ that can be released from a given reaction wherein metabolite concentration is allowed to vary from 0.1 mM to 100 mM. In this work, we considered NADH/NAD and NADPH/NADP as distinct cofactor pairs since they are bound by different concentration ratios.

[ATP]/ [ADP] = 10

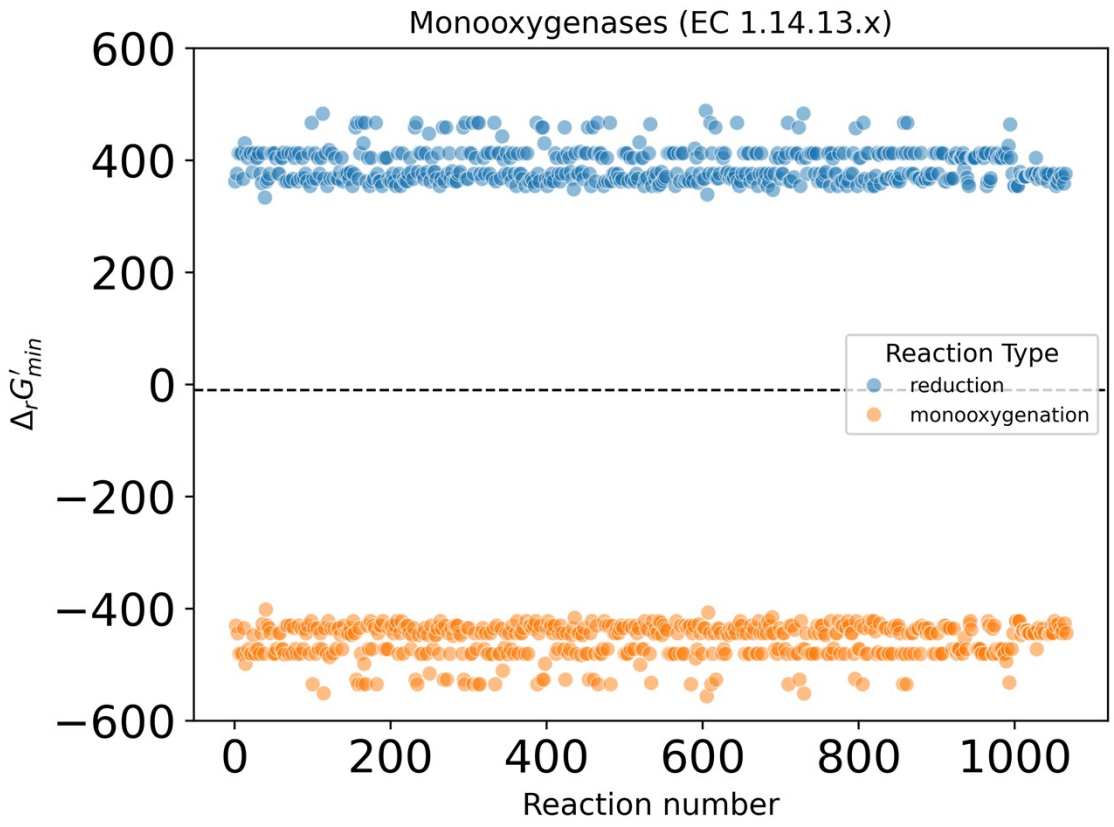
[ADP]/ [AMP] = 1

[NADH]/ [NAD+] = 0.1

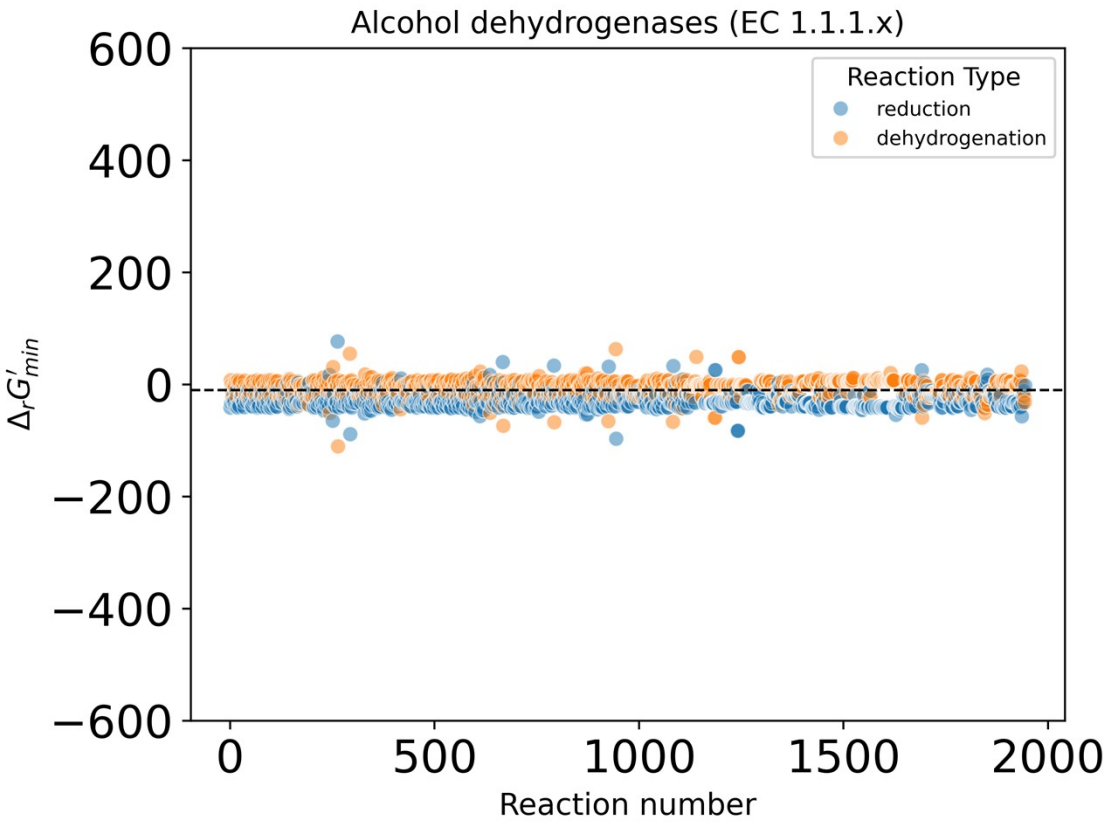
[NADPH]/ [NADP+] = 10

3.2 Distribution of $\Delta_r G'_{min}$ values for six select classes of enzymatic reactions.

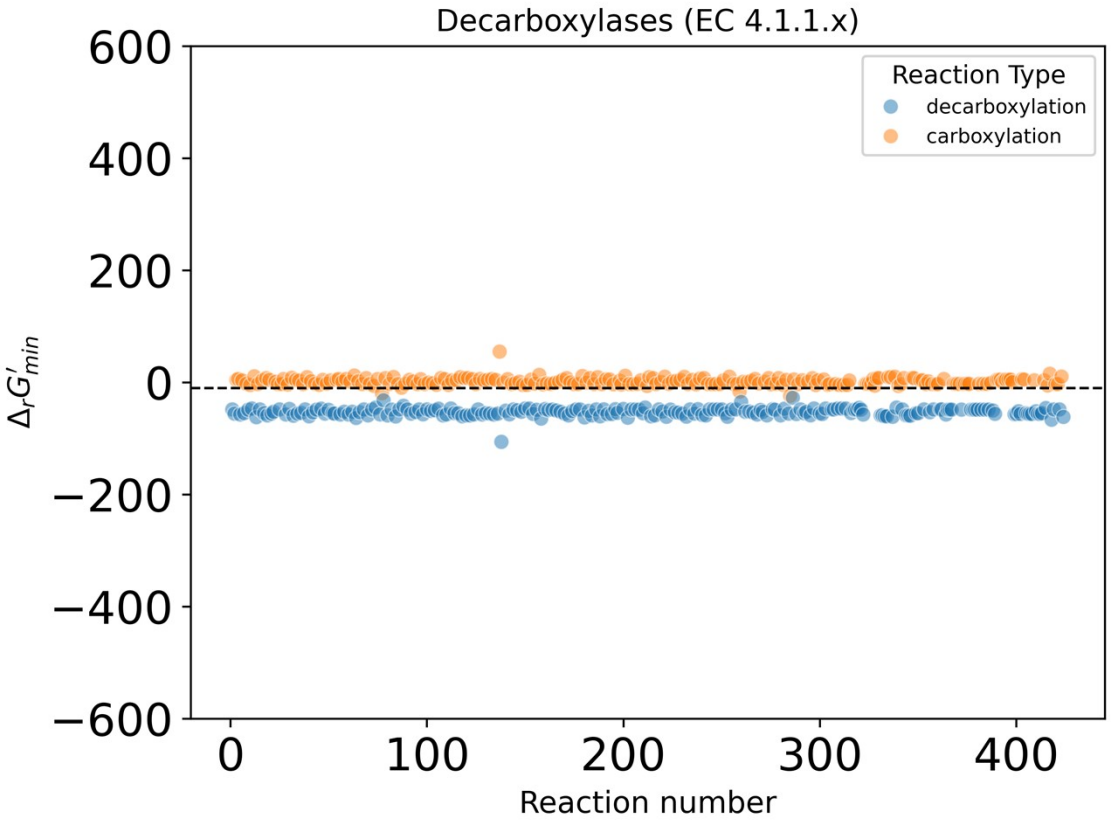
Here, we present the distribution of $\Delta_r G'_{min}$ values for six select classes of enzymatic reactions. These six classes are: (1) monooxygenases (EC 1.14.13.x) (SI Fig. 2), (2) alcohol dehydrogenases (EC 1.1.1.x) (SI Fig. 3), (3) decarboxylases (EC 4.1.1.x) (SI Fig. 4), (4) aldehyde dehydrogenases (EC 1.2.1.x) (SI Fig. 5), (5) phosphoryltransferases (EC 2.7.1.x) (SI Fig. 6), and (6) glucosyltransferases (EC 2.4.1.x) (SI Fig. 7). In each plot, $\Delta_r G'_{min}$ values are shown for both directions of the generalized transformation. For example, within the monooxygenase class of enzymes (EC 1.14.13.x), $\Delta_r G'_{min}$ values are far more downhill (i.e., more negative) in the monooxygenation direction wherein an oxygen is added to the substrate than in the reduction direction wherein an oxygen is removed from the substrate (SI Fig. 2).



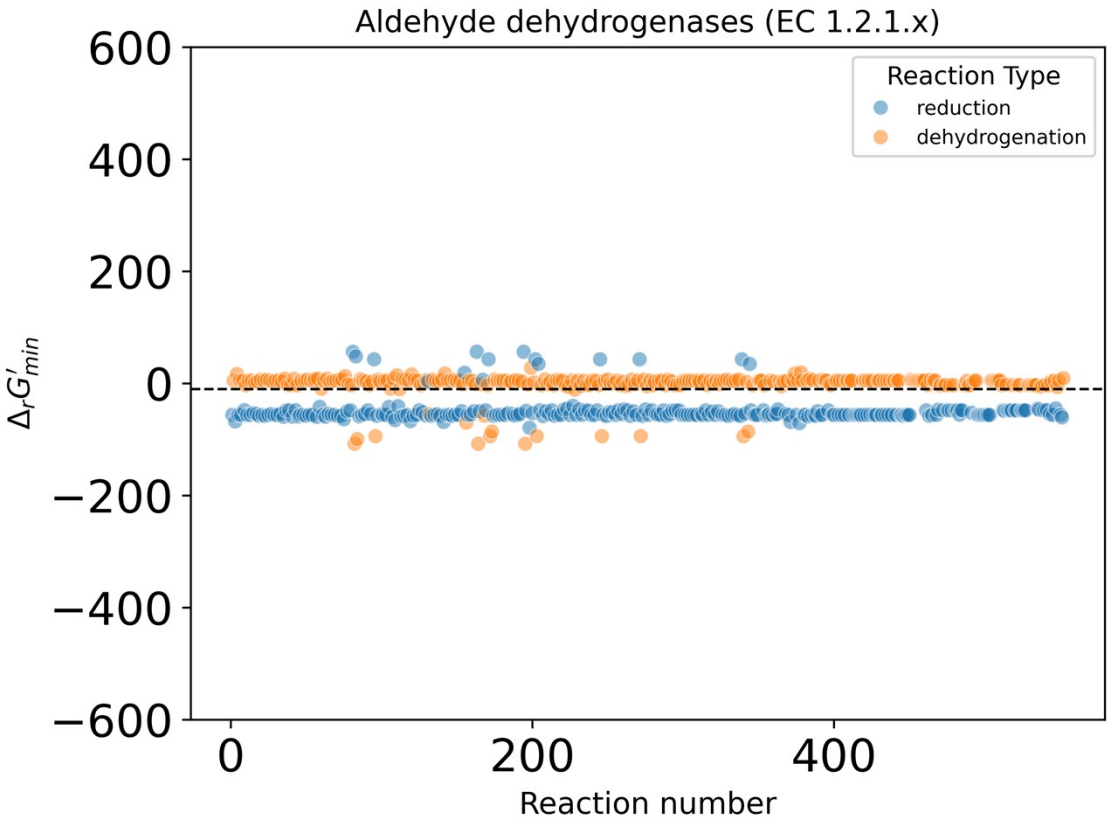
SI Fig. 2 Distribution of $\Delta_r G'_{min}$ values in both directions for monooxygenases (EC 1.14.13.x). The dashed line represents our thermodynamic feasibility threshold of -10 kJ/mol.



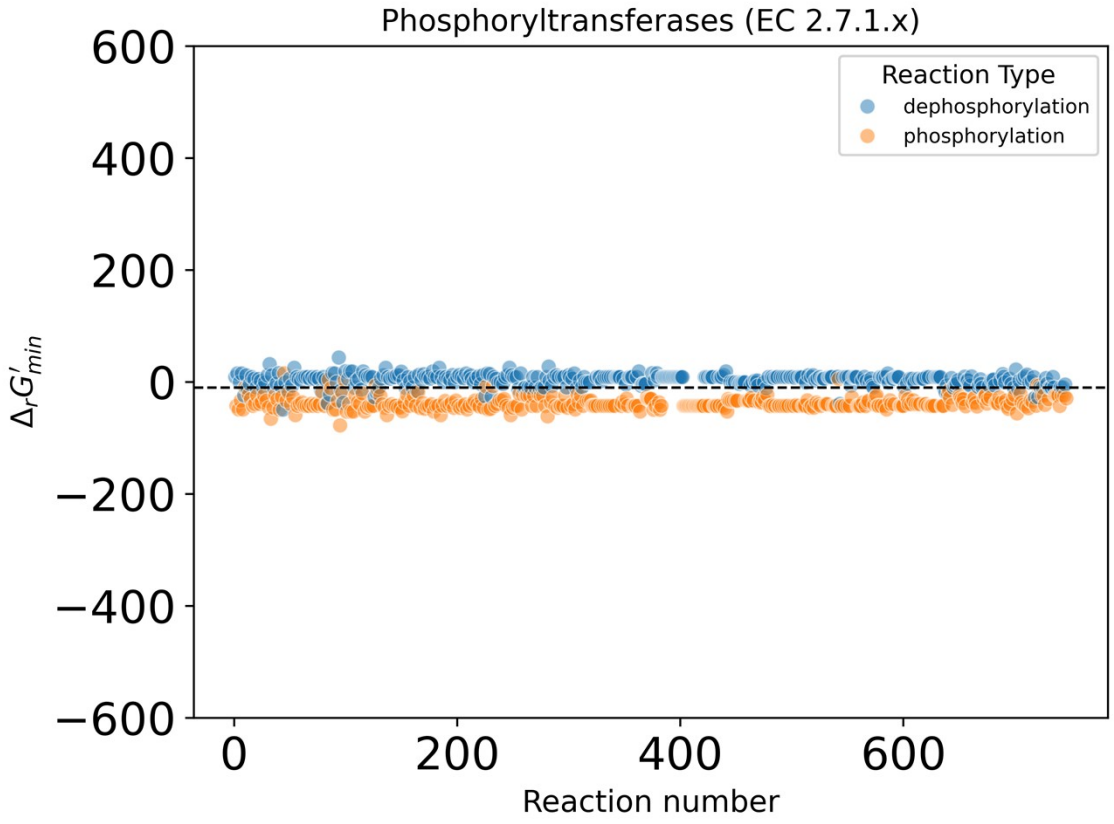
SI Fig. 3 Distribution of $^{\Delta_r G'_{min}}$ values in both directions for alcohol dehydrogenases (EC 1.1.1.x). The dashed line represents our thermodynamic feasibility threshold of -10 kJ/mol.



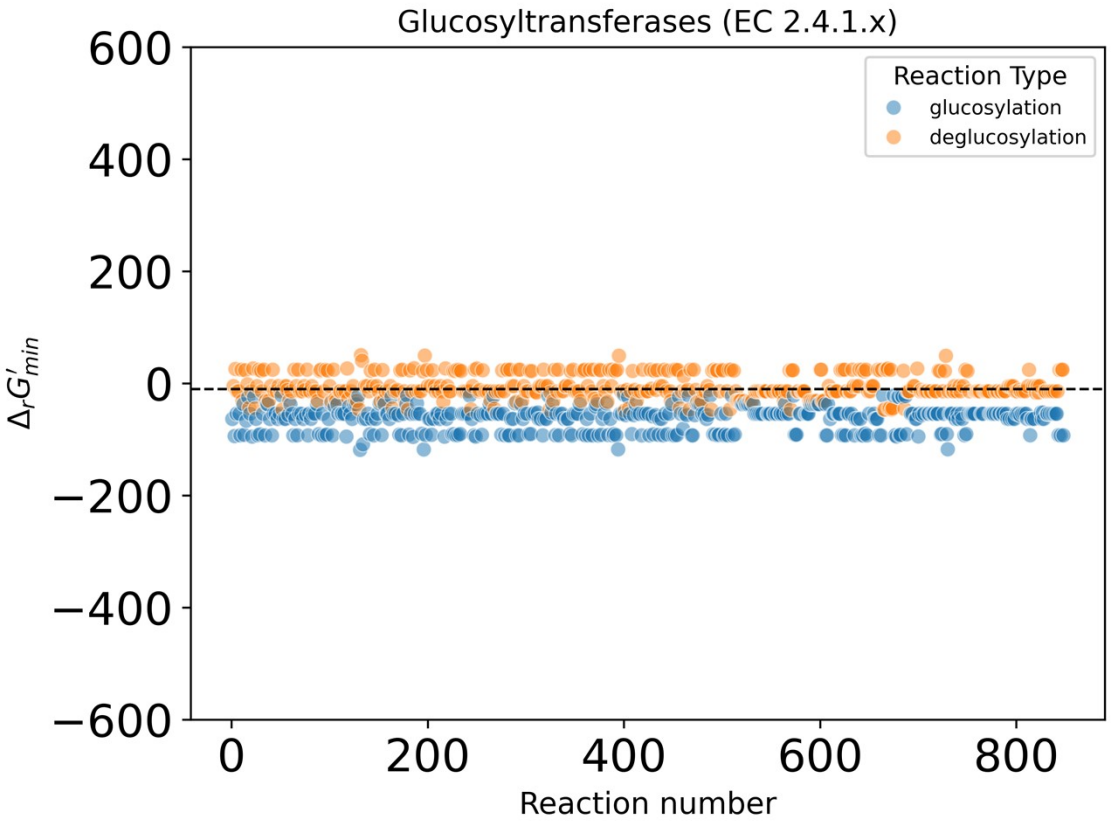
SI Fig. 4. Distribution of $^{\Delta_r G'_{min}}$ values in both directions for decarboxylases (EC 4.1.1.x). The dashed line represents our thermodynamic feasibility threshold of -10 kJ/mol.



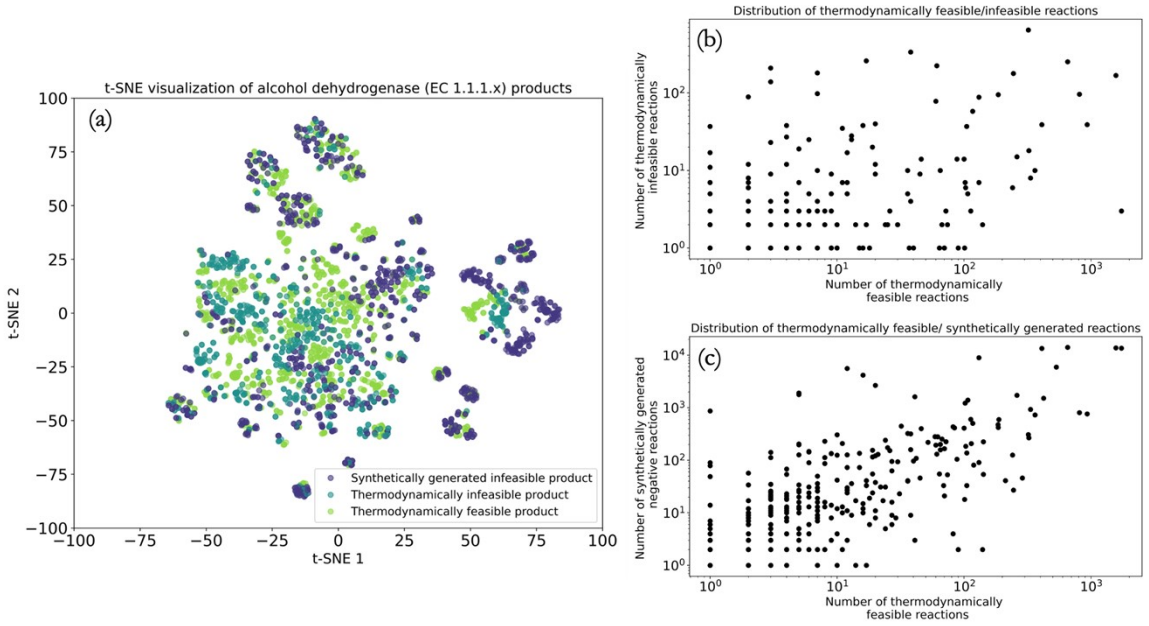
SI Fig. 5 Distribution of $^{\Delta_r G'_{min}}$ values in both directions for aldehyde dehydrogenases (EC 1.2.1.x). The dashed line represents our thermodynamic feasibility threshold of -10 kJ/mol.



SI Fig. 6 Distribution of $^{\Delta_r G_{min}}$ values in both directions for phosphoryltransferases (EC 2.7.1.x). The dashed line represents our thermodynamic feasibility threshold of -10 kJ/mol.



SI Fig. 7 Distribution of $^{\Delta_r G'_{min}}$ values in both directions for glucosyltransferases (EC 2.4.1.x)

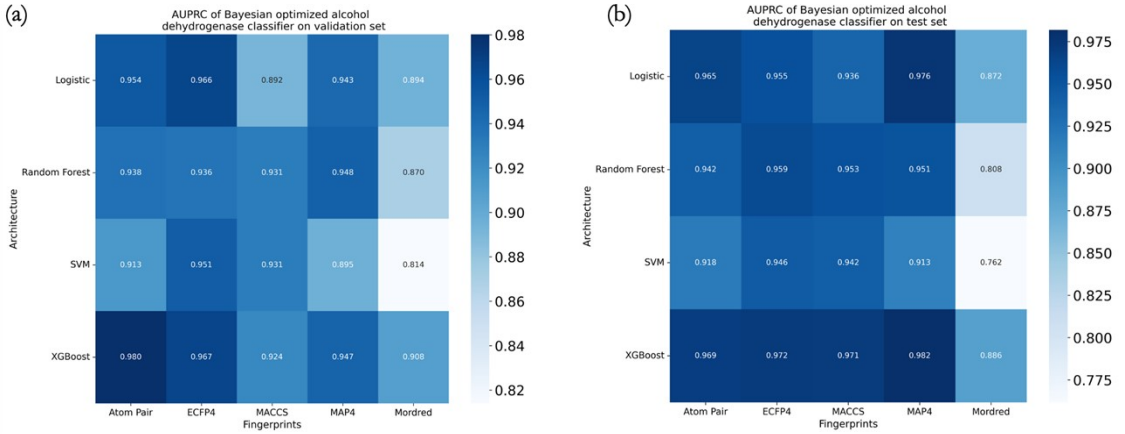


SI Fig. 8 (a) t-stochastic neighbors' estimation (t-SNE) of thermodynamically feasible, thermodynamically infeasible, and synthetically generated negative products that fall under the alcohol dehydrogenase transformation; **(b)** number of all thermodynamically infeasible reactions versus thermodynamically feasible reactions; **(c)** number of all synthetically generated negative reactions versus thermodynamically feasible reactions.

3.3 Statistics for feasible and infeasible reactions in the training dataset.

4. EVALUATING MODEL PERFORMANCE

4.1 Prototyping architecture – fingerprint combinations on training an alcohol dehydrogenase reaction feasibility classifier.



SI Fig. 9 (a) The average area under the precision-recall curve (AUPRC) of alcohol dehydrogenase classifiers deployed on a validation set and **(b)** test set of alcohol dehydrogenase reactions. With 1254 feasible and 1759 infeasible alcohol dehydrogenase reactions in total, a stratified train/validation/test split in an 80/10/10 ratio was performed to extract training, validation and test sets. All model hyperparameters were optimized on the validation set using a Bayesian hyperparameter approach. Reaction fingerprints are created by arranging molecular fingerprints in the order [substrate, NAD, product, NADH] for alcohol dehydrogenase reactions in the oxidation direction and [substrate, NADH, product, NAD] for alcohol dehydrogenase reactions in the reduction direction.

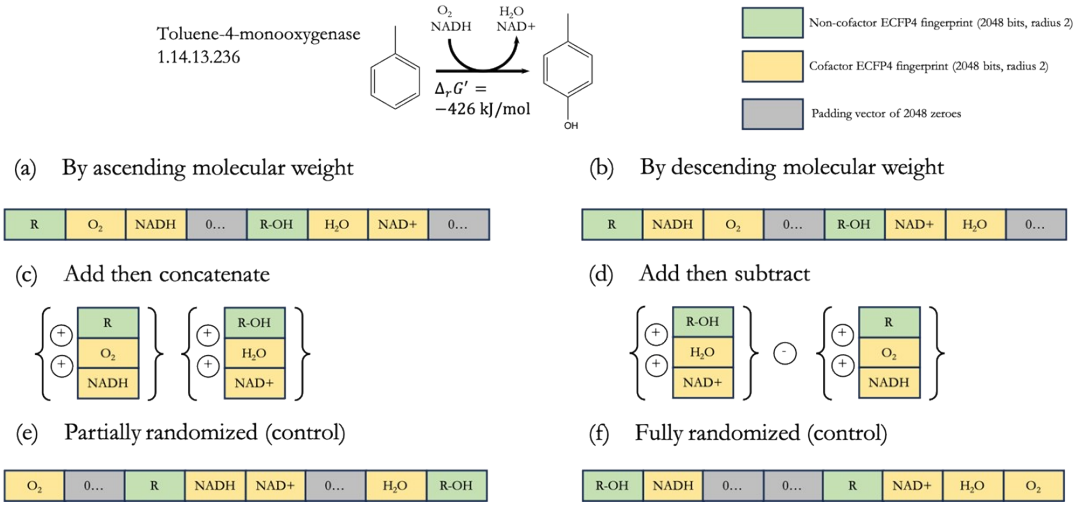
4.2 Exploring various arrangements of molecular fingerprints along reaction feature vectors

In this study, we explore different methods to arrange molecular fingerprints of species along reaction feature vectors. The simplest of these configurations involves simply concatenating molecular fingerprints of substrates, cofactors on the reactants' side, products, and cofactors on the products' side in ascending as well as descending molecular weights within each category. Since different reactions involve different numbers of species, each reaction vector is padded with zeros to 16,384 bits, which is the total number of elements present within the reaction vector representing the longest reaction within our curated database (SI Fig. 10(a) and 10(b)). In addition to concatenating all fingerprints together, we also explored performing simple operations onto molecular fingerprints to represent enzymatic reactions. In one of these configurations, "add then concatenate", the element-wise sum of all reactant fingerprints is taken and concatenated with that of product fingerprints (SI Fig. 10(c)). In another configuration, "add then subtract", the element wise of reactant fingerprints is subtracted from that of product fingerprints (SI Fig. 10(d)). In both of these configurations, paddings were not required.

Two additional fingerprinting configurations were also implemented to serve as negative controls (SI Fig. 10(e) and (f)). These controls seek to determine if arranging molecular fingerprints

in a standardized manner – as opposed to doing so randomly – truly optimizes model performance. In the first of these negative controls, the positions of reactant fingerprints are randomized along the first four 'slots' of a reaction's feature vector. This randomization is then repeated for product fingerprints, along the next four slots of the reaction feature vector to give a "partially randomized" feature vector (SI Fig. 10(e)). In the final negative control, the positions of all fingerprints are completely randomized throughout the reaction feature vector. For both of these negative controls, if model performance were to degrade, this decline would then confirm that there does, in fact, exist some dependency and value to the order in which molecular fingerprints are arranged to create reaction feature vectors.

Another form of negative control was then implemented in which we randomly scrambled assigned feasibility labels within our training data only. Models were then trained on this augmented dataset to determine if they would perform well against a test dataset within which the labels had not been altered. This would again confirm if our classifiers are performing well by chance or if they are truly learning to capture subtle differences within the training data. These negative controls are crucial for imbalanced datasets such as ours. Through such rigorous controls,

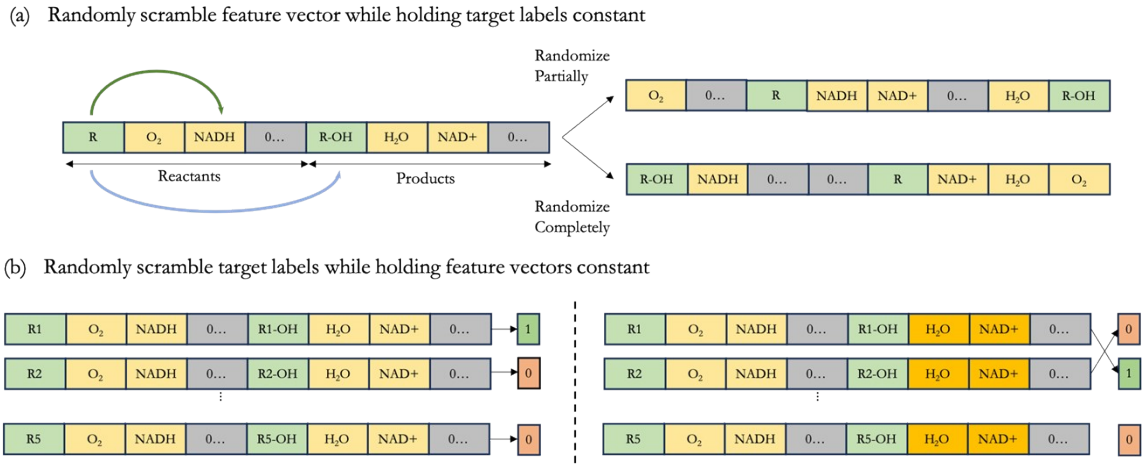


SI Fig. 10 Six different configurations for arranging molecular fingerprints along a reaction feature vector are explored in this study and depicted here through the example of the monooxygenation of toluene catalyzed by the enzyme toluene-4-monooxygenase (EC 1.14.13.236). In configurations **(a)** and **(b)**, primary reactant, primary product, and cofactor fingerprints are arranged in terms of ascending and descending molecular weights within categories. In configuration **(c)**, the fingerprints of all reactant structures are added in an element-wise fashion and concatenated with the element-wise sum of product fingerprints. In configuration **(d)**, the element-wise sum of product fingerprints is subtracted from that of reaction fingerprints. Configurations **(e)** and **(f)** serve as negative controls to confirm that there in indeed value to the order in which molecular fingerprints are arranged along a feature vector and that models are not just performing well by random chance.

we validate that the performance of our models is not merely by chance.

4.3 Additional negative control experiments

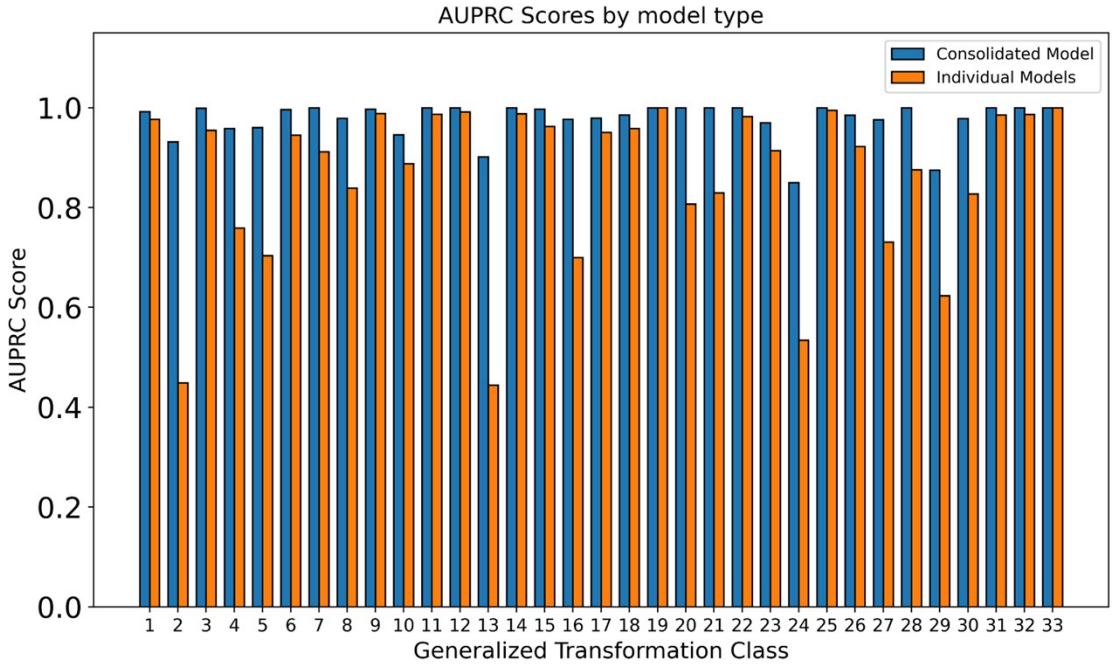
We implemented another negative control experiment to rigorously confirm the optimal performance of our feasibility classifier models. Here, assigned feasibility labels (as determined by thermodynamics and the synthetic generation of negative data) on reactions within our consolidated training set were randomly scrambled (SI Fig. 11 (a) and (b)). Models were then trained on this augmented dataset to determine if they would perform well against a test dataset within which feasibility labels had not been altered. This would again confirm if DORA-XGB was performing well solely by chance or if it were truly learning to capture subtle differences within reaction data. Such controls are crucial for imbalanced datasets such as ours because any model that simply predicts negative labels by default would be accurate 7/8 times anyway. Thus, these controls can ascertain that models are not being overfit to training data and not merely predicting reactions as



SI Fig. 11 In order to confirm that the performance of trained feasibility classification models is not merely by chance, we performed two types of negative control experiments where we expected model performance to decline: **(a)** in the first type of negative control, molecular fingerprints arranged along a reaction's feature vector are randomly scrambled within only the 'slots' allocated to reactants and products (partially randomized) as well as throughout the entire feature vector (fully randomized); **(b)** in the second type of control experiment, the configuration in which reaction feature vectors are constructed is held constant while the target feasibility labels within the training set are mutated. Feasibility models are then trained on these augmented labels to confirm that they will perform poorly on a test dataset in which assigned labels have not been mutated.

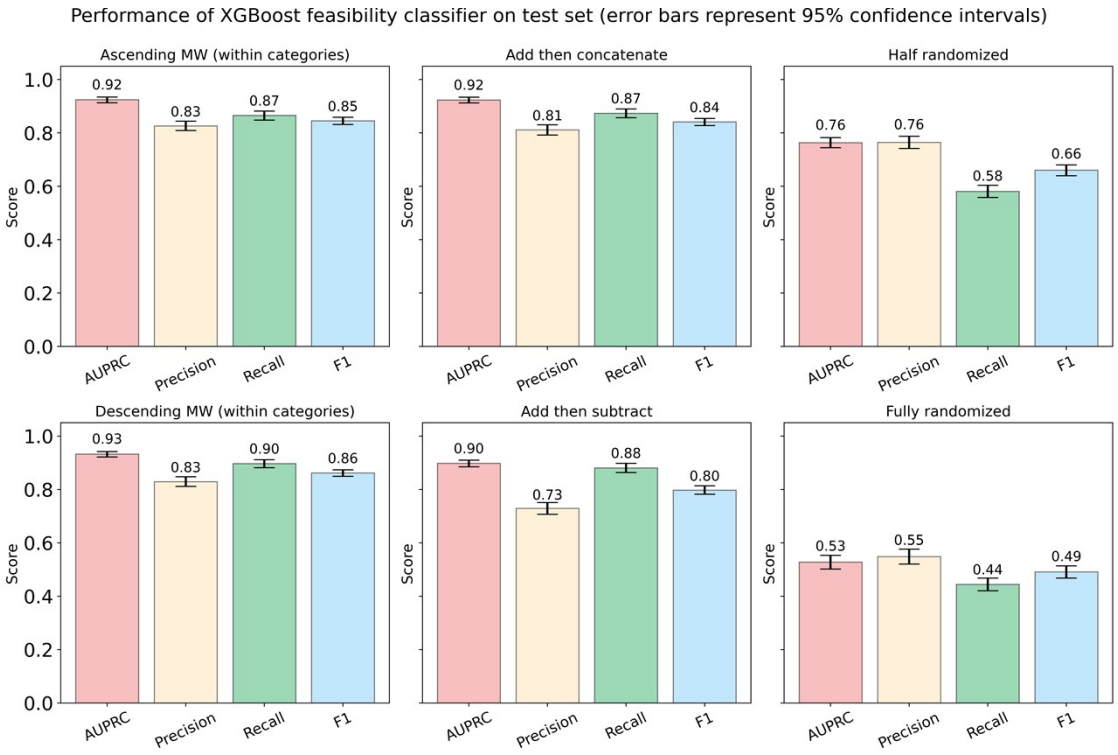
infeasible by default.

4.4 Comparing the performance of individual vs. consolidated classifiers



SI Fig. 12 In order to determine if it would be more effective to train multiple individual feasibility classifiers specific to each class or to train a single, consolidated feasibility classifier for all enzymatic reaction classes, we computed the average area under the precision-recall curve (AUPRC) between individual classifiers trained on 33 classes of enzymatic reactions and our consolidated classifier. The average AUPRC from individual classifiers was found to be lower than that of the consolidated classifier. The top 33 classes of generalized transformations make up for 64.3% of the reactions in our dataset.

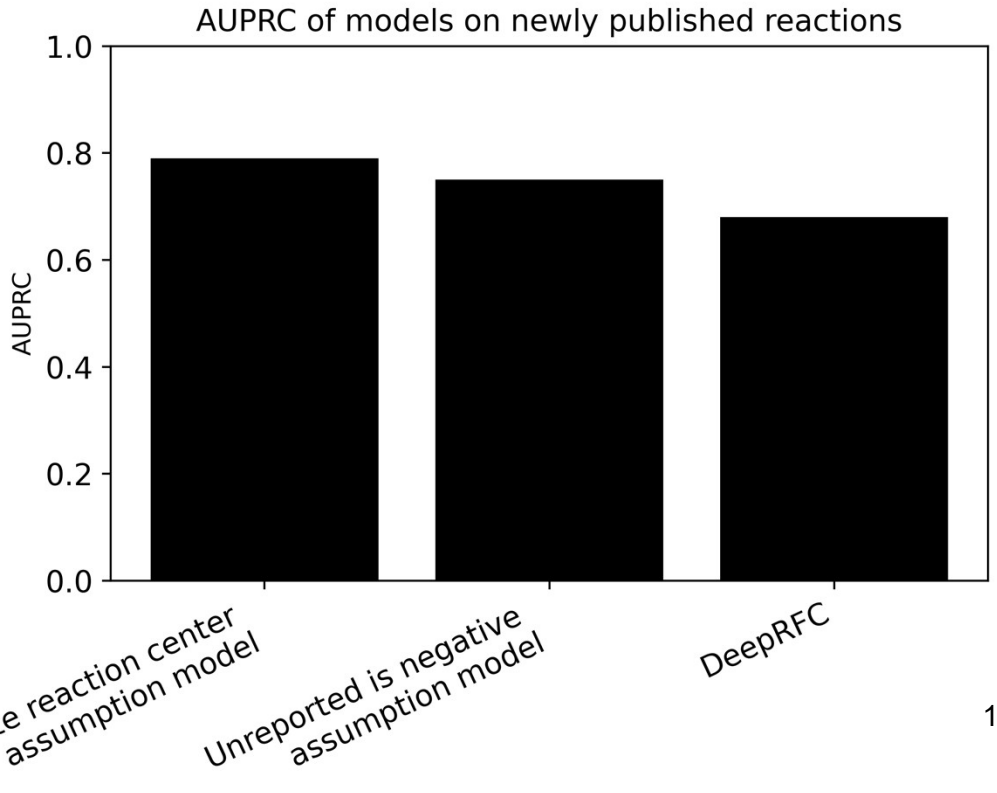
4.5 Performance of consolidated classifiers trained with various cofactor configurations against the test set.



SI Fig. 13 The average area under the precision-recall curve (AUPRC), precision, recall, and F1 scores of six consolidated feasibility classifiers trained on all enzymatic reactions with various feature vector configurations.

4.6 Comparing the performance of our alternate reaction center assumption model with the unreported is negative assumption model.

SI Fig. 14 Our reaction feasibility classifier trained on the "alternate reaction center" assumption receives a higher AUPRC score than DeepRFC, another deep-learning based classifier trained with negative data generated under the "unreported is negative" assumption. Our in-house "unreported is negative" assumption dataset led to a sharp decline in model performance in contrast to our model trained under the "alternate reaction center" assumption.



4.7 Performance on newly discovered *Escherichia coli* reactions.

A

		Actual: Positive	Actual: Negative
Predicted	Positive	28	0
Predicted	Negative	2	0

B

		Actual: Positive	Actual: Negative
Predicted	Positive	25	0
Predicted	Negative	15	0

C

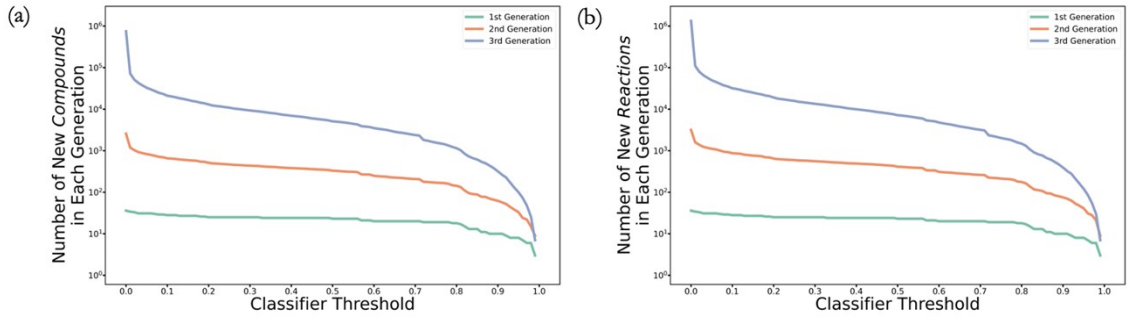
		Actual: Positive	Actual: Negative
Predicted	Positive	0	1581
Predicted	Negative	0	2955

D

		Actual: Positive	Actual: Negative
Predicted	Positive	0	4526
Predicted	Negative	0	12827

SI Table 1. **(a)** DORA-XGB's recovery of novel, experimentally validated reactions (30 total) obtained from an *e. coli* nontargeted metabolomics dataset; **(b)** recovery of predicted, plausibly positive and novel *e. coli* reactions (40 total) obtained from the same dataset; **(c)** prediction of 4536 total plausibly negative reactions that were synthetically generated from the 30 experimentally validated positive *e. coli* reactions; **(d)** prediction of 17353 plausibly negative reactions that were synthetically generated from the 40 plausibly positive *e. coli* reactions.

4.8 Filtering out infeasible compounds and reactions in a network expansion

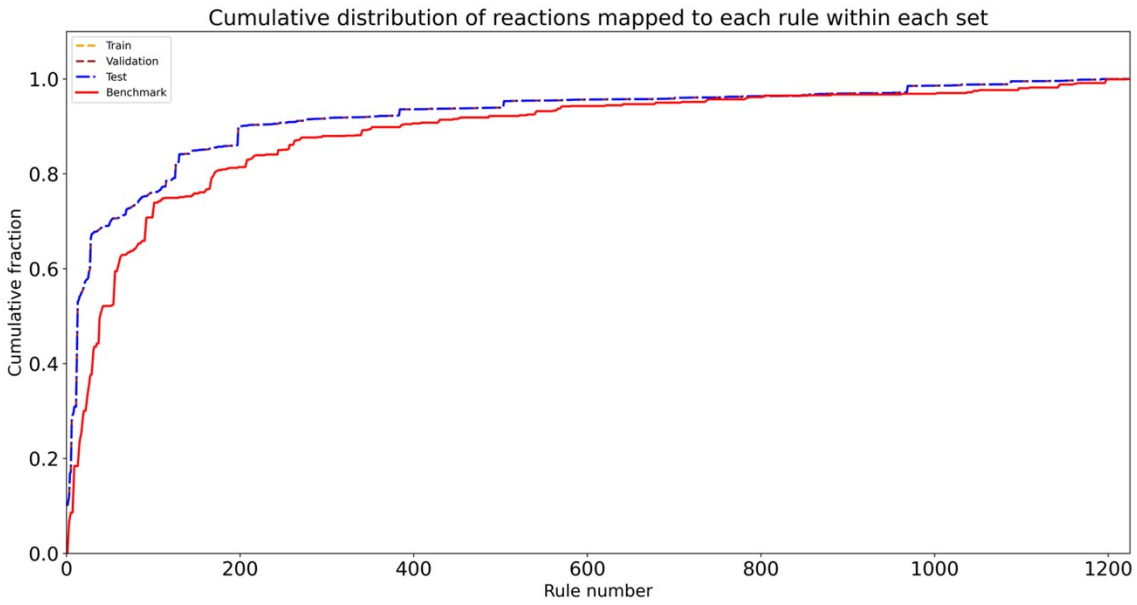


SI Fig. 15 (a) Number of new reactions and **(b)** compounds remaining after each generation of a three-step network expansion performed by DORA-XGB starting from pyruvic acid. With DORA-XGB, users can either set custom thresholds or use the ones reported in this work. A higher threshold would lead to the prediction of few higher confidence pathways within short computational runtimes, but this efficiency comes at the cost of filtering out several other potential candidate pathways. Meanwhile, a lower threshold would return a larger space of candidate pathways but with longer runtimes and greater computational expense.

4.9 Examining performance drops between test and benchmarking AUPRC of DORA-XGB

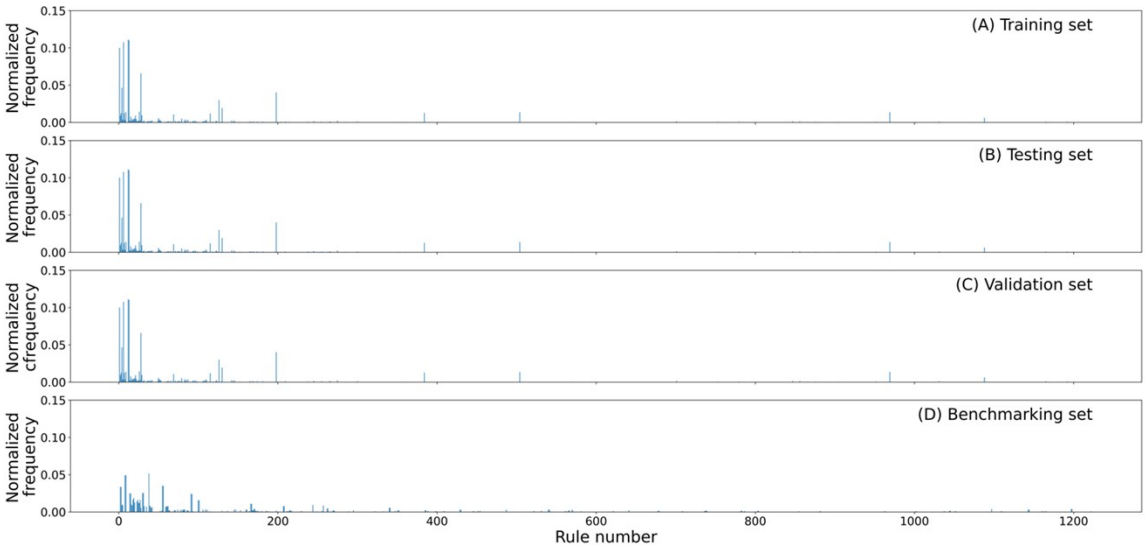
We realize that there is a considerable performance drop in terms of AUPRC between the external benchmarking set and the testing set (0.79 vs. 0.92). We highlight that this performance drop arises because enzymatic transformations that are frequent in the benchmarking set are not commonly observed in the training, validation, or the testing sets. We encode for enzymatic transformations in terms of our publicly available reaction rules (or templates). There exist 1224 such unique rules, and they are ordered in terms of the number of known reactions mapped to each rule, i.e., rule0001 has far more reactions mapped to it (1236 reactions in BRENDA) than rule1224 (only 2 reactions in BRENDA). Consequently, lower-numbered rules represent more common transformations (e.g., the dehydrogenation of alcohols as encoded in rule0002) than higher-numbered rules (e.g., the hydrolysis of nitrile-containing substrates encoded in rule0243). When we consider the proportion of reactions that are cumulatively represented by each rule within the four total sets (training, validation, testing, and benchmarking), we find that this distribution is nearly identical for the training, testing, and validation sets but distinct for the benchmarking set. This is expected since we performed our train/ validation/ test splits iteratively on a rule-by-rule basis and with stratification such that the distribution of positive (feasible) to negative (infeasible) reactions is retained for each rule within the training, validation, and testing sets. Meanwhile, the external benchmarking set represents an out-of-distribution sample so it is again expected that that the cumulative distribution of reactions mapped to each rule in the benchmarking set would be

different than that in the training, validation, and testing sets. More crucially, visualizing the cumulative distribution reveals that there is actually a skew in the benchmarking set towards higher-numbered rules. Put differently, reactions that exhibit rarer transformations occur more frequently in the benchmarking set than in the training set. Given fewer opportunities to learn such rare transformations during training, our model performance drops when it confronts these transformations in benchmarking.



SI Fig. 16 Cumulative distribution of reactions mapped to each rule within the training (dashed orange line), validation (dashed maroon line), testing (dashed blue line), and benchmarking (solid red line) sets.

Instead of a cumulative distribution for each set, we can also plot the distribution of the number of reactions mapped to each rule normalized by the total number of reactions in that set. This further confirms that the benchmarking set not only comprises a different frequency distribution of reactions mapped to each rule when compared to the training, validation, and testing sets but also a skew towards reactions mapped to higher-numbered rules, i.e., rarer transformation types.



SI Fig. 17 Normalized frequency of reactions mapped to each reaction rule in the (a) training, (b) testing, (c) validation, and (d) benchmarking sets.