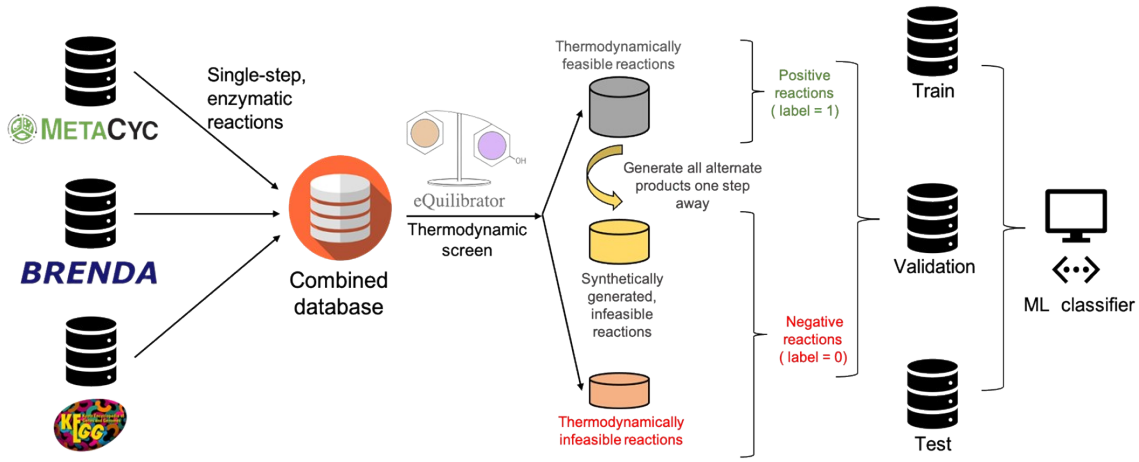
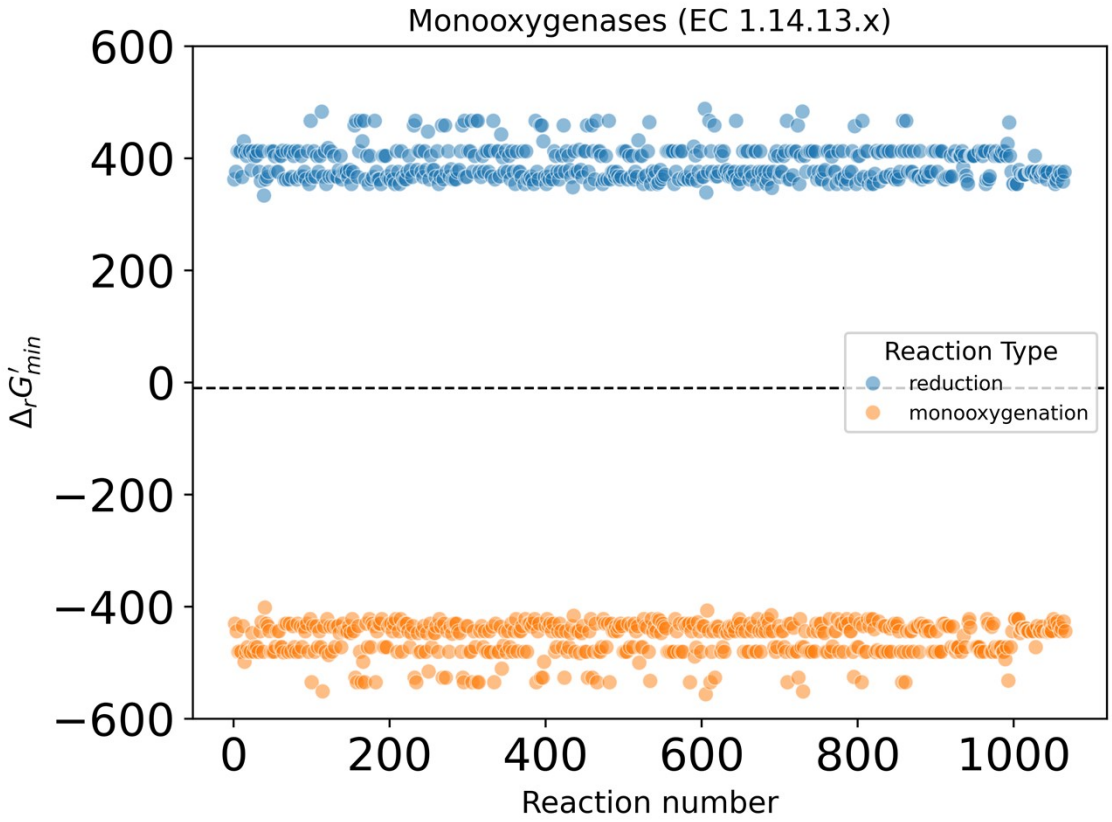
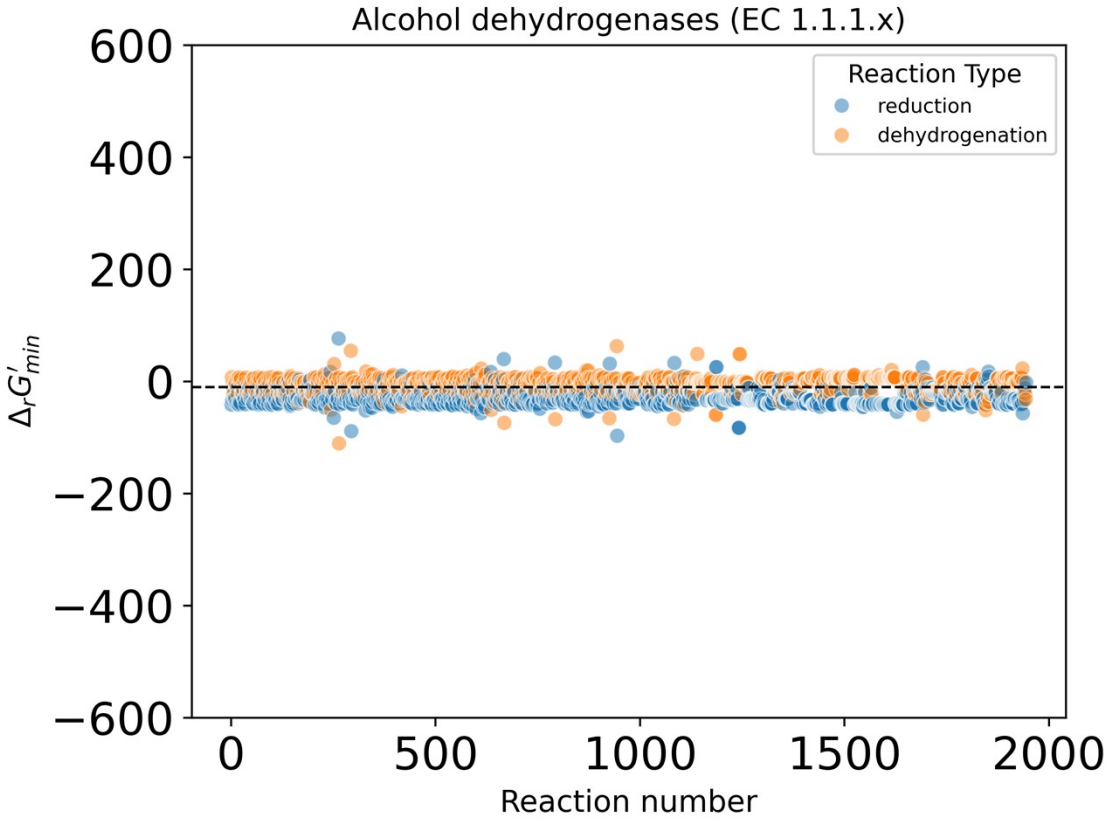


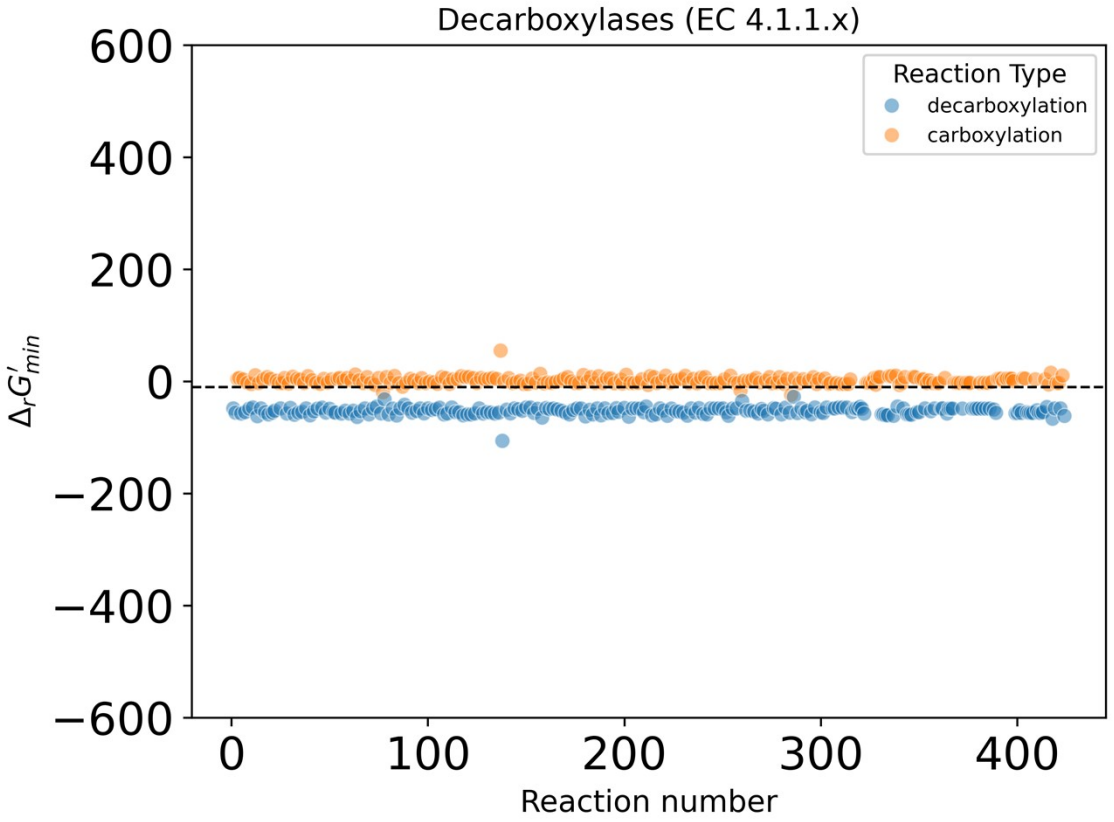
SI Fig. 1 Overall workflow for developing DORA-XGB



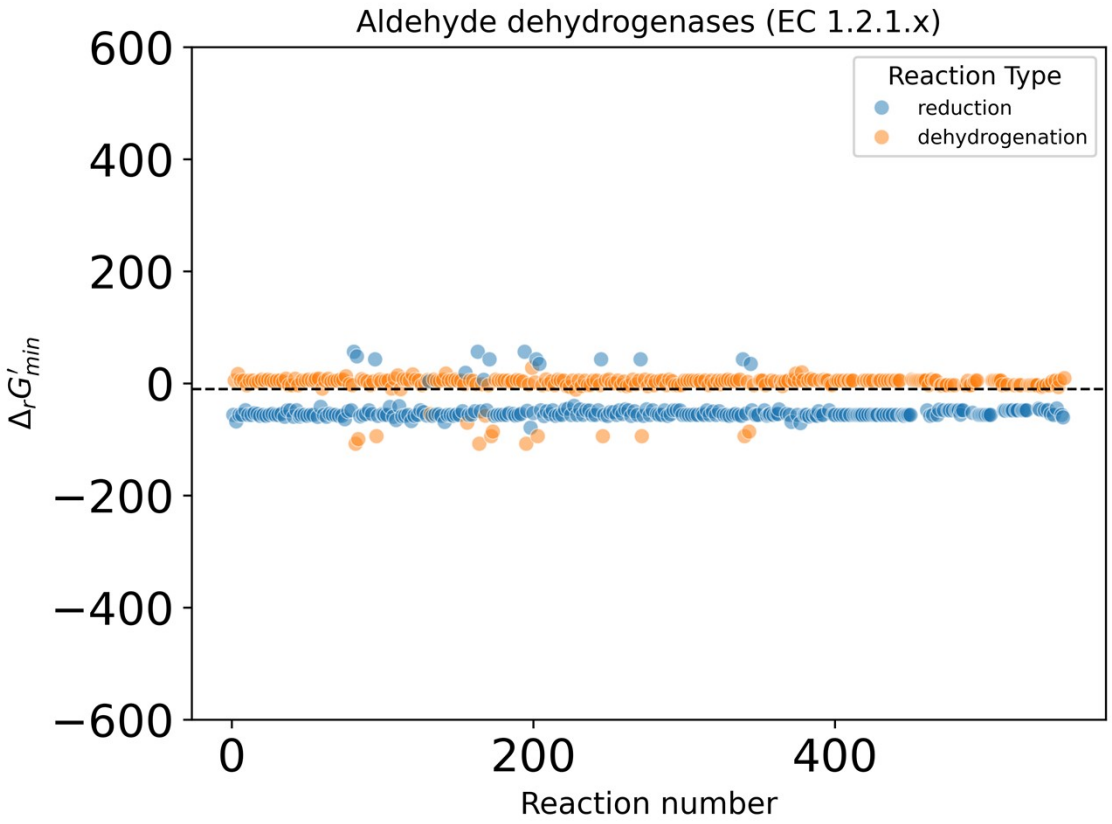
SI Fig. 2 Distribution of $^{\Delta_r G_{min}}$ values in both directions for monooxygenases (EC 1.14.13.x). The dashed line represents our thermodynamic feasibility threshold of -10 kJ/mol.



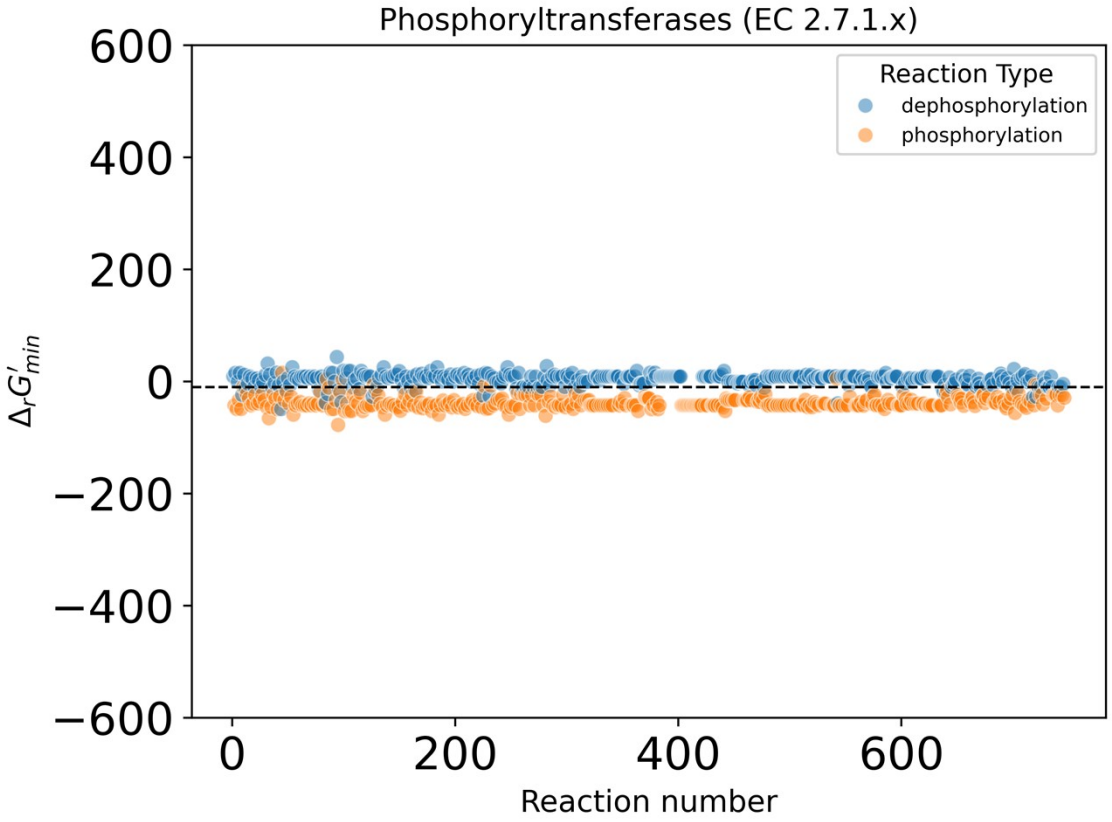
SI Fig. 3 Distribution of $^{\Delta_r G_{min}}$ values in both directions for alcohol dehydrogenases (EC 1.1.1.x). The dashed line represents our thermodynamic feasibility threshold of -10 kJ/mol.



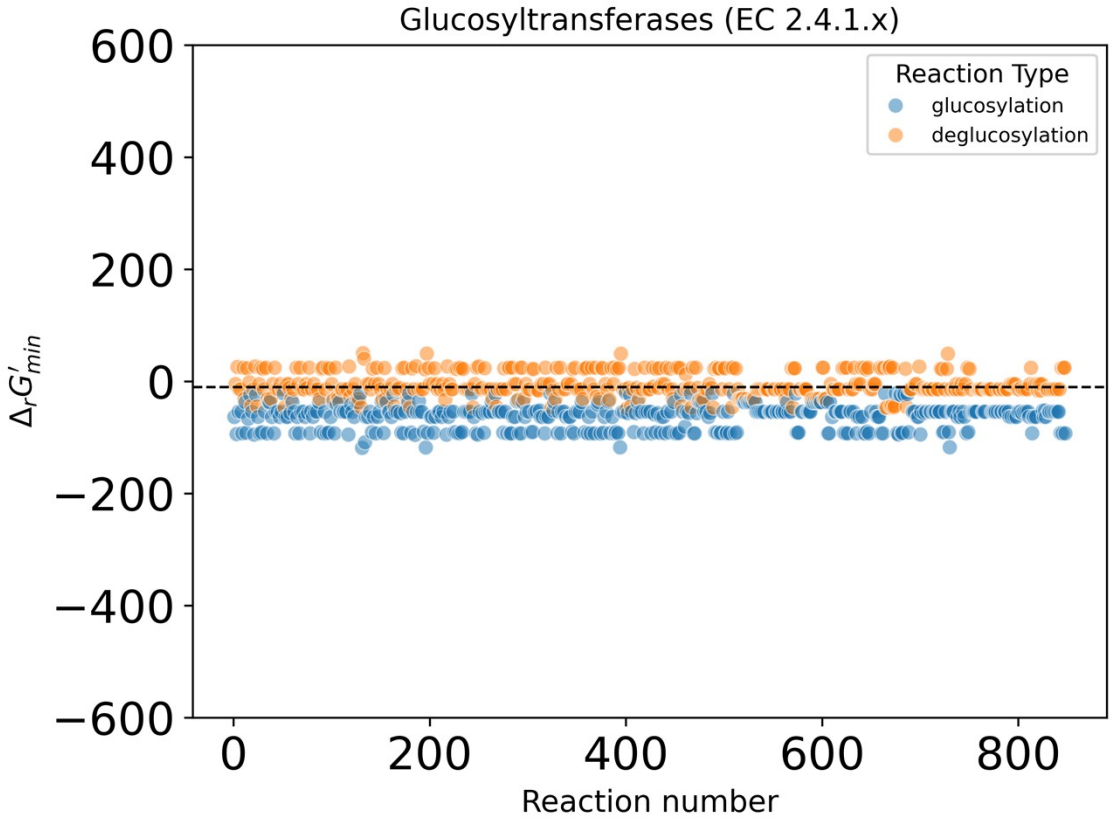
SI Fig. 4. Distribution of $\Delta_r G'_{min}$ values in both directions for decarboxylases (EC 4.1.1.x). The dashed line represents our thermodynamic feasibility threshold of -10 kJ/mol.



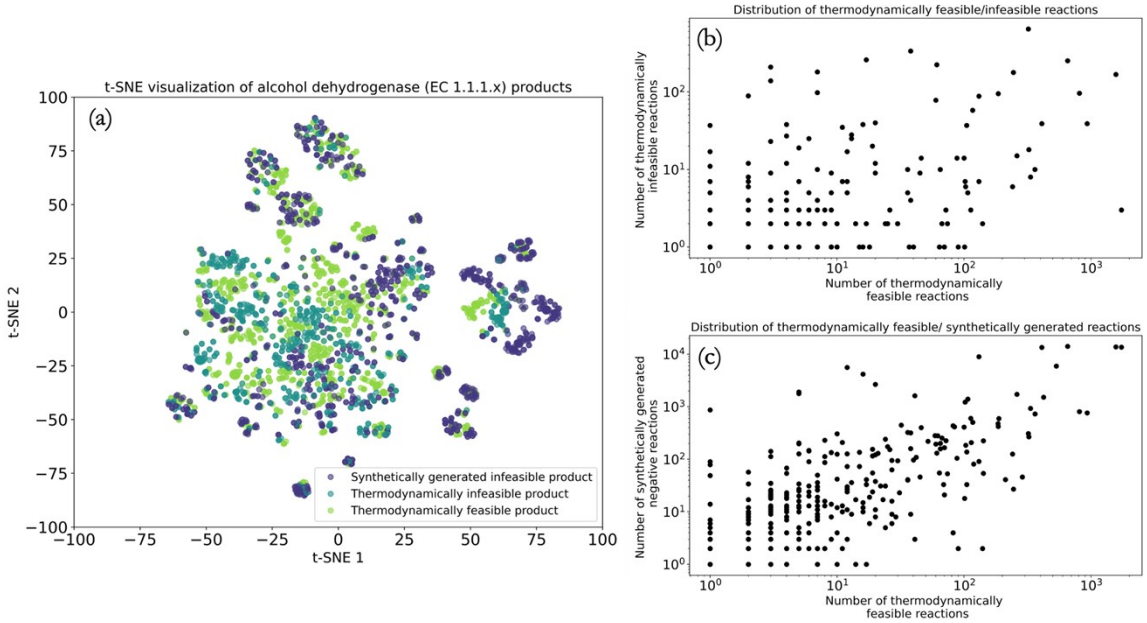
SI Fig. 5 Distribution of $^{\Delta_r G_{min}}$ values in both directions for aldehyde dehydrogenases (EC 1.2.1.x). The dashed line represents our thermodynamic feasibility threshold of -10 kJ/mol.



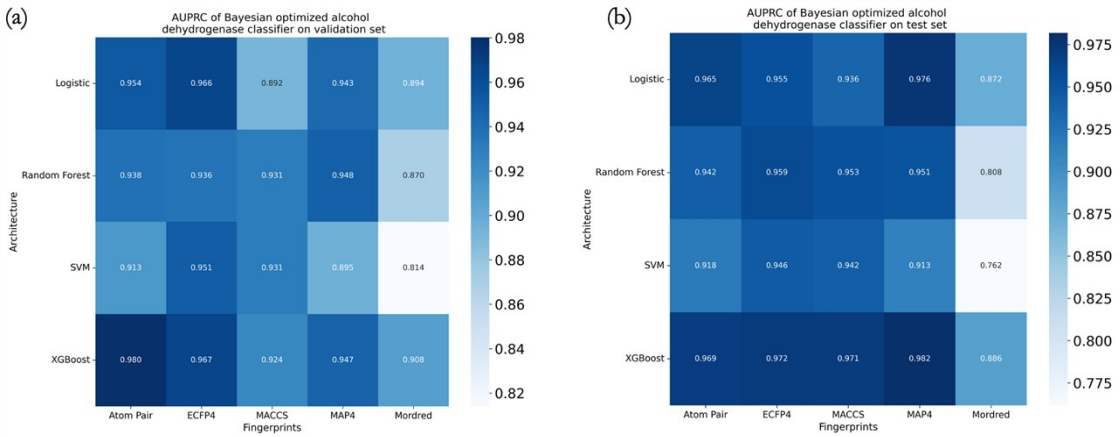
SI Fig. 6 Distribution of $^{\Delta_r G_{min}}$ values in both directions for phosphoryltransferases (EC 2.7.1.x). The dashed line represents our thermodynamic feasibility threshold of -10 kJ/mol.



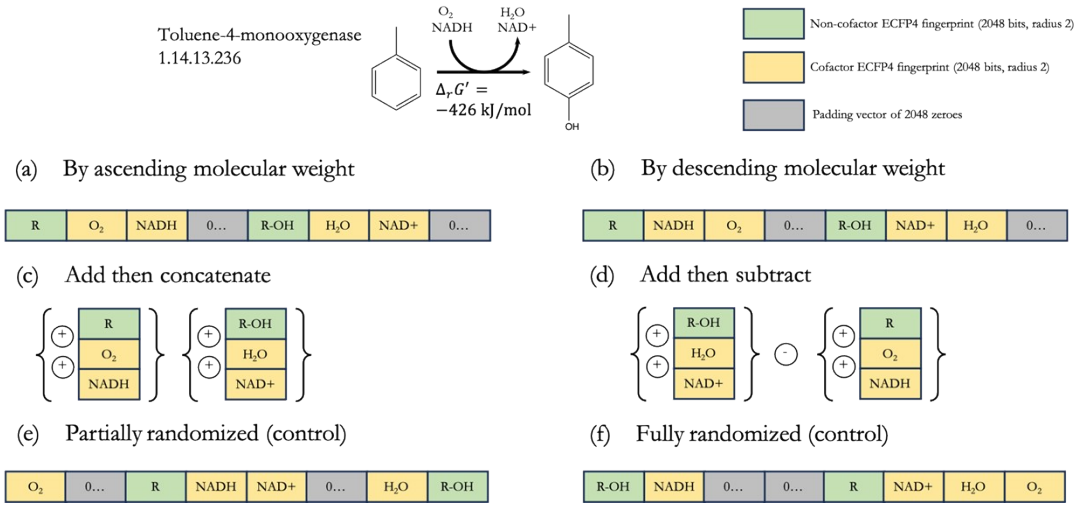
SI Fig. 7 Distribution of $^{\Delta_r G_{min}}$ values in both directions for glucosyltransferases (EC 2.4.1.x)



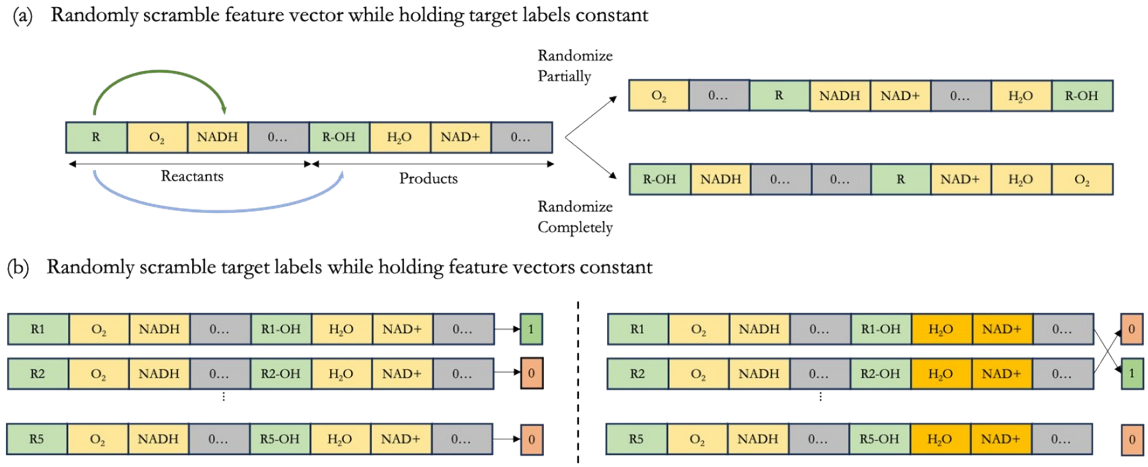
SI Fig. 8 (a) t-stochastic neighbors' estimation (t-SNE) of thermodynamically feasible, thermodynamically infeasible, and synthetically generated negative products that fall under the alcohol dehydrogenase transformation; **(b)** number of all thermodynamically infeasible reactions versus thermodynamically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



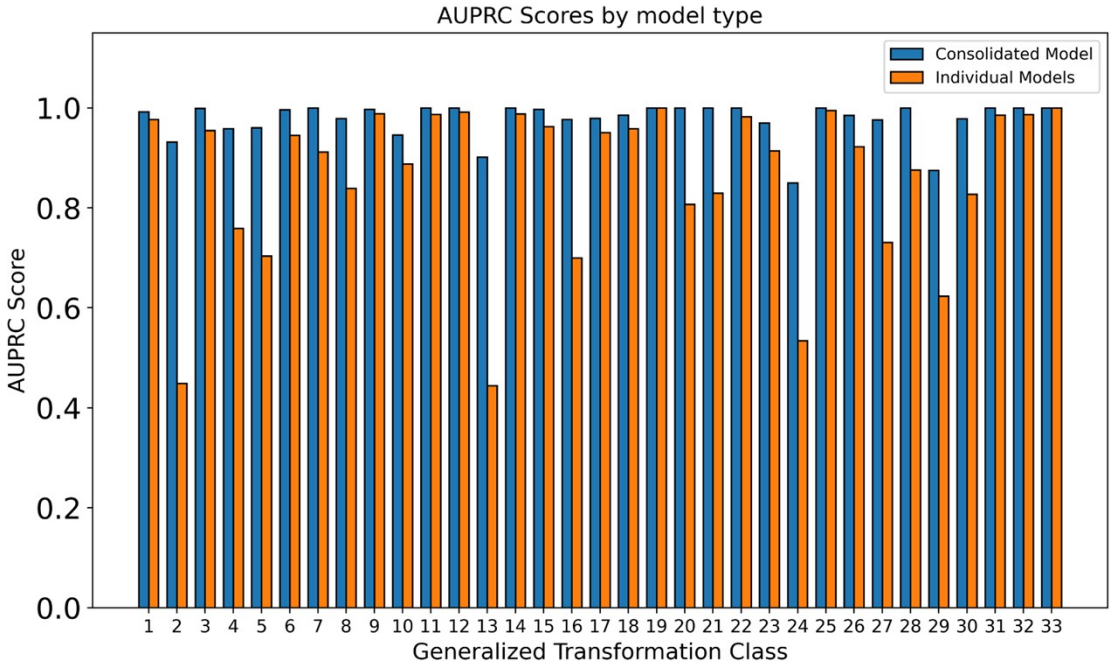
SI Fig. 9 (a) The average area under the precision-recall curve (AUPRC) of alcohol dehydrogenase classifiers deployed on a validation set and **(b)** test set of alcohol dehydrogenase reactions. With 1254 feasible and 1759 infeasible alcohol dehydrogenase reactions in total, a stratified train/validation/test 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.



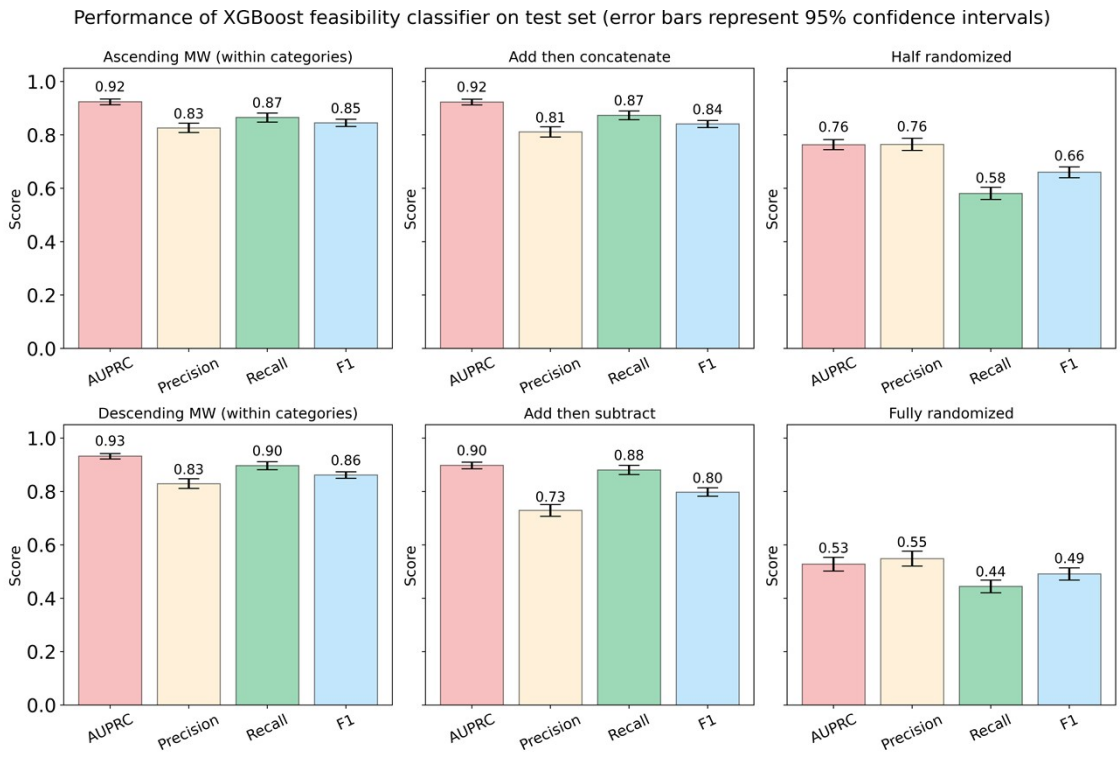
SI Fig. 10 Six different configurations for arranging molecular fingerprints along a reaction feature vector are explored in this study and depicted here through the example of the monooxygenation of toluene catalyzed by the enzyme toluene-4-monooxygenase (EC 1.14.13.236). In configurations **(a)** and **(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.



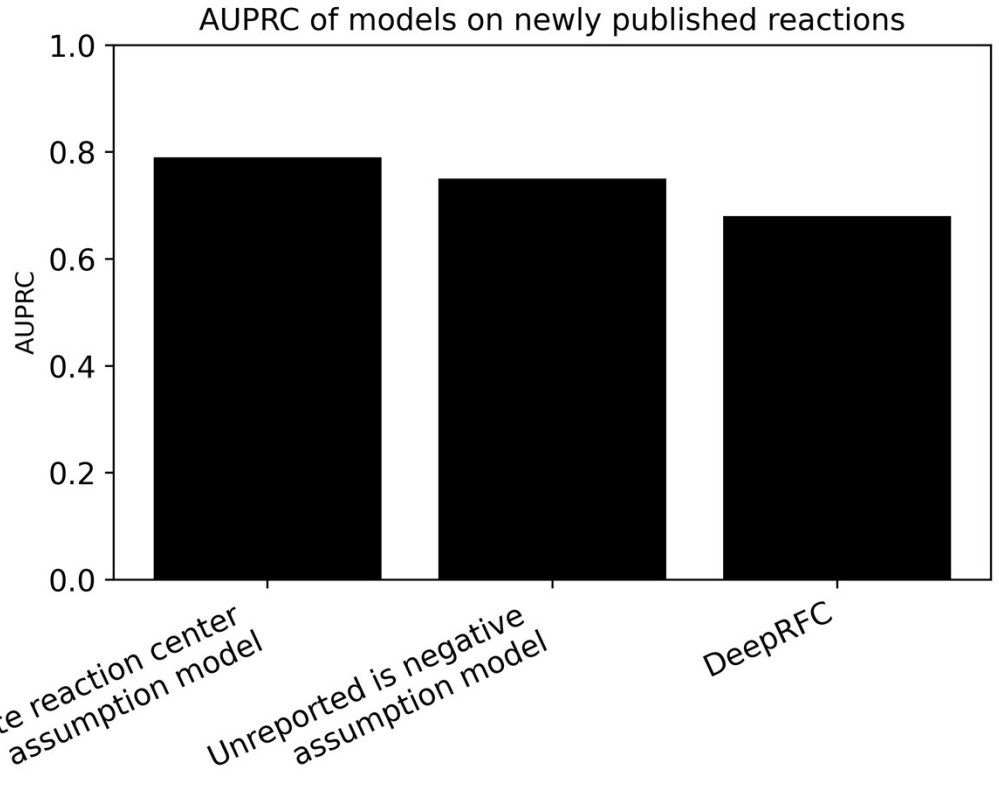
SI Fig. 11 In order to confirm that the performance of trained feasibility classification models is not merely by chance, we performed two types of negative control experiments where we expected model performance to decline: **(a)** in the first type of negative control, molecular fingerprints arranged 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.



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.



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.



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.

A

		Actual	
		Positive	Negative
Predicted	Positive	28	0
	Negative	2	0

B

		Actual	
		Positive	Negative
Predicted	Positive	25	0
	Negative	15	0

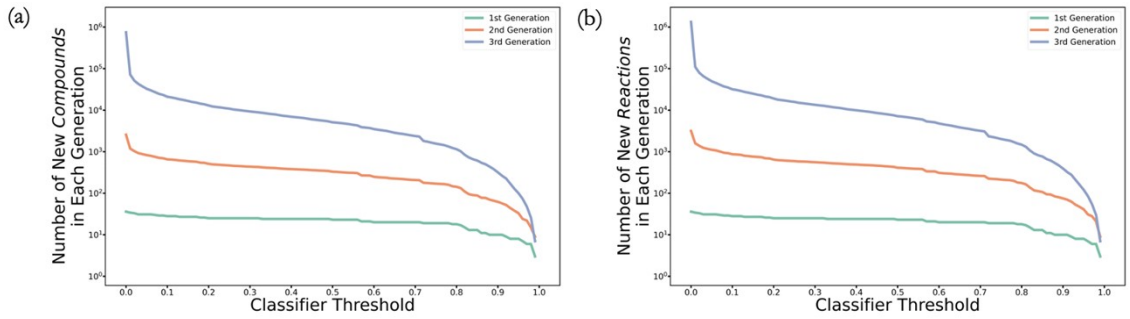
C

		Actual	
		Positive	Negative
Predicted	Positive	0	1581
	Negative	0	2955

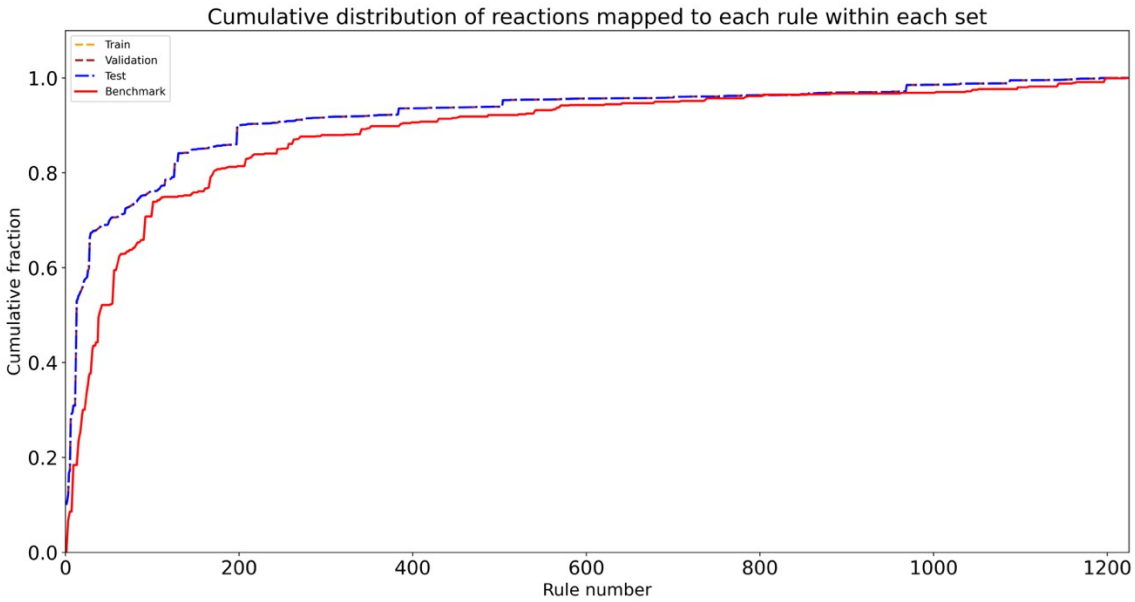
D

		Actual	
		Positive	Negative
Predicted	Positive	0	4526
	Negative	0	12827

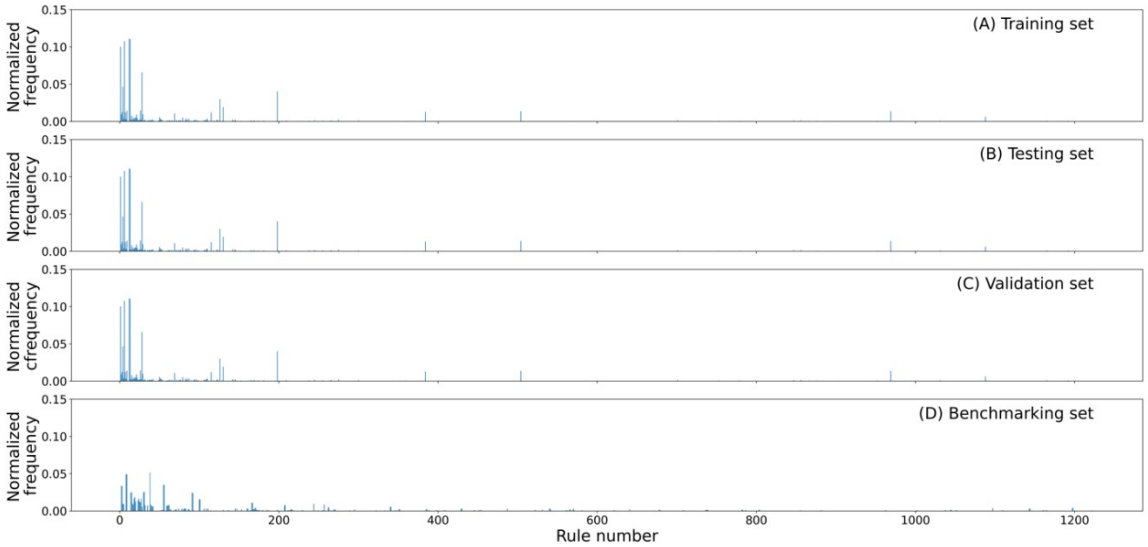
SI Table 1. **(a)** Recovery of novel, experimentally validated reactions obtained from an *E. coli* nontargeted metabolomics dataset; **(b)** recovery of predicted, novel *e. coli* reactions obtained from the same dataset; **(c)** prediction of 4536 total plausibly negative reactions that were synthetically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



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.



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.



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