Molecular Transformer-aided Biocatalysed Synthesis Planning

Daniel Probst,* Matteo Manica, Yves Gaëtan Nana Teukam, Alessandro Castrogiovanni, Federico Paratore, and Teodoro Laino

IBM Research Europe, CH-8803 Rüschlikon, Switzerland

E-mail: dpr@zurich.ibm.com

Abstract

Enzyme catalysts are an integral part of green chemistry strategies towards a more sustainable and resource-efficient chemical synthesis. However, the use of enzymes on unreported substrates and their specific stereo- and regioselectivity are domain-specific knowledge factors that require decades of field experience to master. This makes the retrosynthesis of given targets with biocatalysed reactions a significant challenge. Here, we use the molecular transformer architecture to capture the latent knowledge about enzymatic activity from a large data set of publicly available biochemical reactions, extending forward reaction and retrosynthetic pathway prediction to the domain of biocatalysis. We introduce the use of a class token based on the EC classification scheme that allows to capture catalysis patterns among different enzymes belonging to the same hierarchical families. The forward prediction model achieves an accuracy of 49.6% and 62.7%, top-1 and top-5 respectively, while the single-step retrosynthetic model shows a round-trip accuracy of 39.6% and 42.6%, top-1 and top-10 respectively. Trained models and curated data are made publicly available with the hope of promoting enzymatic catalysis and making green chemistry more accessible through the use of digital technologies.

Introduction

Chemistry fostered the unprecedented rise of overall human wealth and well-being during the past two centuries, and it is today our trump card to avert and mitigate global crisis while reshaping our lives towards a more responsible use of natural resources. Innovation in synthetic chemistry will be critical to make chemical processes and products more sustainable, resource-efficient and CO2-neutral. The design and development of catalysts is identified as the heart of greening chemistry. However, biocatalysis together with chemoinformatics and artificial intelligence have the power to accelerate the adoption of existing sustainable catalytic processes already today.

At the core of biocatalysis are enzymes, an integral part of all living organisms used in important industrial processes thanks to the multiple key advantages provided over conventional chemical reagents. In addition to their extremely high catalytic activity, enzymes catalyse chemo-, regio-, and stereo-selective reactions and are both reusable and allow for an easy recovery of products when immobilised.[1] Enzyme-catalysed reactions are usually performed in water under mild conditions and significantly reduce waste. Moreover, enzymes themselves are fully degradable in the environment, and as such, they represent an important strategy towards greener chemistry.[2] These advantages provide a large number of opportunities documented by the increased corpus of scientific and patents publications related to enzymes, as well as their increased use in industrial applications.[3,4] It is no surprise if enzymes are one of the key enablers of sustainable chemical processes, with a growing interest in their use at an industrial scale to convert waste into valuable raw materials.[5] Although the ability to use enzymatic reactions to catalyse organic synthesis of chemical compounds gained widespread attention for large scale production,[6–8] enzymes are still far from being widely adopted in daily synthetic laboratory works. The narrow substrate scope available from enzymatic databases, the difficulty in identifying patterns within enzymes classes that would extend the range of applicability to unreported substrates, and the specific stereo- and/or regioselectivity are domain-specific knowledge factors that make the adoption of en-

zymatic processes a challenging problem for synthetic chemists.[9]

The lack of a qualitative rationale for the latent relationships between enzymes and substrates makes the construction of retrosynthetic routes using biocatalysed a hard problem. The knowledge gap between large corpora of enzymatic chemical reactions data and the human understanding of the structure-activity relationship hinders the possibility to predict successful routes[10–12] when the substrates of interests are not directly associated with an enzyme. In the last years, the use of machine learning and data-driven approaches proved to be a very effective way to capture patterns from complex chemistry knowledge collections.[13] The extraction of chemical reaction rules from large data sets of traditional organic chemistry reactions[14] is one of the most successful examples[15] of providing transparency and explainability with AI applications in chemistry.

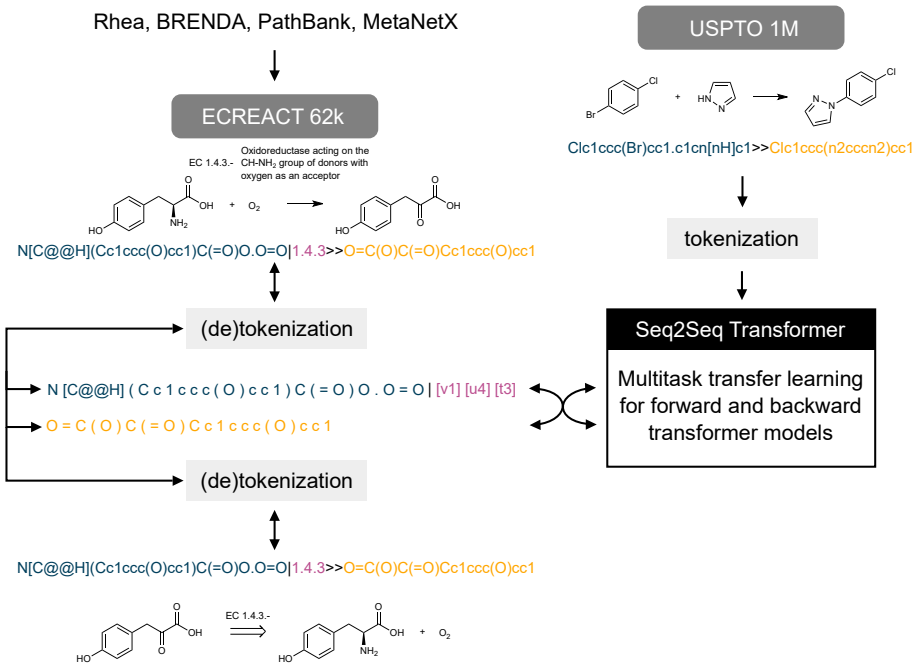

Figure 1: Introducing enzymes as green catalysts to data-driven template-free chemical synthesis. The molecular transformer was trained on chemical reactions extracted from the USPTO data set and the new *ECREACT* data set using multitask transfer learning.

While traditional synthetic organic chemistry went through its renaissance period thanks to recent development in machine learning and the availability of public chemical reaction datasets, the impact in biochemistry remained mostly bounded to the context of metabolic

pathways prediction.[16,17] Computer-aided synthesis planning tools using biocatalytic reactions are still in their infancy compared to the same developments for traditional synthetic organic chemistry. RetroBioCat,[18] one of the pioneering approaches, uses a set of expertly encoded reaction rules and, when available, a system for retrieving enzymes records with the correct substrate specificity from a database. Like the original efforts in traditional synthetic chemistry, this approach suffers from the human curation process: the lack of scalability in processing the large amount of data collected yearly and the failure to capture the effects of long-range substituents in reaction rules. Nevertheless, RetroBioCat has the merit of having pioneered the first chemoinformatic approach for easing the adoption of biocatalytic reactions in chemical reaction tasks. More recently, Kreutter et al.[19] presented a forward reaction prediction model based on the Molecular Transformer.[20] This approach exploits a multitask transfer learning to train a Molecular Transformer architecture, originally trained with chemical reactions from the US Patent Office (USPTO) data set, with 32,000 enzymatic transformations, each one annotated with the corresponding enzyme name. The enzymatic transformer model predicts the products formed from a given substrate and enzyme in the forward prediction task reaching an accuracy of 54% when using the enzyme name information only and 62% when using the complete enzyme information as full sentence (often including also the organism name). The approach solves some of the concerns around scalability and data curation of reaction templates, but the use of enzyme names as reaction tokens adds an additional level of challenge when trying to learnchemical reactivity patterns among enzymes with different names but belonging to closely related families.

Here, we generalise the use of the Molecular Transformer by adopting a tokenisation system based on enzyme classes and introducing an extension of the retrosynthetic algorithm by Schwaller et al.[21] to biocatalysis. The use of a backward model allows to predict substrates and catalysing enzyme classes given a target product. Unlike previous work, we incorporate the EC (enzyme commission) number into the reaction SMILES, rather than encoding enzymes with their natural language name. Enzymatic reactions and the accompa-

nying EC numbers were extracted from four databases, namely Rhea, BRENDA, PathBank, and MetaNetX and merged into a new data set, named *ECREACT*, containing enzyme-catalysed reactions with the respective EC number as shown in Figure 1. The resulting data set contains more than 62,000 unique enzymatic reactions. The forward and backward (substrate + EC → product and product → substrate + EC, respectively) models were then trained using multitask transfer learning on the *ECREACT* data set. The baseline model was trained on the USPTO data set (see Data Sets), containing 1 million reactions without enzymatic information, acting as a training set for learning the general knowledge on chemical reactions and the SMILES grammar.

The forward prediction model achieves an accuracy of 49.6% and 62.7%, top-1 and top-5 respectively, while the single-step retrosynthetic model shows a round-trip accuracy of 39.6% and 42.6%, top1 and top-10 respectively. An extensive analysis of the data set pinpoints the performance to the number of training samples available for each enzyme class, with the forward prediction model ranging from an accuracy of 18.6% and round-trip accuracy of 1.7% in isomerases to a forward prediction accuracy of 64.4% and a round-trip prediction accuracy of 60.5% in transferases.

Results and Discussion

Data Sets and Models Training

Enzymatic reactions with related EC (enzyme commission) numbers were extracted from Rhea,[22] BRENDA,[23] PathBank,[24] and MetaNetX.[25] The resulting collection of data sets, named *ECREACT*, contains 5 token schemes for each enzymatic reaction with different levels of enzyme information: *EC0* (no enzyme information); *EC1* (EC-level 1, corresponding to the enzyme class); *EC2* (EC-levels 1-2); *EC3* (EC-levels 1-3); and *EC4* (EC-levels 1-4). Given the low specificity of enzyme information in *EC1* and *EC2* tokens, and the insufficient sampling for *EC4*, which is often confined to one enzyme-substrate example only,

the *EC3* data set remains the only one containing sufficient variability in terms of enzyme-substrate examples across individual tokens. Figure 2 shows the composition of the data set with token scheme *EC3*, containing 62,403 unique enzymatic reactions. At EC-level 1, which corresponds to enzyme classes, EC 2.x.x.x (transferases) account for 53.5% of total entries, EC 1.x.x.x (oxidoreductases) for 24.5%, EC 3.x.x.x (hydrolases) for 10.7%, EC 4.x.x.x (lysases) for 6.3%, EC 6.x.x.x (ligases) for 2.3%, EC 5.x.x.x (isomerases) for 2.2%, and EC 7.x.x.x (translocases) for 0.4%. The high fraction of transferase-catalysed reactions is a consequence of the large number of non-primary lipid pathways stored in PathBank. Among transferases, the most represented subclasses at EC-level 2 are EC 2.7.x.x (transferases transferring phosphorus-containing groups) at 24.5%, EC 2.3.x.x (acetyltransferases) at 16.8% and EC 2.1.x.x (transferases transferring one-carbon groups) 7.5%. The complete information on the distribution of samples across EC-levels 2 and 3 is provided in the supplementary information (Tables S3 and S4, with a breakdown of the data set by data source shown in Figure S2).

This distribution of the available data reveals a heavy imbalance in the distribution of the enzyme-substrate examples. Whereas transferase-catalyzed reactions encompass few subclasses at EC-level 3 with a large sample size, the oxireductase- and hydrolase-catalyzed reactions are divided into many subclasses at EC-level 3 with a small sample size. Although lyases, isomerases, ligases, and translocases are split into fewer subclasses at EC-level 3, most of them contain very few samples. Therefore, the evaluation of the performance of the data-driven models will need to consider the different population of each EC-level 3 subclass for a proper assessment.

A further property of interest regarding the reaction is the distribution of reactants and products within and across the enzyme classes at EC-level 1. The data set created with the *EC3* token scheme contains 141,051 (56,017 unique) reactants and 62,403 (53,658 unique) products. In Figure 3, we show the distribution of the compounds in the substrate and product chemical spaces using the 2048-dimensional binary MAP4 fingerprints and embed-

Figure 2: The distribution of samples at EC-levels 1 (corresponding to enzyme classes) and 2, as well as EC-levels 2 and 3 of oxidoreductases (class 1), transferases (class 2), hydrolases (class 3), lysases (class 4), isomerases (class 5), ligases (class 6), and translocases (class 7), in the *ECREACT EC3* data set.

ding them using TMAP.[26,27] The data points, coloured by enzyme class corresponding to EC-level 1, highlight the different distributions within the substrate set (containing cofactors) and the product set (where co-enzymes and common byproducts have been removed). Substrates and products of transferase- (class 2), lyase- (class 4), and to a lesser degree, hydrolase-catalysed (class 3) reactions populate regions of the chemical space peculiar to

each class (homogeneous), with little overlap with other classes. The chemical space covered by the molecules belonging to the remaining classes is non-specific (heterogeneous), with wide areas shared among different classes. The location of the substrates and products in homogeneous regions acts as an implicit feature, reducing the importance of the EC number token (explicit feature) during training. Instead, the lack of implicit features in substrates and products belonging to heterogeneous regions requires the use of explicit tokens (EC numbers) during training to learn the chemical transformation rules.

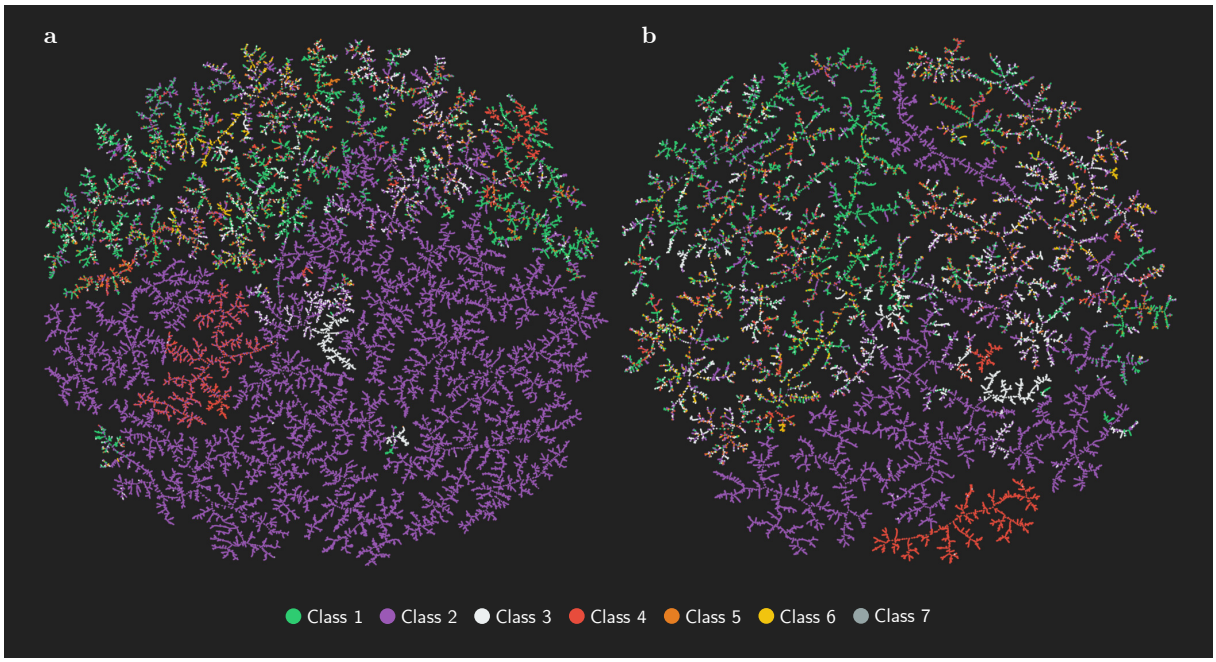

Figure 3: TMAPS visualising the distribution of MAP4-encoded (**a**) reactants and (**b**) products in the *ECREACT EC3* subset coloured by enzyme class corresponding to EC-level 1. Distributions of molecular distances (MAP4) per class are shown in Figure S1. While molecules associated with hydrolase- and lyase-catalysed reactions (class 3 and class 4, respectively) follow a similar pattern of populating homogeneous regions, molecules involved with other classes are found in predominantly heterogeneous regions.

A forward and backward model was trained for each of the *ECREACT* token schemes, with the *EC0* acting as a control on the influence of including enzymatic information in the reaction. The trained models were evaluated for forward, backward, round-trip, and EC number prediction accuracy using 5% test splits with the condition that a product in the test split must not occur as a product in the training split (Figure 4). The results show that the

EC3 token scheme has better performance than *EC0* and *EC2*, yet performs slightly worse than *EC1* and *EC4* (Figure 4a). In the backward prediction task, *EC3* performs slightly worse than *EC0*, *EC1*, and *EC2* , but significantly better than *EC4*; this is most likely due to the low number of samples in each *EC4* category (Figure 4b, Table S5). Regarding the round-trip accuracy,[21] *EC3* performs better than both *EC2* and *EC4* (Figure 4c). When solely focusing on the prediction of the correct EC number in the backward prediction, the models perform better the less detailed information they have to predict (Figure 4d). These data show that the inclusion of enzymatic information in the form of the EC number does not affect the prediction performance negatively as long as each EC category has a sufficient number of training samples, which restricts the use of *EC4* (Figure S10). The *EC1* token, although performing well across different metrics, averages across reaction classes with different schemes and for this reason is of little interest for retrosynthetic purposes. The *EC3* token scheme balances the specificity of enzyme information with performance compared to the other *ECREACT* token schemes, resulting in prediction performance similar or better to *EC1* and *EC2* while retaining detailed information of the reaction-specific enzyme. Therefore, the relative performance among the five *ECREACT* token schemes *EC0*, *EC1*, *EC2*, *EC3*, and *EC4*, identifies *EC3* as the one with the most rich amount of statistically significant information.

Forward Prediction

We split the *EC3* data into a test and a training set, enforcing a zero overlap between product molecules distribution in the two ensembles, i.e. each product molecule present in the test set does not appear in the training set. This is a strict requirement that reduces to zero the chance that forward and, most important, backward predictions are affected by memorization of reaction records rather than by learning enzyme-substrate patterns. The zero product overlap between tests and training set penalizes the measure of the performance of the forward model when compared to a random splitting.[20,21] Despite the various

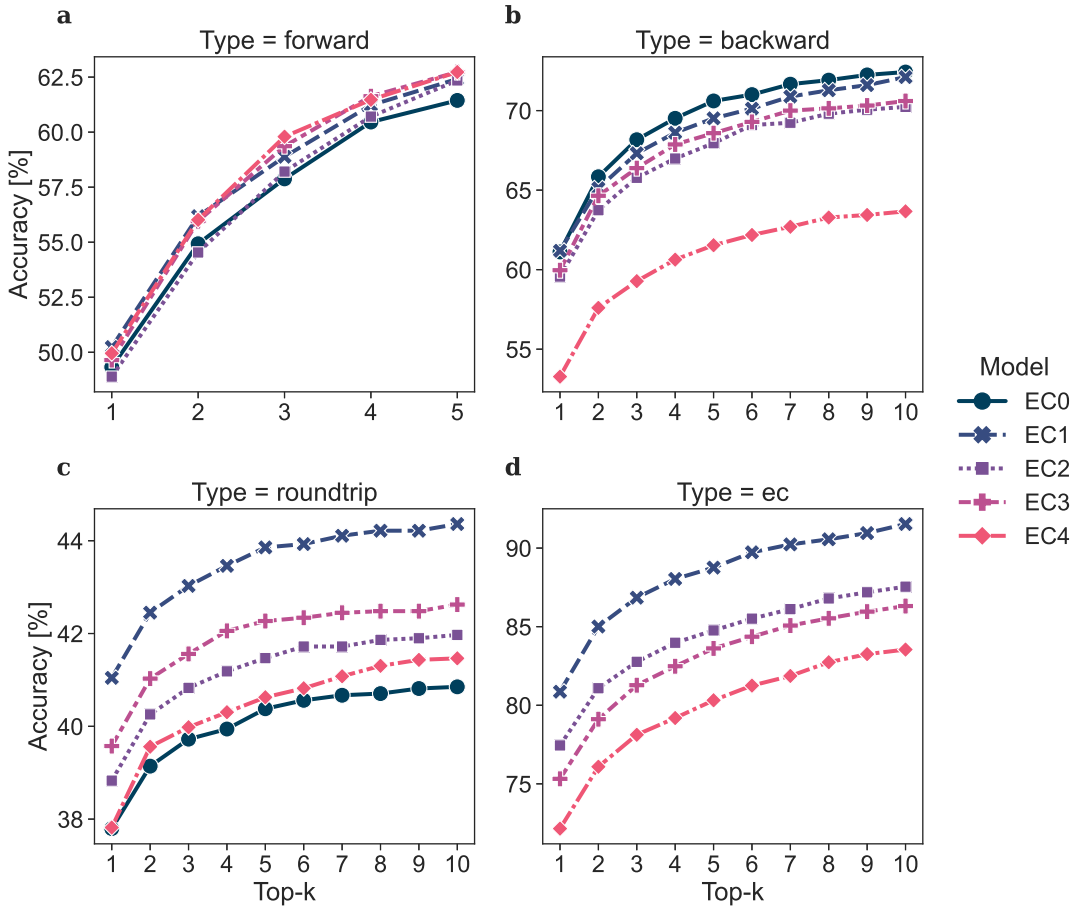

Figure 4: Overall accuracy of models based on different *ECREACT* token schemes (*EC0*, *EC1*, *EC2*, *EC3*, and *EC4*) for (**a**) forward prediction, (**b**) backward prediction, (**c**) round-trip prediction (a forward prediction followed by a backward prediction), and (**d**) backward EC number only prediction. Top-n indicates the accuracy when checking the top n predictions for the correct one.

similarities between the use of EC number and the use of catalysts' token in a chemical reaction, we decided to assess the impact on learning of including EC number tokens in chemical reaction representation. We randomized the EC numbers in the test set within and across classes (corresponding to EC-level 1) and measured the performance of the forward prediction models in different scenarios. The resulting overall accuracy for evaluation tests in which the EC tokens were not randomized, randomized within the same class, and randomized across different classes was 49.6%, 41.3%, and 38.3%, respectively. Although, the inclusion of EC numbers may seem of limited benefits, a more detailed analysis point

to the real gains (Table 1 and Figures 5a, S4a, and S5a). In fact, the value of including the EC number becomes apparent upon grouping the test samples by class and linking each to their own sample size (Figures 5b-d, S4b-d, and S5b-d). Tests belonging to EC-level 3 subclasses containing a large number of samples keep performing relatively well even with incorrect EC number. This reflects the data set imbalance identified for oxidoreductase- (class 1) and transferase- (class 2), in which the larger set size or the homogeneity of the chemical space covered by substrates/products make the presence of the EC number non-essential to determine the outcome of the chemical transformation. Instead, the accuracy among small and medium-sized classes drops, inversely correlating with the number of test and training samples for each EC-level 3 category over all classes (Figure S9). These results suggest the model can successfully predict the reaction outcome without relying on the EC number when the sample size of specific enzyme-catalyzed reactions is large enough or the space mapped by the substrate/product molecules is specific of the enzymatic reaction (homogeneous). The performance on the ligases (class 6) shows a marginal increase from 32.3% to 33.9% when EC numbers are randomized within the same class and drops to 8.1% when EC numbers are randomized across different classes, suggesting that the attention is focused on the class-level of the EC number. As a general trend over all classes and experiments, the accuracy increases increasing the number of predictions to match (top-k, $k \in \{1, 2, 3, 4, 5\}$), with the biggest effect between $k = 1$ and $k = 2$.

Table 1: Forward model accuracies with non-randomized and randomized EC numbers.

Class	Top-1 Accuracy [%]		
	Non-Randomized	Randomized within Class	Randomized
Oxidoreductases (1)	28.0	18.5	18.6
Transferases (2)	64.4	55.8	54.8
Hydrolases (3)	39.7	32.6	18.0
Lysases (4)	28.8	22.0	16.9
Isomerases (5)	18.6	8.5	0.0
Ligases (6)	32.3	33.9	8.1
Translocases (7)	100.0	100.0	100.0
Overall	49.6	41.3	38.3

Figure 6 and Figure 7 show a selection of few successful and unsuccessful predictions extracted from the test data set. The set covers reactions catalysed by an oxidoreductase (1), two transferases (2, 3), two hydrolases (4, 5), a lyase (6), an isomerase (7), and a

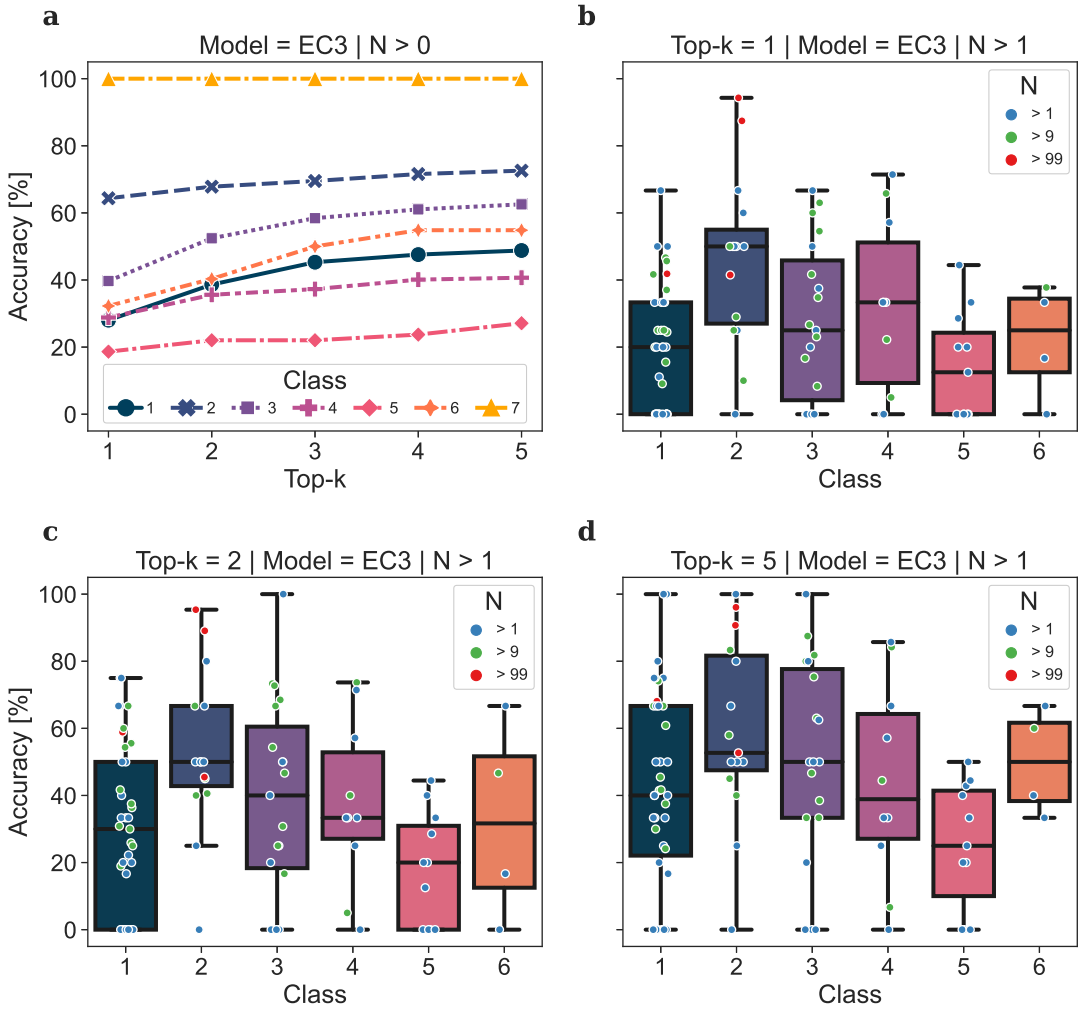

Figure 5: Class-wise accuracy for the forward model trained on *EC3*. (**a**) The top-k prediction accuracies for each class show significant differences among classes caused by the number of available samples per EC-level 3 category. The accuracy of (**b**) top-1, (**c**) top-2, and (**d**) top-5 predictions per EC-level 3 category. Each dot represents an EC-level 3 subclass coloured by the number of test samples N. Large EC-level 3 subclasses (red) greatly influence the performance of predicting transferase-catalyzed reaction (class 2) outcomes. Oxidoreductase-catalyzed reactions (class 1) are distributed among many EC-level 3 subclasses, causing a lower performance compared to other classes with fewer samples overall.

ligase (8). These successful examples reflect the models' capability to predict enzymatic reaction outcomes across all enzyme classes. Instead, the analysis of the incorrectly predicted reaction outcomes (Figure 7) highlights peculiar patterns. Reactions (1) and (2) are both catalysed by an oxidoreductase acting on the CH-NH_2 group of donors. The predicted

reaction (1) contains an excessive number of carbon atoms. The inferred product of reaction (2) is equivalent to the ground truth, as the linear and cyclic forms are in equilibrium. (3) shows an example of the model correcting an error in the data set and predicting the correct stereochemistry. In addition, it highlights the possibility of false negatives due to the prediction of zwitterions. The products of the phosphoric diester hydrolase-catalysed reaction (4) and the intramolecular lyase (5) are predicted incorrectly because the training set contains an enzymatic reaction with identical substrate and, on an EC-level 1-3, identical EC number. (6) is an example of the model failing to predict the correct stereochemistry of a product. Prediction of correct stereochemistry has been reported by Schwaller et al.[20] as a major challenge for the molecular transformer and is linked to the lack of coherent stereochemical information in the USPTO data set.[28] Similarly, the correct stereochemical prediction, affected by the limited data coverage on stereochemical examples, remains the major challenge for the model, especially when predicting reactions catalysed by isomerases (class 5, Figure 5a); the removal of all stereochemical information from the predicted products increases the accuracy of isomerase-catalysed reaction prediction by a factor of two (Figure S3).

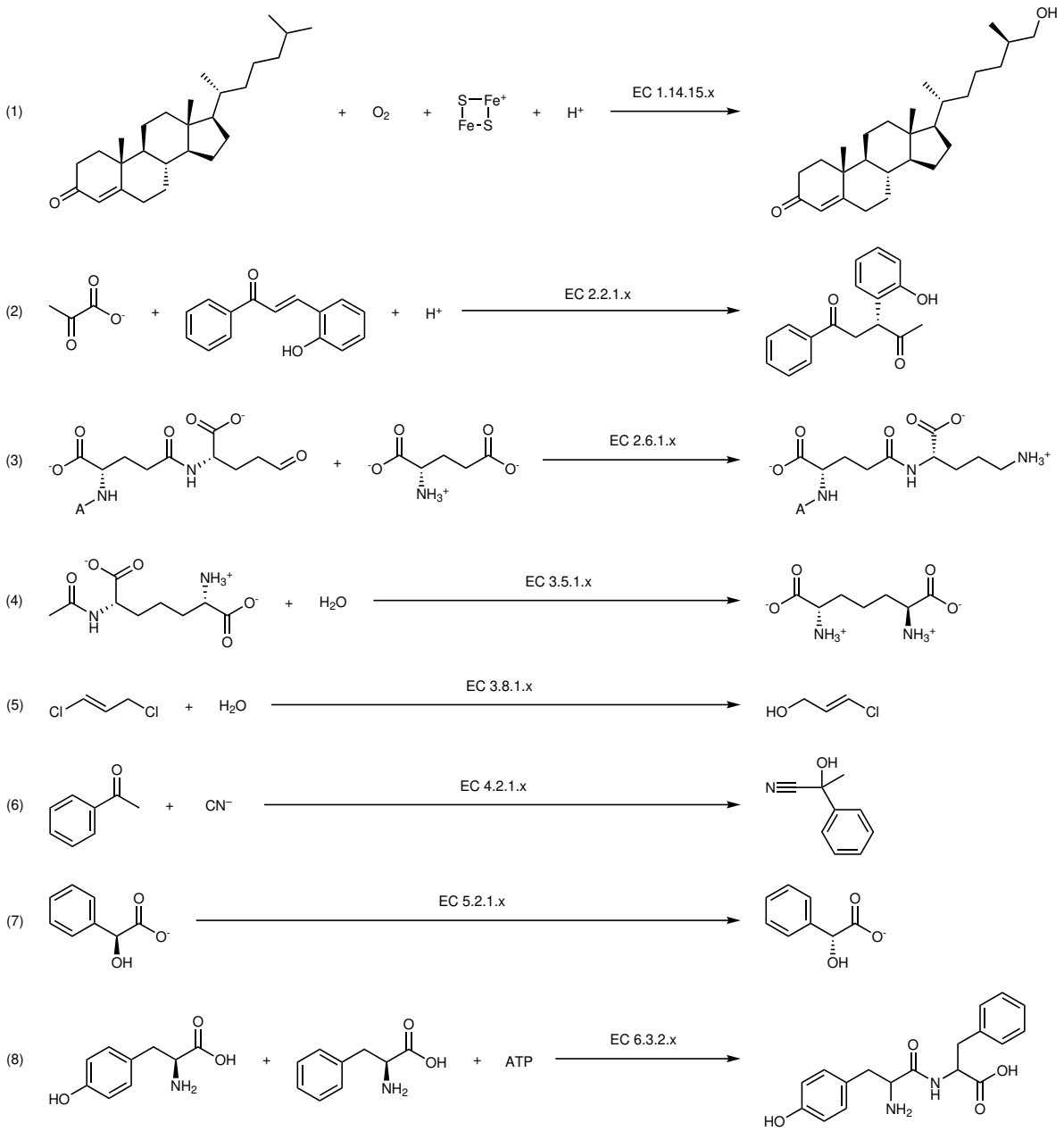

Figure 6: Examples of successful forward predictions. The reactions are catalysed by (1) an oxydoreductase with reduced iron-sulfur protein as one donor, and incorporation of one atom of oxygen, (2) aldehyde transferase, (3) an acetylornithine transaminase, (4) a *N*-acetyldiaminopimelate deacetylase, (5) a haloalkane dehalogenase, (6) an (*R*)-mandelonitrile lyase, (7) a mandelate racemase, and (8) an L-alanine-L-anticapsin ligase.

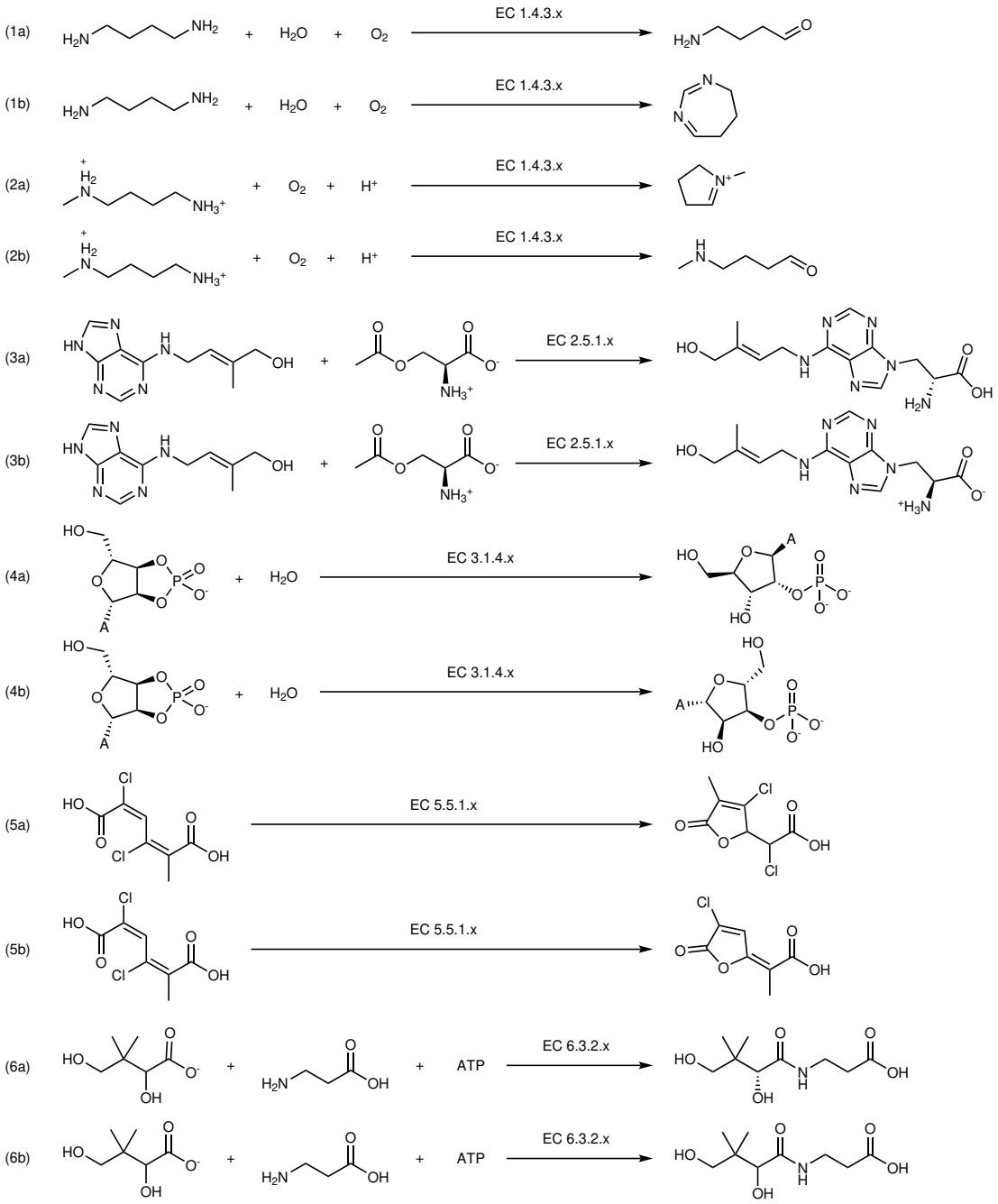

Figure 7: Incorrect forward predictions. For each reaction, (a) is the ground truth while (b) is the prediction. The reactions are catalysed by (1, 2) oxidoreductases acting on the CH-NH_2 group of donors with oxygen as acceptor, (3) a zeatin 9-aminocarboxyethyltransferase, (4) a cyclic-CMP phosphodiesterase, (5) a chloromuconate cycloisomerase, (6) and a pantothenate synthetase.

Backward Prediction

The assessment of the backward model for any given target molecule requires the use of the round-trip accuracy[21] to evaluate the correctness of possible disconnections differing from the one reported in the test data set. In fact, the top-k accuracy when applied to backward model predictions comes down to assessing how good the model is memorising one specific disconnection compared to other meaningful ones. Here, the test and training sets have been constructed by enforcing a zero overlap between product molecules. As the model has not encountered a product of a reaction in the test set during training, the top-k performance can be used to assess how the model is matching a specific disconnection scheme. Figure 4 shows the backward model performance (round-trip and top-k accuracy) for the EC-level 3. With a top-1 accuracy of 60%, the backward model has a behaviour similar to the forward model in the performance between and within classes (Figure 8), as well as in the correlation between the size of the training samples and accuracy (Figure S10). In addition to the substrates, the model also predicts the enzyme EC-level 3 token; the accuracy of predicting EC numbers only is shown in Figure S6. The analysis of the model shows an exceptional performance on transferase-catalysed reactions (class 2), traceable to two large EC-level 3 subclasses EC 2.3.1.x, and EC 2.7.8.x., which contain 17%, and 20% of all available samples, respectively (Figure 8b-d). This analysis further explains the comparatively low prediction accuracy on the class of oxidoreductases (class 1) as it contains a large number of EC-level 3 classes, each small in size (Figure 8, Table S4). Translocases are involved in catalysing the movement of molecules or ions across membranes. This specific function, together with the limited set of reaction records (191), causes the substrates and products to have lower diversity than in other classes. The constraint preventing product molecules from being present both in the training and learning data set reduces further the variability of the population in the translocases data sets. Because of the limited data, there is no statistical significance for the class of translocases, and we have thus opted to discard this class from a detailed analysis.

The confusion matrix (Figure 9) provides further insight into the backward prediction

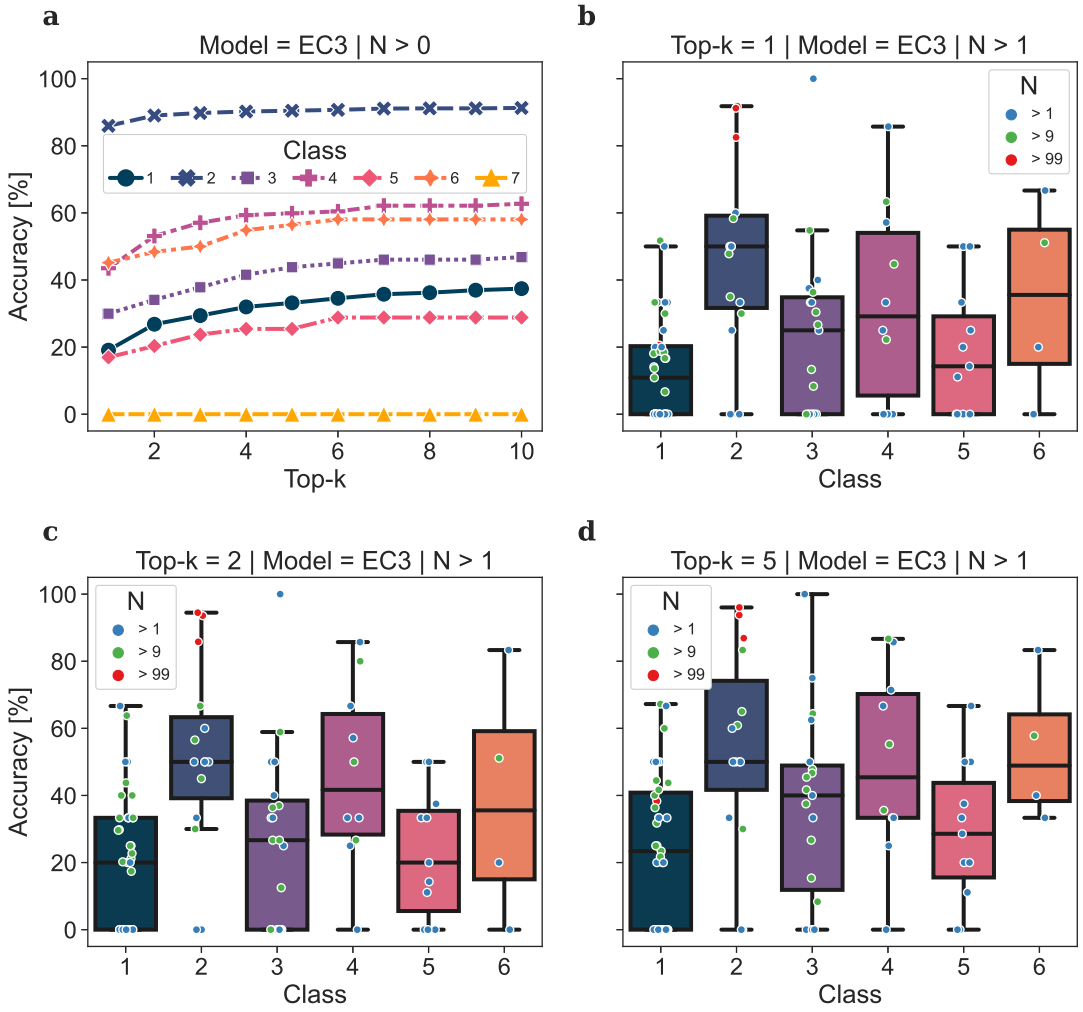

Figure 8: Class-wise accuracy for the backward model trained on EC3. (**a**) The top-k prediction accuracies for each class (corresponding to EC-level 1) show significant differences among classes caused by the number of available samples per EC-level 3 category. The accuracy of (**b**) top-1, (**c**) top-2, and (**d**) top-5 predictions per EC-level 3 category. Each dot represents an EC-level 3 category coloured by the number of test samples N. Large EC-level 3 subclasses (red) greatly influence the performance of predicting transferase-catalyzed reaction (class 2) outcomes. Oxidoreductase-catalyzed reactions (class 1) are distributed among many EC-level 3 subclasses, causing a lower performance compared to other classes with fewer samples overall.

performance. The model's ability to assign a product to the correct enzymatic class differs significantly between classes and is again influenced by the cohort of each class. Despite the larger population of the oxidoreductases, the extreme split in EC-level 3 subclasses causes the backward model to perform worse in predicting substrates for oxidoreductase-catalyzed

reactions than for hydrolase-, lyase-, and ligase-catalyzed reactions (Figure 8a). The prediction of the enzyme class shows high accuracy (71.97%) for the class of oxidoreductases. A challenge in terms of predicting the correct enzyme class are the isomerases (Class 5), as they encompass intramolecular oxidoreductases, transferases, and lyases; this is reflected in the relatively high misassignment of isomerases to oxidoreductases, transferases, and lyases (classes 1, 2, and 4).

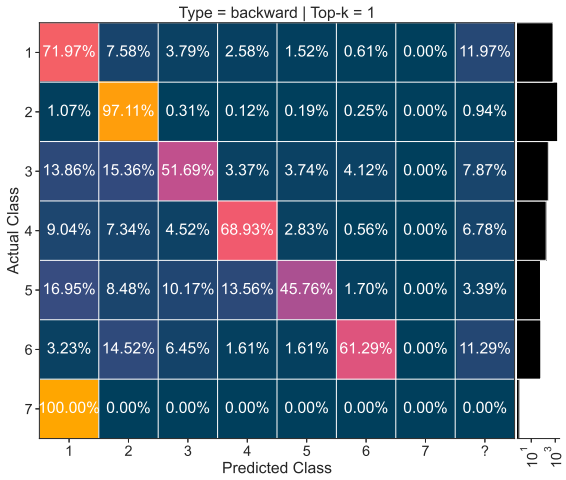

Figure 9: The confusion matrix based on predicted EC numbers by the backward model. The bars right of the plot show the number of samples per class.

In Figure 10 and Figure 11 we show successful and unsuccessful backward predictions. Among the successful examples, we report enzymes belonging to the oxidoreductase (1), a transferase (2), a hydrolase (3), two lyases (4, 5), an isomerase (6), and a ligase (7). The model did not predict any translocase-catalysed reactions because of the statistically insignificant data set. Figure 11 shows a selection of incorrectly predicted backward reactions, together with their ground truth. In the example (1), the model predicts a different EC-level 3 token to catalyse the reaction. Whereas the ground truth reaction is catalysed by an oxidoreductase acting on the CH-OH group of donors with oxygen as an acceptor (1.1.3.x), the prediction suggests the reaction to be catalysed by an oxidoreductase acting on the CH-OH group of donors with NAD^+ as an acceptor (1.1.1.x). This choice may reflect the respective number of training samples for the two classes (3,220 and 176 for 1.1.3.x and 1.1.1.x, re-

spectively) and may be considered as a viable alternative to the ground truth. Example (2) shows a correct EC-level 3 prediction (hexokinase). However, the substrate did not match the ground truth because the model predicted the acyclic rather than the linear form of aldehydo-D-galactose. (3) is an example of the model adding stereochemistry information missing in the test data set. In the ground truth, only L-tyrosine is represented by an isomeric SMILES, while PAPS (3'-Phosphoadenosine-5'-phosphosulfate) and the product are represented in their racemic form. The model predicts both L-tyrosine and PAPS with the correct stereochemistry. In (4), the model predicts an alternative way to synthesise 2-fluorobenzoate. Rather than hydrolysing a coenzyme A thioester using a thioesterase, the model suggests an aldehyde dehydrogenase acting on the -CHO group of 2-fluorobenzaldehyde with NAD^+ as an acceptor. In contrast with the ground truth, the carboxylic acid can be obtained by mild oxidation of a commercially available substrate. Finally (5), the model fails to predict an enzymatic reaction for the synthesis of 3,5-dichloro-2-methylmuconolactone and falls back to a reaction learned from the USPTO data set.

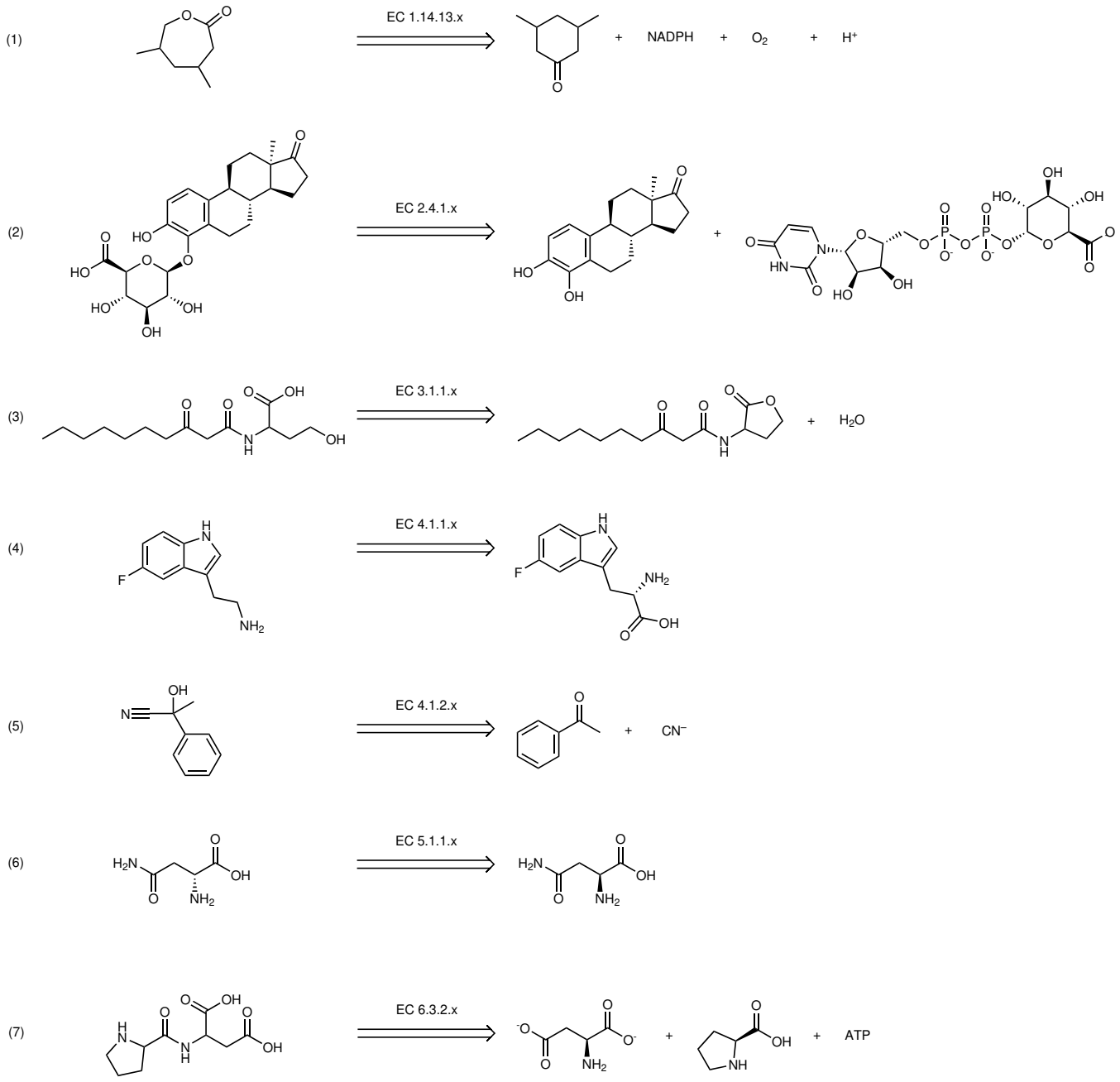

Figure 10: Successful backwards predictions. The reactions are catalyzed by (1) a cyclohexanone monooxygenase, (2) a glucuronosyltransferase, (3) a quorum-quenching *N*-acyl-homoserine lactonase, (4) an aromatic-L-amino-acid decarboxylase, (5) an aldehyde-lyase, (6) an asparagine racemase, and (7) a peptide synthase.

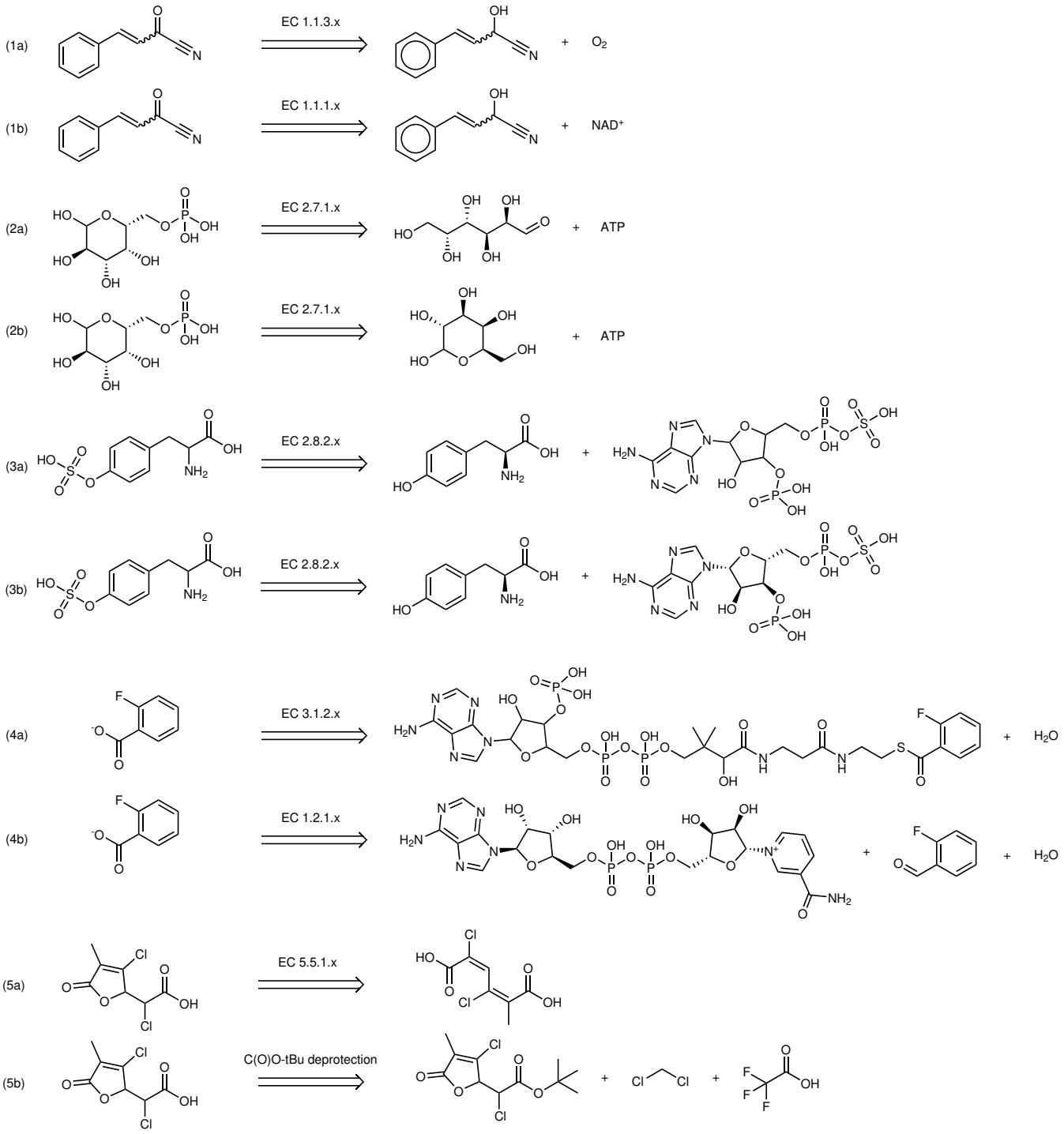

Figure 11: Incorrect backward predictions. For each reaction, (a) is the ground truth while (b) is the prediction. The model predicted (1, 4) different enzyme-catalyzed reactions leading to the same product, (2) predicted a substrate with a different isomer, (3) corrected an erroneous data set entry, and (5) was not able to predict an enzymatic reaction and fell back on a reaction learned from USPTO data.

Attention Analysis

The analysis of the patterns in the attention weights of the Molecular Transformer provides insights on the interpretability of these complex models and on potential biases.[29] In the case of reaction SMILES, attention weights have shown to uncover complex reaction information with no supervision, such as atom mappings.[14]

Here, inspired by the work of Schwaller et al.[14] we unboxed the forward prediction model to understand how it exploits enzyme information. We considered all the reactions included in our test set and inspected the attention weights in all the heads, considering the mean weights over EC tokens after discarding values lower than a threshold (details in the Methods Section). We analyzed the attention patterns across all reactions (see Figure S11) and for the most three representative enzymatic reaction classes: oxidoreductases, transferases and hydrolases (see Figure S13).

Specific heads focus their attention on the different levels of the EC token, while others attend the complete enzymatic information, attributing comparable weights to all levels of the token. On average, the heads pay more attention to the first two EC number levels and less to the third, causing the level 1 and 2 of the token to be primarily responsible for forward reaction prediction. The comparison of the mean attention for oxidoreductases, transferases and hydrolases reactions (see Figure S13) reveals that the model captures variations in enzymatic reactions, focusing on different EC number levels based on the reaction type. Figure 12 shows few representative examples of enzymatic reactions and the attention relationship between the EC token levels and the tokens of the product. In all examples, the EC tokens are related to the centre of the enzymatic reaction. Example (a) shows how level 3 of the EC token focuses on key features of the enzymatic reaction: the centre subject to nucleophilic substitution and the token related to the configurational information. Example (b) reveals the connection between the EC token and the centre of the nucleophilic addition as well as the introduced nucleophile. Finally, example (c) reveals the connection of the EC token with the stereochemical centre undergoing inversion of configuration. The analysis of

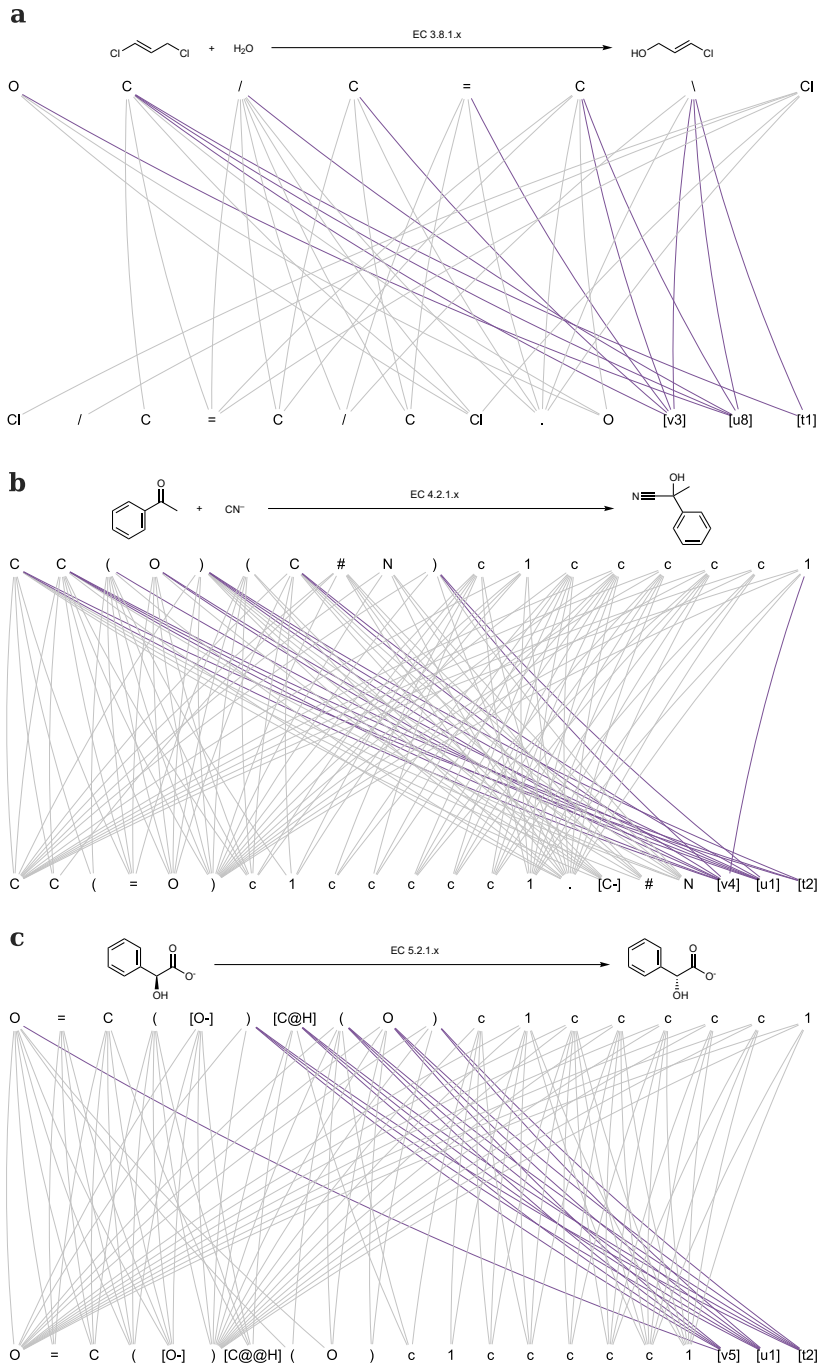

Figure 12: Analysis of the attention weights in the forward prediction models on reactions (5), (6) and (7) from Figure 6 ((a), (b) and (c) respectively). For each reaction, the attention mapping between tokens representing EC numbers is highlighted in purple (reactant atom tokens are connected using grey curves). The curve thickness is proportional to the attention weight computed by the forward Molecular Transformer.

the attention weights confirms the capacity of the forward Molecular Transformer to use the EC token for discerning the enzymatic reaction centre while capturing enzymatic reaction

rules.

Overall, oxidoreductases exhibit higher values on the enzymatic tokens compared to the others. In contrast, transferases present low values, except for head 3, where the EC number class receives in general higher weights in respect to the average. This explains why transferases data sets can be predicted with a slight loss of accuracy even when paired with wrong EC numbers. Hydrolases show more variation in attention values, with the highest weight given by head 3 to the EC-level 2. Besides these differences, head 3 always receives the highest attention values, while head 2 receives the lowest in all the reaction classes considered.

In an attempt to capture similarities in attention patterns, we extended our analysis to consider average correlations between the attention heads (see Figure S12, details on the correlation analysis can be found in the Methods Section). Attention weights for heads 3, 6 and 7 tends to focus on single tokens (i.e., atoms and EC levels) and exhibit highly significant correlation values ($\rho_{3,6} = 0.78$, $\rho_{3,7} = 0.65$, $\rho_{6,7} = 0.66$), providing the inherent mapping between tokens/atoms in the reactants and the ones in product. Heads 2 and 4, which tend to focus on the structurally larger group of tokens, e.g., representing branches, show a weakly positive correlation ($\rho_{2,4} = 0.33$). This suggests that the two heads are capturing distinct aspects of the enzymatic reactions while attending similar token lengths. The remaining heads are uncorrelated, highlighting the existence of more complex attention patterns captured by the model.

Retrosynthesis Use-Cases

The trained forward and backward models allow us to extend the approach for template-free retrosynthesis prediction by Schwaller et al. [21] to enzymatic reactions, introducing the first template-free biocatalysed synthesis planning tool (see the Methods Section for details). Here, we present the predicted pathways for a selected number of target molecules and compare them to classical organic synthesis routes. We selected the target molecules from

the RetroBioCat's curated set of biocatalyzed pathways[30] based on the intersection between chemistry coverage in our data set *ECREACT* and the data set of RetroBioCat. In fact, the encoding of *ECREACT* and the RetroBioCat test set using rxnfp[31] shows that the RetroBioCat test set reactions are forming distinct clusters in the TMAP-embedded reaction space (Figure 13a), in which the fraction of nearest neighbors from the set itself is consistently higher compared to reactions from *ECREACT* (13b). This analysis highlights the different chemistry captured by the datasets and anticipates a poor performance for those Retrobiocat examples poorly covered in the *ECREACT* data set (more details can be found in the Method Section, see Figure S14 for a depiction of the reaction classes' statistics from Finnigan[30]).

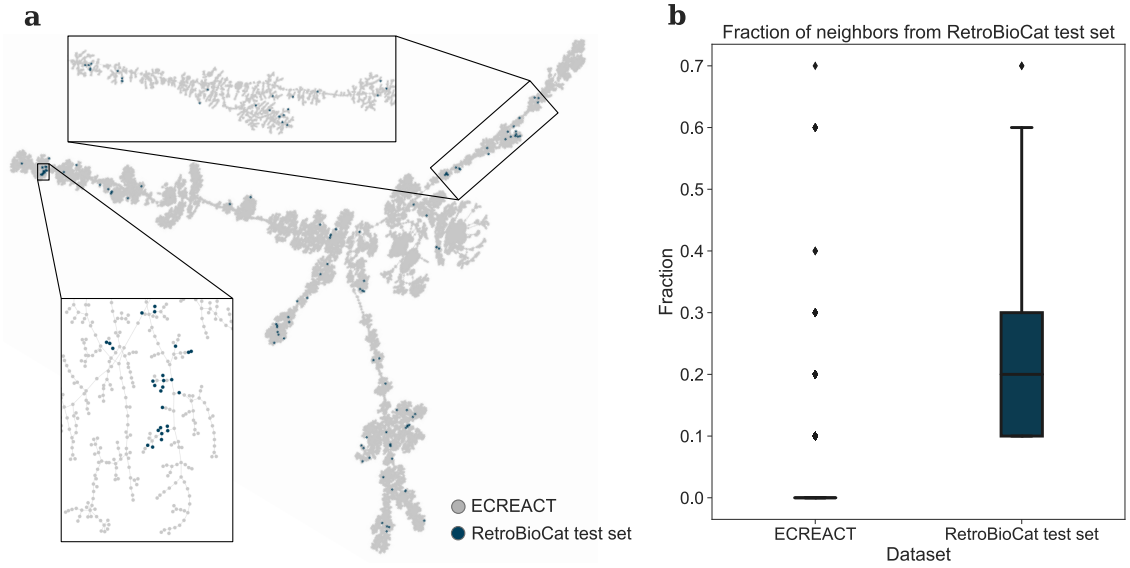

Figure 13: Distribution of rxnfp fingerprints for the reactions in the combined space of *ECREACT* (grey) and RetroBioCat test set reactions (blue), embedded with TMAP. (**a**) The reactions from the RetroBioCat test set are forming distinct clusters in the combined reaction space. (b) For RetroBioCat test set (blue) reactions, the fraction of nearest neighbors ($k = 10$) from the set itself is consistently higher compared to reactions from *ECREACT* (grey).

In Figure 14 we report the synthesis of the target molecules as recommended by the model using enzymatic transformations in mild conditions. Aminoalcohol (1) can be synthesized by regioselective transamination of the precursor dione, followed by reduction of the aminoketone with NADH as the hydride source. This approach represents an alternative to gaseous hydrogen or other solid hydride sources typically employed in the reduction of

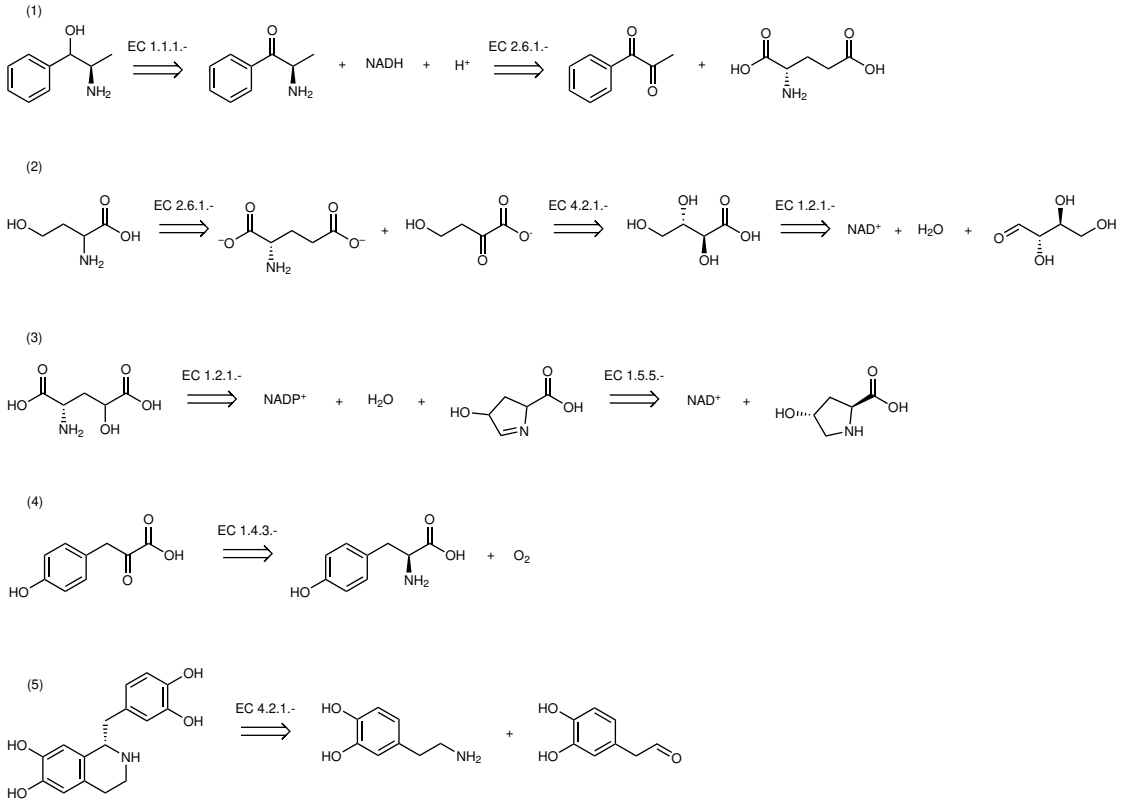

Figure 14: Enzyme-catalysed synthesis of synthetically useful compounds under mild conditions. (1) Aminoalcohol, (2) Homoaspartate, (3) 4-hydroxy-L-glutamic acid, (4) β-ketoacid, and (5) (*S*)-norlaudanosoline.

carbonyls, which can often represent a safety concern when employed already on gram-scale. Homoaspartate (2) can be accessed by a series of chemoselective enzymatic transformations of L-erythrose to the corresponding carboxylic acid, followed by regioselective dehydration to the α-ketoacid. Finally, the model infers that a transamination with glutamate on the newly-introduced keto functionality ensures the delivery of the target amino acid. Given the similar reactivity of the -OH groups within the substrate, such a series of transformations would require considerable effort to be achieved by non-enzymatic approaches.[32] As the third example, the model predicts that 4-hydroxy-L-glutamic acid (3) can be obtained from oxidation of inexpensive L-hydroxyproline in the presence of NAD^+ (catalyzed by EC 1.5.5.x), followed by a further oxidation of the aldehyde intermediate with EC 1.2.1.x and $NADP^+$. Enzymatic reactions enable oxidations to be carried out also in the presence of

O_2, as exemplified by the prediction for the synthesis of α-ketoacid (4). The chemoselective oxidation of the amino group of L-tyrosine leaves the sensitive and electron-rich aromatic moiety unaltered and obviates the use of stronger oxidizing agents. Lastly, the model predicts that an enzymatic Pictet-Spengler reaction catalyzed by EC 4.2.1.x, can convert dopamine and the corresponding aldehyde to the alkaloid (*S*)-norlaudanosoline (5) enantioselectively, which typically requires the presence of organocatalysts or transition metals.[33–35] It is interesting to compare the routes suggested by our model with the ones from RetroBioCat.[18] In reaction (1) the starting substrate is styrene, which undergoes epoxidation in presence of an epoxidase, followed by epoxide opening, partial oxidation of the primary alcohol to aldehyde and transamination. Homoaspartate (2) is instead shown to be obtained via aldol addition of sodium pyruvate on formaldehyde, followed by transamination with alanine. Similarly, RetroBioCat shows that 4-hydroxy-L-glutamic acid (3) can be prepared by treatment of pyruvic acid with glyoxylic acid in presence of an aldolase, delivering the target compound after transamination with an amino donor. Pyruvic acid is also the substrate suggested for the synthesis of α-ketoacid (4), which is delivered upon reaction with phenol in presence of a lyase. Alkaloid (*S*)-norlaudanosoline (5) is synthesized in a similar fashion as suggested by our model with a norcoclaurine synthase, which acts on the same primary amine and aldehyde substrates shown by the Molecular Transformer. With the exception of route (5), which is highly substrate-specific, one can appreciate the dissimilarity of the synthetic pathways suggested by our model when compared with RetroBioCat, which open the way to synthetically useful compounds from a variety of different inexpensive substrates.

Conclusion

We presented a molecular transformer trained on enzyme-catalysed reactions extended with EC (enzyme commission) numbers. Our results show that the Molecular Transformer performs well in predicting products based on EC number and substrates, predicting substrates and EC number based on a product, and predicting the EC number based on product. The enzymatic models reach an overall top-1 accuracy of 49.6%, 60%, and 39.6% in forward, backward and round-trip accuracy, respectively. The accuracies correlate heavily with the amount of training data in each token class, presenting a major challenge given the limited availability of data. In addition to applying the Molecular Transformer to biocatalysed reactions, we introduced an aggregated data set, *ECREACT*, containing preprocessed enzyme-catalysed reactions sourced from different publicly available databases. With the increase of the quantity and quality of available training data and the experimental validation of proposed synthetic routes, the research community will be able to build on the legacy of the present work to retrain models with higher accuracy and broader scope without the limitation of humanly curating reaction rules. Finally, we presented few use-cases based on well-understood pathways that showed how template-free machine learning models trained on enzymatic reactions can play an essential role in promoting the adoption of greener chemistry strategies in daily laboratory work.

Data and Models

The *ECREACT* data set will be made publicly available upon acceptance of this manuscript at the URL: `https://github.com/rxn4chemistry/green_cat_rxn`. The trained models are made publicly available as part of IBM RXN for Chemistry (`https://rxn.res.ibm.com/`).

Acknowledgements and Funding

This work has been carried out within the framework of the National Centre of Competence in Research **Catalysis** supported by the Swiss National Science Foundation. The authors acknowledge the financial support of the SNSF.

Methods

Data Sets

The enzymatic reaction data set with related EC (enzyme commission) numbers was created by merging entries extracted from Rhea (n = 8659), BRENDA (n = 11130), PathBank (n = 31047), and MetaNetX (n = 34485).[22–25] This data set was then further processed by (1) removing products that occur as reactants in the same reaction, (2) removing known co-enzymes and common byproducts from the products in reactions that exceed 1 product (Tables S1 and S2), (3) removing molecules with a heavy atom count < 4 from the products, and (4) removing reactions with > 1 or < 1 products or no reactants. The resulting data set contains 62,222 unique reaction–EC number combinations. The data set is available in 5 different token schemes: With no EC number (*EC0*, n = 55115), only EC-level 1 (*EC1*, n = 55707), EC-levels 1-2 (*EC2*, n = 56222), EC-levels 1-3 (*EC3*, n = 56579), EC-levels 1-4 (*EC4*, n = 62222). The different token schemes result in different set sizes as the removal of EC-levels leads to duplication and removal of extended reaction SMILES

We used the USPTO data set, which contains 1 million organic chemical reactions, together with the more specific enzymatic reaction data set to train the molecular transformer using multi-task transfer learning. This approach was previously successfully applied to carbohydrate reactions by Pesciullesi et al.[28]. The reactions in the USPTO data set are encoded as so-called reaction SMILES, using the same convention of Schwaller et al.[20]. An example is the reaction SMILES `CC(=O)O.OCC>OS(=O)(=O)O>CC(=O)OCC.O` encoding a Fischer esterification.

Preprocessing

The standard definition of a reaction SMILES was extended to include EC numbers (e.g. the reaction catalysed by the maltose alpha-D-glucosyltransferase is written as `A|5.4.99>>B`, where the SMILES for D-maltose and α,α-trehalose have been replaced by A and B for

brevity). We denote this extension to reaction SMILES enzymatic reaction SMILES.

We adapted the tokenisation operation used by Schwaller et al. [20] for the molecular transformer to handle enzymatic reaction SMILES. EC-levels 1-3 are treated as unique tokens to enable the transformer to learn the hierarchical structure of the EC numbering scheme. Because digits are already used to represent ring closures in SMILES, a number prefix is added to each level (v for EC-level 1, u for EC-level 2, and t for EC-level 3) during tokenisation. In addition, each EC token is encapsulated in brackets to simplify the tokenisation and detokenisation process. An example tokenisation of an enzymatic reaction SMILES is shown in Figure 15.

Finally, the resulting tokenised data set was split into a training, validation and test set (90%, 5%, and 5%, respectively). The training set was sampled so that none of the products contained within can be found in the training and the validation data set.

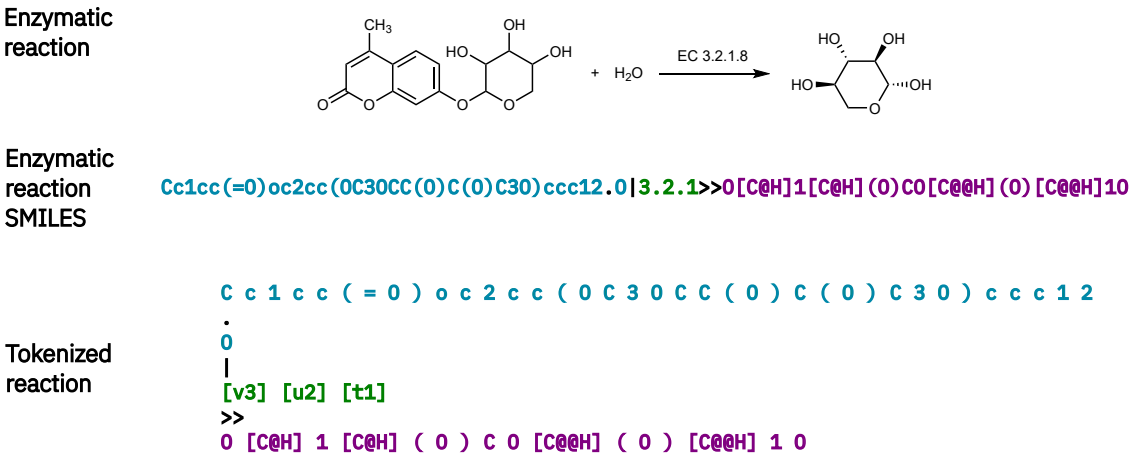

Figure 15: Step-wise description of the tokenisation process. Starting from an enzymatic reaction (top), a reaction SMILES representation is extracted (middle). The enzymatic reaction SMILES is finally tokenised both at the atom level and at the EC level (bottom).

Transfer Learning

The Molecular Transformer models have been implemented following the protocol introduced by Schwaller et al. [20]. The main conceptual difference lies in extending the reaction SMILES tokeniser to handle enzymatic reactions represented with the EC number as detailed in

the Preprocessing section . Multitask transfer learning has been implemented, as described by Pesciullesi et al.[28], using a convex weighting scheme for USPTO and *ECREACT*, 9 and 1 respectively. All the models have been trained using a version of OpenNMT[36] adapted for the Molecular Transformer.[37]

Attention Analysis

In the forward fine-tuned molecular transformer, the connection between the reactants and enzyme components and the products is modelled via self-attention and multi-head attention in the encoder/decoder layers. Since the probability distribution over all prediction candidates is computed based on the current translation state summarised by the last multi-head attention and the output layer, we focused our analysis on this last part of the decoder by considering only its attention weights.

We used relevant examples from the test set to analyse the patterns emerging from the mean attention over the heads. Using these examples, we investigated attention weights focusing on EC-levels 1-3 of the different heads. We started by analysing all reactions in our test set, focusing at a later stage on the three most frequent enzymatic reaction classes (oxidoreductases, transferases, and hydrolases). Finally, we analysed the correlation between the heads' attention weights to inspect redundancy.

In EC-level analysis, we filtered weights greater than a *noise* threshold. The threshold has been set at $\frac{1}{N}$, where N indicates the number of tokens in the input; the value has been determined by considering a baseline where each output token attend uniformly all the input tokens, i.e., no specific focus. By masking certain values, we have an appropriate metric to evaluate attention focus. If a token received weights lower or equal than the threshold, its value has been automatically excluded from contributing to the mean calculation. For the correlation analysis, we randomly selected 20 reactions for each class from which we extracted the corresponding head weights. For each reaction, we computed pairwise Pearson correlations[38] between the heads' flattened attention matrices. The correlation matrices for

each reaction have been aggregated by averaging the Fisher-transformed[39] correlation values. The resulting averaged correlation matrix was then derived by anti-transforming the values using a hyperbolic tangent.

Retrosynthesis routes prediction

We adapted the methodology proposed by Schwaller et al.[21], extending the retrosynthetic routes' prediction to handle enzyme information using the EC number format. The hypergraph exploration algorithm, at each step, is proposing disconnections using the backward model and computing a score for each prediction, in a Bayesian sense, based on the confidence of the forward model reweighted by the SCScore[40] measured on the precursors. The pathways are then prioritized, exploiting the score using beam search until a terminating condition is satisfied, i.e., commercial availability of the precursors (see Figure 16).

For the analysis of the targets from[18] we used an interactive version of the approach, where the backward Molecular Transformer allowed us to explore the synthetic routes iteratively until reaching commercially available precursors and proposing, at the same time, enzymes (represented up to EC level 3) that catalyze the corresponding reaction.

We based the selection of the targets on a comparative analysis of the coverage of the chemistry embedded in the reactions from the RetroBioCat[18] test set and the *ECREACT* data set. We annotated reaction SMILES for each step of the biocatalytic cascades considered in the test set from Finnigan et al.[18], excluding solvents information. For each reaction SMILES we extracted fingerprints using rxnfp[31] and we computed among the k-nearest neighbors (k=10), the fraction of neighbors belonging to RetroBioCat test set. The visualization of the embedded reactions was generated using TMAP.[26]

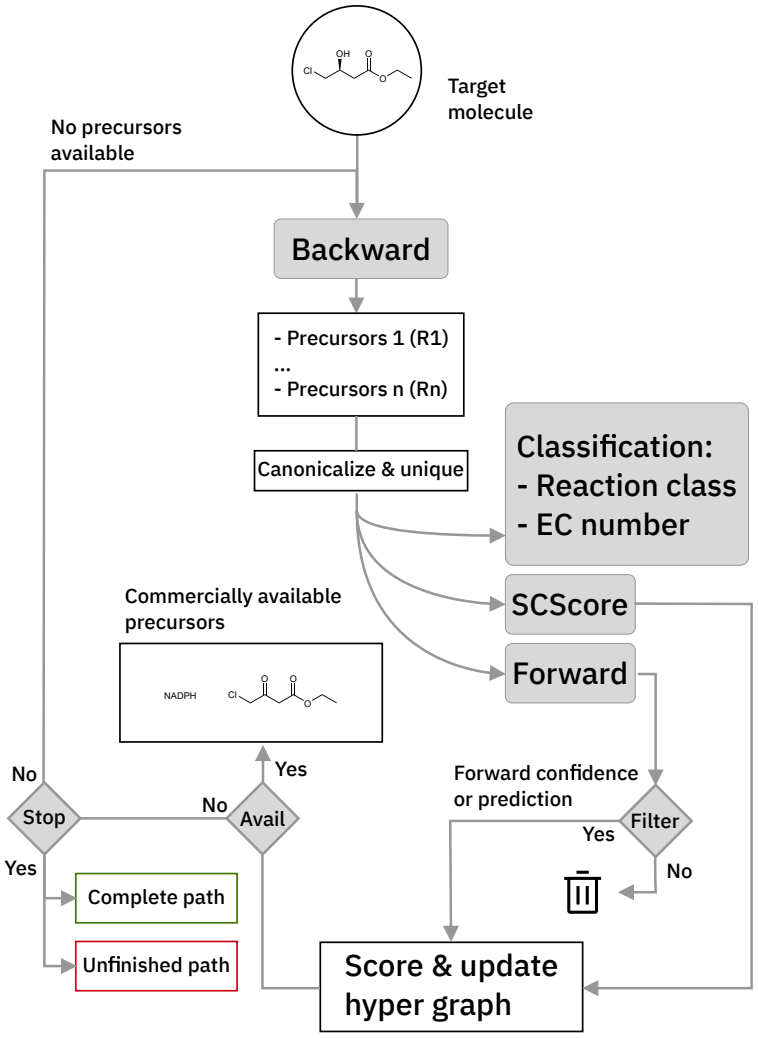

Figure 16: Detailed workflow of the retrosynthesis algorithm adapted from.[21] The hypergraph exploration algorithm combining two Molecular Transformer models for forward and backward predictions is extended to handle EC level information at each disconnection predicted by the model encoding it as a reaction class.

Acknowledgement

References

(1) Homaei, A. A.; Sariri, R.; Vianello, F.; Stevanato, R. Enzyme immobilization: An update. 2013; https://link.springer.com/article/10.1007/s12154-013-0102-9.

(2) Sheldon, R. A.; Woodley, J. M. Role of Biocatalysis in Sustainable Chemistry. 2018; https://pubs.acs.org/doi/abs/10.1021/acs.chemrev.7b00203.

(3) Hecht, K.; Meyer, H.-P.; Wohlgemuth, R.; Buller, R. Biocatalysis in the Swiss Manufacturing Environment. *Catalysts* **2020**, *10*, 1420.

(4) Wu, S.; Snajdrova, R.; Moore, J. C.; Baldenius, K.; Bornscheuer, U. T. Biocatalysis: Enzymatic Synthesis for Industrial Applications. 2021.

(5) Andler, S. M.; Goddard, J. M. Transforming food waste: how immobilized enzymes can valorize waste streams into revenue streams. *npj Science of Food* **2018**, *2*, 1–11.

(6) Sheldon, R. A.; Brady, D.; Bode, M. L. The Hitchhiker's guide to biocatalysis: recent advances in the use of enzymes in organic synthesis. *Chem. Sci.* **2020**, *11*, 2587–2605.

(7) Winkler, C. K.; Schrittwieser, J. H.; Kroutil, W. Power of Biocatalysis for Organic Synthesis. *ACS central science* **2021**, *7*, 55–71.

(8) Strohmeier, G. A.; Pichler, H.; May, O.; Gruber-Khadjawi, M. Application of designed enzymes in organic synthesis. *Chemical Reviews* **2011**, *111*, 4141–4164.

(9) Sheldon, R. A.; Pereira, P. C. Biocatalysis engineering: The big picture. 2017; https://pubs.rsc.org/en/content/articlehtml/2017/cs/c6cs00854bhttps://pubs.rsc.org/en/content/articlelanding/2017/cs/c6cs00854b.

(10) Turner, N. J.; O'reilly, E. Biocatalytic retrosynthesis. *Nature chemical biology* **2013**, *9*, 285–288.

(11) de Souza, R. O.; Miranda, L. S.; Bornscheuer, U. T. A retrosynthesis approach for biocatalysis in organic synthesis. *Chemistry–A European Journal* **2017**, *23*, 12040–12063.

(12) Hönig, M.; Sondermann, P.; Turner, N. J.; Carreira, E. M. Enantioselective chemo- and biocatalysis: partners in retrosynthesis. *Angewandte Chemie International Edition* **2017**, *56*, 8942–8973.

(13) Jorner, K.; Tomberg, A.; Bauer, C.; Sköld, C.; Norrby, P.-O. Organic reactivity from mechanism to machine learning. *Nature Reviews Chemistry* **2021**, *5*, 240–255.

(14) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances* **2021**, *7*.

(15) Madzhidov, T.; Lin, A. I.; Nugmanov, R.; Dyubankova, N.; Gimadiev, T.; Wegner, J. K.; Rakhimbekova, A.; Akhmetshin, T.; Ibragimova, Z.; Varnek, A.; Suleymanov, R.; Ceulemans, H.; Verhoeven, J. Atom-to-Atom Mapping: A Benchmarking Study of Popular Mapping Algorithms and Consensus Strategies. **2020**,

(16) Baranwal, M.; Magner, A.; Elvati, P.; Saldinger, J.; Violi, A.; Hero, A. O. A deep learning architecture for metabolic pathway prediction. *Bioinformatics* **2020**, *36*, 2547–2553.

(17) ATLAS of Biochemistry: A Repository of All Possible Biochemical Reactions for Synthetic Biology and Metabolic Engineering Studies. *ACS Synth. Biol.* **2016**, *5*, 1155–1166.

(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. *Nature Catalysis* **2021**, *4*, 98–104.

(19) Kreutter, D.; Schwaller, P.; Reymond, J. L. Predicting Enzymatic Reactions with a Molecular Transformer. 2020.

(20) 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*, 1572–1583.

(21) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical Science* **2020**, *11*, 3316–3325.

(22) Alcántara, R.; Axelsen, K. B.; Morgat, A.; Belda, E.; Coudert, E.; Bridge, A.; Cao, H.; De Matos, P.; Ennis, M.; Turner, S.; Owen, G.; Bougueleret, L.; Xenarios, I.; Steinbeck, C. Rhea - A manually curated resource of biochemical reactions. *Nucleic Acids Res.* **2012**, *40*.

(23) Schomburg, I.; Chang, A.; Schomburg, D. BRENDA, enzyme data and metabolic information. *Nucleic Acids Res.* **2002**, *30*, 47–49.

(24) Wishart, D. S. et al. PathBank: A comprehensive pathway database for model organisms. *Nucleic Acids Res.* **2020**, *48*, D470–D478.

(25) Ganter, M.; Bernard, T.; Moretti, S.; Stelling, J.; Pagni, M. MetaNetX.org: a website and repository for accessing, analysing and manipulating metabolic networks. *Bioinformatics* **2013**, *29*, 815–816.

(26) Probst, D.; Reymond, J. L. Visualization of very large high-dimensional data sets as minimum spanning trees. *J. Cheminform.* **2020**, *12*, 12.

(27) Capecchi, A.; Probst, D.; Reymond, J. L. One molecular fingerprint to rule them all: Drugs, biomolecules, and the metabolome. *J. Cheminform.* **2020**, *12*, 43.

(28) Pesciullesi, G.; Schwaller, P.; Laino, T.; Reymond, J. L. Transfer learning enables the molecular transformer to predict regio- and stereoselective reactions on carbohydrates. *Nat. Commun.* **2020**, *11*, 1–8.

(29) Hoover, B.; Strobelt, H.; Gehrmann, S. exbert: A visual analysis tool to explore learned representations in transformers models. *arXiv preprint arXiv:1910.05276* **2019**,

(30) Finnigan, W. RetroBioCat database files. **2020**,

(31) Schwaller, P.; Probst, D.; Vaucher, A. C.; Nair, V. H.; Kreutter, D.; Laino, T.; Reymond, J.-L. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence* **2021**, *3*, 144–152.

(32) Dimakos, V.; Taylor, M. S. Site-Selective Functionalization of Hydroxyl Groups in Carbohydrate Derivatives. *Chemical Reviews* **2018**, *118*, 11457–11517.

(33) Klausen, R. S.; Kennedy, C. R.; Hyde, A. M.; Jacobsen, E. N. Chiral Thioureas Promote Enantioselective Pictet–Spengler Cyclization by Stabilizing Every Intermediate and Transition State in the Carboxylic Acid-Catalyzed Reaction. *Journal of the American Chemical Society* **2017**, *139*, 12299–12309.

(34) Glinsky-Olivier, N.; Yang, S.; Retailleau, P.; Gandon, V.; Guinchard, X. Enantioselective Gold-Catalyzed Pictet–Spengler Reaction. *Organic Letters* **2019**, *21*, 9446–9451.

(35) Huang, D.; Xu, F.; Lin, X.; Wang, Y. Highly Enantioselective Pictet–Spengler Reaction Catalyzed by SPINOL-Phosphoric Acids. *Chemistry – A European Journal* **2012**, *18*, 3148–3152.

(36) Klein, G.; Kim, Y.; Deng, Y.; Senellart, J.; Rush, A. OpenNMT: Open-Source Toolkit for Neural Machine Translation. Proceedings of ACL 2017, System Demonstrations. Vancouver, Canada, 2017; pp 67–72.

(37) ONMT adaptation for rxn4chemistry. `https://github.com/rxn4chemistry/OpenNMT-py`.

(38) Pearson, K.; Galton, F. VII. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* **1895**, *58*, 240–242.

(39) Fisher, R. A. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **1915**, *10*, 507–521.

(40) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: synthetic complexity learned from a reaction corpus. *Journal of chemical information and modeling* **2018**, *58*, 252–261.

(41) Fuglede, B.; Topsøe, F. Jensen-Shannon divergence and Hubert space embedding. IEEE Int. Symp. Inf. Theory - Proc. 2004; p 31.

Supplemental Materials: Enzymes as Green Catalysts for Data-driven Template-free Chemical Synthesis

Table S1: SMARTS patterns of co-enzymes that were removed from the products.

Name	SMARTS
Coenzyme A	O=C(NCC*)CCNC(=O)C(O)C(C)(C)COP(=O)(*)OP(=O)(*)OC*3O*(n2cnc1c(ncnc12)N)*(O)*3OP(=O)(*)*
Nicotinamide adenine dinucleotides	**1*(*)*(COP(*)(=O)OP(*)(=O)OC*2O*(*)*(*)*2*)O*1*
Nucleoside phosphates	**1*(*)*(O*1COP(*)(=O)O)[R]
Nucleoside phosphates isomers	*P(*)(=O)O*1*(*)*(*)O[*]1COP(*)(*)=O
Sulfonium betaines	**1*(*)*(O*1CS*)[R]
Flavines	**1**2**3*(**(=O)**3=O)*(*)*2**1*
Hemes	*~1~*~*~2~*~*~1~*~*~1~*~*~*(~*~*~3~*~*~*(~*~*~4~*~*~*(~*~2)~*~4)~*~3)~*~1
Iron-sulfur cluster(s)	S1[Fe]S[Fe]1

Table S2: SMILES of common byproducts that were removed from the products.

Name	SMILES
Phosphate trianion	O=P([O-])([O-])[O-]
Hydrogen phosphate dianion	O=P([O-])([O-])O
(2-hydroxyethyl)trimethylammonium	C[N+](C)(C)CCO
Ethanolamine	NCCO
Diphosphate	O=P([O-])([O-])OP(=O)([O-])[O-]
Hydrogen diphosphate trianion	O=P([O-])([O-])OP(=O)([O-])O
2-Oxoglutarate dianion	O=C([O-])CCC(=O)C(=O)[O-]
Acetate ion	CC(=O)[O-]
Pyruvate	CC(=O)C(=O)[O-]

Table S3: The data set composition by EC-level 2.

EC number	Count	% (of total)	EC number	Count	% (of total)	EC number	Count	% (of total)	EC number	Count	% (of total)
1.-.x.x	75	0.120	1.7.x.x	191	0.306	3.2.x.x	888	1.423	5.2.x.x	36	0.058
1.1.x.x	4481	7.181	1.8.x.x	248	0.397	3.3.x.x	148	0.237	5.3.x.x	451	0.723
1.10.x.x	80	0.128	1.9.x.x	16	0.026	3.4.x.x	419	0.671	5.4.x.x	360	0.577
1.11.x.x	191	0.306	1.97.x.x	19	0.030	3.5.x.x	1082	1.734	5.5.x.x	198	0.317
1.12.x.x	34	0.054	2.-.x.x	8	0.013	3.6.x.x	1372	2.199	5.6.x.x	8	0.013
1.13.x.x	665	1.066	2.1.x.x	4420	7.083	3.7.x.x	115	0.184	5.99.x.x	5	0.008
1.14.x.x	4800	7.692	2.10.x.x	3	0.005	3.8.x.x	109	0.175	6.-.x.x	4	0.006
1.16.x.x	43	0.069	2.2.x.x	55	0.088	3.9.x.x	14	0.022	6.1.x.x	158	0.253
1.17.x.x	201	0.322	2.3.x.x	10369	16.616	3.A.x.x	21	0.034	6.2.x.x	517	0.828
1.18.x.x	43	0.069	2.4.x.x	2015	3.229	4.-.x.x	6	0.010	6.3.x.x	534	0.856
1.19.x.x	11	0.018	2.5.x.x	580	0.929	4.1.x.x	2024	3.243	6.4.x.x	47	0.075
1.2.x.x	1404	2.250	2.6.x.x	175	0.280	4.2.x.x	1762	2.824	6.5.x.x	26	0.042
1.20.x.x	19	0.030	2.7.x.x	14973	23.994	4.3.x.x	188	0.301	6.6.x.x	5	0.008
1.21.x.x	65	0.104	2.8.x.x	419	0.671	4.4.x.x	196	0.314	7.1.x.x	33	0.053
1.22.x.x	3	0.005	2.9.x.x	10	0.016	4.5.x.x	19	0.030	7.2.x.x	56	0.090
1.23.x.x	13	0.021	3.-.x.x	6	0.010	4.6.x.x	39	0.062	7.3.x.x	17	0.027
1.3.x.x	1362	2.183	3.1.x.x	2861	4.585	4.7.x.x	2	0.003	7.4.x.x	45	0.072
1.4.x.x	443	0.710	3.10.x.x	3	0.005	4.99.x.x	65	0.104	7.5.x.x	36	0.058
1.5.x.x	376	0.603	3.11.x.x	4	0.006	5.-.x.x	9	0.014	7.6.x.x	66	0.106
1.6.x.x	260	0.417	3.13.x.x	20	0.032	5.1.x.x	359	0.575			

Table S4: The data set composition by EC-level 3.

EC number	Count	% (of total)	EC number	Count	% (of total)	EC number	Count	% (of total)	EC number	Count	% (of total)
1.-.-.x	75	0.120	1.21.1.x	12	0.019	2.4.99.x	92	0.147	3.7.1.x	115	0.184
1.1.-.x	30	0.048	1.21.21.x	1	0.002	2.5.1.x	580	0.929	3.8.1.x	109	0.175
1.1.1.x	4023	6.447	1.21.3.x	35	0.056	2.6.-.x	1	0.002	3.9.1.x	14	0.022
1.1.2.x	45	0.072	1.21.4.x	2	0.003	2.6.1.x	167	0.268	3.A.1.x	19	0.030
1.1.3.x	245	0.393	1.21.98.x	6	0.010	2.6.99.x	7	0.011	3.A.3.x	2	0.003
1.1.4.x	1	0.002	1.21.99.x	8	0.013	2.7.-.x	2	0.003	4.-.-.x	6	0.010
1.1.5.x	33	0.053	1.22.1.x	3	0.005	2.7.1.x	1187	1.902	4.1.-.x	1	0.002
1.1.7.x	1	0.002	1.23.1.x	11	0.018	2.7.10.x	5	0.008	4.1.1.x	1594	2.554
1.1.9.x	1	0.002	1.23.5.x	2	0.003	2.7.11.x	31	0.050	4.1.2.x	214	0.343
1.1.98.x	18	0.029	1.3.-.x	20	0.032	2.7.12.x	10	0.016	4.1.3.x	107	0.171
1.1.99.x	84	0.135	1.3.1.x	936	1.500	2.7.13.x	7	0.011	4.1.4.x	1	0.002
1.10.1.x	1	0.002	1.3.2.x	18	0.029	2.7.14.x	3	0.005	4.1.99.x	107	0.171
1.10.2.x	7	0.011	1.3.3.x	121	0.194	2.7.2.x	77	0.123	4.2.1.x	996	1.596
1.10.3.x	68	0.109	1.3.5.x	11	0.018	2.7.3.x	29	0.046	4.2.2.x	41	0.066
1.10.5.x	3	0.005	1.3.7.x	40	0.064	2.7.4.x	163	0.261	4.2.3.x	701	1.123
1.10.99.x	1	0.002	1.3.8.x	134	0.215	2.7.6.x	29	0.046	4.2.99.x	24	0.038
1.11.-.x	4	0.006	1.3.98.x	17	0.027	2.7.7.x	766	1.228	4.3.-.x	2	0.003
1.11.1.x	135	0.216	1.3.99.x	65	0.104	2.7.8.x	12645	20.263	4.3.1.x	119	0.191
1.11.2.x	52	0.083	1.4.-.x	3	0.005	2.7.9.x	18	0.029	4.3.2.x	33	0.053
1.12.1.x	12	0.019	1.4.1.x	138	0.221	2.7.99.x	1	0.002	4.3.3.x	26	0.042
1.12.2.x	2	0.003	1.4.13.x	4	0.006	2.8.1.x	39	0.062	4.3.99.x	8	0.013
1.12.5.x	3	0.005	1.4.2.x	6	0.010	2.8.2.x	225	0.361	4.4.1.x	196	0.314
1.12.7.x	2	0.003	1.4.3.x	250	0.401	2.8.3.x	141	0.226	4.5.1.x	19	0.030
1.12.98.x	11	0.018	1.4.4.x	3	0.005	2.8.4.x	7	0.011	4.6.1.x	39	0.062
1.12.99.x	4	0.006	1.4.5.x	6	0.010	2.8.5.x	7	0.011	4.7.1.x	2	0.003
1.13.-.x	7	0.011	1.4.7.x	5	0.008	2.9.1.x	10	0.016	4.99.1.x	65	0.104
1.13.11.x	527	0.845	1.4.9.x	3	0.005	3.-.-.x	6	0.010	5.-.-.x	9	0.014
1.13.12.x	115	0.184	1.4.99.x	25	0.040	3.1.-.x	4	0.006	5.1.-.x	3	0.005
1.13.99.x	16	0.026	1.5.-.x	2	0.003	3.1.1.x	1006	1.612	5.1.1.x	124	0.199
1.14.-.x	118	0.189	1.5.1.x	258	0.413	3.1.11.x	6	0.010	5.1.2.x	23	0.037
1.14.11.x	291	0.466	1.5.3.x	64	0.103	3.1.13.x	5	0.008	5.1.3.x	182	0.292
1.14.12.x	191	0.306	1.5.5.x	7	0.011	3.1.14.x	2	0.003	5.1.99.x	27	0.043
1.14.13.x	1611	2.582	1.5.7.x	7	0.011	3.1.15.x	1	0.002	5.2.-.x	4	0.006
1.14.14.x	1312	2.102	1.5.8.x	10	0.016	3.1.2.x	326	0.522	5.2.1.x	32	0.051
1.14.15.x	338	0.542	1.5.99.x	28	0.045	3.1.21.x	6	0.010	5.3.-.x	1	0.002
1.14.16.x	2	0.003	1.6.-.x	1	0.002	3.1.22.x	1	0.002	5.3.1.x	208	0.333
1.14.17.x	7	0.011	1.6.1.x	2	0.003	3.1.26.x	6	0.010	5.3.2.x	35	0.056
1.14.18.x	141	0.226	1.6.2.x	25	0.040	3.1.27.x	8	0.013	5.3.3.x	155	0.248
1.14.19.x	442	0.708	1.6.3.x	26	0.042	3.1.3.x	1167	1.870	5.3.4.x	3	0.005
1.14.20.x	70	0.112	1.6.4.x	3	0.005	3.1.4.x	222	0.356	5.3.99.x	49	0.079
1.14.21.x	43	0.069	1.6.5.x	176	0.282	3.1.5.x	2	0.003	5.4.1.x	13	0.021
1.14.3.x	1	0.002	1.6.6.x	5	0.008	3.1.6.x	43	0.069	5.4.2.x	75	0.120
1.14.99.x	233	0.373	1.6.98.x	1	0.002	3.1.7.x	36	0.058	5.4.3.x	43	0.069
1.16.1.x	37	0.059	1.6.99.x	21	0.034	3.1.8.x	20	0.032	5.4.4.x	67	0.107
1.16.3.x	1	0.002	1.7.-.x	2	0.003	3.10.1.x	3	0.005	5.4.99.x	162	0.260
1.16.5.x	1	0.002	1.7.1.x	91	0.146	3.11.1.x	4	0.006	5.5.1.x	198	0.317
1.16.8.x	3	0.005	1.7.2.x	44	0.071	3.13.1.x	20	0.032	5.6.1.x	8	0.013
1.16.9.x	1	0.002	1.7.3.x	29	0.046	3.2.-.x	1	0.002	5.99.-.x	1	0.002
1.17.-.x	2	0.003	1.7.5.x	6	0.010	3.2.1.x	803	1.287	5.99.1.x	4	0.006
1.17.1.x	69	0.111	1.7.6.x	2	0.003	3.2.2.x	84	0.135	6.-.-.x	4	0.006
1.17.2.x	8	0.013	1.7.7.x	9	0.014	3.3.1.x	10	0.016	6.1.-.x	2	0.003
1.17.3.x	47	0.075	1.7.99.x	8	0.013	3.3.2.x	138	0.221	6.1.1.x	142	0.228
1.17.4.x	25	0.040	1.8.1.x	137	0.220	3.4.-.x	9	0.014	6.1.2.x	7	0.011
1.17.5.x	13	0.021	1.8.2.x	26	0.042	3.4.11.x	108	0.173	6.1.3.x	7	0.011
1.17.7.x	16	0.026	1.8.3.x	30	0.048	3.4.13.x	80	0.128	6.2.1.x	516	0.827
1.17.8.x	4	0.006	1.8.4.x	16	0.026	3.4.14.x	12	0.019	6.2.2.x	1	0.002
1.17.9.x	2	0.003	1.8.5.x	17	0.027	3.4.15.x	6	0.010	6.3.-.x	2	0.003
1.17.98.x	6	0.010	1.8.7.x	9	0.014	3.4.16.x	28	0.045	6.3.1.x	69	0.111
1.17.99.x	9	0.014	1.8.98.x	6	0.010	3.4.17.x	42	0.067	6.3.2.x	306	0.490
1.18.-.x	1	0.002	1.8.99.x	7	0.011	3.4.18.x	1	0.002	6.3.3.x	28	0.045
1.18.1.x	32	0.051	1.9.3.x	7	0.011	3.4.19.x	50	0.080	6.3.4.x	84	0.135
1.18.4.x	1	0.002	1.9.6.x	4	0.006	3.4.21.x	29	0.046	6.3.5.x	45	0.072
1.18.6.x	5	0.008	1.9.98.x	3	0.005	3.4.22.x	18	0.029	6.4.1.x	47	0.075
1.18.99.x	4	0.006	1.9.99.x	2	0.003	3.4.23.x	7	0.011	6.5.-.x	1	0.002
1.19.1.x	8	0.013	1.97.1.x	19	0.030	3.4.24.x	21	0.034	6.5.1.x	25	0.040
1.19.6.x	3	0.005	2.-.-.x	8	0.013	3.4.25.x	6	0.010	6.6.1.x	5	0.008
1.2.-.x	5	0.008	2.1.1.x	4352	6.974	3.4.99.x	2	0.003	7.1.1.x	25	0.040
1.2.1.x	1157	1.854	2.1.2.x	28	0.045	3.5.-.x	1	0.002	7.1.2.x	4	0.006
1.2.2.x	9	0.014	2.1.3.x	38	0.061	3.5.1.x	606	0.971	7.1.3.x	4	0.006
1.2.3.x	94	0.151	2.1.4.x	1	0.002	3.5.2.x	91	0.146	7.2.1.x	8	0.013
1.2.4.x	51	0.082	2.1.5.x	1	0.002	3.5.3.x	73	0.117	7.2.2.x	36	0.058
1.2.5.x	27	0.043	2.10.1.x	3	0.005	3.5.4.x	161	0.258	7.2.4.x	12	0.019
1.2.7.x	48	0.077	2.2.1.x	55	0.088	3.5.5.x	96	0.154	7.3.2.x	17	0.027
1.2.98.x	4	0.006	2.3.-.x	4	0.006	3.5.99.x	54	0.087	7.4.2.x	45	0.072
1.2.99.x	9	0.014	2.3.1.x	10174	16.304	3.6.-.x	1	0.002	7.5.2.x	36	0.058
1.20.1.x	9	0.014	2.3.2.x	101	0.162	3.6.1.x	551	0.883	7.6.1.x	1	0.002
1.20.2.x	4	0.006	2.3.3.x	90	0.144	3.6.2.x	6	0.010	7.6.2.x	65	0.104
1.20.4.x	2	0.003	2.4.-.x	5	0.008	3.6.3.x	794	1.272			
1.20.9.x	4	0.006	2.4.1.x	1732	2.776	3.6.4.x	16	0.026			
1.21.-.x	1	0.002	2.4.2.x	186	0.298	3.6.5.x	4	0.006			

Table S5: The data set composition by size n of sub set and EC level.

	n = 1		**n < 5**		**n < 10**		**n < 100**		**n ≥ 100**		**Total**
	count	%	count	%	count	%	count	%	count	%	count
EC1	0	0.0	0	0.0	0	0.0	0	0.0	7	100	7
EC2	2	2.6	7	9.0	12	15.4	42	53.8	36	46.2	78
EC3	33	10.7	81	26.2	129	41.7	248	80.3	61	19.7	309
EC4	1593	25.3	4473	71.1	5571	88.6	6251	99.4	38	0.6	6460

Table S6: The data set composition after train/test split at EC-level 2.

EC number	Train Samples	Test Samples	EC number	Train Samples	Test Samples	EC number	Train Samples	Test Samples	EC number	Train Samples	Test Samples
1.-.x.x	58	6	1.8.x.x	209	7	3.3.x.x	106	12	5.3.x.x	341	21
1.1.x.x	3569	161	1.9.x.x	12	1	3.4.x.x	151	8	5.4.x.x	252	17
1.10.x.x	66	4	1.97.x.x	17	0	3.5.x.x	694	51	5.5.x.x	159	8
1.11.x.x	148	7	2.-.x.x	8	0	3.6.x.x	1207	14	5.6.x.x	6	0
1.12.x.x	26	1	2.1.x.x	3759	184	3.7.x.x	53	3	5.99.x.x	2	0
1.13.x.x	419	65	2.10.x.x	3	0	3.8.x.x	75	11	6.-.x.x	4	0
1.14.x.x	3793	259	2.2.x.x	48	1	3.9.x.x	10	0	6.1.x.x	101	1
1.16.x.x	42	0	2.3.x.x	9041	567	3.A.x.x	19	0	6.2.x.x	416	6
1.17.x.x	167	6	2.4.x.x	1570	79	4.-.x.x	6	0	6.3.x.x	527	54
1.18.x.x	44	0	2.5.x.x	577	20	4.1.x.x	1640	50	6.4.x.x	37	0
1.19.x.x	10	0	2.6.x.x	129	4	4.2.x.x	1201	109	6.5.x.x	25	1
1.2.x.x	1126	31	2.7.x.x	13217	726	4.3.x.x	119	8	6.6.x.x	4	0
1.20.x.x	15	0	2.8.x.x	338	13	4.4.x.x	142	7	7.1.x.x	30	1
1.21.x.x	48	7	2.9.x.x	9	0	4.5.x.x	10	1	7.2.x.x	36	0
1.23.x.x	10	2	3.-.x.x	5	0	4.6.x.x	22	1	7.3.x.x	8	0
1.3.x.x	1077	60	3.1.x.x	2092	149	4.7.x.x	2	0	7.4.x.x	32	0
1.4.x.x	332	18	3.10.x.x	1	0	4.99.x.x	46	1	7.5.x.x	30	0
1.5.x.x	280	13	3.11.x.x	1	0	5.-.x.x	5	0	7.6.x.x	55	0
1.6.x.x	216	5	3.13.x.x	10	2	5.1.x.x	264	13			
1.7.x.x	186	6	3.2.x.x	387	17	5.2.x.x	28	0			

Table S7: The data set composition after train/test split at EC-level 3.

EC number	Train Samples	Test Samples	EC number	Train Samples	Test Samples	EC number	Train Samples	Test Samples	EC number	Train Samples	Test Samples
1.-.-.x	58	6	1.21.-.x	1	0	2.4.-.x	4	0	3.6.4.x	11	1
1.1.-.x	25	0	1.21.1.x	10	0	2.4.1.x	1358	69	3.6.5.x	3	0
1.1.1.x	3220	141	1.21.21.x	1	0	2.4.2.x	141	6	3.7.1.x	53	3
1.1.2.x	40	0	1.21.3.x	25	3	2.4.99.x	67	4	3.8.1.x	75	11
1.1.3.x	176	15	1.21.4.x	2	0	2.5.1.x	577	20	3.9.1.x	10	0
1.1.4.x	1	0	1.21.98.x	3	2	2.6.1.x	122	3	3.A.1.x	17	0
1.1.5.x	33	0	1.21.99.x	6	2	2.6.3.x	2	0	3.A.3.x	2	0
1.1.7.x	1	0	1.23.1.x	9	2	2.6.99.x	5	1	4.-.-.x	6	0
1.1.9.x	1	0	1.23.5.x	1	0	2.7.-.x	2	0	4.1.-.x	1	0
1.1.98.x	7	0	1.3.-.x	18	1	2.7.1.x	964	10	4.1.1.x	1359	38
1.1.99.x	65	5	1.3.1.x	718	46	2.7.10.x	5	0	4.1.2.x	134	7
1.10.2.x	8	0	1.3.2.x	15	2	2.7.11.x	21	2	4.1.3.x	64	2
1.10.3.x	56	4	1.3.3.x	91	6	2.7.12.x	8	0	4.1.99.x	82	3
1.10.5.x	2	0	1.3.5.x	10	0	2.7.13.x	7	0	4.2.1.x	731	45
1.11.-.x	3	0	1.3.7.x	44	1	2.7.14.x	3	0	4.2.2.x	25	3
1.11.1.x	104	3	1.3.8.x	120	3	2.7.2.x	60	0	4.2.3.x	435	60
1.11.2.x	41	4	1.3.98.x	9	0	2.7.3.x	22	1	4.2.99.x	10	1
1.12.1.x	10	0	1.3.99.x	52	1	2.7.4.x	135	2	4.3.-.x	1	0
1.12.2.x	1	0	1.4.-.x	3	0	2.7.6.x	23	0	4.3.1.x	79	3
1.12.5.x	2	0	1.4.1.x	101	3	2.7.7.x	669	5	4.3.2.x	18	1
1.12.7.x	2	0	1.4.13.x	2	2	2.7.8.x	11278	706	4.3.3.x	17	4
1.12.98.x	8	1	1.4.2.x	6	0	2.7.9.x	19	0	4.3.99.x	4	0
1.12.99.x	3	0	1.4.3.x	182	12	2.7.99.x	1	0	4.4.1.x	142	7
1.13.-.x	3	1	1.4.4.x	5	1	2.8.1.x	38	1	4.5.1.x	10	1
1.13.11.x	339	58	1.4.5.x	6	0	2.8.2.x	179	12	4.6.1.x	22	1
1.13.12.x	64	5	1.4.7.x	4	0	2.8.3.x	108	0	4.7.1.x	2	0
1.13.99.x	13	1	1.4.9.x	1	0	2.8.4.x	7	0	4.99.1.x	46	1
1.14.-.x	99	9	1.4.99.x	22	0	2.8.5.x	6	0	5.-.-.x	5	0
1.14.11.x	55	3	1.5.-.x	2	0	2.9.1.x	9	0	5.1.-.x	3	0
1.14.12.x	150	10	1.5.1.x	202	12	3.-.-.x	5	0	5.1.1.x	86	7
1.14.13.x	1329	94	1.5.3.x	32	1	3.1.-.x	1	1	5.1.2.x	15	1
1.14.14.x	1123	85	1.5.5.x	7	0	3.1.1.x	746	46	5.1.3.x	140	5
1.14.15.x	303	12	1.5.7.x	6	0	3.1.11.x	9	0	5.1.99.x	20	0
1.14.16.x	2	0	1.5.8.x	10	0	3.1.12.x	1	0	5.2.-.x	4	0
1.14.17.x	7	1	1.5.99.x	21	0	3.1.13.x	4	1	5.2.1.x	24	0
1.14.18.x	124	5	1.6.-.x	1	0	3.1.14.x	1	0	5.3.-.x	1	0
1.14.19.x	375	22	1.6.1.x	2	0	3.1.15.x	1	0	5.3.1.x	168	5
1.14.20.x	3	0	1.6.2.x	23	1	3.1.16.x	2	0	5.3.2.x	21	2
1.14.21.x	36	2	1.6.3.x	21	1	3.1.2.x	223	15	5.3.3.x	113	9
1.14.3.x	1	0	1.6.4.x	2	0	3.1.21.x	5	0	5.3.4.x	3	0
1.14.99.x	186	16	1.6.5.x	150	3	3.1.26.x	5	0	5.3.99.x	35	5
1.16.1.x	37	0	1.6.6.x	1	0	3.1.27.x	8	0	5.4.1.x	10	0
1.16.3.x	1	0	1.6.98.x	1	0	3.1.3.x	951	73	5.4.2.x	57	3
1.16.5.x	1	0	1.6.99.x	15	0	3.1.4.x	104	8	5.4.3.x	29	3
1.16.8.x	2	0	1.7.-.x	2	0	3.1.6.x	7	0	5.4.4.x	52	2
1.16.9.x	1	0	1.7.1.x	70	5	3.1.7.x	20	4	5.4.99.x	104	9
1.17.-.x	2	0	1.7.2.x	48	1	3.1.8.x	4	1	5.5.1.x	159	8
1.17.1.x	58	1	1.7.3.x	23	0	3.10.1.x	1	0	5.6.1.x	6	0
1.17.2.x	8	0	1.7.5.x	16	0	3.11.1.x	1	0	5.99.1.x	2	0
1.17.3.x	36	2	1.7.6.x	2	0	3.13.1.x	10	2	6.-.-.x	4	0
1.17.4.x	22	0	1.7.7.x	10	0	3.2.1.x	358	12	6.1.-.x	1	0
1.17.5.x	13	1	1.7.99.x	15	0	3.2.2.x	29	5	6.1.1.x	89	0
1.17.7.x	13	0	1.8.1.x	117	4	3.3.1.x	6	0	6.1.2.x	5	1
1.17.8.x	3	0	1.8.2.x	21	0	3.3.2.x	100	12	6.1.3.x	6	0
1.17.9.x	1	0	1.8.3.x	18	2	3.4.-.x	5	0	6.2.1.x	416	6
1.17.98.x	5	0	1.8.4.x	10	0	3.4.11.x	43	1	6.3.-.x	22	3
1.17.99.x	6	2	1.8.5.x	17	1	3.4.13.x	17	0	6.3.1.x	58	5
1.18.-.x	1	0	1.8.7.x	12	0	3.4.14.x	4	1	6.3.2.x	325	45
1.18.1.x	32	0	1.8.98.x	8	0	3.4.15.x	3	2	6.3.3.x	22	0
1.18.4.x	1	0	1.8.99.x	6	0	3.4.16.x	10	0	6.3.4.x	69	1
1.18.6.x	6	0	1.9.3.x	4	0	3.4.17.x	12	1	6.3.5.x	31	0
1.18.99.x	4	0	1.9.6.x	4	1	3.4.19.x	13	1	6.4.1.x	37	0
1.19.1.x	8	0	1.9.98.x	3	0	3.4.21.x	18	1	6.5.1.x	25	1
1.19.6.x	2	0	1.9.99.x	1	0	3.4.22.x	5	1	6.6.1.x	4	0
1.2.-.x	5	0	1.97.1.x	17	0	3.4.23.x	6	0	7.1.1.x	21	1
1.2.1.x	921	27	2.-.-.x	8	0	3.4.24.x	10	0	7.1.2.x	4	0
1.2.2.x	8	0	2.1.1.x	3698	183	3.4.25.x	5	0	7.1.3.x	5	0
1.2.3.x	69	1	2.1.2.x	24	0	3.5.-.x	1	0	7.2.1.x	7	0
1.2.4.x	47	2	2.1.3.x	35	1	3.5.1.x	378	24	7.2.2.x	22	0
1.2.5.x	22	0	2.1.4.x	1	0	3.5.2.x	59	5	7.2.4.x	7	0
1.2.7.x	45	0	2.1.5.x	1	0	3.5.3.x	31	2	7.3.2.x	8	0
1.2.98.x	1	1	2.10.1.x	3	0	3.5.4.x	105	15	7.4.2.x	32	0
1.2.99.x	8	0	2.2.1.x	48	1	3.5.5.x	83	3	7.5.2.x	30	0
1.20.1.x	7	0	2.3.-.x	4	0	3.5.99.x	37	2	7.6.1.x	1	0
1.20.2.x	4	0	2.3.1.x	8887	561	3.6.1.x	440	13	7.6.2.x	54	0
1.20.4.x	2	0	2.3.2.x	84	4	3.6.2.x	2	0			
1.20.9.x	2	0	2.3.3.x	66	2	3.6.3.x	751	0			

EC-level 1 analysis

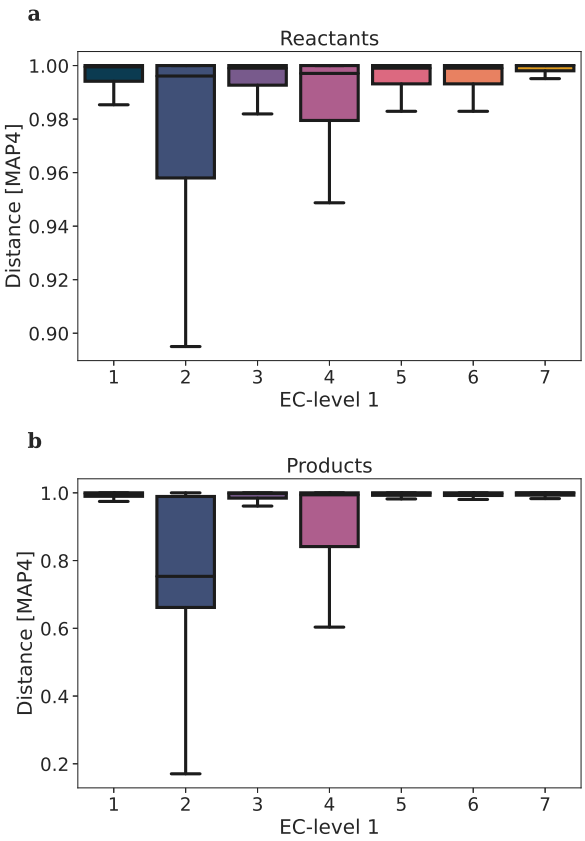

Figure S1: Sampled (10%) intra-class MAP4 distances of unique reactants (**a**) and products (**b**) participating in reactions in *EC3*. Transferases (2), lyases (4), and to a lesser degree hydrolases (3) show lower mean distances compared to other classes. This confirms the findings of the visual inspection carried out on the TMAP in Figure 3. The existence of homogeneous clusters of molecules within a class acts as an implicit feature, reducing the importance of the EC number token (explicit feature) during training and might increase accuracy compared to other classes.

EC-level 2 analysis

The most represented subclasses at EC-level 2 are EC 2.7.x.x (transferases transferring phosphorus-containing groups) at 24.5%, EC 2.3.x.x (acetyltransferases) at 16.8%, EC 1.1.x.x (oxidoreductases acting on the CH-OH group of donors) at 8%, EC 2.1.x.x (transferases transferring one-carbon groups) 7.5%, EC 1.14.x.x (oxidoreductases acting on paired donors, with incorporation or reduction of molecular oxygen) 7.3%, EC 3.1.x.x (hydrolases acting on ester bonds) 4.5%, EC 4.1.x.x (carbon-carbon lyases) 3%, EC 2.4.x.x (glycosyltransferases) 2.9%, EC 4.2.x.x (carbon-oxygen lyases) at 2.5%, and EC 3.6.x.x (hydrolases acting on acting on acid anhydrides) 2.5%.

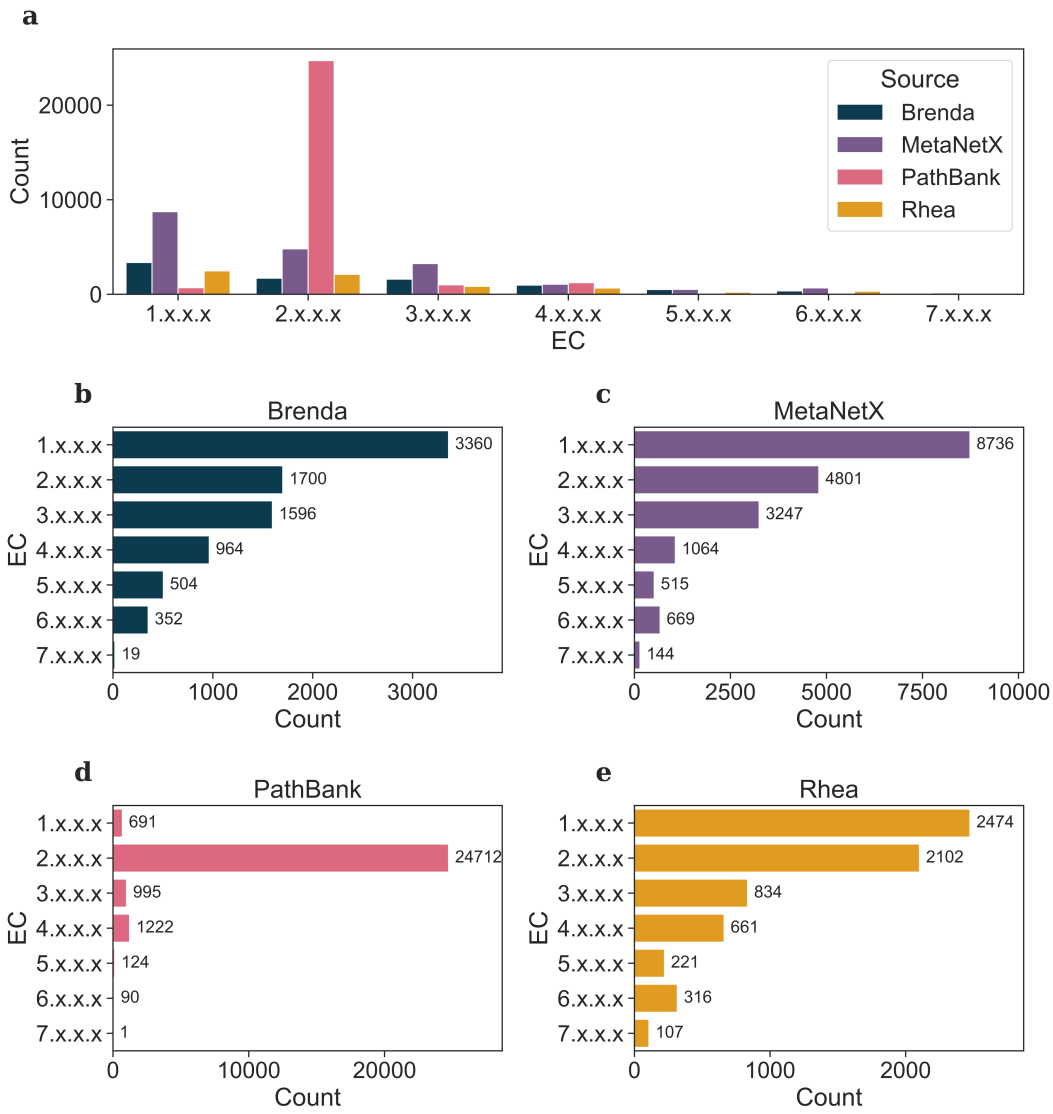

Figure S2: Data sources for the *ECREACT* data set. (**a**) The overall composition of the data set by EC-level 1 and source. (**b-e**) Composition of the data by EC-level 1 imported from Brenda, MetaNetX, PathBank, and Rhea, respectively. The large number of transferase-catalysed reactions imported from PathBank reflects the high number of lipid pathways stored in the database.

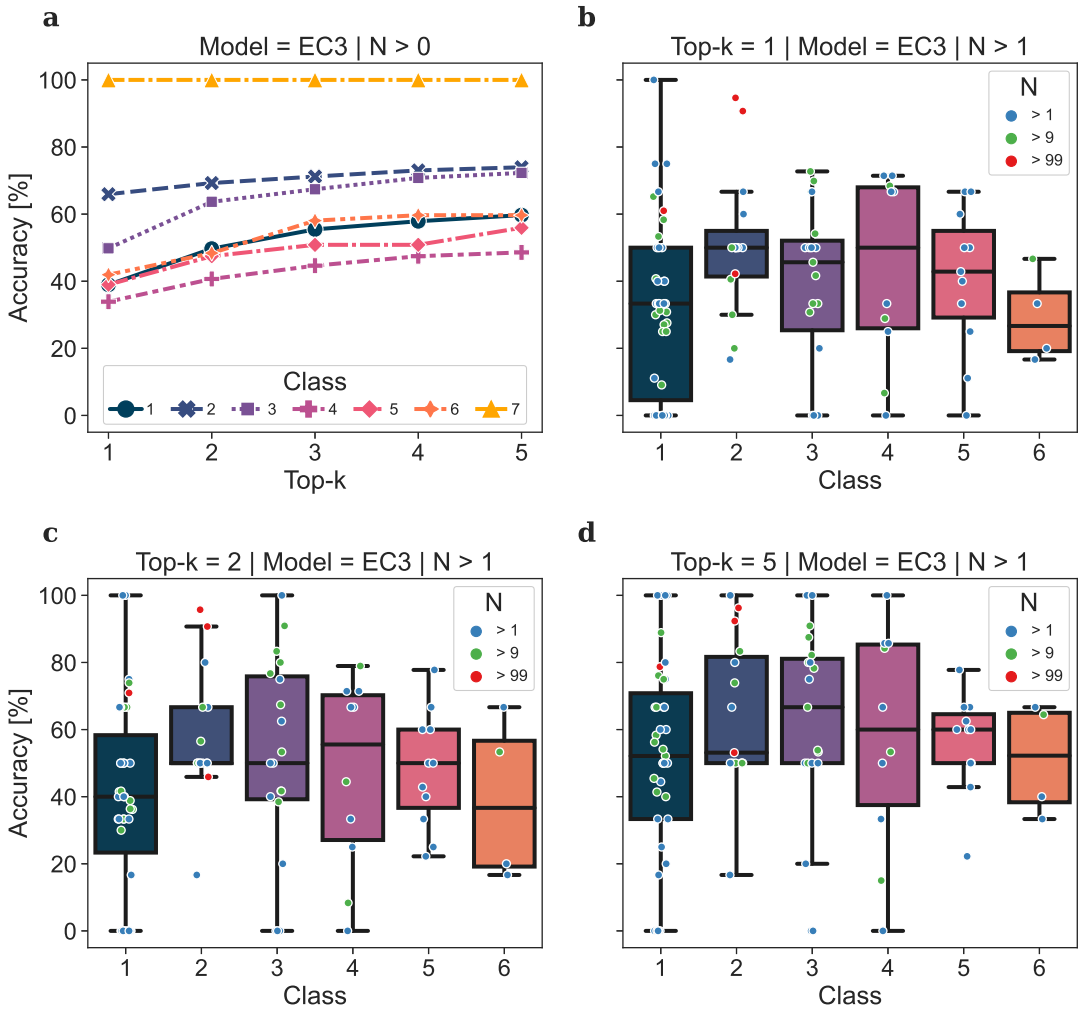

Figure S3: Class-wise accuracy for the forward model trained on token scheme *EC3* with **stereochemistry information removed**. (**a**) The top-k prediction accuracy for each class. The accuracy of (**b**) top-1, (**c**) top-2, and (**d**) top-5 predictions shown in detail. Each dot represents an EC-level 3 category with a number of test samples > 1. The EC-level 3 subclasses are further stratified by test sample size N. Removing all information related to stereochemistry leads to an increase in overall accuracy from 49.6% to 55%. With the highest increase among isomerase-catylsed reactions (class 5) of 18.6% to 40%.

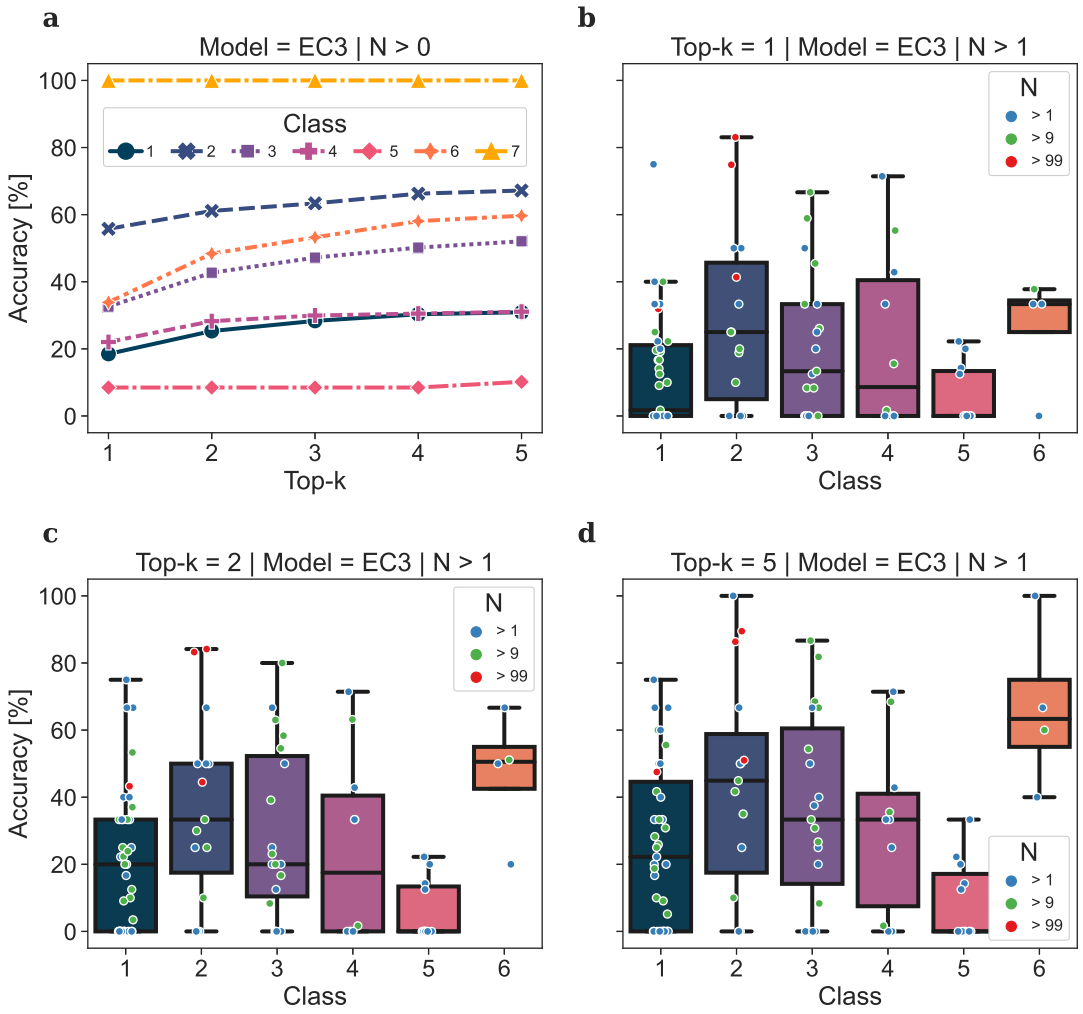

Figure S4: Class-wise accuracy for the forward model trained on token scheme *EC3* with EC numbers **randomized within classes**. (**a**) The top-k prediction accuracy for each class. The accuracy of (**b**) top-1, (**c**) top-2, and (**d**) top-5 predictions shown in detail. Each dot represents an EC-level 3 category with a number of test samples > 1. The EC-level 3 subclasses are further stratified by test sample size N.

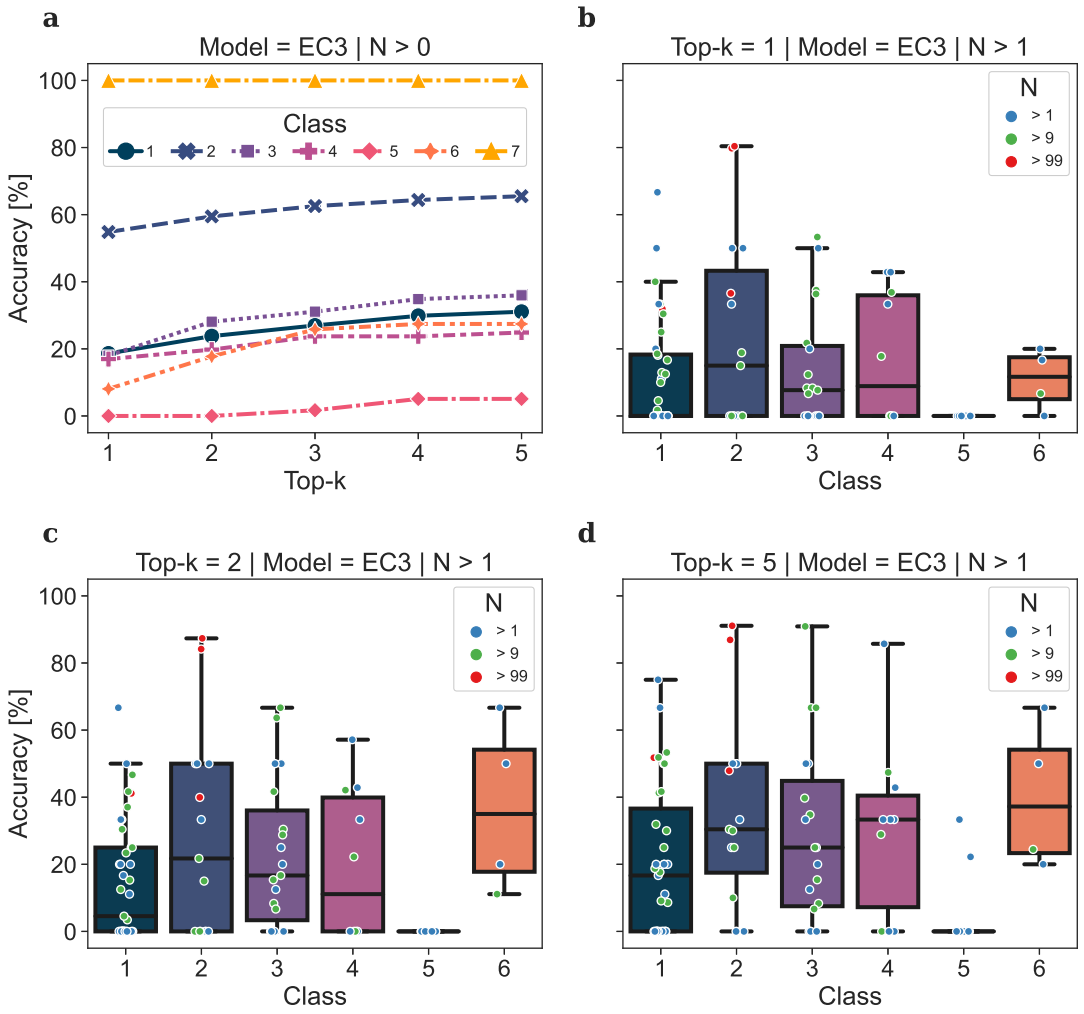

Figure S5: Class-wise accuracy for the forward model trained on token scheme *EC3* with EC **numbers randomized across classes**. (**a**) The top-k prediction accuracy for each class. The accuracy of (**b**) top-1, (**c**) top-2, and (**d**) top-5 predictions shown in detail. Each dot represents an EC-level 3 category with a number of test samples > 1. The EC-level 3 subclasses are further stratified by test sample size N.

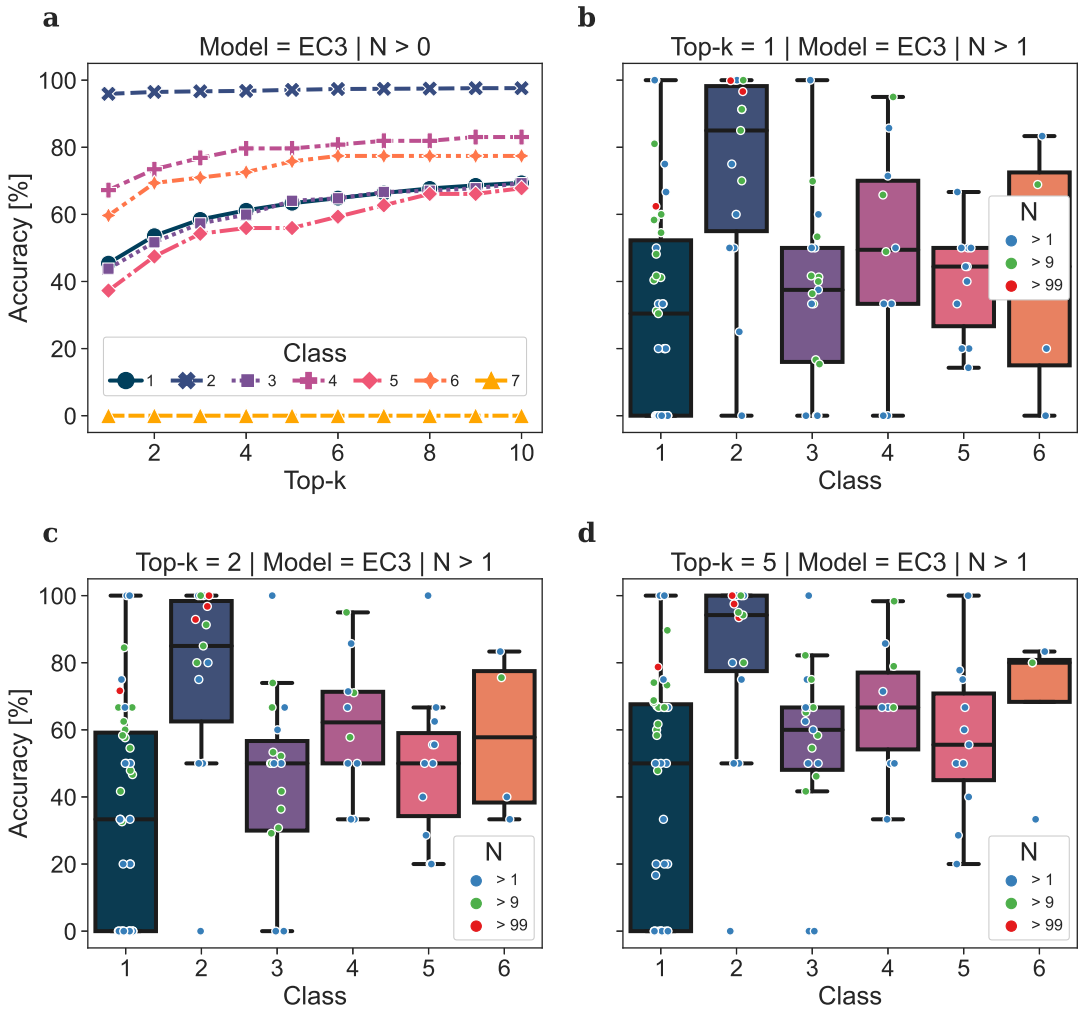

Figure S6: Class-wise accuracy for the backward model trained on token scheme *EC3* predicting the **EC number only**. (**a**) The top-k prediction accuracy for each class. The accuracy of (**b**) top-1, (**c**) top-2, and (**d**) top-5 predictions shown in detail. Each dot represents an EC-level 3 category with a number of test samples > 1. The EC-level 3 subclasses are further stratified by test sample size N. Given the high number of subclasses for oxidoreductases (class 1) on EC-level 3, it's relatively performance is in line with previous assumptions.

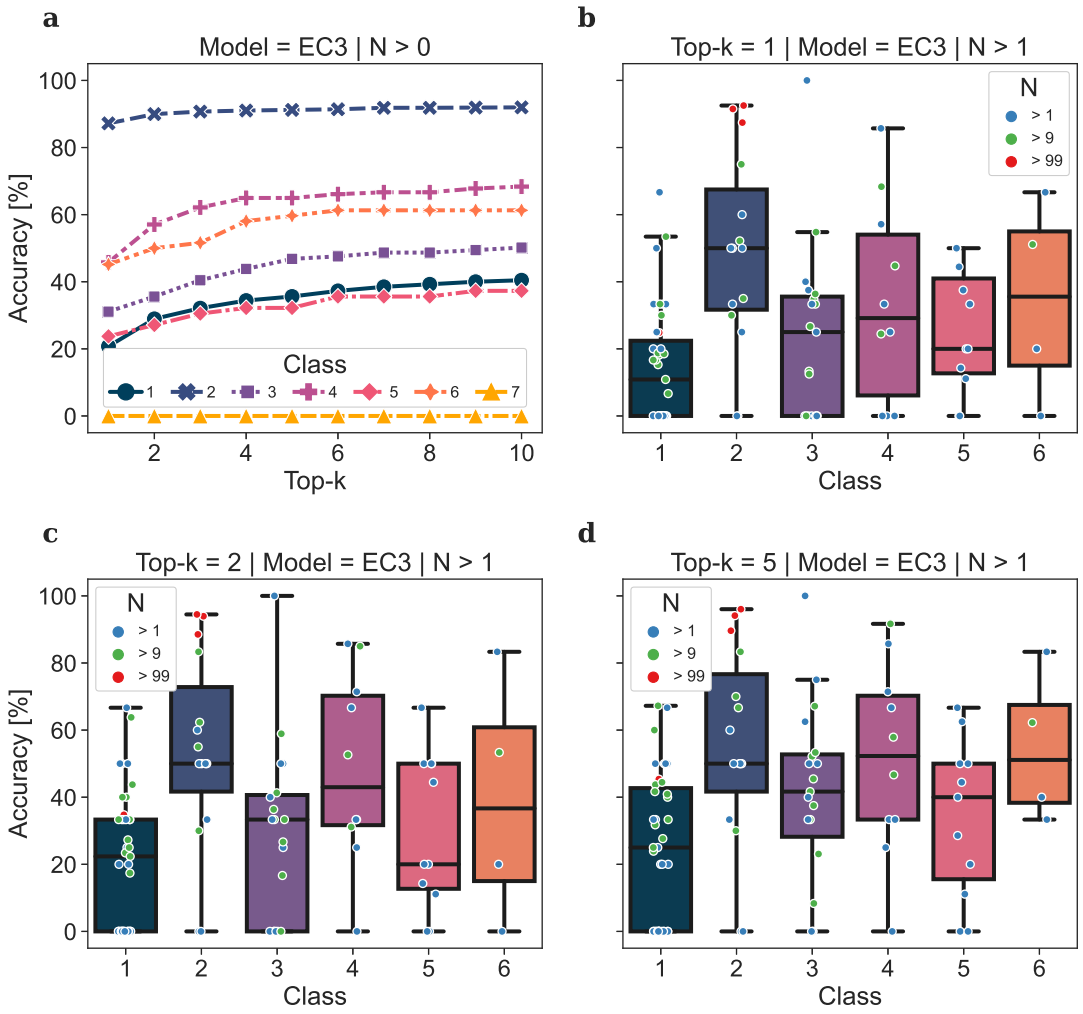

Figure S7: Class-wise accuracy for the backward model trained on token scheme *EC3* with **stereochemistry information removed**. (**a**) The top-k prediction accuracy for each class. The accuracy of (**b**) top-1, (**c**) top-2, and (**d**) top-5 predictions shown in detail. Each dot represents an EC-level 3 category with a number of test samples > 1. The EC-level 3 subclasses are further stratified by test sample size N.

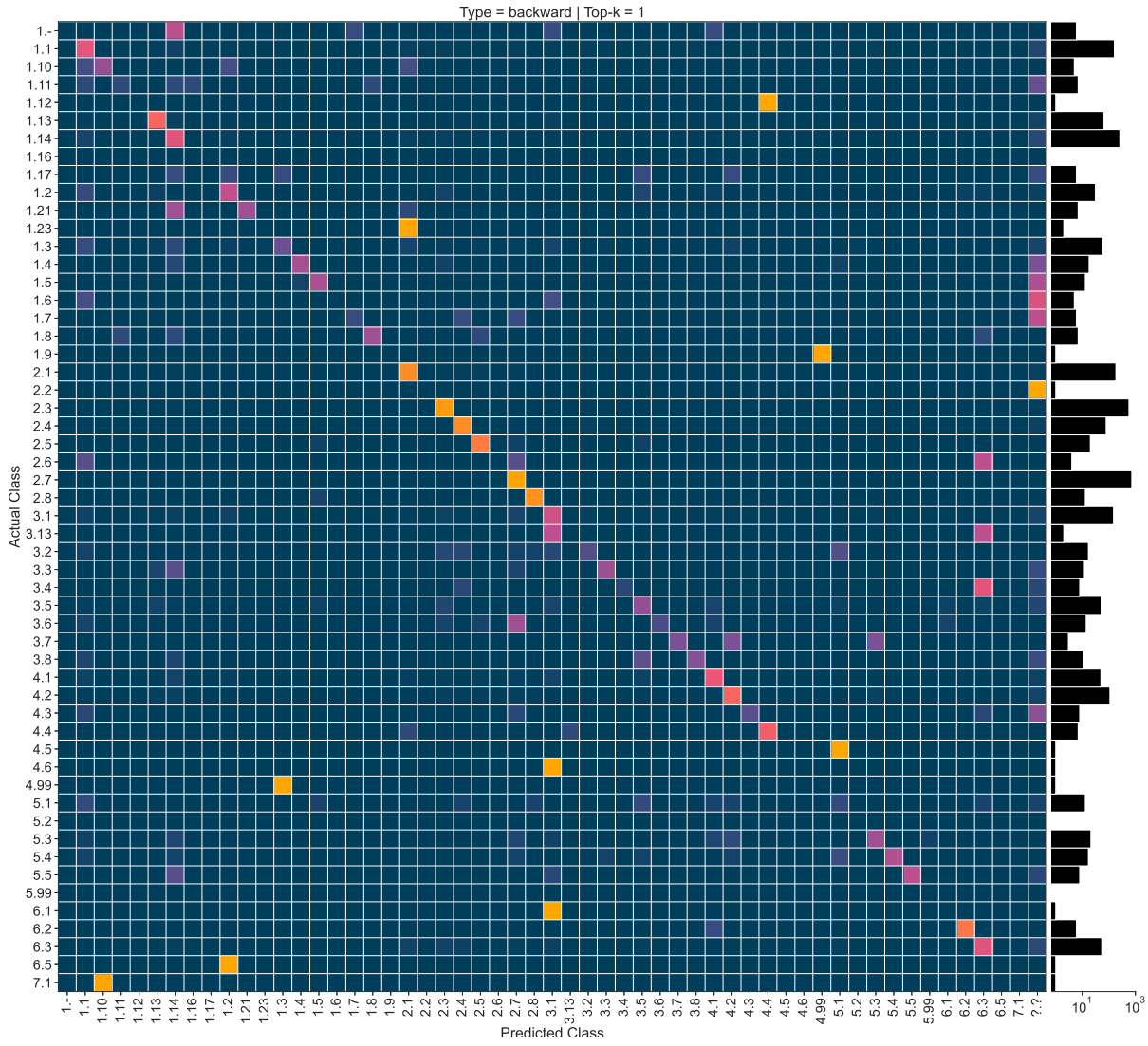

Figure S8: The confusion matrix based on predicted EC numbers by the backward model for EC-level 2. The bars right of the plot show the number of samples per EC-level 2 category. Comparing the sample sizes with the respective accuracies shows the established pattern of subclasses with high sample count having also higher accuracy.

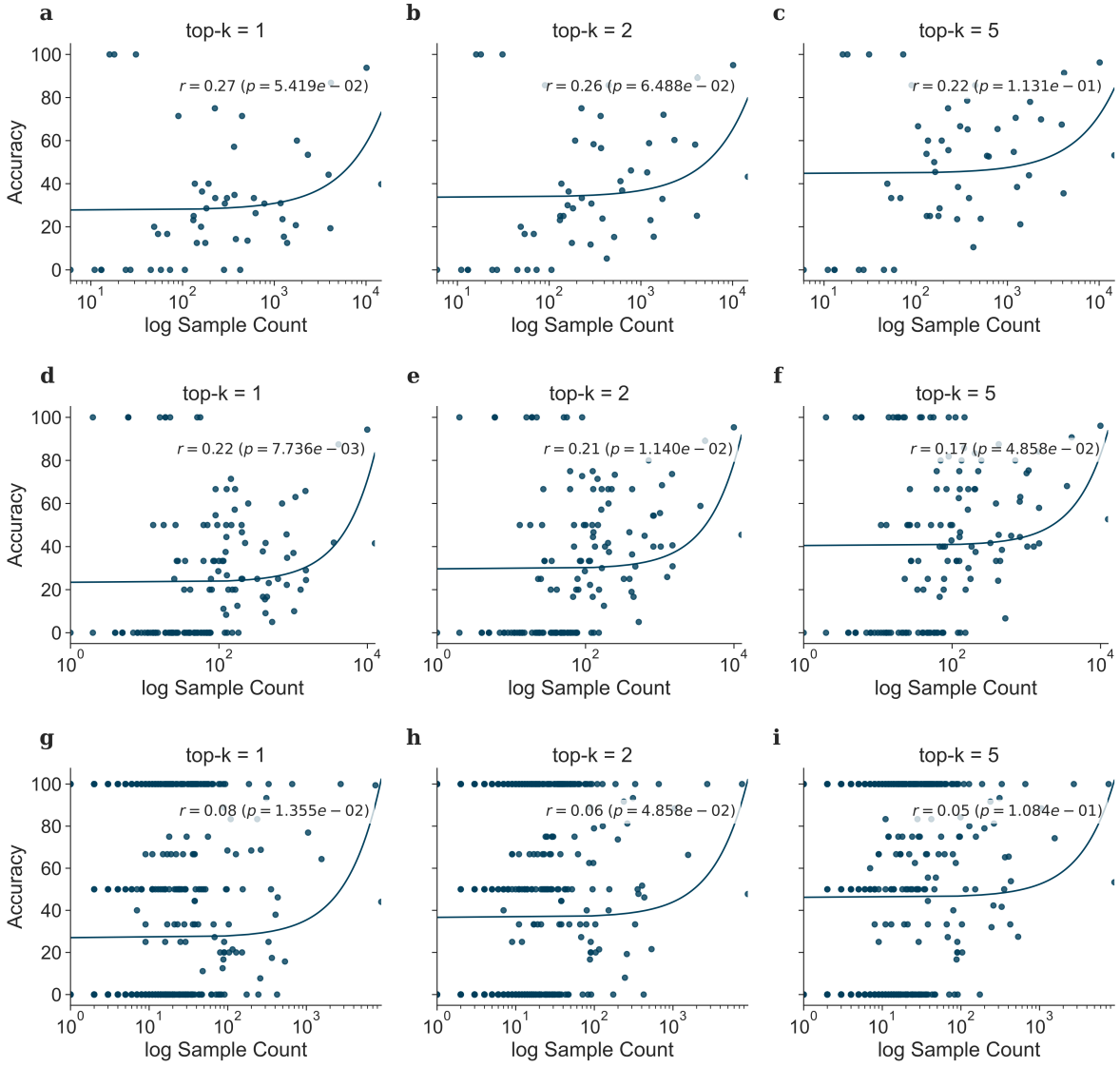

Figure S9: Correlation between forward prediction accuracy and sample count in *EC2* (**a**, **b**), *EC3* (**c**, **d**), and *EC4* (**e**, **f**). We observe a significant correlation between sample size in token schemes *EC2* and *EC3*. The trend towards lower correlations in higher EC-level token schemes is caused by a further reduction in test cases due to the selection of unique test products not found in the training sets and the resulting hit-or-miss accuracies appearing as bands at 0 and 100% accuracy, respectively. Increasing k results not only in increasing the accuracy but also in lowering the correlation.

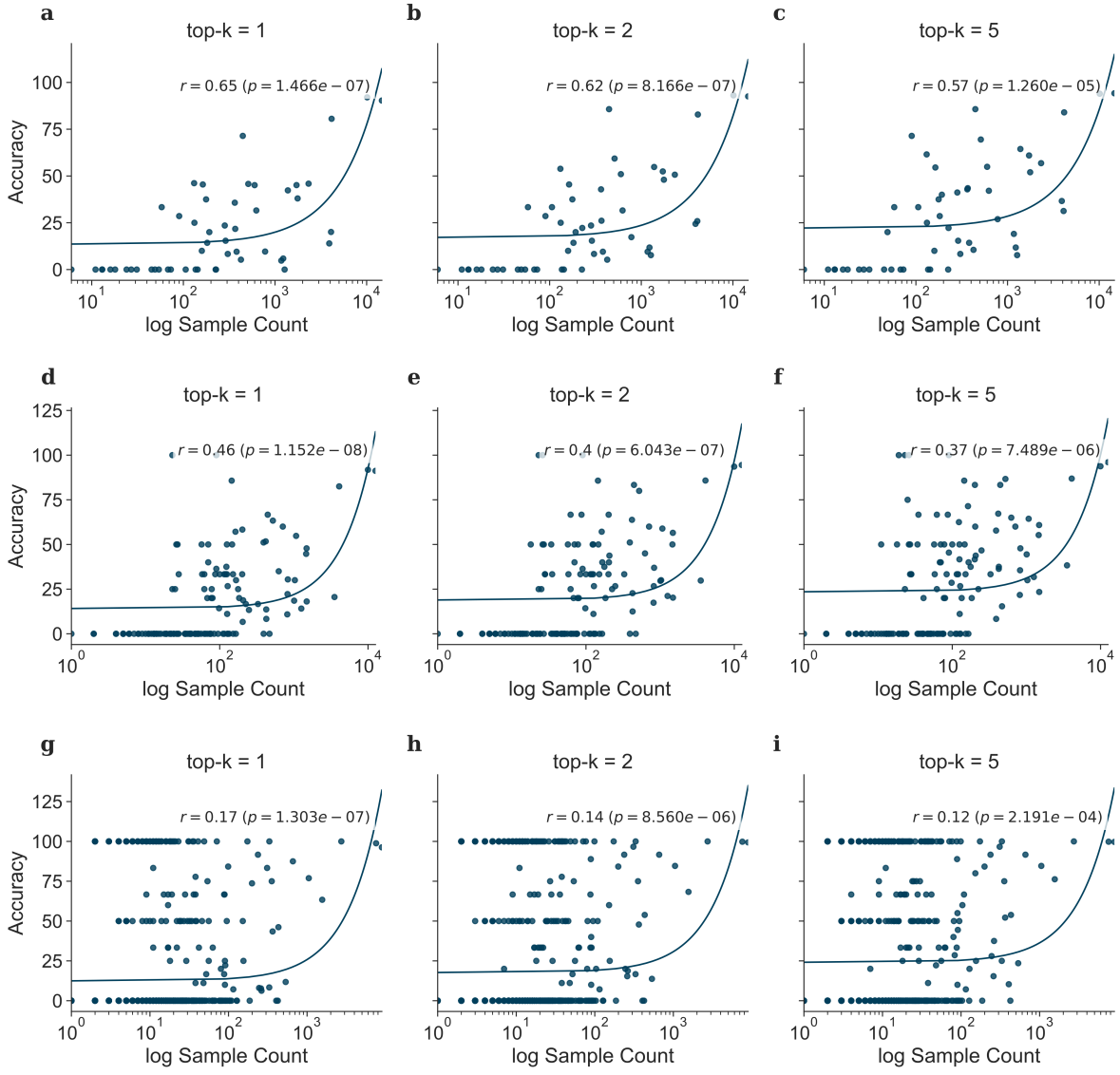

Figure S10: Correlation between backward prediction accuracy and sample count in *EC2* (**a**, **b**), *EC3* (**c**, **d**), and *EC4* (**e**, **f**). The trend towards lower correlations in higher EC-level token schemes is caused by a further reduction in test cases due to the selection of unique test products not found in the training sets and the resulting hit-or-miss accuracies appearing as bands at 0 and 100% accuracy, respectively. Increasing k results not only in increasing the accuracy but also in lowering the correlation.

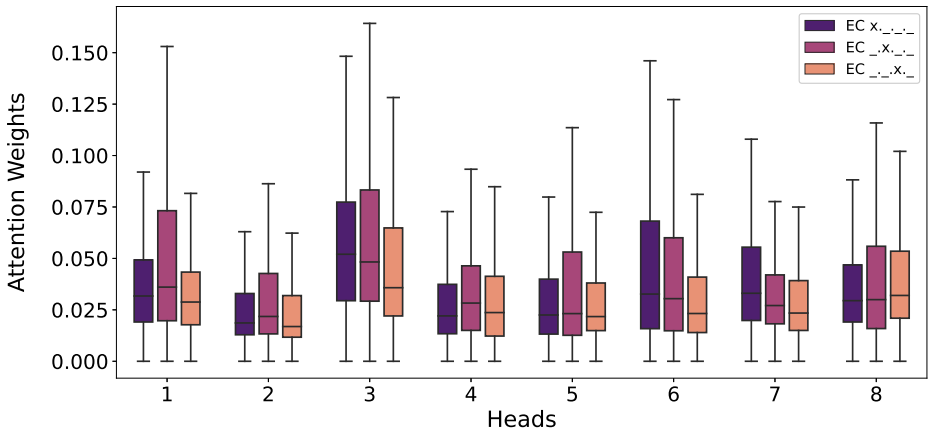

Figure S11: Average attention received EC-level 1-3 tokens for each head in the last decoder layer of the forward model considering all reactions in the test set. Although some heads focus on EC-level 3, the majority focuses on EC-levels 1 and 2 stressing their importance in the prediction of the enzymatic reaction outcome. The consistently high values computed for head 3 suggest its importance in the prediction.

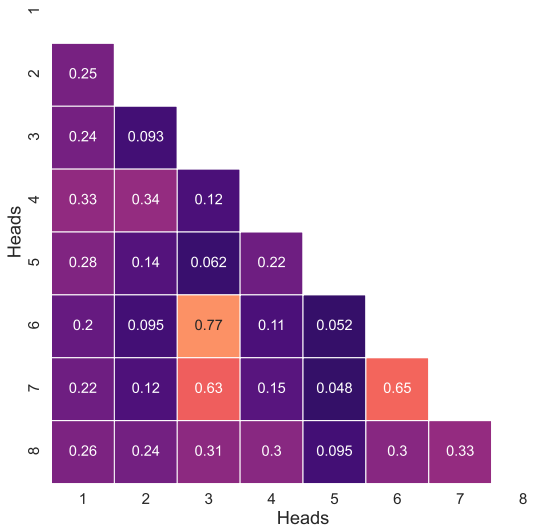

Figure S12: The correlation heatmap shows the similarity of the average attention weight received by the heads on the last layer of the decoder of the forward model. Three highly correlated heads (3, 6 and 7) emerge, highlighting preserved patterns among them (attention on single tokens). Other heads, e.g., 2 and 4, show weakly positive correlations highlighting additional preserved patterns (attention on larger groups of tokens). The remaining lower weights indicate the presence of specific patterns that are captured only by specific heads.

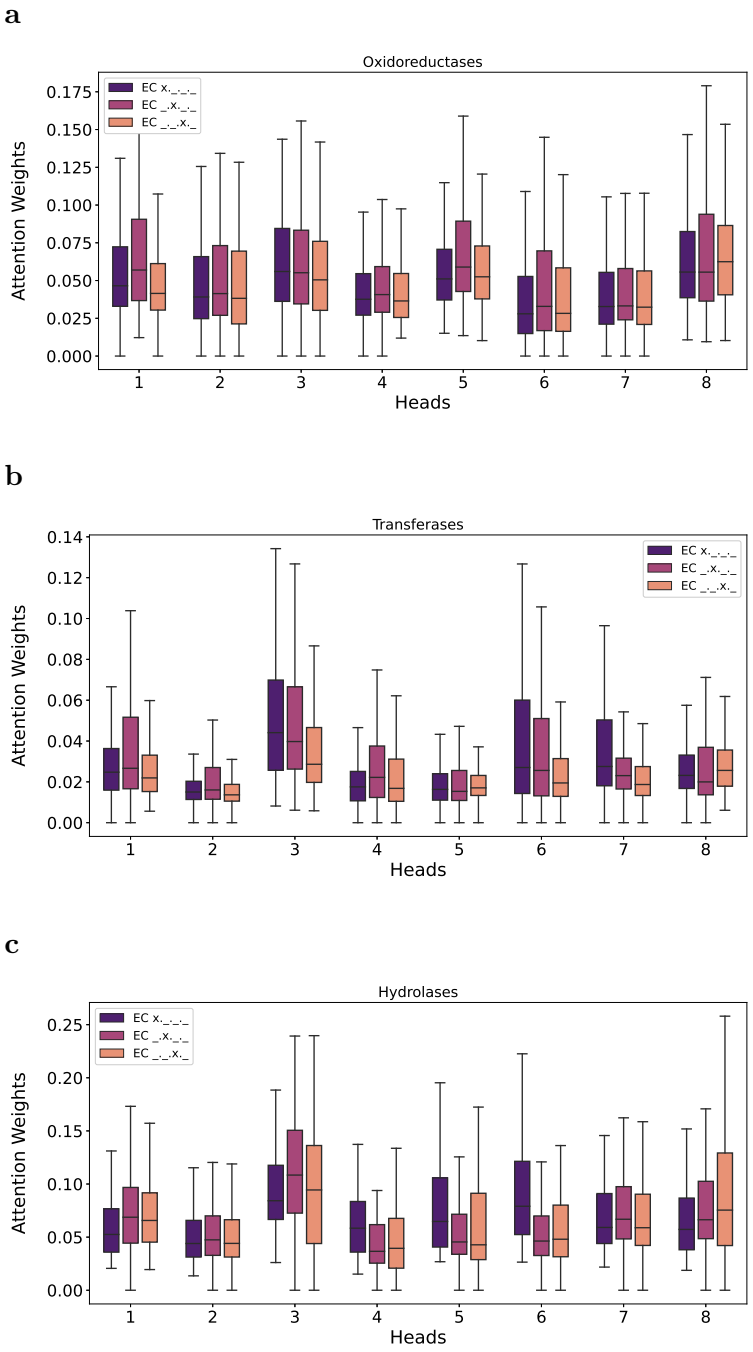

Figure S13: Average attention on the EC-level 1-3 tokens for each head using test reactions from the three most represented enzyme classes in the forward model: oxidoreductases (top), transferases (middle), hydrolases (bottom). The difference in the distributions highlight peculiar aspects of each class: oxidoreductases exhibit higher values, transferases relatively low ones, while hydrolases exhibit a more pronounced variability. In general we can appreciate how head 3 shows consistently larger values, unveiling its role in capturing enzymatic information.

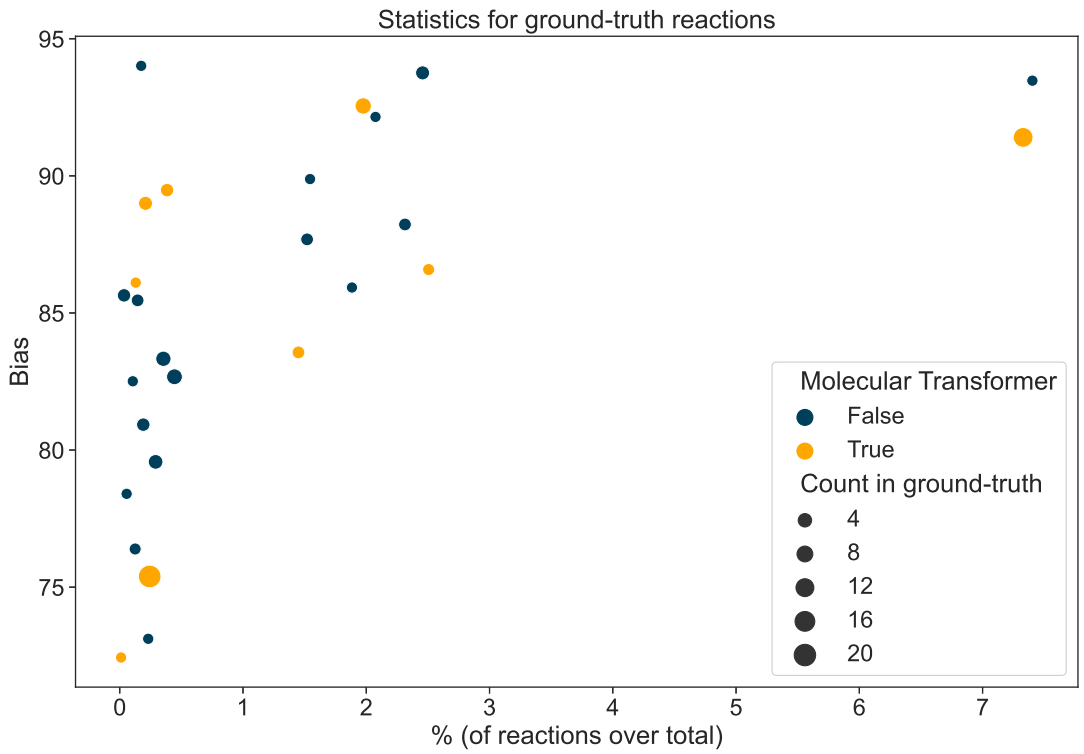

Figure S14: Summarized depiction of the most relevant statistics for the curated biocatalysed pathways from Finnigan[30]. In the scatter plot, each enzymatic reaction subclass at EC-level 3 is represented as a point. On the x-axis, we report the percentage of reactions in *ECREACT* belonging to the class. On the y-axis, we report a biased measure (between 0 and 100) for the EC-level 3 subclass, calculated using the Jensen-Shannon divergence[41] in base 2 between the distribution of EC-level 4 reaction subclasses and a baseline, defined as a uniform distribution of reactions in the EC-level 3 subclass. The bias measure the diversity in the EC-level 3 subclass considered. The point size encodes the number of EC-level 3 reaction subclasses reported in the set of enzymatic reactions from Finnigan[30]. Points are coloured based on the capability of the Molecular Transformer to find a successful route for at least one of the product considered. The depiction shows the high diversity of the reaction subclasses considered in the datasets (bias higher than 70 for all subclasses) and the low sample size for most of the reactions.

Graphical TOC Entry