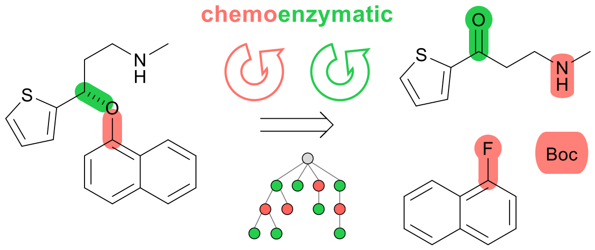
Chemoenzymatic Multistep Retrosynthesis with Transformer Loops

David Kreutter[a] and Jean-Louis Reymond*[a]

[a] *Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland*

e-mails: david.kreutter@unibe.ch
jean-louis.reymond@unibe.ch



Abstract

Integrating enzymatic reactions into computer-aided synthesis planning (CASP) should help devise more selective, economical, and greener synthetic routes. Herein we report the triple-transformer loop algorithm with biocatalysis (TTLAB) as a new CASP tool for chemo-enzymatic multistep retrosynthesis. Single-step retrosyntheses are performed using two triple transformer loops (TTL), one trained with chemical reactions from the US Patent Office (USPTO-TTL), the second one obtained by multitask transfer learning combining the USPTO dataset with preparative biotransformations from the literature (ENZR-TTL). Each TTL performs single-step retrosynthesis independently by tagging potential reactive sites in the product, predicting for each site possible starting materials (T1) and reagents or enzymes (T2), and validating the predictions via a forward transformer (T3). TTLAB combines predictions from both TTLs to explore multistep sequences using a heuristic best-first tree search and propose short routes from commercial building blocks including enantioselective biocatalytic steps. TTLAB can be used to assist chemoenzymatic route design.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Introduction

Computer-aided synthesis planning (CASP), originally proposed by E. J. Corey in the 1960’s, uses computational approaches, including rule-based systems as well as various types of neural networks, to exploit synthetic methodology as recorded in the scientific literature to propose multistep syntheses of target molecules from commercial precursors.[1–27] Integrating enzyme-catalyzed reactions would enable CASP to participate in the global effort towards more selective, economical, and greener chemical manufacturing processes. However, the task is challenging due to the sparsity and very different nature of biotransformations compared to chemical reactions.[28–33] Both template-based and transformer-based CASP tools for biocatalysis were recently reported,[34–37] which make use of biochemical reaction data describing mostly metabolic pathways as collected in databases such as BRENDA, KEGG, MetaCyc, Rhea, PathBank, MetaNetX or EzCatDB.[38–44] However, these biochemical pathway datasets only partly reflect the use of enzymes in organic synthesis, where enzymes or enzyme preparations (extracts, whole cells, etc.) are used under non-natural conditions, such as in immobilized form and at very high substrate concentrations, and to convert molecules often quite different from the natural substrate.[31]

We recently showed that CASP tools based on transformer models,[17,18] trained on SMILES descriptions[45,46] of chemical reactions of starting materials (SM) with a set of reagents (R) to form a product (P) as collected in the public USPTO dataset,[47,48] can be adapted to specific reaction subclasses by transfer learning.[49] Extending on this opportunity, we then showed that literature information on a few ten thousand biotransformations extracted from Reaxys,[50] for which the reagent set R is substituted with a text description of the enzyme or enzyme preparation, can be combined with the USPTO dataset to train a transformer model by multi-task transfer learning (MTL).[51] The resulting enzymatic transformer performed forward predictions of enzymatic reactions as used in typical

https://doi.org/10.26434/chemrxiv-2024-svr99 ORCID: https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

preparative biotransformations, including enantioselective processes such as kinetic resolution with lipases or enantioselective ketone reduction and reductive aminations with 71% top-2 accuracy, approaching the typical performance of forward transformer models.

Herein we report the integration of our enzymatic transformer model into our recently reported triple transformer loop algorithm (TTLA) for multistep chemical retrosynthesis,[52] to obtain a triple transformer loop algorithm with biocatalysis (TTLAB, **Figure 1**). The triple transformer loop (TTL) performs single-step retrosynthesis exploring diverse bond disconnections by tagging potential reactive sites in P to produce a series of P*,[53] and for each P* applying a first transformer T1 to predict SM, a second transformer T2 to predict a suitable R for the proposed transformation SM→P, and finally a third transformer T3 to predict P from the predicted SM and R, thereby potentially validating the retrosynthetic step. TTLAB combines the original triple transformer loop trained on USPTO (USPTO-TTL) with a second TTL obtained by MTL of USPTO reactions combined with 57,176 biotransformations collected from the literature (ENZR-TTL), which explores diverse enzymatic disconnections via a similar reactive site tagging approach. TTLAB considers single-step predictions from both USPTO-TTL and ENZR-TTL to explore multistep retrosyntheses using a heuristic best-first tree search. Possible routes are ranked with the route penalty score (RPscore),[52] combining the simplicity of all SM along the route,[54,55] with the confidence score of each retrosynthetic step, as well as route length. TTLAB can be used to assist chemoenzymatic route design.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

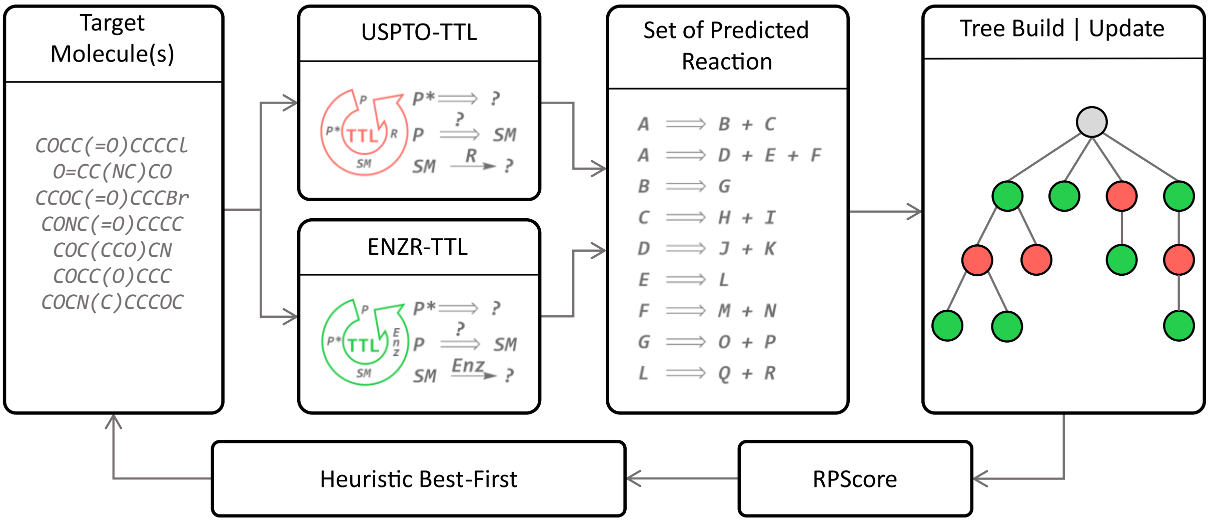


Figure 1. Concept of the TTLAB multistep search operating organic (USPTO-TTL) and enzymatic (ENZR-TTL) catalysis in parallel.

Methods

Chemical reaction dataset

The same United States Patent and Trademark Office (USPTO) chemical reaction dataset as in our previous report was used.[56] It is a version curated by Thakkar *et al.*[53] derived from the data mining work of Lowe.[47,48]

Triple Transformer Loop models for chemical reactions (USPTO-TTL)

The models trained on the USPTO dataset are identical as in our previous study and available on Zenodo,[56,57] and herein named USPTO-TTL. AutoTag is a tagging model predicting tagged product P* from the target product P. T1 is a disconnection-aware retrosynthesis model predicting starting materials SM from the target tagged product P*. T2 is a reaction condition model predicting reagents R, including catalyst and solvent, from the reaction SM→P. T3 is a forward validation model predicting P from SM + R.[58]

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Enzymatic dataset

The enzymatic reaction dataset, herein named ENZR, was extracted from Reaxys using the API accessible under a commercial license.[50] We first isolated reactions labelled as “enzymatic reaction” in the “other conditions” field (“RXD.COND”). Next, we compiled a list of reagents, catalysts, and solvents typically associated with enzymatic reactions. This involved identifying components with the "ase" suffix in the text fields "RXD.RGT," "RXD.CAT," and "RXD.SOL,". Additionally, we manually selected keywords corresponding to enzymatic transformations, such as "P450," "NADP," "CAL-B," "flavin mononucleotide," and others, from the most frequently occurring reagents and catalysts in the initial data retrieval. Finally, we queried these enzymatic components individually in the Reaxys database and retrieved the associated reactions. This process resulted in a raw dataset consisting of 107,865 enzymatic reactions.

Enzymatic dataset: cleaning

The process of cleaning the ENZR dataset involved several steps, wherein the RDKit library was used across various stages.[59] Initially, multistep reactions and those lacking any reactant or product were excluded, leaving 95,389 reactions. Next, reactions were mapped using RxnMapper,[60] for which 1,333 reactions failed and were removed. Reactions with unspecified atomic symbols (“*”) were also removed. Unmapped reactant molecules were removed for each reaction. A significant number of reactions (32,527) with more than one product were removed. The remaining reactions were tagged with reactive atoms as described previously,[56] and reactions with no tagged atoms, or with more than 10 tagged atoms, were removed. This cleaning process results in a final enzymatic dataset of 57,176 unique reactions SMILES[45,46] associated with textual descriptions of each reagent, including cofactors, enzymes, and solvent.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Enzymatic AutoTag and Triple Transformer Loop (ENZR-TTL) models

Enzymatic transformer models for the ENZR-TTL, including the AutoTag to tag reactive sites, and T1, T2 and T3 in the TTL itself, were trained using the USPTO and the ENZR dataset through MTL, similar to our previous Enzymatic Transformer model with identical training hyperparameters.[51] The split ratio 90:5:5 was applied as in the USPTO dataset resulting in 51,459 : 2,859 : 2,858 reactions in the training, validation, and test set respectively. The dataset split was done such that reactions resulting in identical products belong to the same splitting set.

During the MTL processes detailed below for all ENZR models, we incorporated instruction tokens. These tokens, "ENZYME" for the ENZR dataset and "USPTO" for the USPTO dataset, were inserted at both the start and end of the SMILES inputs. This addition aimed to provide additional context to the model and enable it to focus on specific prediction types as needed.

The ENZR-AutoTag model was trained to predict the tagged SMILES of the product (P*) from the product SMILES (P), in a similar manner to the USPTO-AutoTag model. The ENZR-T1 was trained to predict SM from P* for enzymatic retrosynthesis. In contrast, the ENZR-T2 model differs significantly from its USPTO-T2 counterpart by predicting a textual description of the enzyme (TDE) rather than reagents (R) in SMILES format from the theoretical reaction SMILES (SM→P). The ENZR-T3, previously reported as the Enzymatic Transformer,[51] serves as forward validation, it was trained from SM+TDE to predict P, now retrained using the new ENZR dataset.

Disconnection-aware automatic tagging strategy

In our previous study,[52] the USPTO-TTL employed a combination of three tagging strategies: (1) a systematic tagging procedure, tagging 1 to 3 neighbouring atoms,

https://doi.org/10.26434/chemrxiv-2024-svr99 ORCID: https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

(2) tagging templates of reactive sites with a conditional structure radius of 2 atoms, and (3) the AutoTag Transformer model with a beam size of 50.

The ENZR-TTL uses a specific tagging strategy combining only an AutoTag model[53] and templates, excluding the systematic tagging approach. The dedicated ENZR-AutoTag was trained from the ENZR dataset and USPTO by MTL. ENZR reactive site templates were extracted from ENZR exclusively with a radius of 2 atoms.

Chemoenzymatic multistep tree search algorithm

In parallel to the existing single-step USPTO-TTL, we added the ENZR-TTL, which the multistep algorithm uses systematically and independently. The prediction outcomes of both TTLs are provided to the heuristic best-first tree search, elaborating routes mixing the predictions of both TTLs. Confidence scores of both TTLs behaving differently, the confidence scores of ENZR-T3 were adapted by polynomial fit to the USPTO-T3 distribution (**Figure S1**) to ensure a fair scoring across TTLs. The RPScore, based on molecular simplicity[54,55] and confidence scores of T3 distinguishes which routes are the best to explore further, and functions the same as reported in our previous study.[52]

Our previous report of the Enzymatic Transformer model, herein named ENZR-T3, demonstrated that a confidence score threshold was required to filter unreasonable enzymatic reactions. A similar evaluation using the round-trip evaluation of the ENZR-TTL was performed and a threshold of 90% confidence of ENZR-T3 was defined for considering ENZR-TTL predictions for multistep retrosynthesis search.

Building block (BB) set

We combined MolPort (www.molport.com) and Enamine (www.enamine.net) databases to build a database of 534,058 commercially available compounds as the building block (BB) set.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Results and Discussion

Realizing the triple transformer loop algorithm with biocatalysis (TTLAB) for chemoenzymatic retrosynthesis required first to select a suitable dataset of enzymatic reactions, second to adapt our previous chemical reaction TTL to these enzymatic reactions, and finally to combine the enzymatic reaction TTL with the chemical reaction TTL in a multistep search algorithm. These steps are described in the following subsections.

Chemical and enzymatic reaction datasets and their comparison

We used the USPTO reaction dataset, which lists one million chemical reactions taken from the patent literature, as a broadly accepted selection of chemical reactions used in organic synthesis.[47,48] In terms of enzymatic reactions, we selected 57,176 enzymatic reactions from the scientific literature using the Reaxys API,[50] forming an enlarged version of our earlier enzymatic reaction dataset (ENZR, see methods for details).[51] The composition of this enlarged ENZR dataset is comparable to its smaller version and reflects the practice of biocatalysis in preparative organic chemistry as reported in the scientific literature, with lipases and dehydrogenases forming the largest class of enzymes (**Figure S2**).

In view of training transformer models for a combined chemoenzymatic retrosynthesis, we analyzed whether the 57,176 enzyme-catalyzed reactions in our ENZR dataset contained starting materials and products comparable to those in USPTO. We also analyzed the EREACT data,[37] which lists 62,222 enzyme-catalyzed reactions associated with their respective enzyme commission (EC) number, aggregated from the biochemical reaction pathways datasets Rhea, BRENDA, PathBank, and MetaNetX (**Table 1**).[38,42–44] ENZR listed fewer reactions than EREACT but more molecules, indicating a larger diversity of molecules tested in preparative biocatalysis compared to biochemical intermediates. Furthermore, ENZR shared a larger number of molecules with USPTO than EREACT, and only shared a

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

small number of molecules with EREACT. A similar distribution was observed when focusing only on reaction products, with only 2,470 molecules and 816 product molecules being shared between all three datasets (**Figure 2a/b**).

Table 1. Dataset information.

	USPTO	ENZR	ECREACT
Number of reactions	1,266,734	57,176	62,222
Number of unique molecules	1,493,418	76,645	45,944
Number (%) of molecules shared with USPTO	-	12,035 (15.7%)	3,502 (7.6%)
Number (%) of molecules shared with EREACT	3,502 (0.27%)	4,236 (7.4%)	-
Number of chiral molecules	271,504	45,277	34,177

To compare the three datasets in terms of molecule types, we selected 10,000 molecules randomly across starting materials and products in each dataset and constructed a TMAP,[61] employing the MinHashed atom-pair fingerprint MAP4 as similarity measure, which considers substructures and their relative position in molecules.[62] Areas of the TMAP covered by molecules from USPTO (green) also contained molecules from ENZR (orange), and to a lesser extent from EREACT (blue), showing a certain level of overlap in structural types between the three datasets (**Figure 2c**). Nevertheless, parts of the map were dominated by one of three datasets. Predominantly green areas (USPTO) contained drug-like heteroaromatic molecules, while predominantly orange areas (ENZR) featured glycosides and peptides. Furthermore, one fourth of the TMAP was standing out because it was entirely blue (EREACT) and was populated by phospholipids and triglycerides apparently completely absent from the other two datasets, probably reflecting the difficulty to work with such molecules in terms of preparative organic synthesis.

Histograms further highlighted similarities and differences between molecules composing the three datasets. A histogram of molecular size as heavy atom count (HAC) showed that ENZR and USPTO contained molecules of comparable size (10 ≤ HAC ≤ 40), while more than half of EREACT contained larger molecules (HAC > 40) (**Figure 2d**).

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Furthermore, a histogram of the fraction of cyclic bonds showed that USPTO contained mostly cyclic molecules, while ENZR contained similarly cyclic molecules but also a sizable fraction of entirely acyclic molecules, and EREACT was almost entirely composed of acyclic molecules (**Figure 2e**). The difference in molecule properties between the three datasets was also visible in scatter plots using molecular weight, the fraction of carbon atoms and the fraction of cyclic bonds as molecular descriptors (**Figure S3**). Note that 47.9% of ECREACT molecules contained a phosphate functional group, compared to 8.2% in ENZR molecules and only 0.5% in USPTO molecules, further highlighting the different nature of molecules involved in biochemical reaction pathways compared to those in use for synthetic chemistry.

Taken together, these comparisons showed that molecules in ENZR and USPTO datasets showed a significant level of overlap and might be useful for a transformer model approach for combined chemoenzymatic retrosynthesis. By contrast, the differences between EREACT and USPTO were more pronounced and suggested that these two datasets were almost incompatible with each other.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

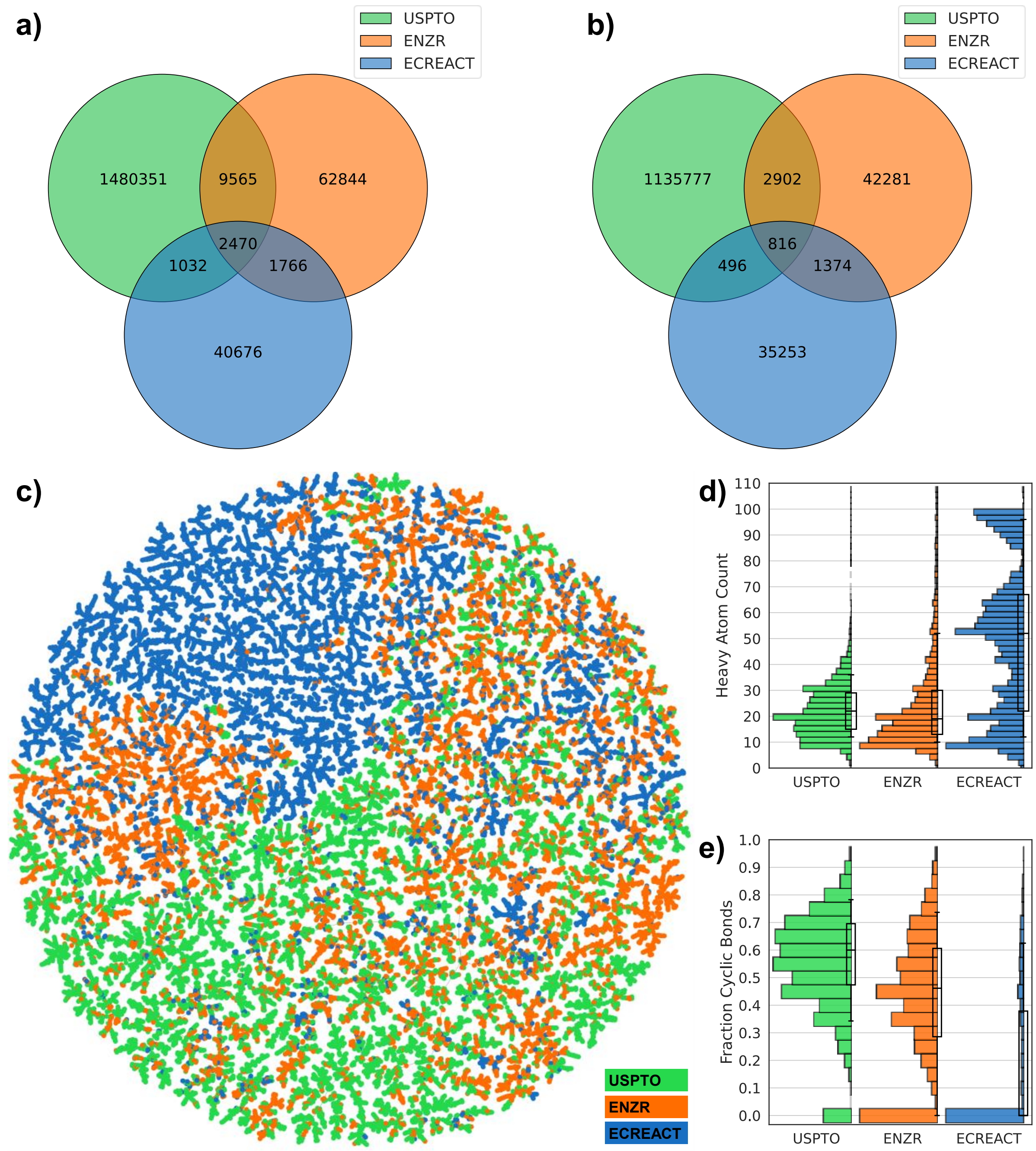


Figure 2. Comparative analysis of USPTO, ENRZ and EREACT datasets. **(a)** Venn diagram of all molecules in the USPTO, ENZR and the ECREACT datasets. (**b**) Venn diagram for only products (P) of reactions. **(c)** TMAP of 3×10,000 randomly chosen molecules from USPTO, ENZR and ECREACT datasets with similarities computed with the MAP4 fingerprint. The interactive map is available at https://tm.gdb.tools/TTLA/EnzymeDB.html. **(d)** Number of heavy atoms distribution for molecules in each dataset. **(e)** Fraction of cyclic bond distribution for molecules in each dataset.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Enzymatic triple transformer loop (ENZR-TTL)

Our TTL approach for single-step retrosynthesis consists of tagging potential reactive sites in the product molecule P to form a series of tagged P*, and for each P* to apply three subsequent transformer models predicting SM from P* (T1), reagents R from SM→P (T2), and finally product P from SM+R (T3). T3 validates the retrosynthetic step if the predicted P is identical to the input P, and the confidence score of the T3 prediction is used to compute the route penalty score (RPscore) for the multistep search.[52]

In our approach, potential reactive sites in the product molecule are first tagged to mark potential reactive sites. Our chemical reaction TTL used a combination of a transformer model, templates and systematic tagging. Due to the much higher substrate specificity of enzymes compared to chemical reagents, we removed the systematic tagging approach for our enzymatic TTL and only considered tagging with a transformer model and with templates. Reactive sites in product molecules of the ENZR dataset were identified from atom-mapping and labelled as previously described for the USPTO.[52] An ENZR-AutoTag transformer was then trained by MTL combining the tagged and untagged datasets of ENZR and USPTO. Enzymatic templates were extracted from the atom-mapped ENZR dataset considering only templates with a radius of two bonds around reacting atoms to take enzyme specificity into account, an aspect which was also reflected by the much smaller number of ENZR templates (18,083) compared to the number of USPTO templates (281,153).

To complement the transformer models for the chemical TTL trained with the USPTO dataset (here named USPTO-TTL), we used MTL of USPTO with the ENZR dataset using the previously described parameters[51] to obtain models for the enzymatic TTL (here named ENZR-TTL). To help the transformers to learn the differences between chemical and enzymatic reactions, all entries for MTL were labelled before and after the SMILES with "ENZYME" for ENZR data, and with "USPTO" for USPTO data. These labels helped to

https://doi.org/10.26434/chemrxiv-2024-svr99 ORCID: https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

avoid task ambiguity between USPTO vs. ENZR caused by the substitution of reagent SMILES with enzyme names in text format for T2 (SMILES→SMILES vs. SMILES→text) and T3 (SMILES→SMILES vs. SMILES+text→SMILES). The influence of the instruction tokens "ENZYME" and "USPTO" added before and after each input was well visible in the case of ENZR-T2, for which the fraction of textual enzyme description produced increased from 85.3% for an uninstructed model to 99.7% for the instructed model.

In terms of single-step round-trip accuracy,[58] the ENZR-TTL achieved 59.0% top-1 accuracy on the ENZR test set, somewhat below the 81.3% top-1 accuracy of the USPTO-TTL on the USPTO test set. In both cases, the top-1 round-trip accuracy measured the percentage of cases where P predicted by T3 matched the input P, which also included cases with different SM and R compared to the ground truth in the test sets (see details in **Table S1 and S2**). In both TTLs, the round-trip accuracy decreased as function of the number of tagged atoms. ENZR-TTL Top-3 round-trip accuracies were as high as 76.2% and 76.9% for single and double atom tags, compared to 94.1% and 92.8% in the case of USPTO-TTL (**Figure 3a**). The lower performance of ENZR-TTL compared to USPTO-TTL probably reflects the smaller training set of enzymatic reactions learned by transfer learning, and a more difficult task associated with the prediction of enzyme names in T2. As for the USPTO-T3, the confidence score of ENZR-T3 was correlated with the round-trip accuracy (**Figure 3b**). Analysis of test cases showed that a cut-off value of 90% had to be applied to select meaningful validated enzymatic retrosynthetic steps.

https://doi.org/10.26434/chemrxiv-2024-svr99 ORCID: https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

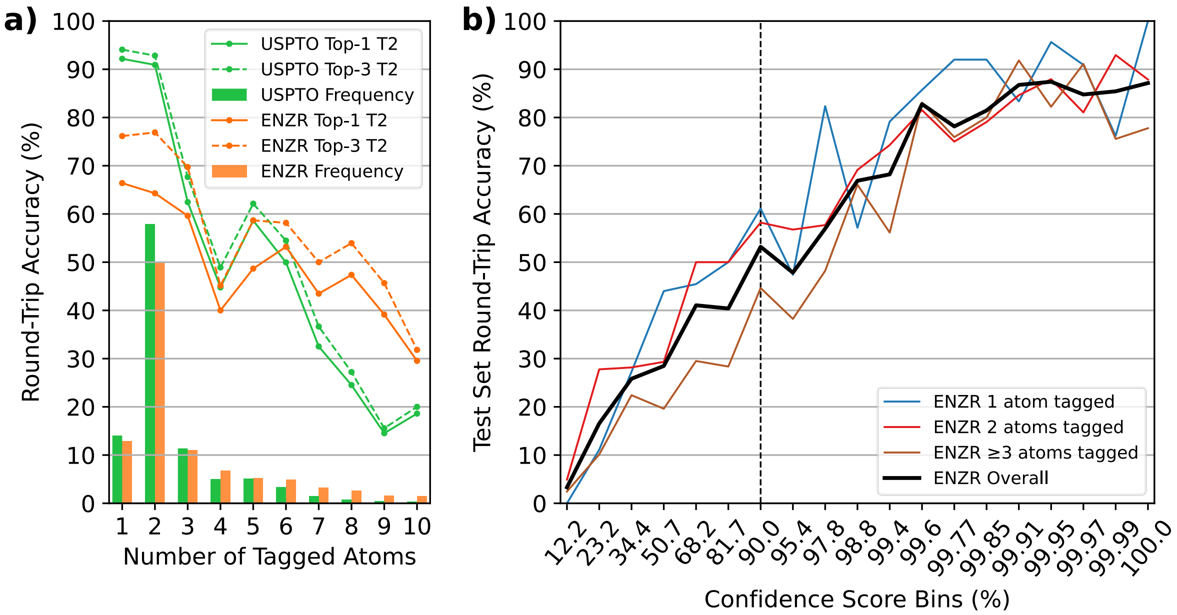


Figure 3. (a) Round-trip accuracies of ENZR-TTL and USPTO-TTL as function of the number of tagged atoms on the target molecules from the ENZR and USPTO test sets respectively. The top-N represents the round-trip accuracy considering multiple examples of enzyme textual descriptions predicted by ENZR-T2 or reagents predicted by USPTO-T2. The bar plots show the frequency fractions as function of the number of tagged atoms for both test sets. **(b)** Round-trip accuracy of ENZR-TTL as function of confidence scores of ENZR-T3. The vertical dashed bar represents the chosen confidence score cut-off. Bins were selected to equally distribute predictions.

Reaction examples illustrate the performance of ENZR-TTL in terms of single-step retrosynthesis. In many cases, T1 predicts the same SM as recorded in the ENZR dataset, T2 predicts the identical or almost identical enzyme description (with enzyme name, additive and solvents), and T3 predicts the correct P (**Figure 4** and **S4**). These include enantioselective reactions with non-biochemical substrates (reaction (**1**)),[63] cofactors (reaction (**2**))[64] and cofactor regeneration systems (reaction (**3**),[65] here with a different T2 output), as well as lipase-catalyzed reactions such as kinetic resolutions by acylation (reaction (**4**))[66] and heterocycle formations exploiting the catalytic promiscuity of lipases (reaction (**5**)).[67]

https://doi.org/10.26434/chemrxiv-2024-svr99 ORCID: https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

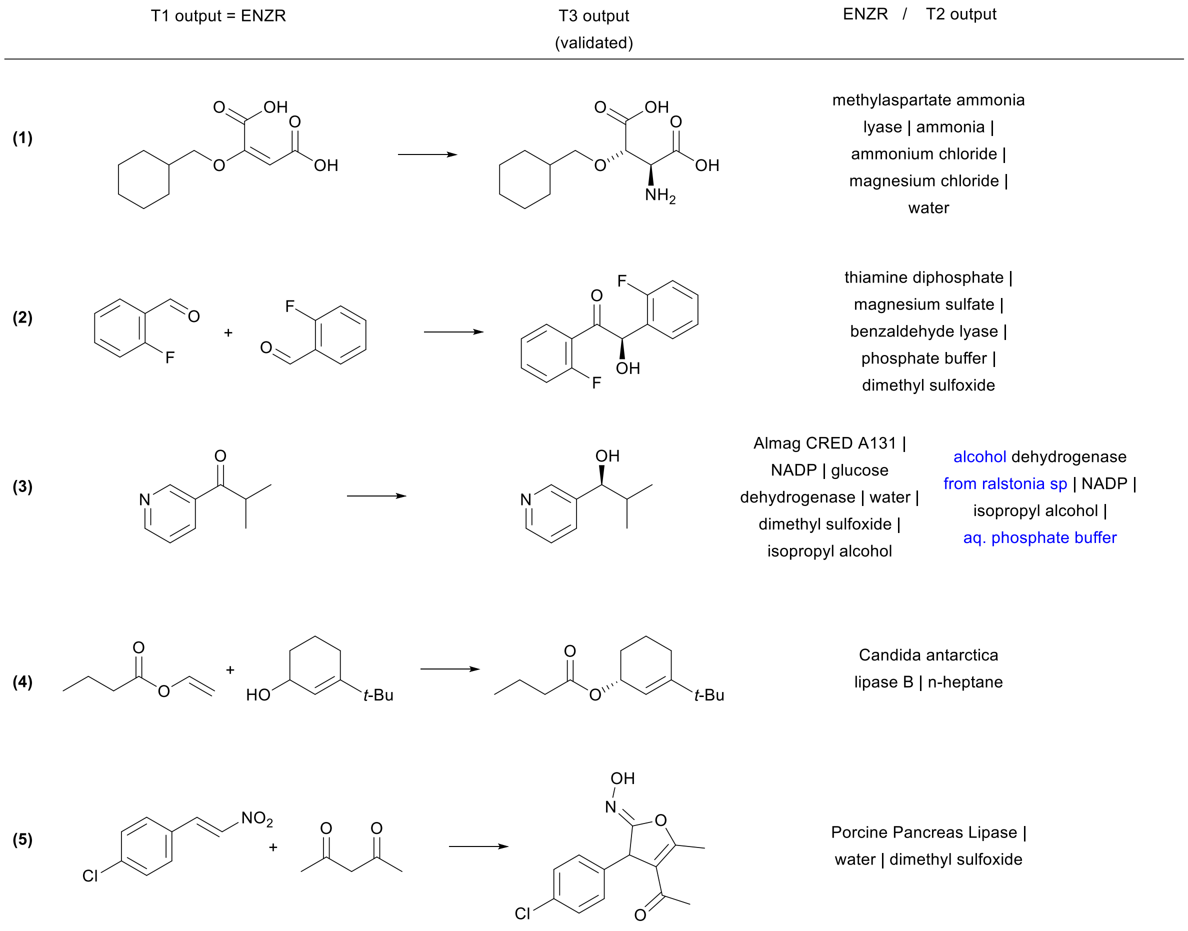

T1 output = ENZR
T3 output (validated)
ENZR / T2 output
(1)
methylaspartate ammonia lyase | ammonia | ammonium chloride | magnesium chloride | water
(2)
thiamine diphosphate | magnesium sulfate | benzaldehyde lyase | phosphate buffer | dimethyl sulfoxide
(3)
Almag CRED A131 | NADP | glucose dehydrogenase | water | dimethyl sulfoxide | isopropyl alcohol
alcohol dehydrogenase from ralstonia sp | NADP | isopropyl alcohol | aq. phosphate buffer
(4)
t-Bu
Candida antarctica lipase B | n-heptane
(5)
Porcine Pancreas Lipase | water | dimethyl sulfoxide

Figure 4. Examples of correctly predicted enzymatic single-step retrosynthesis by ENZR-TTL. The confidence scores of T3 are >99.5% in all cases. Enzyme names from the T2 output that differ from the database entry are highlighted in blue.

Validated retrosyntheses by ENZR-TTL include cases where the SM output by T1 and sometimes the enzyme name output by T2 are different from those recorded in ENZR, with interesting cases of reactions involving ketones and aldehydes as SM or P (**Figure 5** and **S5**). In one case, the T1 output specifies alcohol chirality for a fatty acid alcohol dehydrogenase reported to be non-enantioselective (although without providing primary data, reaction (**6**)),[68] whereby T1 probably infers alcohol chirality from other alcohol dehydrogenases. In another case, a chiral cyclobutanol is proposed by ENZR-TTL to be obtained by reduction of the parent ketone by a microbial dehydrogenase, while the database case involves baker's yeast and a ketal precursor of the cyclobutanone in aqueous pH 2, under which conditions the ketal spontaneously hydrolyzes to give the ketone (reaction (**7**)).[69] Furthermore, a (2-

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

chlorophenyl)-ketoacid recorded in ENZR to be formed by enzymatic oxidation of the corresponding mandelic acid,[70] is predicted by ENZR-TTL to stem from a transaminase reaction from the parent phenylglycine, a known type of biotransformation (reaction (**8**)).[71]

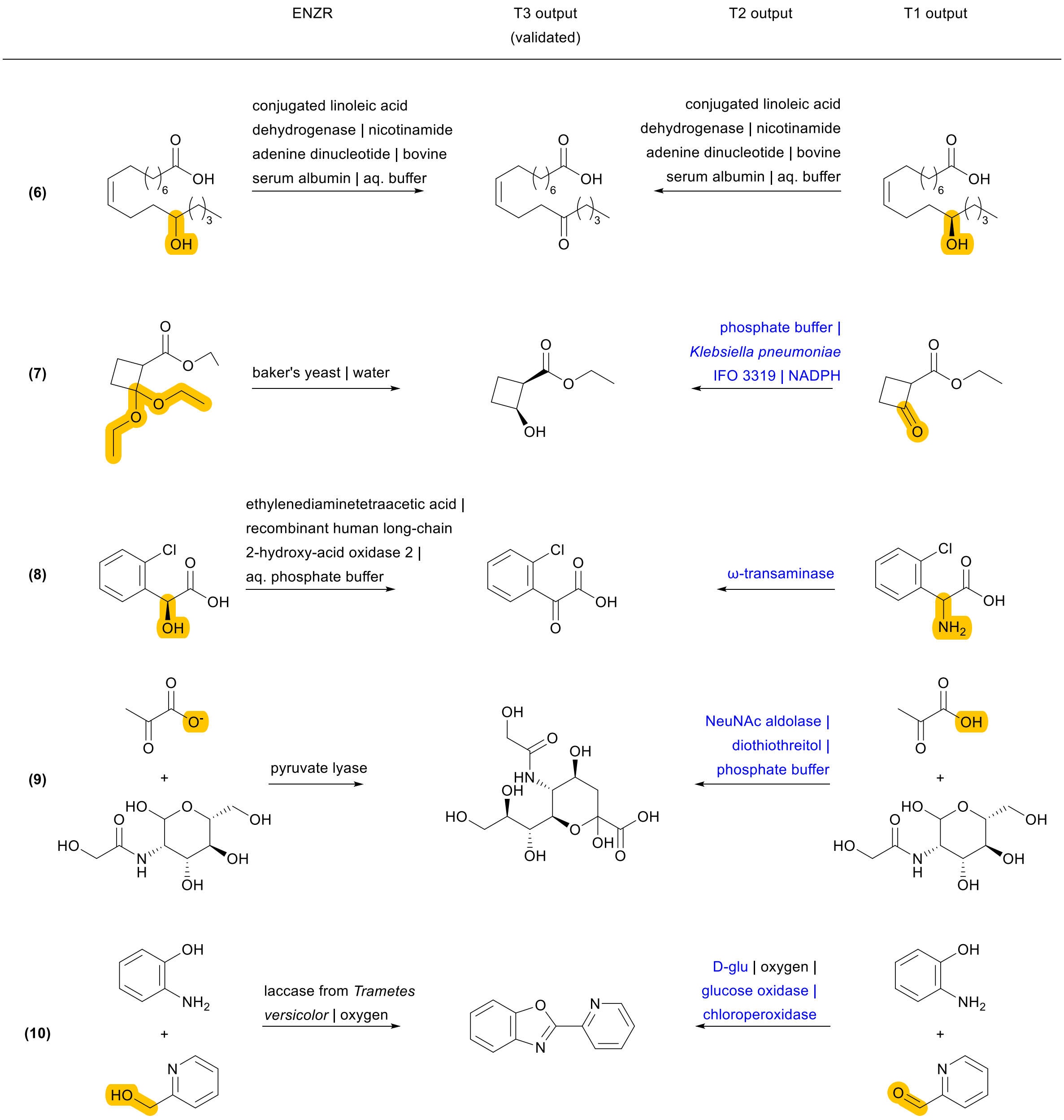

ENZR
T3 output (validated)
T2 output
T1 output
(6)
conjugated linoleic acid dehydrogenase | nicotinamide adenine dinucleotide | bovine serum albumin | aq. buffer
conjugated linoleic acid dehydrogenase | nicotinamide adenine dinucleotide | bovine serum albumin | aq. buffer
(7)
baker's yeast | water
phosphate buffer | Klebsiella pneumoniae IFO 3319 | NADPH
(8)
ethylenediaminetetraacetic acid | recombinant human long-chain 2-hydroxy-acid oxidase 2 | aq. phosphate buffer
ω-transaminase
(9)
pyruvate lyase
NeuNAc aldolase | diothiothreitol | phosphate buffer
(10)
laccase from Trametes versicolor | oxygen
D-glu | oxygen | glucose oxidase | chloroperoxidase

Figure 5. Examples of ENZR-TTL retrosynthetic steps validated by T3 involving different precursors and/or enzymes than those in ENZR. Structural differences between SM database entry and T1 output are highlighted in orange and enzyme names from T2 output that differ from the database entry are highlighted in blue.

https://doi.org/10.26434/chemrxiv-2024-svr99 ORCID: https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

Some discrepancies between ENZR data and ENZR-TTL output are caused by database entry mistakes and illustrate the self-correcting ability of the transformer model approach. For example, *N*-Acetylneuraminic acid is incorrectly recorded in ENZR as involving a "pyruvate lyase" due to an enzyme naming mistake in the corresponding publication (reaction (**9**)).[72] For this reaction ENZR-TTL correctly predicts that the enzymatic conversion of SM (*N*-acetyl-mannosamine and pyruvic acid) is carried out by the enzyme NeuNAc aldolase.[73] Similarly, the oxidative condensation of 2-pyridylmethanol with 2-aminophenol listed in Reaxys as an enzymatic process and recorded in ENZR (reaction (**10**)) actually involves TEMPO (2,2,6,6-tetramethylpiperidine-1-oxyl) as a chemical oxidant, which is recycled by air oxidation using laccase as enzyme but was not recorded in Reaxys.[74] Here, ENZR-TTL proposes pincolinaldehyde and 2-aminophenol as SM and a true enzymatic process using glucose oxidase and chloroperoxidase. This bi-enzymatic process has been reported for the related oxidative condensation of benzaldehyde and several *para*-substituted benzaldehydes with 2-aminophenol to form benzoxazoles.[75]

Finally, some incorrect cases involve a correct SM prediction by T1, but a different choice of enzyme by T2, resulting in a valid biotransformation but a different product P predicted by T3, and a non-validated reaction in terms of round-trip accuracy of ENZR-TTL (**Figure 6** and **S6**). For example, the correct phenolic SM is predicted by T1 for the formation of an *O*-methylated macrolactone (reaction (**11**)). However, T2 selects a different *O*-methyl transferase enzyme with a different regioselectivity, and therefore T3 predicts a different regioselectivity for the methylation. Note however that the proposed product is the correct one for the selected enzyme, as recorded in the same original publication focusing on tuning *O*-methylation regioselectivity.[76] In a related case of a chiral propargylic alcohol stemming from reduction of the corresponding ketone by an alcohol dehydrogenase, T1 predicts the correct SM but a change of enzyme choice by T2 results in a T3 prediction of P with the opposite enantioselectivity, which is correct for the selected enzyme but incorrect relative to

https://doi.org/10.26434/chemrxiv-2024-svr99 ORCID: https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

database entry (reaction (**12**)).[77] A similar different enzyme choice by T2 resulting in an enantiomeric P correctly predicted by T3 also occurs for the addition of hydrogen cyanide to cyclohexane carbaldehyde catalyzed by two different hydroxynitrile lyases (reaction (**13**)).[78,79]

In a related case involving tryptophan synthase, T1 predicts the correct SM, T2 the correct enzyme, and T3 the correct L-enantiomer, however the database entry lists the D-enantiomer, which was obtained by coupling tryptophan synthase with a stereoinversion cascade involving two enzymes that were not listed in the database entry (reaction (**14**)).[80,81] In a similar enzymatic cascade yielding 2-(2-naphthyl)propylamine from an epoxide precursor, T1 predicts the correct epoxide SM but combines styrene oxide isomerase with a different transaminase producing the (*R*)-enantiomeric P. By contrast, the database entry for P has an undefined stereochemistry, probably because the parent publications tested various transaminases with different enantioselectivities (reaction (**15**)).[82,83]

Taken together, the above analysis showed that biocatalytic retrosynthesis predictions by ENZR-TTL were generally relevant and sometimes even corrected inaccuracies in database entries. Encouraged by these data, we moved on to test multi-step chemoenzymatic retrosyntheses with our TTL approach.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

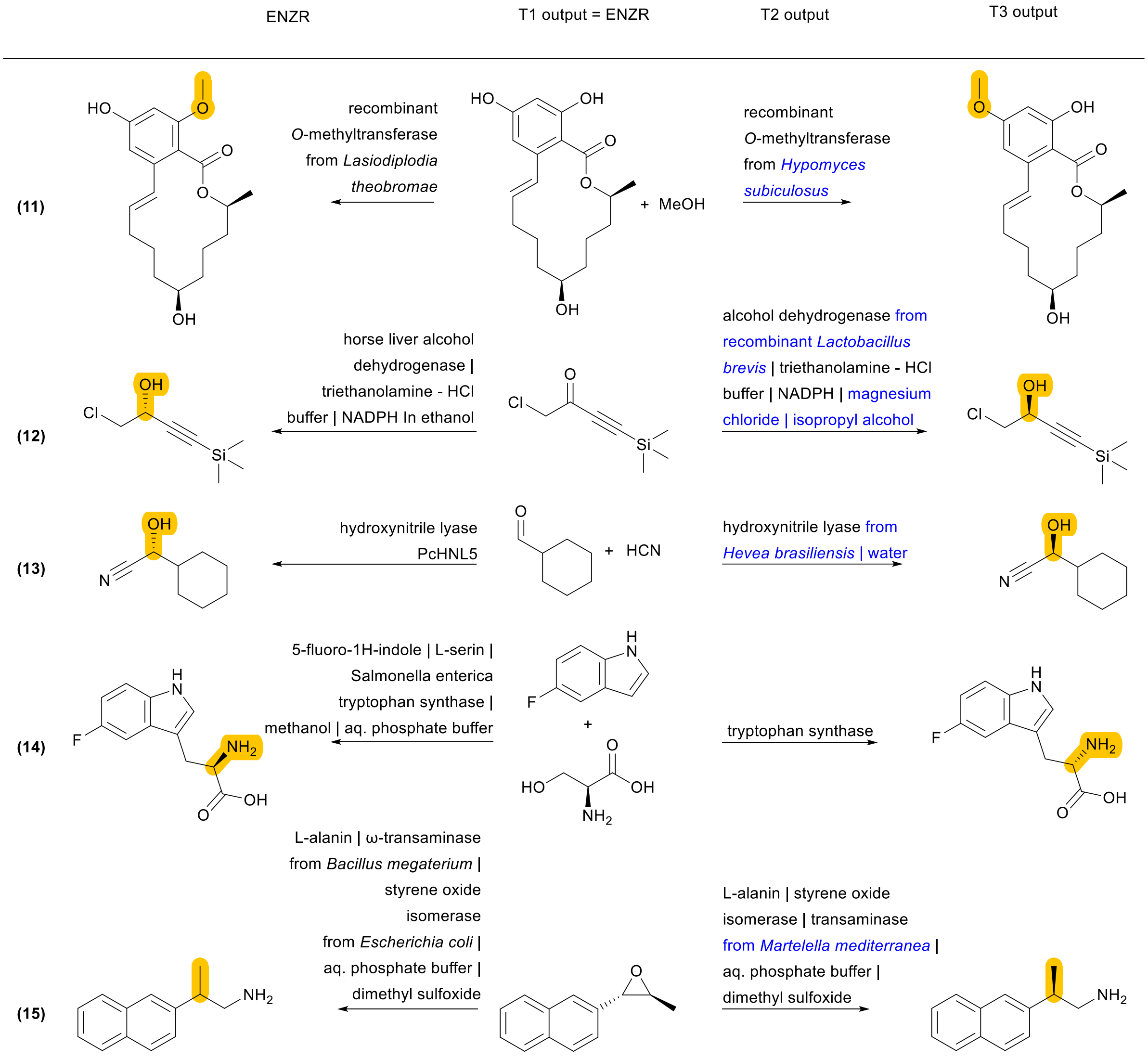

ENZR
T1 output = ENZR
T2 output
T3 output
(11)
recombinant O-methyltransferase from Lasiodiplodia theobromae
+ MeOH
recombinant O-methyltransferase from Hypomyces subiculosus
(12)
horse liver alcohol dehydrogenase | triethanolamine - HCl buffer | NADPH In ethanol
alcohol dehydrogenase from recombinant Lactobacillus brevis | triethanolamine - HCl buffer | NADPH | magnesium chloride | isopropyl alcohol
(13)
hydroxynitrile lyase PcHNL5
+ HCN
hydroxynitrile lyase from Hevea brasiliensis | water
(14)
5-fluoro-1H-indole | L-serin | Salmonella enterica tryptophan synthase | methanol | aq. phosphate buffer
tryptophan synthase
(15)
L-alanin | ω-transaminase from Bacillus megaterium | styrene oxide isomerase from Escherichia coli | aq. phosphate buffer | dimethyl sulfoxide
L-alanin | styrene oxide isomerase | transaminase from Martelella mediterranea | aq. phosphate buffer | dimethyl sulfoxide

Figure 6. Examples of ENZR-TTL prediction involving a correct SM prediction by T1 but a different enzyme choice by T2 and therefore a different product P compared to the database entry. Structural differences between P from database entry and T3 output are highlighted in orange and enzyme names from T2 output that differ from the database entry are highlighted in blue.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Chemoenzymatic multistep retrosynthesis with TTLAB

Integrating ENZR-TTL alongside the previously reported USPTO-TTL provided the chemo-enzymatic retrosynthesis prediction system, named TTLAB (**Figure 1**). To ensure the reliability of the enzymatic steps selected by TTLAB, a confidence score filter of 90% was applied to ENZR-T3. This filter eliminated chemically incorrect enzymatic retrosynthetic steps which would otherwise be selected by the tree-search because they achieved a high RPScore due to a high degree of molecular simplification.

We challenged TTLAB to propose retrosyntheses for 100 product molecules from the USPTO test set and 80 product molecules from the ENZR test set. A retrosynthesis was judged successful whenever the reaction sequence went back to a SM molecule available in the BB set, which consisted of 534,058 commercially available compounds (see Methods for details). TTLAB proposed synthetic routes for 88 of the 100 USPTO test set product molecules and 61 of the 80 ENZR test set product molecules, and in almost all cases at least one of the proposed routes contained at least an enzymatic step (**Table S3**). For TTLAB-predicted syntheses of USPTO molecules, approximately 8% of the proposed steps were enzymatic. This percentage ranged from 17% to 50% for TTLAB predicted syntheses of ENZR molecules considering either all proposed syntheses or only Top-scoring ones (**Table S4**). The ability of TTLAB to identify short chemo-enzymatic synthetic routes was well visible when analyzing the number of steps per route as well as the number of enzymatic steps per route among the Top-5, Top-50, Top-500 or all routes for both USPTO and ENZR molecules (**Figure 7**).

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

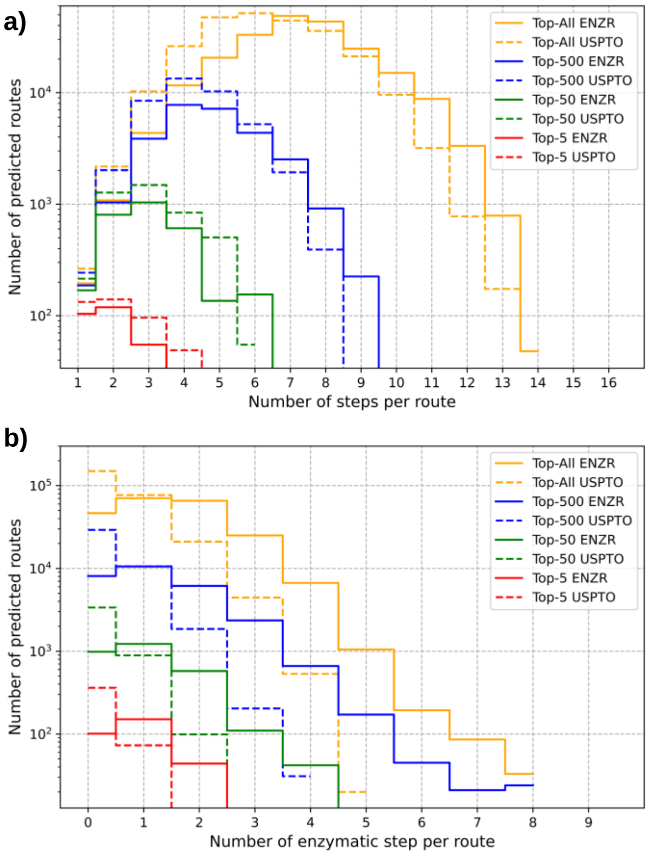


Figure 7. Analysis of synthetic routes predicted by TTLAB on product molecules from the USPTO and ENZR test sets. The route count as function of (**a**) the number of steps per route or (**b**) the number of enzymatic steps per route is given for the different Top-N categories.

The chemoenzymatic routes predicted by TTLAB are well illustrated by three examples from the ENZR test set, for which we show in each case the best RPScoring route including at least one enzymatic step (**Figure 8**). The first example is the predicted synthesis of the chiral cyanocarboxylic acid **1**, which was reported as the product of the enantioselective mono-hydrolysis of the prochiral dinitrile **2** by a mutant nitrilase enzyme.[84] TTLAB predicts the identical biotransformation as the first retrosynthetic operation, and proposes to assemble dinitrile **2** by Michael addition of cyanoacetic acid to unsaturated nitrile **3** and decarboxylation. Finally, TTLAB proposes to prepare nitrile **3** from the parent aldehyde **4**, which is a well-known type of transformation however using different reagents.[85]

https://doi.org/10.26434/chemrxiv-2024-svr99 ORCID: https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

The second example is the predicted synthesis of the phospha-C-peptide **5**, which was reported to be formed by coupling L-methionine ethyl ester with ethyl phosphinate **6** catalyzed by a phosphordiesterase.[86] TTLAB proposes the identical last step using the same enzyme. Since phosphinate **6** is not present in the commercial BB set, TTLAB further proposes a synthesis from vinyl glycine **7** by *N*-acetylation and esterification, done as a single step, followed by addition of ethyl methylphosphinate to the double bond. The latter reaction had been reported to prepare L-phosphinothricin, a naturally occurring herbicidal amino acid, however TTLAB omits to list the required radical initiator *tert*-butyl *per*-2-ethylhexanoate.[87]

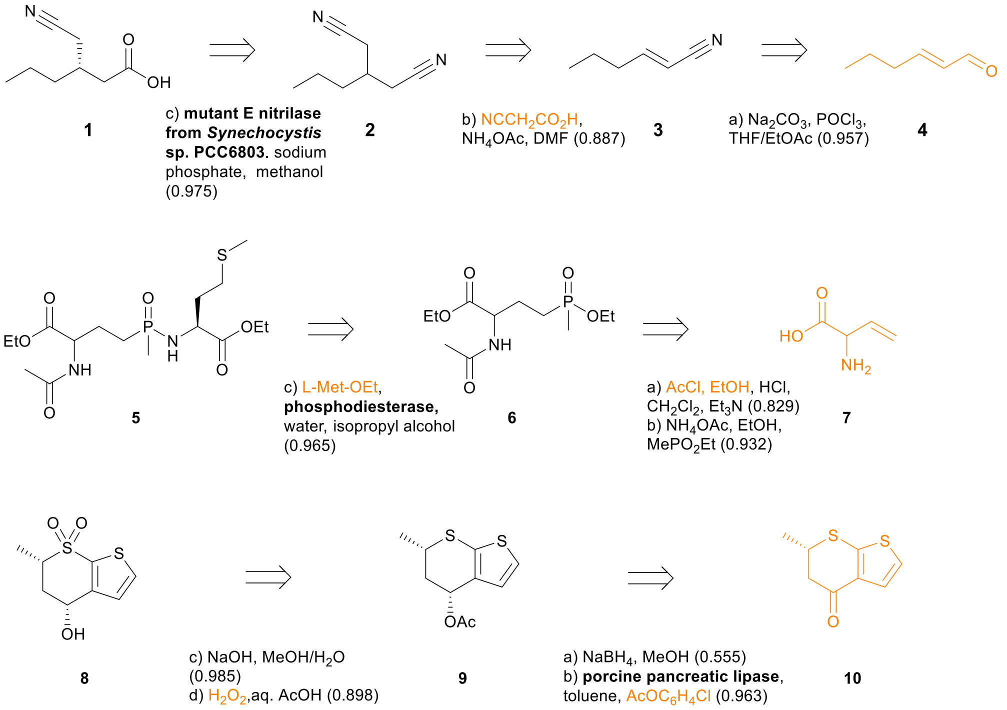

1
c) mutant E nitrilase from Synechocystis sp. PCC6803. sodium phosphate, methanol (0.975)
2
b) NCCH$_2$CO$_2$H, NH$_4$OAc, DMF (0.887)
3
a) Na$_2$CO$_3$, POCl$_3$, THF/EtOAc (0.957)
4
5
c) L-Met-OEt, phosphodiesterase, water, isopropyl alcohol (0.965)
6
a) AcCl, EtOH, HCl, CH$_2$Cl$_2$, Et$_3$N (0.829)
b) NH$_4$OAc, EtOH, MePO$_2$Et (0.932)
7
8
c) NaOH, MeOH/H$_2$O (0.985)
d) H$_2$O$_2$,aq. AcOH (0.898)
9
a) NaBH$_4$, MeOH (0.555)
b) porcine pancreatic lipase, toluene, AcOC$_6$H$_4$Cl (0.963)
10

Figure 8. Top RPScoring retrosyntheses predicted by TTLAB including at least one enzymatic step for three ENZR test set products. The confidence score of each predicted step is indicated in parentheses. Starting materials in the commercial BB set are written in orange.

The third example is chiral sulfone **8**, which TTLAB would prepare by deacetylation and sulfide oxidation of intermediate **9** using known chemistry.[88] Intermediate **9** would be formed by diastereoselective enzymatic acetylation of the parent alcohol by porcine pancreatic lipase using *p*-chlorophenylacetate as acylating agent, a biotransformation reaction known from the

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

test set.[89] This parent alcohol would be formed by non-stereoselective reduction of ketone **10** using sodium borohydride. This reduction is predicted with low confidence by TTLAB because this reaction can in fact be performed stereoselectively using $LiAlH_4$.[90] Indeed, when the condition of an enzymatic step is not imposed, TTLAB readily proposes, as the second best RPScoring route, a two-step chemical synthesis of **8** from **10** by stereoselective reduction followed by thioether oxidation to the sulfone.

We further exemplify TTLAB in the prediction of chemoenzymatic retrosyntheses for three drugs with known chemoenzymatic routes (**Figure 9**). In these cases, TTLAB often identifies steps that are part of the training sets. For the first case of the cholesterol-lowering drug atorvastatin **11**, our algorithm proposes as best RPScoring route the acidic deprotection of the corresponding *tert*-butyl ester, which is a commercial building block. Imposing at least one enzymatic step results in a four-step sequence from a linear chiral keto-ester precursor **12**, for which the first step is an enzymatic reduction by an aldo-keto reductase which was evolved precisely for this purpose and is present in the TTLAB training set.[91] The overall TTLAB route design is similar to the chemoenzymatic process developed for this drug involving an enzymatic enantioselective reduction of ethyl cyanoacetoacetate as initial step.[92]

In the second case of the antidepressant (*S*)-duloxetine **13**, the top-RPScoring route with at least one enzymatic step predicted by TTLAB is the single-step demethylation of the commercial *N*,*N*-dimethyl analog **14** catalyzed by a laccase, and the second best is a three-step sequence involving Boc protection of the achiral ketone precursor **15a**, followed by enantioselective reduction with an alcohol dehydrogenase and arylation of the resulting alcohol with fluoronaphthalene. This route is similar to the published chemoenzymatic synthesis of this drug starting with *N*,*N*-dimethylketone **15b**,[93] also proposed by the chemo-enzymatic ASKOS CASP tool with the help of manual intervention.[94]

https://doi.org/10.26434/chemrxiv-2024-svr99 ORCID: https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

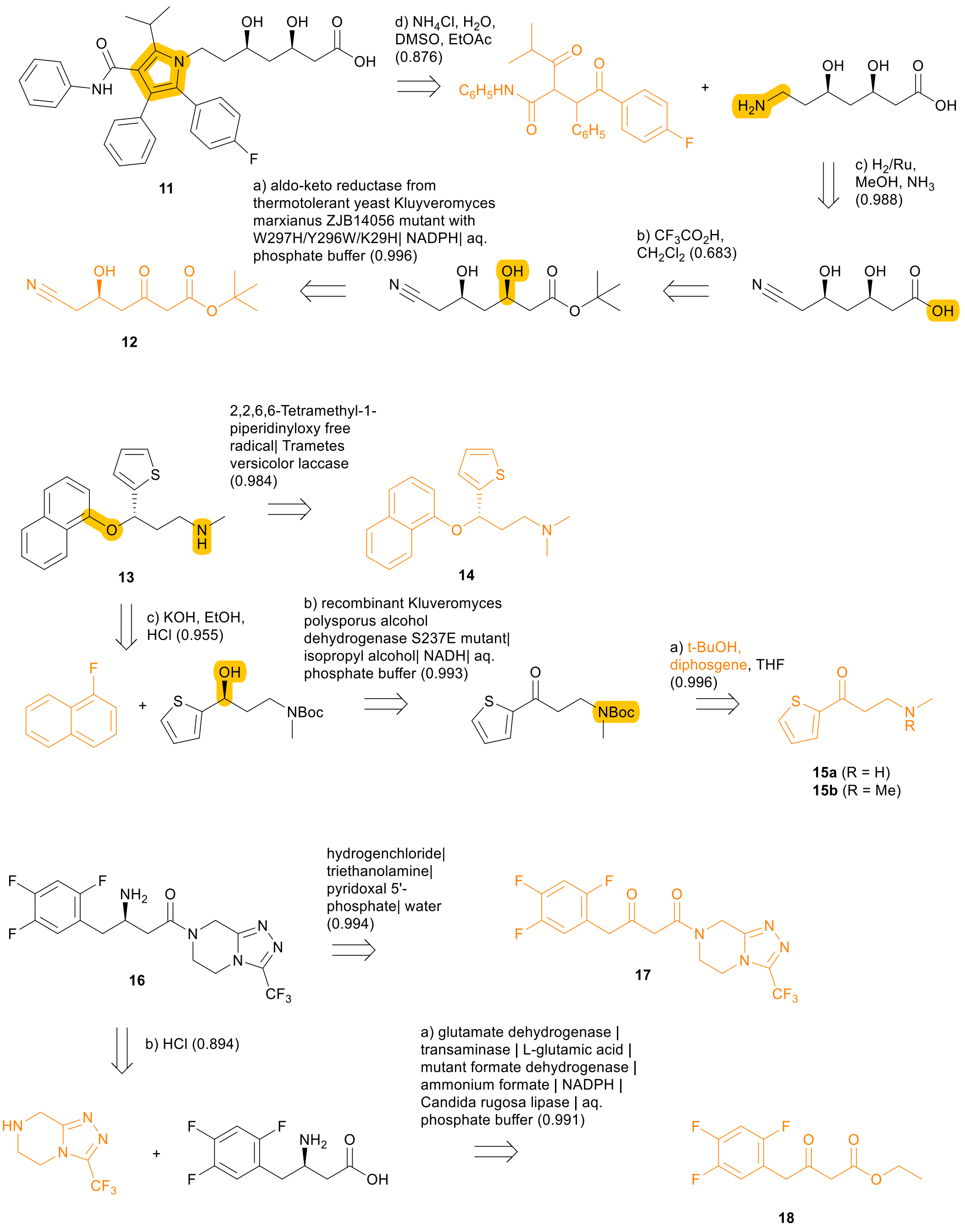

11
d) NH_4Cl, H_2O, DMSO, EtOAc (0.876)
c) H_2/Ru, MeOH, NH_3 (0.988)
b) CF_3CO_2H, CH_2Cl_2 (0.683)
a) aldo-keto reductase from thermotolerant yeast Kluyveromyces marxianus ZJB14056 mutant with W297H/Y296W/K29H| NADPH| aq. phosphate buffer (0.996)
12
13
2,2,6,6-Tetramethyl-1-piperidinyloxy free radical| Trametes versicolor laccase (0.984)
14
c) KOH, EtOH, HCl (0.955)
b) recombinant Kluveromyces polysporus alcohol dehydrogenase S237E mutant| isopropyl alcohol| NADH| aq. phosphate buffer (0.993)
a) t-BuOH, diphosgene, THF (0.996)
15a (R = H)
15b (R = Me)
16
hydrogenchloride| triethanolamine| pyridoxal 5'-phosphate| water (0.994)
17
b) HCl (0.894)
a) glutamate dehydrogenase | transaminase | L-glutamic acid | mutant formate dehydrogenase | ammonium formate | NADPH | Candida rugosa lipase | aq. phosphate buffer (0.991)
18

Figure 9. Retrosyntheses of atorvastatin (**11**), (*S*)-duloxetine (**13**) and sitagliptin (**16**) proposed by TTLAB. Reactive bonds and starting materials in the commercial BB set are drawn in orange. The confidence scores of individual retrosynthetic steps are indicated in parentheses after the predicted reagents.

In the third case of the DDP4 inhibitor sitagliptin (**16**) used to treat type II diabetes, TTLAB identifies a single-step enzymatic enantioselective retrosynthesis from the commercial β-ketoamide **17** using a transaminase. Although TTLAB only names the PLP cofactor in the reagents, this step is present in the ENZR training set using a transaminase that has been

https://doi.org/10.26434/chemrxiv-2024-svr99 ORCID: https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

engineered for the synthesis of this drug.[95] The second best RPScoring route is a similar two step sequence from the commercial ketoester **18** involving an enzymatic enantioselective reductive amination followed by amide bond formation. Note that the enzymatic step is part of the ENZR training set and uses the exact same combination of four enzymes for this biotransformation,[96] illustrating that transformer model ENZR-T2 memorizes enzyme textual description with high accuracy.

The above analysis and application examples show that TTLAB is able to propose short chemoenzymatic retrosyntheses for various target molecules. It should be noted that enzymatic steps are selected by TTLAB only when the reaction is closely related to a training set reaction, reflecting the fact that biocatalytic reactions are often highly specific for certain types of starting materials and are intrinsically poorly generalizable.

Conclusion

In summary, our work integrates biocatalysis in a computer-assisted synthesis planning (CASP) system, going towards greener and more sustainable chemistry. We achieved this by introducing a dual multistep retrosynthesis prediction system, integrating both chemical and biocatalytic steps. Trained on experimental enzymatic reactions from Reaxys, the enzymatic triple transformer loop operates in parallel to the chemocatalytic loop. The competitive framework, driven by the route penalty score (RPScore), drives the selection of optimal steps by our best-first tree search, incorporating both catalytic steps to generate mixed synthesis routes. Our results not only showcase the tool's capabilities in proposing viable solutions for drug-like molecules but also establish it as a valuable resource for synthesis design. Furthermore, the continuous enrichment of data in Reaxys promises ongoing enhancements in enzymatic capabilities, progressively going towards enzymatic synthesis.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Availability of data and materials

Code and instructions to compute multistep retrosynthesis as well as the code to tag reactive sites are available on our GitHub repository: https://github.com/reymond-group/MultiStepRetrosynthesisTTL

The original USPTO dataset can be found at https://doi.org/10.6084/m9.figshare.5104873.v1. The derived version of the USPTO dataset of Thakkar *et al*. can be found in their preprint.[53] The Reaxys enzymatic dataset is a licensed commercial database that cannot be made available.

Author contributions

DK designed and carried out the study and wrote the paper, JLR designed and supervised the study and wrote the paper.

Competing interests

The authors declare that they have no competing interests.

Acknowledgements

This work was supported financially by Novartis. We would like to thank Dr. Thierry Schlama for guidance and insightful advice, Dr. Daniel Kaufmann for his support, Dr. Frederic Stanger for internal testing of the retrosynthesis tool at Novartis, Dr. Radka Snajdrova, Dr. John Lopez, and Dr. Thomas Ruch for their helpful discussions, and Reaxys for access to the API for retrieving enzymatic reactions,. Calculations were performed on UBELIX (http://www.id.unibe.ch/hpc), the HPC cluster at the University of Bern.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

References

(1) Corey, E. J. General Methods for the Construction of Complex Molecules. *Pure and Applied Chemistry* **1967**, *14* (1), 19–38. https://doi.org/doi:10.1351/pac196714010019.

(2) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, *166* (3902), 178–192. https://doi.org/10.1126/science.166.3902.178.

(3) PENSAK, D. A.; COREY, E. J. LHASA—Logic and Heuristics Applied to Synthetic Analysis. In *Computer-assisted organic synthesis*; pp 1–32. https://doi.org/10.1021/bk-1977-0061.ch001.

(4) Corey, E. J.; Long, A. K.; Rubenstein, S. D. Computer-Assisted Analysis in Organic Synthesis. *Science* **1985**, *228* (4698), 408–418. https://doi.org/10.1126/science.3838594.

(5) Johnson, P. Y.; Burnstein, I.; Crary, J.; Evans, M.; Wang, T. Designing an Expert System for Organic Synthesis. In *Expert System Applications in Chemistry*; ACS Symposium Series; American Chemical Society, 1989; Vol. 408, pp 102–123. https://doi.org/10.1021/bk-1989-0408.ch009.

(6) Ihlenfeldt, W.-D.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angewandte Chemie International Edition in English* **1996**, *34* (23–24), 2613–2633. https://doi.org/10.1002/anie.199526131.

(7) Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y. Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation. *J. Chem. Inf. Model.* **2009**, *49* (3), 593–602. https://doi.org/10.1021/ci800228y.

(8) Christ, C. D.; Zentgraf, M.; Kriegl, J. M. Mining Electronic Laboratory Notebooks: Analysis, Retrosynthesis, and Reaction Based Enumeration. *J. Chem. Inf. Model.* **2012**, *52* (7), 1745–1756. https://doi.org/10.1021/ci300116p.

(9) Bøgevig, A.; Federsel, H.-J.; Huerta, F.; Hutchings, M. G.; Kraut, H.; Langer, T.; Löw, P.; Oppawsky, C.; Rein, T.; Saller, H. Route Design in the 21st Century: The ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction. *Org. Process Res. Dev.* **2015**, *19* (2), 357–368. https://doi.org/10.1021/op500373e.

(10) Nam, J.; Kim, J. Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions. *CoRR* **2016**, *abs/1612.09529*.

(11) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angewandte Chemie International Edition* **2016**, *55* (20), 5904–5937. https://doi.org/10.1002/anie.201506101.

(12) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chemistry – A European Journal* **2017**, *23* (25), 5966–5971. https://doi.org/10.1002/chem.201605499.

(13) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information*

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Processing Systems 30; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; pp 5998–6008.

(14) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3* (10), 1103–1113. https://doi.org/10.1021/acscentsci.7b00303.

(15) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3* (5), 434–443. https://doi.org/10.1021/acscentsci.7b00064.

(16) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698, 7698), 604–610. https://doi.org/10.1038/nature25978.

(17) Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. "Found in Translation": Predicting Outcomes of Complex Organic Chemistry Reactions Using Neural Sequence-to-Sequence Models. *Chem. Sci.* **2018**, *9* (28), 6091–6098. https://doi.org/10.1039/C8SC02339E.

(18) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Cent. Sci.* **2019**, *5* (9), 1572–1583. https://doi.org/10.1021/acscentsci.9b00576.

(19) Lee, A. A.; Yang, Q.; Sresht, V.; Bolgar, P.; Hou, X.; Klug-McLeod, J. L.; Butler, C. R. Molecular Transformer Unifies Reaction Prediction and Retrosynthesis across Pharma Chemical Space. *Chem. Commun.* **2019**, *55* (81), 12152–12155. https://doi.org/10.1039/C9CC05122H.

(20) Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. Datasets and Their Influence on the Development of Computer Assisted Synthesis Planning Tools in the Pharmaceutical Domain. *Chem. Sci.* **2019**, *11* (1), 154–168. https://doi.org/10.1039/C9SC04944D.

(21) Karpov, P.; Godin, G.; Tetko, I. V. A Transformer Model for Retrosynthesis. In *Artificial Neural Networks and Machine Learning – ICANN 2019: Workshop and Special Sessions*; Tetko, I. V., Kůrková, V., Karpov, P., Theis, F., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, 2019; pp 817–830. https://doi.org/10.1007/978-3-030-30493-5_78.

(22) Lin, K.; Xu, Y.; Pei, J.; Lai, L. Automatic Retrosynthetic Route Planning Using Template-Free Models. *Chem. Sci.* **2020**, *11* (12), 3355–3364. https://doi.org/10.1039/C9SC03666K.

(23) Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; Yang, Y. Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks. *J. Chem. Inf. Model.* **2020**, *60* (1), 47–55. https://doi.org/10.1021/acs.jcim.9b00949.

(24) Duan, H.; Wang, L.; Zhang, C.; Guo, L.; Li, J. Retrosynthesis with Attention-Based NMT Model and Chemical Analysis of "Wrong" Predictions. *RSC Adv.* **2020**, *10* (3), 1371–1378. https://doi.org/10.1039/C9RA08535A.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

(25) Strieth-Kalthoff, F.; Sandfort, F.; Segler, M. H. S.; Glorius, F. Machine Learning the Ropes: Principles, Applications and Directions in Synthetic Chemistry. *Chem. Soc. Rev.* **2020**, *49* (17), 6154–6168. https://doi.org/10.1039/C9CS00786E.

(26) Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. AiZynthFinder: A Fast, Robust and Flexible Open-Source Software for Retrosynthetic Planning. *Journal of Cheminformatics* **2020**, *12* (1), 70. https://doi.org/10.1186/s13321-020-00472-1.

(27) Thakkar, A.; Johansson, S.; Jorner, K.; Buttar, D.; Reymond, J.-L.; Engkvist, O. Artificial Intelligence and Automation in Computer Aided Synthesis Planning. *React. Chem. Eng.* **2021**, *6* (1), 27–51. https://doi.org/10.1039/D0RE00340A.

(28) Turner, N. J.; O'Reilly, E. Biocatalytic Retrosynthesis. *Nat. Chem. Biol.* **2013**, *9* (5), 285–288. https://doi.org/10.1038/nchembio.1235.

(29) Arnold, F. H. Directed Evolution: Bringing New Chemistry to Life. *Angew. Chem., Int. Ed. Engl.* **2018**, *57* (16), 4143–4148. https://doi.org/10.1002/anie.201708408.

(30) Sheldon, R. A.; Woodley, J. M. Role of Biocatalysis in Sustainable Chemistry. *Chem. Rev.* **2018**, *118* (2), 801–838. https://doi.org/10.1021/acs.chemrev.7b00203.

(31) Wu, S.; Snajdrova, R.; Moore, J. C.; Baldenius, K.; Bornscheuer, U. T. Biocatalysis: Enzymatic Synthesis for Industrial Applications. *Angew. Chem., Int. Ed. Engl.* **2020**, *59*, 2–34. https://doi.org/10.1002/anie.202006648.

(32) Bell, E. L.; Finnigan, W.; France, S. P.; Green, A. P.; Hayes, M. A.; Hepworth, L. J.; Lovelock, S. L.; Niikura, H.; Osuna, S.; Romero, E.; Ryan, K. S.; Turner, N. J.; Flitsch, S. L. Biocatalysis. *Nat Rev Methods Primers* **2021**, *1* (1), 1–21. https://doi.org/10.1038/s43586-021-00044-z.

(33) Gröger, H.; Gallou, F.; Lipshutz, B. H. Where Chemocatalysis Meets Biocatalysis: In Water. *Chem. Rev.* **2023**, *123* (9), 5262–5296. https://doi.org/10.1021/acs.chemrev.2c00416.

(34) 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* (2, 2), 98–104. https://doi.org/10.1038/s41929-020-00556-z.

(35) Zheng, S.; Zeng, T.; Li, C.; Chen, B.; Coley, C. W.; Yang, Y.; Wu, R. Deep Learning Driven Biosynthetic Pathways Navigation for Natural Products with BioNavi-NP. *Nat Commun* **2022**, *13* (1, 1), 3342. https://doi.org/10.1038/s41467-022-30970-9.

(36) Levin, I.; Liu, M.; Voigt, C. A.; Coley, C. W. Merging Enzymatic and Synthetic Chemistry with Computational Synthesis Planning. *Nat Commun* **2022**, *13* (1), 7747. https://doi.org/10.1038/s41467-022-35422-y.

(37) Probst, D.; Manica, M.; Nana Teukam, Y. G.; Castrogiovanni, A.; Paratore, F.; Laino, T. Biocatalysed Synthesis Planning Using Data-Driven Learning. *Nat Commun* **2022**, *13* (1), 964. https://doi.org/10.1038/s41467-022-28536-w.

(38) Schomburg, I.; Chang, A.; Schomburg, D. BRENDA, Enzyme Data and Metabolic Information. *Nucleic Acids Res* **2002**, *30* (1), 47–49.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

(39) Kanehisa, M. The KEGG Database. In *'In Silico' Simulation of Biological Processes*; John Wiley & Sons, Ltd, 2002; pp 91–103. https://doi.org/10.1002/0470857897.ch8.

(40) Karp, P. D.; Riley, M.; Paley, S. M.; Pellegrini-Toole, A. The MetaCyc Database. *Nucleic Acids Research* **2002**, *30* (1), 59–61. https://doi.org/10.1093/nar/30.1.59.

(41) Nagano, N. EzCatDB: The Enzyme Catalytic-Mechanism Database. *Nucleic Acids Research* **2005**, *33* (suppl_1), D407–D412. https://doi.org/10.1093/nar/gki080.

(42) 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* (Database issue), D754-760. https://doi.org/10.1093/nar/gkr1126.

(43) 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* (6), 815–816. https://doi.org/10.1093/bioinformatics/btt036.

(44) Wishart, D. S.; Li, C.; Marcu, A.; Badran, H.; Pon, A.; Budinski, Z.; Patron, J.; Lipton, D.; Cao, X.; Oler, E.; Li, K.; Paccoud, M.; Hong, C.; Guo, A. C.; Chan, C.; Wei, W.; Ramirez-Gaona, M. PathBank: A Comprehensive Pathway Database for Model Organisms. *Nucleic Acids Research* **2020**, *48* (D1), D470–D478. https://doi.org/10.1093/nar/gkz861.

(45) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28* (1), 31–36. https://doi.org/10.1021/ci00057a005.

(46) Weininger, D.; Weininger, A.; Weininger, J. L. SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29* (2), 97–101. https://doi.org/10.1021/ci00062a008.

(47) Lowe, D. M. Extraction of Chemical Structures and Reactions from the Literature. PhD Thesis, University of Cambridge, 2012. https://doi.org/10.17863/CAM.16293.

(48) Lowe, Daniel. Chemical Reactions from US Patents (1976-Sep2016). *figshare. dataset.* **2017**. https://doi.org/10.6084/m9.figshare.5104873.v1.

(49) 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), 4874. https://doi.org/10.1038/s41467-020-18671-7.

(50) Lawson, A. J.; Swienty-Busch, J.; Géoui, T.; Evans, D. The Making of Reaxys—Towards Unobstructed Access to Relevant Chemistry Information. In *The Future of the History of Chemical Information*; ACS Symposium Series; American Chemical Society, 2014; Vol. 1164, pp 127–148. https://doi.org/10.1021/bk-2014-1164.ch008.

(51) Kreutter, D.; Schwaller, P.; Reymond, J.-L. Predicting Enzymatic Reactions with a Molecular Transformer. *Chem. Sci.* **2021**, *12* (25), 8648–8659. https://doi.org/10.1039/D1SC02362D.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

(52) Kreutter, D.; Reymond, J.-L. Multistep Retrosynthesis Combining a Disconnection Aware Triple Transformer Loop with a Route Penalty Score Guided Tree Search. *Chemical Science* **2023**, *14* (36), 9959–9969. https://doi.org/10.1039/D3SC01604H.

(53) Thakkar, A.; Vaucher, A. C.; Byekwaso, A.; Schwaller, P.; Toniato, A.; Laino, T. Unbiasing Retrosynthesis Language Models with Disconnection Prompts. *ACS Cent. Sci.* **2023**. https://doi.org/10.1021/acscentsci.3c00372.

(54) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58* (2), 252–261. https://doi.org/10.1021/acs.jcim.7b00622.

(55) Schwaller, P.; Petraglia, R.; Zullo, V.; H. Nair, V.; Andreas Haeuselmann, R.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting Retrosynthetic Pathways Using Transformer-Based Models and a Hyper-Graph Exploration Strategy. *Chem. Sci.* **2020**, *11* (12), 3316–3325. https://doi.org/10.1039/C9SC05704H.

(56) Kreutter, D.; Reymond, J.-L. Multistep Retrosynthesis Combining a Disconnection Aware Triple Transformer Loop with a Route Penalty Score Guided Tree Search. *Chem. Sci.* **2023**. https://doi.org/10.1039/D3SC01604H.

(57) Kreutter, D.; Reymond, J.-L. Transformer Models for Disconnection-Aware Triple Transformer Loop, 2023. https://doi.org/10.5281/zenodo.8160148.

(58) Schwaller, P.; Petraglia, R.; Nair, V. H.; Laino, T. Evaluation Metrics for Single-Step Retrosynthetic Models. *Second Workshop on Machine Learning and the Physical Sciences (NeurIPS 2019)* **2019**.

(59) Landrum, G. RDKit: Open-Source Cheminformatics. **2006**.

(60) 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. *Nat Mach Intell* **2021**, *3* (2), 144–152. https://doi.org/10.1038/s42256-020-00284-w.

(61) Probst, D.; Reymond, J.-L. Visualization of Very Large High-Dimensional Data Sets as Minimum Spanning Trees. *Journal of Cheminformatics* **2020**, *12* (1), 12. https://doi.org/10.1186/s13321-020-0416-x.

(62) Capecchi, A.; Probst, D.; Reymond, J.-L. One Molecular Fingerprint to Rule Them All: Drugs, Biomolecules, and the Metabolome. *Journal of Cheminformatics* **2020**, *12* (1), 43. https://doi.org/10.1186/s13321-020-00445-4.

(63) Fu, H.; Zhang, J.; Tepper, P. G.; Bunch, L.; Jensen, A. A.; Poelarends, G. J. Chemoenzymatic Synthesis and Pharmacological Characterization of Functionalized Aspartate Analogues As Novel Excitatory Amino Acid Transporter Inhibitors. *J. Med. Chem.* **2018**, *61* (17), 7741–7753. https://doi.org/10.1021/acs.jmedchem.8b00700.

(64) Demir, A. S.; Şeşenoglu, Ö.; Eren, E.; Hosrik, B.; Pohl, M.; Janzen, E.; Kolter, D.; Feldmann, R.; Dünkelmann, P.; Müller, M. Enantioselective Synthesis of α-Hydroxy Ketones via Benzaldehyde Lyase-Catalyzed C−C Bond Formation Reaction. *Advanced Synthesis & Catalysis* **2002**, *344* (1), 96–103. https://doi.org/10.1002/1615-4169(200201)344:1<96::AID-ADSC96>3.0.CO;2-Z.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

(65) Rowan, A. S.; Moody, T. S.; Howard, R. M.; Underwood, T. J.; Miskelly, I. R.; He, Y.; Wang, B. Preparative Access to Medicinal Chemistry Related Chiral Alcohols Using Carbonyl Reductase Technology. *Tetrahedron: Asymmetry* **2013**, *24* (21), 1369–1381. https://doi.org/10.1016/j.tetasy.2013.09.015.

(66) Wagner, B.; Binder, F. P. C.; Jiang, X.; Mühlethaler, T.; Preston, R. C.; Rabbani, S.; Smieško, M.; Schwardt, O.; Ernst, B. A Structural-Reporter Group to Determine the Core Conformation of Sialyl Lewisx Mimetics. *Molecules* **2023**, *28* (6), 2595. https://doi.org/10.3390/molecules28062595.

(67) Wu, M.-Y.; Li, K.; He, T.; Feng, X.-W.; Wang, N.; Wang, X.-Y.; Yu, X.-Q. A Novel Enzymatic Tandem Process: Utilization of Biocatalytic Promiscuity for High Stereoselective Synthesis of 5-Hydroxyimino-4,5-Dihydrofurans. *Tetrahedron* **2011**, *67* (14), 2681–2688. https://doi.org/10.1016/j.tet.2011.01.060.

(68) Takeuchi, M.; Kishino, S.; Park, S.-B.; Kitamura, N.; Ogawa, J. Characterization of Hydroxy Fatty Acid Dehydrogenase Involved in Polyunsaturated Fatty Acid Saturation Metabolism in *Lactobacillus Plantarum* AKU 1009a. *J. Mol. Catal. B Enzym.* **2015**, *117*, 7–12. https://doi.org/10.1016/j.molcatb.2015.03.020.

(69) Buisson, D.; Azerad, R. Preparation and Use of (*S*)-*O*-Acetyllactyl Chloride (Mosandl's Reagent) as a Chiral Derivatizing Agent. *Tetrahedron: Asymmetry* **1999**, *10* (15), 2997–3002. https://doi.org/10.1016/S0957-4166(99)00285-2.

(70) Zhang, Y.; Su, C.; Lei, J.; Chen, L.; Hu, H.; Zeng, S.; Yu, L. Studies on the L-2-Hydroxy-Acid Oxidase 2 Catalyzed Metabolism of *S*-Mandelic Acid and Its Analogues. *DMPK* **2019**, *34* (3), 187–193. https://doi.org/10.1016/j.dmpk.2019.02.003.

(71) Zhu, D.; Hua, L. Biocatalytic Asymmetric Amination of Carbonyl Functional Groups – a Synthetic Biology Approach to Organic Chemistry. *Biotechnol. J.* **2009**, *4* (10), 1420–1431. https://doi.org/10.1002/biot.200900110.

(72) Pearce, O. M. T.; Varki, A. Chemo-Enzymatic Synthesis of the Carbohydrate Antigen *N*-Glycolylneuraminic Acid from Glucose. *Carbohydr. Res.* **2010**, *345* (9), 1225–1229. https://doi.org/10.1016/j.carres.2010.04.003.

(73) Li, Y.; Yu, H.; Cao, H.; Lau, K.; Muthana, S.; Tiwari, V. K.; Son, B.; Chen, X. Pasteurella Multocida Sialic Acid Aldolase: A Promising Biocatalyst. *Appl. Microbiol. Biotechnol.* **2008**, *79* (6), 963–970. https://doi.org/10.1007/s00253-008-1506-2.

(74) Mogharabi-Manzari, M.; Kiani, M.; Aryanejad, S.; Imanparast, S.; Amini, M.; Faramarzi, M. A. A Magnetic Heterogeneous Biocatalyst Composed of Immobilized Laccase and 2,2,6,6-Tetramethylpiperidine-1-Oxyl (TEMPO) for Green One-Pot Cascade Synthesis of 2-Substituted Benzimidazole and Benzoxazole Derivatives under Mild Reaction Conditions. *Adv. Synth. Catal.* **2018**, *360* (18), 3563–3571. https://doi.org/10.1002/adsc.201800459.

(75) Kumar, A.; Sharma, S.; Maurya, R. A. Bienzymatic Synthesis of Benzothia/(Oxa)Zoles in Aqueous Medium. *Tet. Lett.* **2010**, *51* (48), 6224–6226. https://doi.org/10.1016/j.tetlet.2010.06.012.

(76) Wang, X.; Wang, C.; Duan, L.; Zhang, L.; Liu, H.; Xu, Y.; Liu, Q.; Mao, T.; Zhang, W.; Chen, M.; Lin, M.; Gunatilaka, A. A. L.; Xu, Y.; Molnár, I. Rational Reprogramming

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

of O-Methylation Regioselectivity for Combinatorial Biosynthetic Tailoring of Benzenediol Lactone Scaffolds. *J. Am. Chem. Soc.* **2019**, *141* (10), 4355–4364. https://doi.org/10.1021/jacs.8b12967.

(77) Schubert, T.; Hummel, W.; Müller, M. Highly Enantioselective Preparation of Multifunctionalized Propargylic Building Blocks. *Angew. Chem., Int. Ed. Engl.* **2002**, *41* (4), 634–637. https://doi.org/10.1002/1521-3773(20020215)41:4<634::AID-ANIE634>3.0.CO;2-U.

(78) Griengl, H.; Klempier, N.; Pöchlauer, P.; Schmidt, M.; Shi, N.; Zabelinskaja-Mackova, A. A. Enzyme Catalysed Formation of (*S*)-Cyanohydrins Derived from Aldehydes and Ketones in a Biphasic Solvent System. *Tetrahedron* **1998**, *54* (48), 14477–14486. https://doi.org/10.1016/S0040-4020(98)00901-6.

(79) Zheng, Y.-C.; Ding, L.-Y.; Jia, Q.; Lin, Z.; Hong, R.; Yu, H.-L.; Xu, J.-H. A High-Throughput Screening Method for the Directed Evolution of Hydroxynitrile Lyase towards Cyanohydrin Synthesis. *Chembiochem* **2021**, *22* (6), 996–1000. https://doi.org/10.1002/cbic.202000658.

(80) Ge, H. M.; Yan, W.; Guo, Z. K.; Luo, Q.; Feng, R.; Zang, L. Y.; Shen, Y.; Jiao, R. H.; Xu, Q.; Tan, R. X. Precursor-Directed Fungal Generation of Novel Halogenated Chaetoglobosins with More Preferable Immunosuppressive Action. *Chem. Commun.* **2011**, *47* (8), 2321–2323. https://doi.org/10.1039/C0CC04183A.

(81) Parmeggiani, F.; Rué Casamajo, A.; Walton, C. J. W.; Galman, J. L.; Turner, N. J.; Chica, R. A. One-Pot Biocatalytic Synthesis of Substituted d-Tryptophans from Indoles Enabled by an Engineered Aminotransferase. *ACS Catal.* **2019**, *9* (4), 3482–3486. https://doi.org/10.1021/acscatal.9b00739.

(82) Xin, R.; See, W. W. L.; Yun, H.; Li, X.; Li, Z. Enzyme-Catalyzed Meinwald Rearrangement with an Unusual Regioselective and Stereospecific 1,2-Methyl Shift. *Angew. Chem., Int. Ed. Engl.* **2022**, *61* (28), e202204889. https://doi.org/10.1002/anie.202204889.

(83) See, W. W. L.; Li, X.; Li, Z. Biocatalytic Cascade Conversion of Racemic Epoxides to (S)-2-Arylpropionic Acids, (R)- and (S)-2-Arylpropyl Amines. *Advanced Synthesis & Catalysis* **2023**, *365* (1), 68–77. https://doi.org/10.1002/adsc.202201061.

(84) Yu, S.; Li, J.; Yao, P.; Feng, J.; Cui, Y.; Li, J.; Liu, X.; Wu, Q.; Lin, J.; Zhu, D. Inverting the Enantiopreference of Nitrilase-Catalyzed Desymmetric Hydrolysis of Prochiral Dinitriles by Reshaping the Binding Pocket with a Mirror-Image Strategy. *Angew. Chem., Int. Ed. Engl.* **2021**, *60* (7), 3679–3684. https://doi.org/10.1002/anie.202012243.

(85) Quinn, D. J.; Haun, G. J.; Moura-Letts, G. Direct Synthesis of Nitriles from Aldehydes with Hydroxylamine-*O*-Sulfonic Acid in Acidic Water. *Tet. Lett.* **2016**, *57* (34), 3844–3847. https://doi.org/10.1016/j.tetlet.2016.07.047.

(86) Natchev, I. A. Organophosphorus Analogues and Derivatives of the Natural L-Aminocarboxylic Acid and Peptides Vii. Enzyme Synthesis of Phospha-c Peptides. *Tetrahedron* **1991**, *47* (7), 1239–1248. https://doi.org/10.1016/S0040-4020(01)86380-8.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

(87) Zeiss, H.-J. Enantioselective Synthesis of L-Phosphinothricin from L-Methionine and L-Glutamic Acid via L-Vinylglycine. *Tetrahedron* **1992**, *48* (38), 8263–8270. https://doi.org/10.1016/S0040-4020(01)80494-4.

(88) Tempkin, O.; Blacklock, T. J.; Andrew Burke, J.; Anastasia, M. β-Butyrolactone as a Chiral Building Block in Organic Synthesis: A Convenient Synthesis of MK-0507 Keto Sulfone. *Tetrahedron: Asymmetry* **1996**, *7* (9), 2721–2724. https://doi.org/10.1016/0957-4166(96)00350-3.

(89) Turcu, M. C.; Rantapaju, M.; Kanerva, L. T. Applying Lipase Catalysis to Access the Enantiomers of Dorzolamide Intermediates. *Eur. J. Org. Chem.* **2009**, *2009* (32), 5594–5600. https://doi.org/10.1002/ejoc.200900672.

(90) Blacklock, T. J.; Sohar, P.; Butcher, J. W.; Lamanec, T.; Grabowski, E. J. J. An Enantioselective Synthesis of the Topically-Active Carbonic Anhydrase Inhibitor MK-0507: 5,6-Dihydro-(S)-4-(Ethylamino)-(S)-6-Methyl-4H-Thieno[2,3-b]Thiopyran-2-Sulfonamide 7,7-Dioxide Hydrochloride. *J. Org. Chem.* **1993**, *58* (7), 1672–1679. https://doi.org/10.1021/jo00059a013.

(91) Qiu, S.; Cheng, F.; Jin, L.-J.; Chen, Y.; Li, S.-F.; Wang, Y.-J.; Zheng, Y.-G. Co-Evolution of Activity and Thermostability of an Aldo-Keto Reductase *Km*AKR for Asymmetric Synthesis of Statin Precursor Dichiral Diols. *Bioorg. Chem.* **2020**, *103*, 104228. https://doi.org/10.1016/j.bioorg.2020.104228.

(92) K. Ma, S.; Gruber, J.; Davis, C.; Newman, L.; Gray, D.; Wang, A.; Grate, J.; W. Huisman, G.; A. Sheldon, R. A Green-by-Design Biocatalytic Process for Atorvastatin Intermediate. *Green Chemistry* **2010**, *12* (1), 81–86. https://doi.org/10.1039/B919115C.

(93) Chen, X.; Liu, Z.-Q.; Lin, C.-P.; Zheng, Y.-G. Chemoenzymatic Synthesis of (*S*)-Duloxetine Using Carbonyl Reductase from *Rhodosporidium Toruloides*. *Bioorg. Chem.* **2016**, *65*, 82–89. https://doi.org/10.1016/j.bioorg.2016.02.002.

(94) Sankaranarayanan, K.; Jensen, K. F. Computer-Assisted Multistep Chemoenzymatic Retrosynthesis Using a Chemical Synthesis Planner. *Chem. Sci.* **2023**, *14* (23), 6467–6475. https://doi.org/10.1039/D3SC01355C.

(95) Savile, C. K.; Janey, J. M.; Mundorff, E. C.; Moore, J. C.; Tam, S.; Jarvis, W. R.; Colbeck, J. C.; Krebber, A.; Fleitz, F. J.; Brands, J.; Devine, P. N.; Huisman, G. W.; Hughes, G. J. Biocatalytic Asymmetric Synthesis of Chiral Amines from Ketones Applied to Sitagliptin Manufacture. *Science* **2010**, *329* (5989), 305–309. https://doi.org/10.1126/science.1188934.

(96) Kim, G.-H.; Jeon, H.; Khobragade, T. P.; Patil, M. D.; Sung, S.; Yoon, S.; Won, Y.; Sarak, S.; Yun, H. Glutamate as an Efficient Amine Donor for the Synthesis of Chiral β- and γ-Amino Acids Using Transaminase. *ChemCatChem* **2019**, *11* (5), 1437–1440. https://doi.org/10.1002/cctc.201802048.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Supporting Information for

Chemoenzymatic Multistep Retrosynthesis with Transformer Loops

David Kreutter[a)] and Jean-Louis Reymond*[a)]

[a)] *Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland*

e-mails: david.kreutter@unibe.ch
jean-louis.reymond@unibe.ch

https://doi.org/10.26434/chemrxiv-2024-svr99 ORCID: https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

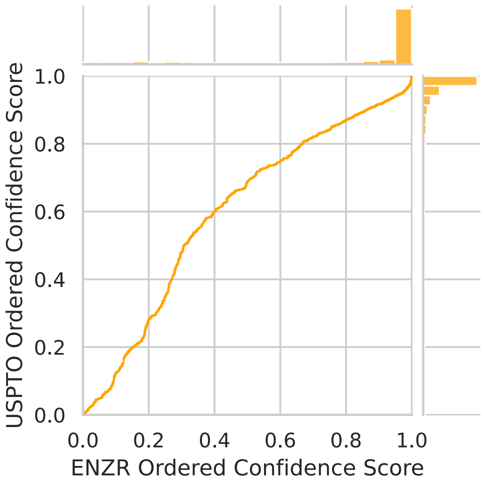


Figure S1. Ordered confidence scores of the ENZR-TTL T3 as function of the ordered confidence scores of USPTO-TTL T3 on the ENZR test set and USPTO test set respectively.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

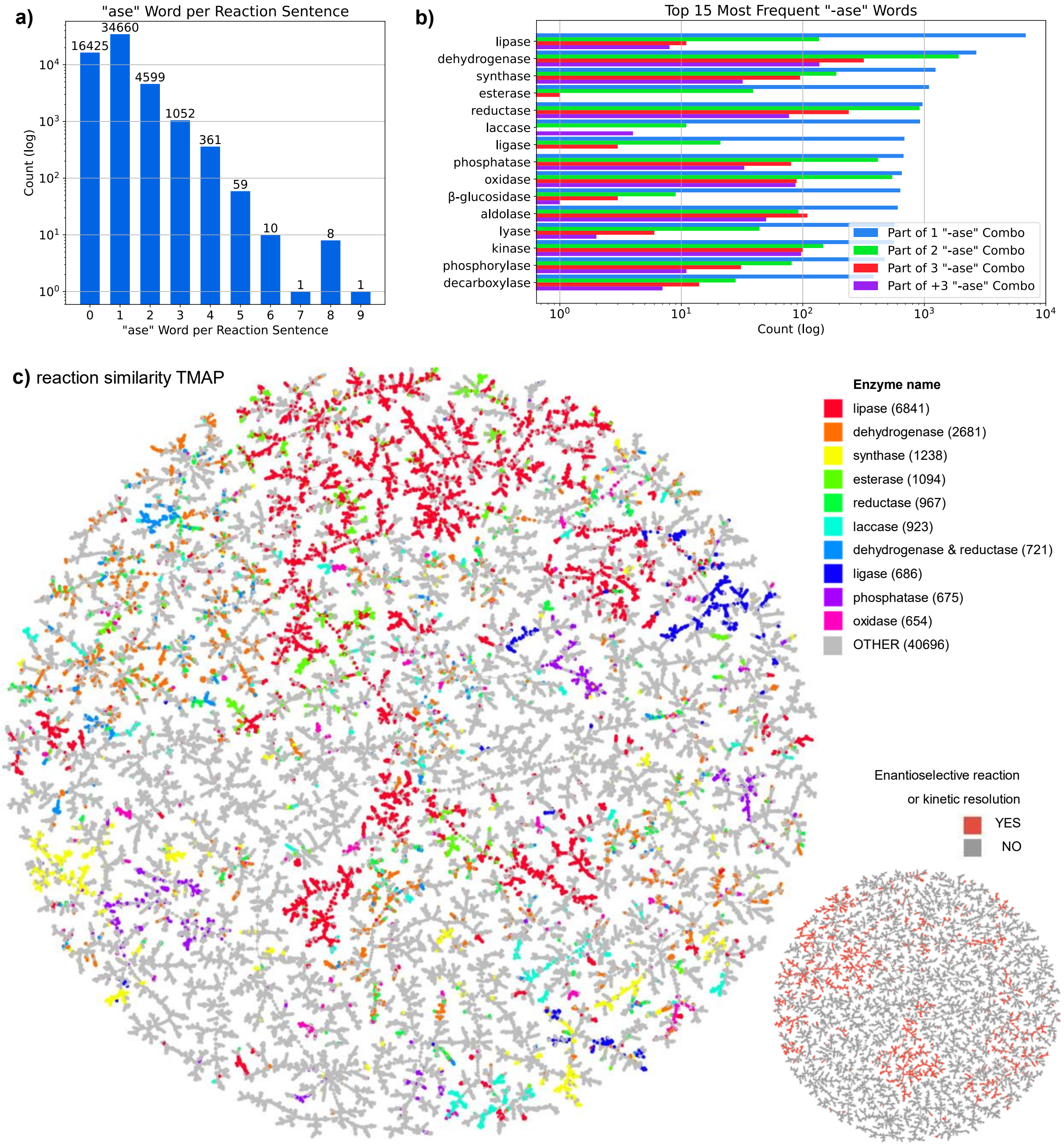


Figure S2. Analysis of the ENZR dataset. (**a**) Number of reactions depending on how many “-ase” words are present in the sentence for a given reaction. (**b**) Frequency of the top 15 “-ase” words depending on the count of enzyme name per reaction. (**c**) TMAP of reactions similarity color-coded by the 10 most frequent "-ase" words as listed in Fig. 2b. combinations. The “other” category groups reactions with “-ase” words other than the top 10 “-ase” words or reaction containing infrequent “-ase” word combinations. Insert lower right: TMAP highlighting enantioselective and kinetic resolution reactions.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

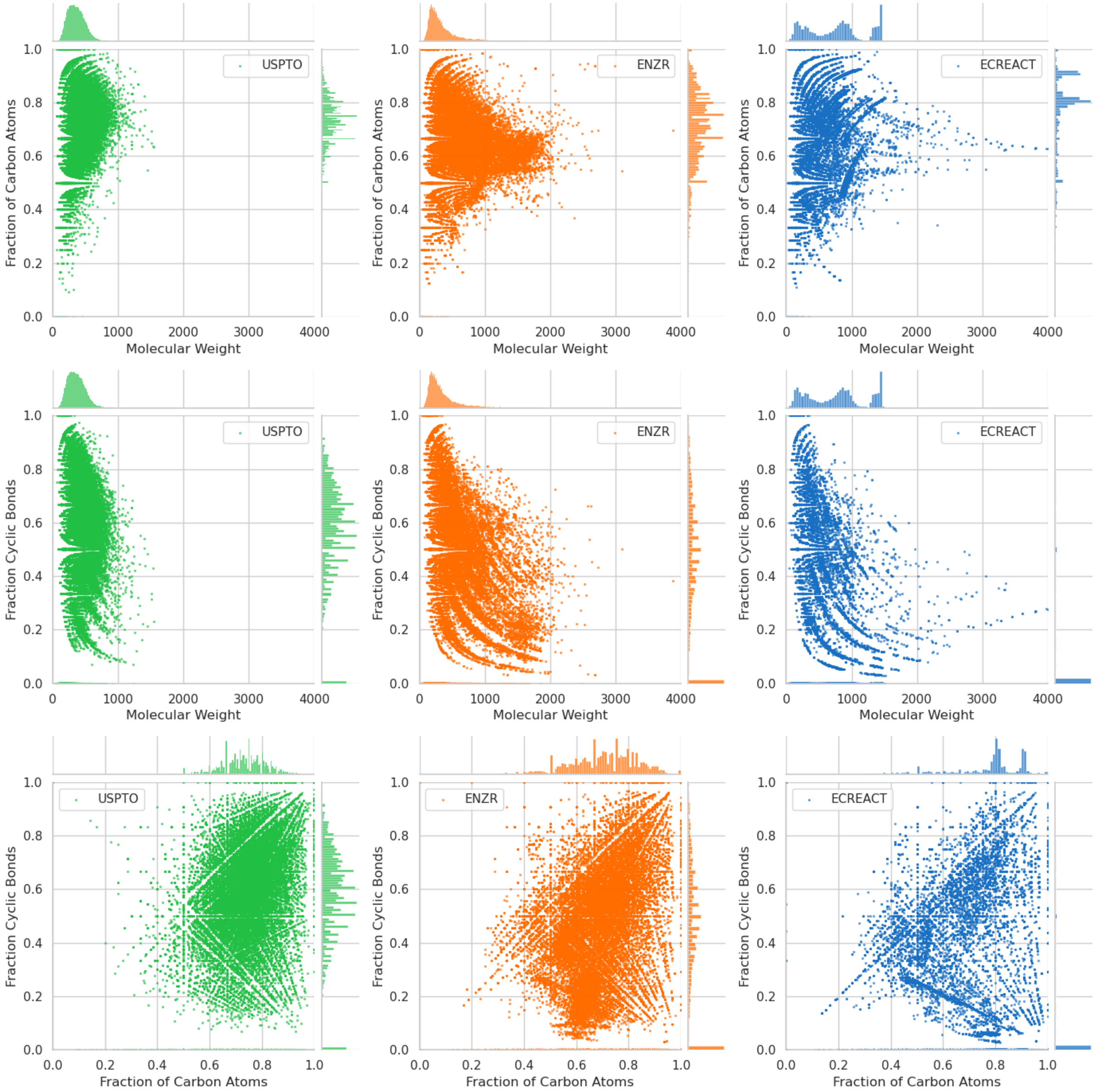


Figure S3. Analysis of the USPTO (green), ENZR (orange) and ECREACT (blue) datasets in forms of scatter plots. First line: Fraction of C-atoms vs. MW. Second line: Fraction cyclic bonds vs. MW. Third line: Fraction cyclic bonds vs. Fraction of C-atoms.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

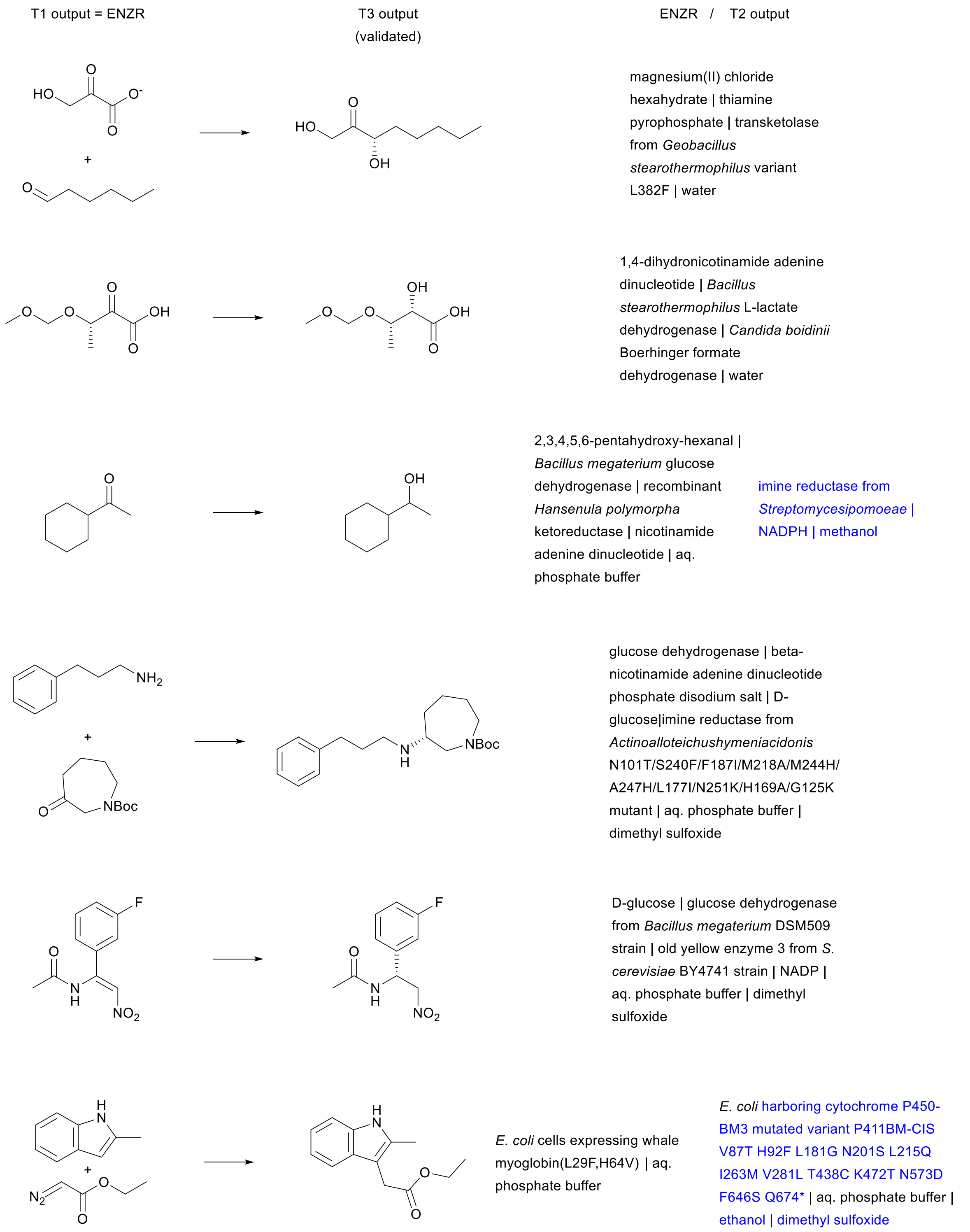
Table S1. Details of top-1 round-trip accuracy by ENZR-TTL single step retrosyntheses on the 2858 molecules of the ENZR test set.

	Round-trip validated by T3	Not validated by T3
Ground-truth predicted SM	49.41%	23.13%
Not ground truth predicted SM	9.55%	17.91%

Table S2. Details of top-1 round-trip accuracy by USPTO-TTL single step retrosyntheses on a sample of 3000 molecules from the USPTO test set.

	Round-trip validated by T3	Not validated by T3
Ground-truth predicted SM	60.57%	6.97%
Not ground truth predicted SM	20.73%	11.73%

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0



T1 output = ENZR

T3 output (validated)

ENZR / T2 output

magnesium(II) chloride hexahydrate | thiamine pyrophosphate | transketolase from *Geobacillus stearothermophilus* variant L382F | water

1,4-dihydronicotinamide adenine dinucleotide | *Bacillus stearothermophilus* L-lactate dehydrogenase | *Candida boidinii* Boerhinger formate dehydrogenase | water

2,3,4,5,6-pentahydroxy-hexanal | *Bacillus megaterium* glucose dehydrogenase | recombinant *Hansenula polymorpha* ketoreductase | nicotinamide adenine dinucleotide | aq. phosphate buffer

imine reductase from *Streptomycesipomoeae* | NADPH | methanol

glucose dehydrogenase | beta-nicotinamide adenine dinucleotide phosphate disodium salt | D-glucose|imine reductase from *Actinoalloteichushymeniacidonis* N101T/S240F/F187I/M218A/M244H/A247H/L177I/N251K/H169A/G125K mutant | aq. phosphate buffer | dimethyl sulfoxide

D-glucose | glucose dehydrogenase from *Bacillus megaterium* DSM509 strain | old yellow enzyme 3 from *S. cerevisiae* BY4741 strain | NADP | aq. phosphate buffer | dimethyl sulfoxide

E. coli cells expressing whale myoglobin(L29F,H64V) | aq. phosphate buffer

E. coli harboring cytochrome P450-BM3 mutated variant P411BM-CIS V87T H92F L181G N201S L215Q I263M V281L T438C K472T N573D F646S Q674* | aq. phosphate buffer | ethanol | dimethyl sulfoxide

Figure S4. Additional examples of correctly predicted enzymatic single step retrosynthesis by ENZR-TTL. The confidence scores of T3 are >99.5% in all cases. Enzyme names from the T2 output that differ from the database entry are highlighted in blue.

https://doi.org/10.26434/chemrxiv-2024-svr99 ORCID: https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. License: CC BY-NC 4.0

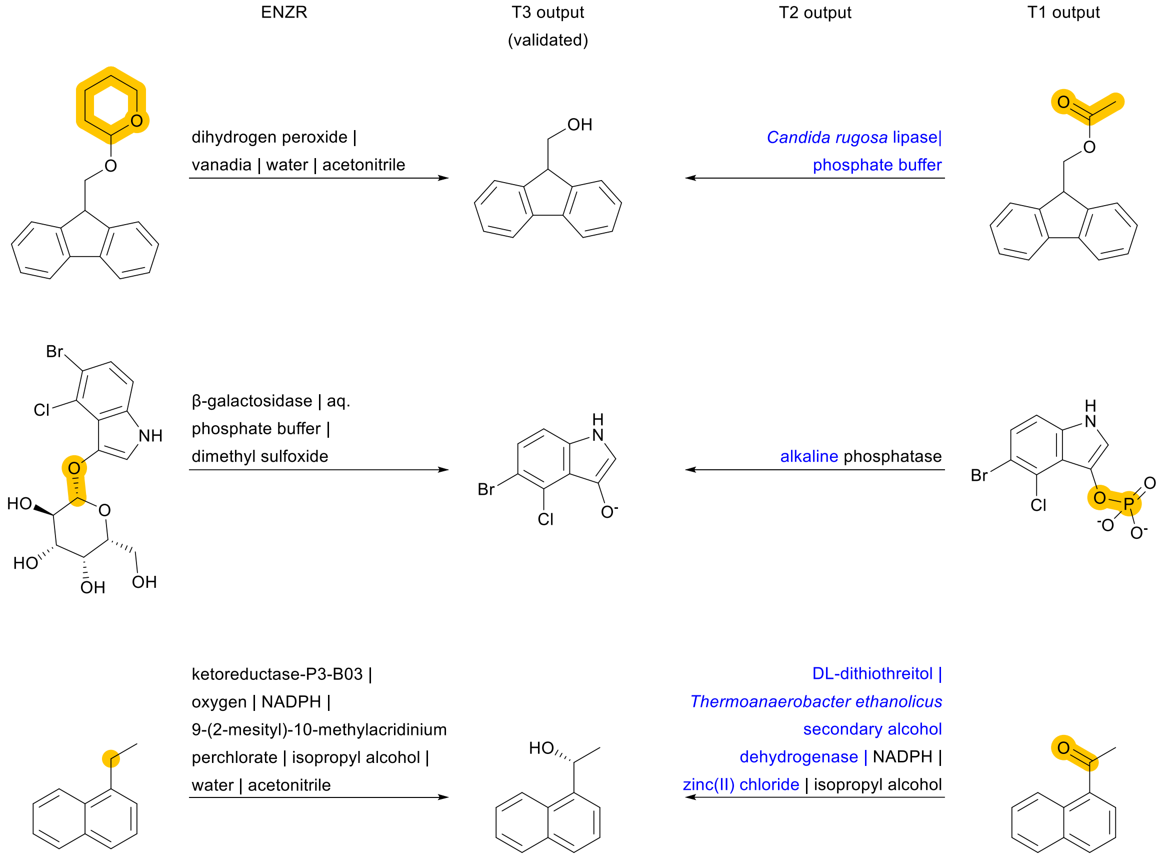


Figure S5. Additional examples of ENZR-TTL retrosynthetic steps validated by T3 involving different precursors and/or enzymes than those in ENZR. Structural differences between SM database entry and T1 output are highlighted in orange and enzyme names from T2 output that differ from the database entry are highlighted in blue.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

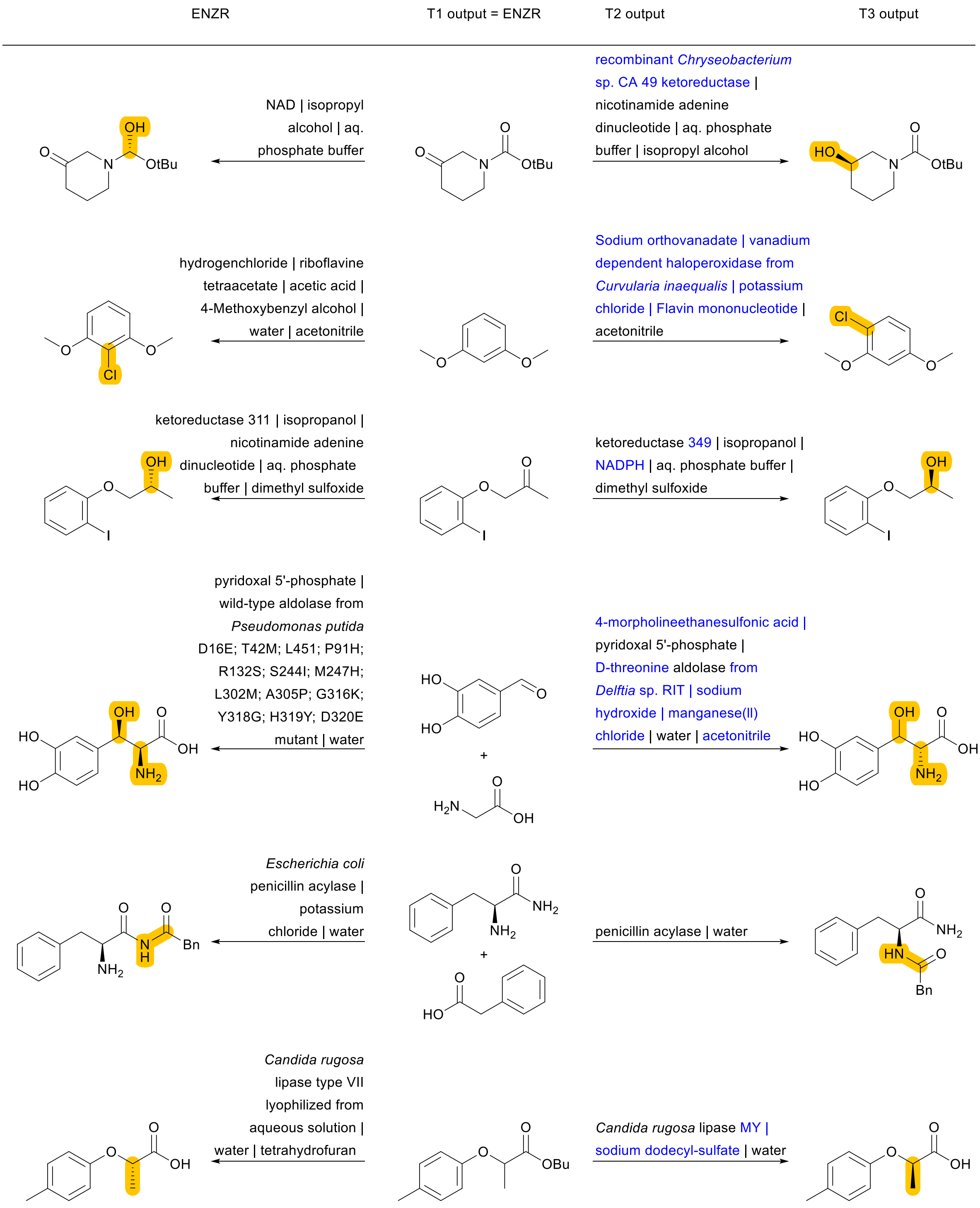

ENZR
T1 output = ENZR
T2 output
T3 output
NAD | isopropyl alcohol | aq. phosphate buffer
recombinant Chryseobacterium sp. CA 49 ketoreductase | nicotinamide adenine dinucleotide | aq. phosphate buffer | isopropyl alcohol
hydrogenchloride | riboflavine tetraacetate | acetic acid | 4-Methoxybenzyl alcohol | water | acetonitrile
Sodium orthovanadate | vanadium dependent haloperoxidase from Curvularia inaequalis | potassium chloride | Flavin mononucleotide | acetonitrile
ketoreductase 311 | isopropanol | nicotinamide adenine dinucleotide | aq. phosphate buffer | dimethyl sulfoxide
ketoreductase 349 | isopropanol | NADPH | aq. phosphate buffer | dimethyl sulfoxide
pyridoxal 5'-phosphate | wild-type aldolase from Pseudomonas putida D16E; T42M; L451; P91H; R132S; S244I; M247H; L302M; A305P; G316K; Y318G; H319Y; D320E mutant | water
4-morpholineethanesulfonic acid | pyridoxal 5'-phosphate | D-threonine aldolase from Delftia sp. RIT | sodium hydroxide | manganese(II) chloride | water | acetonitrile
Escherichia coli penicillin acylase | potassium chloride | water
penicillin acylase | water
Candida rugosa lipase type VII lyophilized from aqueous solution | water | tetrahydrofuran
Candida rugosa lipase MY | sodium dodecyl-sulfate | water
OH
O
N
OtBu
HO
Cl
I
NH$_2$
H$_2$N
+
Bn
HN
H
OBu

Figure S6. Additional examples of ENZR-TTL prediction involving a correct SM prediction by T1 but a different enzyme choice by T2 and therefore a different product P compared to the database entry.

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0

Table S3. Number (percentage) of product molecule from the test set with solved routes for the selection of 80 molecules from the ENZR test set, and for the 100 molecules from the USPTO test set.

	USPTO test set (100 molecules)	ENZR test set (80 molecules)
Molecules with Route solved	88 (88%)	61 (76%)
Molecules with Route solved with at least one route including an enzymatic step	86 (86%)	60 (75.0%)

Table S4. Fraction of enzymatic reaction steps present in the predicted and solved multistep routes among the top-X route unique steps, ranked according to the RPScore. Tested on 100 USPTO test set and 80 ENZR test set molecules.

	for USPTO test set molecules (%)	for ENZR test set molecules (%)
Overall	7.88	16.86
Top-100 RPScoring routes	9.33	21.67
Top-50 RPScoring routes	9.33	24.79
Top-10 RPScoring routes	8.65	33.76
Top-5 RPScoring routes	8.23	38.22
Top-1 RPScoring routes	7.10	50.00

https://doi.org/10.26434/chemrxiv-2024-svr99 **ORCID:** https://orcid.org/0000-0003-2724-2942 Content not peer-reviewed by ChemRxiv. **License:** CC BY-NC 4.0