Chemoenzymatic synthesis planning by evaluating the synthetic potential in biocatalysis and chemocatalysis

Xuan Liu[1,2,3], Hongxiang Li[1,2,3,4], Huimin Zhao[1,2,3,4,5,*]

[1]NSF Molecular Maker Lab Institute, [2]Department of Chemical and Biomolecular Engineering, [3]Carl R. Woese Institute for Genomic Biology, [4]Department of Chemistry, [5]DOE Center for Advanced Bioenergy and Bioproducts Innovation, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.

*Corresponding to: zhao5@illinois.edu

Abstract

Chemoenzymatic synthesis integrates the advantages of chemocatalysis and biocatalysis to design efficient synthesis routes. However, current computer-assisted chemoenzymatic synthesis planning tools lack a heuristic method to unify step-by-step chemoenzymatic synthesis planning and molecule-by-molecule identification of chemo-/biocatalysis opportunities in synthesis routes. Here we develop an asynchronous chemoenzymatic retrosynthesis planning algorithm (ACERetro) which employs a search strategy that prioritizes the exploration of a molecule's promising catalytic methods. The suitability of a molecule to be synthesized via chemo- or biocatalysis is quantitatively evaluated by a data-driven Synthetic Potential Score (SPScore) using a neural network model. Additionally, the SPScore can be used to heuristically identify chemo-/biocatalysis opportunities in synthesis routes. For a given synthesis route, this algorithm uses SPScore to identify the molecules that offer optimization potential when synthesized by an alternative catalytic method, and then ACERetro is used to search synthesis routes. Case studies on synthesis planning for ethambutol and epidiolex demonstrate that our strategy can design concise chemoenzymatic synthesis routes by applying enzymatic steps to introduce stereochemistry and find shortcuts. Moreover, case studies on synthesis route optimization for rivastigmine and (*R,R*)-formoterol demonstrate how our strategy finds bypasses to form alternative, shorter chemoenzymatic synthesis routes. Our findings demonstrate that ACERetro with evaluating the synthetic potential of molecules represents a versatile and effective search framework for chemoenzymatic synthesis planning.

Introduction

Biocatalysis and chemocatalysis span distinct reaction spaces in terms of designing synthesis routes for molecules of interest due to their different characteristics. For biocatalysis, enzymes are green catalysts for reactions with excellent regioselectivity, stereoselectivity, and activity. For chemocatalysis, organic reactions are advantageous due to the broad substrate scope, vast types of reactions, and numerous well-studied cases. Combining these two catalytic methods can capitalize on their unique advantages to build more efficient chemoenzymatic synthesis routes to many compounds[1–5]. A prominent example is the use of an engineered ribosyl-1-kinase for the synthesis of molnupiravir, an antiviral drug, which shortened the original synthesis route by 70% and achieved a sevenfold higher yield[6].

To accelerate the process of discovering novel synthesis routes, computer-aided synthesis planning (CASP) tools assist scientists in swiftly designing synthetic routes by retracing precursors step by step, replacing the previously cumbersome manual design process[7–9]. Based on the method of retracing precursors, CASP tools can be categorized into two types: template-based and template-free. Template-based methods utilize a set of reaction templates, either expert-curated or extracted from reaction databases, to generate precursors for target molecules[10–15]. Using reaction templates makes precursor predictions adhere to the transformation rules between reactants and products. Training a prioritizer to rank reaction templates still requires enough examples in a reaction database[16], while a single occurrence of a reaction in a database is sufficient to extract a reliable template, ensuring the reliability of precursor inference process. For chemocatalysis, Synthia[10] utilizes expert-curated templates. In contrast, AiZynthFinder[13] and ASKCOS[12] employ templates extracted from organic synthesis databases. For biocatalysis, RetroBioCat[15] uses expertly encoded reaction rules, while novoStoic[17] and RetroPath[14,18] extract templates from metabolic pathway databases. Template-free methods fully capitalize on the advanced developments of language models, transforming retrosynthesis into the problem of translating product SMILES strings into precursor SMILES strings[19–25]. Learning chemical transformations directly from reaction data enables template-free methods to predict novel reactions. However, this also implies that training template-free models demands a large amount of data, and models without pre-defined rules may output unfeasible SMILES and reactions[26]. Examples of this strategy include RXN4Chemistry[27,28] and RetroTRAE[29] in chemocatalysis, and RXN4Chemistry (biocatalysis model)[30] and BioNavi-NP[31] in biocatalysis.

Chemoenzymatic synthesis encompasses two distinct catalytic methods: chemocatalysis and biocatalysis. A chemoenzymatic synthesis route searching method is developed by Zhang et al. using historical published chemical and enzymatic reactions to build a combined reaction dataset for pathway searching[32]. This method avoids uncertainty in reaction predictions but cannot be used to design novel chemoenzymatic synthesis routes. To achieve computer-aided chemoenzymatic synthesis planning with predicted reactions, an intuitive approach is employing two retrosynthetic models for chemocatalysis and biocatalysis respectively. These models' results are then integrated at each step when searching precursors for a target molecule, facilitating a hybrid search process. Levin et al.'s hybrid planner[33] adopts this strategy by integrating chemical ('organic') reaction templates with enzymatic reaction templates, and then combines the results of two prioritizers trained in each reaction database to rank the templates. While Levin et al.'s tool effectively predicts chemoenzymatic syntheses, it is not advisable to search for precursors in both catalytic methods simultaneously because the two prioritizers are trained on different datasets and lack proper alignment, leading to potential bias. Especially when the search space grows exponentially with search depth, the cumulative bias significantly affects the search performance. Additionally, Zeng et al. adopted a multitask learning strategy for template-free models, enabling automatic indication of the catalytic method[34]. However, this approach is limited to template-free methods, and there is currently no unified hybrid search algorithm applicable to both template-based and template-free tools. Another strategy in computer-aided chemoenzymatic synthesis planning involves the identification of alternative bypass syntheses: seeking enzymatic reactions as bypasses to existing or predicted chemical syntheses, or similarly, seeking chemical reactions as bypasses to enzymatic syntheses. For example, Sankaranarayanan et al. designed chemoenzymatic routes by exhaustively applying the biocatalytic templates to each intermediate in predicted synthesis routes[35]. However, because chemical reaction templates outnumber enzymatic reaction templates, adopting an exhaustive search strategy to identify bypasses in enzymatic synthesis routes by applying chemical reaction templates will inevitably result in an exponential growth of the search space. Employing heuristic methods is therefore demanded to determine which steps in the synthesis route need optimization, which is crucial to enhance the search efficiency.

Herein, we report a synthetic potential guided asynchronous chemoenzymatic retrosynthesis planning algorithm (ACERetro). ACERetro utilizes the synthetic potential score (SPScore) to infer the promising catalytic method for the synthesis of a given molecule (**Fig. 1a,b**). Additionally, we introduce a synthetic potential guided synthesis route optimization algorithm to generate chemoenzymatic synthesis routes via identifying chemocatalysis/biocatalysis opportunities for enzymatic/organic reactions. This algorithm uses SPScore to find steps with opportunities for improvement that can be catalyzed by alternative catalytic methods (**Fig. 1c**). Evaluating the synthetic potential provides a heuristic and universal approach to the chemoenzymatic synthesis planning, which facilitates the utilization of either template-based or template-free CASP tools. Based on the characteristics of synthetic reaction and enzymatic reaction databases, we employed a template-free retrosynthesis tool for chemocatalysis where data is abundant, and a template-based tool for biocatalysis where database size is limited. We evaluated the performance of the SPScore on both single-step and multi-step retrosynthetic routes. Subsequently, we conducted a benchmark of

ACERetro with the state-of-the-art tool on hybrid searches for 1,001 molecules. Lastly, we applied ACERetro to case studies of two FDA-approved drugs, ethambutol and epidiolex, to demonstrate how ACERetro finds promising chemoenzymatic synthesis routes. In addition, the optimization algorithm with ACERetro was applied to analyze the synthesis routes of another two FDA-approved drugs, rivastigmine and (*R,R*)-formoterol. The results demonstrate the algorithm's capability to optimize chemoenzymatic synthesis routes by identifying promising bypasses.

Results

Developing a synthetic potential scoring function

The synthetic potential of a molecule in the chemo- or biocatalytic synthesis highly depends on the molecule's structure and the development of chemo- and biocatalysis. To make a quantitative evaluation on the synthetic potential of molecules, we employed a data-driven method to train a synthetic potential scoring function on reaction databases. The method based on the premise that if a molecule has documented reactions for its synthesis, the catalytic method of these reported reactions is the molecule's promising catalytic method. A dataset comprising reaction products was extracted from two primary sources: USPTO 480K[36], which contains 484,706 reactions in chemocatalysis, and ECREACT[30], which contains 62,222 reactions in biocatalysis. After removing duplicates and molecules that could not be converted into valid molecular fingerprints for each respective catalytic method, the resulting dataset comprised of 437,781 molecules in chemocatalysis and of 37,939 molecules in biocatalysis, while 515 molecules were present in both catalytic methods.

Molecules were represented by ECFP4[37] (extended connectivity fingerprint, up to four bonds) and MAP4[38] (MinHashed Atom Pair fingerprint, diameter $d = 4$) with three different lengths (*length* = 1024, 2048, and 4096) to train several multi-layer perceptron (MLP) models. Due to the limitation of a reaction corpus that does not cover all the possible reactions of a molecule, it is not advisable to make a binary classification model to predict the promising catalytic method for molecules. Therefore, a Margin Ranking Loss was used as the training objective, which infers the promising catalytic method based on the relative value between predicted synthetic potential score in chemocatalysis (S_{Chem}) and synthetic potential score in biocatalysis (S_{Bio}). The SPScores range from 0 to 1, so they can act as the probability of a molecule being promisingly synthesized by the catalysis of each catalytic method. The concrete idea of margin is when the difference between two SPScores of a molecule is within the margin, both catalytic methods are considered promising for the molecules' synthesis. If a molecule's SPScore of one catalytic method is greater than the other, and the difference is greater than the margin, the field with the larger SPScore is more suitable for the synthesis of the molecule. In the training process, a margin of 0.15 was used, which helps ensure that the three regions have similar areas. A margin that is neither too severe nor too trivial benefits subsequent adjustments to user preferences without the need of model retraining on a different margin. An excessively large number of epochs will cause the model to overfit the distribution of the training data[39]. Since reactions sourced from the USPTO are confined to patents and differ in distribution from those documented in the literature[40], the number of epochs is also considered in the evaluation criteria. Therefore, the best model was obtained by a comprehensive evaluation of precision, recall, and F1 on the validation dataset, as well as the number of epochs (**Supplementary Fig. 2**). As a result, the model that used ECFP4 with a length of 4096 as molecular embedding will be used for the subsequent tasks.

Benchmarking SPScore on single-step retrosynthesis

To evaluate the performance of SPScore, we assessed the benchmark of our scoring model on a dataset comprising 11,003 molecules randomly selected from the "in-vitro" subset of ZINC15[41] that were not in the training dataset. Subsequently, their corresponding SPScores were calculated. The distribution of two SPScores and the difference of SPScores are shown in **Fig. 2a**. Although the margin used to train the model is fixed, a flexible user-defined margin can be adopted for different tolerance of catalytic methods during application. The use of margin also provides an intuitive perspective to understand applications of SPScore.

With the margin increasing from 0.05 to 0.25, more molecules are located in the region where molecules can be promisingly synthesized by both catalytic methods (**Fig. 2b**).

To further explore whether SPScore is representative on predicting a molecule's promising catalytic method, we conducted one-step retrosynthesis in each catalytic method by employing RXN4Chemistry[28] for chemocatalysis, and employing Levin et al.'s enzymatic templates in ref. 33 for biocatalysis. For a given target molecule, RXN4Chemistry predicts possible synthetic reactions ranked by a confidence score, namely backward confidence, while Levin et al.'s enzymatic templates predict possible enzymatic reactions ranked by template score. The average backward confidence of the top-5 predictions increases with the mean synthetic potential score in chemocatalysis predicted by our scoring model (**Fig. 2c**). A similar trend between the average template score and the mean synthetic potential score in biocatalysis is observed (**Fig. 2d**). This suggests that as the predicted probability of molecules being synthesized by a specific catalytic method increases, the corresponding retrosynthesis tool's confidence in its predictions also tends to increase.

To determine whether the relative value of S_{Chem} and S_{Bio} can serve as a representative metric, we analyzed both the average backward confidence and the average template score for top-5 predictions versus the mean SPScore difference ($S_{Chem} - S_{Bio}$), as plotted in **Fig. 2e,f**. The result reveals that when S_{Bio} is larger than S_{Chem}, molecules tend to have relatively high template score to be synthesized by biocatalysis and low backward confidence to be synthesized by chemocatalysis. Collectively, these trends in **Fig. 2c-f** suggest that our scoring function exhibits a good ability to deduce the promising catalytic method for molecules.

Benchmarking SPScore on multi-step retrosynthesis

A multi-step synthesis route dataset is used to evaluate the performance of SPScore on multi-step retrosynthesis. Because there is lack of databases including multi-step synthesis routes, especially a multi-step hybrid synthesis database, the synthesis routes of 493 target molecules searched by Levin et al.'s tool (the hybrid planner in ref. 33) in three minutes were used in this benchmark. The SPScore prediction of molecules is compared with the reaction type (chemical reaction or enzymatic reaction) used in the synthesis routes. Out of 397,040 synthesis routes linked to the 493 target molecules, 26,741 distinct product molecules with their catalytic methods were identified. Specifically, 9,162 (34.3%) molecules synthesized exclusively by chemocatalysis, 10,211 (38.2%) synthesized exclusively by biocatalysis, and 7,368 (27.5%) having both organic reactions and enzymatic reactions to synthesize them are in the overlap part. Furthermore, from the 1,531 shortest synthesis routes related to the 493 target molecules, 1,544 unique product molecules with their respective catalytic methods were identified: 788 (51.0%) in chemocatalysis, 481 (31.2%) in biocatalysis, and 275 (17.8%) spanning both fields.

When using SPScore to guide the search where the margin is set as 0.15, 85.8% of molecules' synthesis field in shortest synthesis routes and 75.0% of molecules' synthesis field in all synthesis routes can be covered. By this way, it can save 40.2% searches in shortest synthesis routes and 33.8% searches in all synthesis routes because SPScore can give the correct prediction that matches the catalytic method in the original synthesis routes (**Fig. 3a,b**). The observed trend indicates that as the margin expands, the catalytic method of a greater number of molecules is encompassed. However, this comes at the expense of a reduced number of saved searches.

The reaction retention rate is determined by counting the instances where the actual catalytic method of a reaction is included in the predicted catalytic method of the reactions' product molecule. When the margin is set as 0.15, 89.9% of reactions in the shortest synthesis routes can be covered, and 86.0% of reaction in the near shortest synthesis routes (route length ≤ shortest route length + 2) can be covered (**Fig. 3c**). Next, we investigate whether SPScore can provide guidance for finding the shortest synthesis route and discovering diversity of shortest routes. Of 493 target molecules, the route retention rate is determined by counting the number of molecules that at least one shortest synthesis routes whose actual catalytic methods can be covered by SPScores' prediction. In scenarios with the same margin of 0.15, 393 (79.7%) of the molecules have at least one shortest synthesis route that can be fully retained, while 109 (73.6%) molecules out of 148 have at least three shortest synthesis routes that can be retained (**Fig. 3d**). The results indicate that, in the context of multi-step retrosynthesis, utilizing SPScore as a guide enables the retention of the majority of favorable synthesis routes.

Hybrid synthesis route search

Hybrid search for chemoenzymatic synthesis route requires predicting precursors in chemocatalysis and biocatalysis. Synchronous methods, like Levin et al.'s tool, search chemocatalytic precursors and biocatalytic precursors simultaneously and then combine the search results. While in this study, we reported an asynchronous method, ACERetro, which prioritizes the search of the most promising catalytic method for a given molecule. To evaluate the performance, a fully hybrid synchronous algorithm (FHSync, **Fig. 4a**) and a SPScore guided synchronous algorithm (SPSync, **Fig. 4b**) are proposed for self-benchmarking to ACERetro, a SPScore guided asynchronous search algorithm (**Fig. 4c**). The single-step precursor prediction tools previously used in the "in-vitro" subset of ZINC15, namely RXN4Chemistry and Levin et al.'s enzymatic templates, were respectively used for chemocatalysis and biocatalysis. The fully hybrid search algorithm without SPScore directly searches both chemocatalysis and biocatalysis, and the results are combined in each step, while the SPScore guided synchronous hybrid search algorithm only searches promising catalytic methods predicted by the SPScore. In the SPScore guided asynchronous hybrid search algorithm, each molecule is weighted by the SPScore associated with the corresponding catalytic method. This allows the algorithm to revert to searching the less promising catalytic method when the data-driven SPScore yields an inappropriate inference. Employing asynchronous search techniques bolsters fault tolerance during the synthesis planning process.

To explore search spaces of these three algorithms, we conducted a comparative evaluation on a set of 1,001 molecules from ZINC, which Levin et al.'s tool had explored in ref. 33 under identical boundary conditions including search time and buyable dataset (see Methods). FHSync found synthesis routes to 597 molecules, SPSync found synthesis routes to 683 molecules, and ACERetro found synthesis routes to 720 moecules (**Fig. 5a**). Compared to the Levin et al.'s tool, which found synthesis routes to only 493 molecules, the FHSync can find synthesis routes to additional 104 (21.1%) molecules. This improvement is mainly attributed to the incorporation of the template-free model, RXN4Chemistry, in chemocatalysis. Moreover, the SPSync and ACERetro found synthesis routes to 190 (38.5%) and 227 (46.0%) more molecules compared with Levin et al.'s tool, respectively. These results underscore that the efficiency of ACERetro surpasses the state-of-the-art method.

In a self-benchmarking analysis, the hybrid search algorithms with SPScore guidance (SPSync and ACERetro) outperform the algorithm without SPScore guidance (FHSync), which could find synthesis routes to 86 and 123 more molecules, emphasizing the pivotal role of SPScore plays in optimizing search efficiency. In the comparsion bewteen SPSync and ACERetro, ACERetro could find synthesis routes to 37 more molecules, which indicates the asynchronous search is more efficient than the synchronous search. Unlike the synchronous search relinquishs the search for molecules' suboptimal catalytic method, the asynchronous search keeps all suboptimal catalytic method of molecules to the queue for later exploration. The algorithm will start to search suboptimal catalytic method of molecules based on a comprehensive consideration inlcuding SPScores, search depth, and molecular complexity (see Methods).

Variations in search spaces and strategies across synthesis planners lead to the prediction of different synthesis routes for molecules. A proficient planner can discover synthesis routes to a greater number of molecules than other planners are able to find. Thus, we conducted a comparative analysis to evaluate the number of molecules whose synthesis routes were exclusively identified by the three algorithms in comparison to Levin et al.'s tool (**Fig. 5b**). It was observed that each algorithm has the capability to discover synthesis routes for molecules that Levin et al.'s tool did not identify. Specifically, out of 1,001 molecules, synthesis routes to 466 could be found by both ACERetro and Levin et al.'s tool. While ACERetro exclusively identified synthesis routes to 254 molecules, Levin et al.'s tool could exclusively find to only

28, indicating that ACERetro discovered approximately 26 times more exclusive molecules than Levin et al.'s tool. These findings imply that ACERetro achieves an expanded search space and better heuristic search strategy than the state-of-the-art tool.

The search quality of the synthesis planning tools can be evaluated by the number of reactions in the predicted synthesis routes through limited context that synthesis planning tools can provide, albeit it is not an exhaustive metric[42]. A smaller number of steps usually implies the use of fewer reagents and fewer purification steps[43]. We compared the length of the shortest synthesis route to 466 molecules found by both ACERetro and Levin et al.'s tool, ACERetro found optimized shortest synthesis route to 167 (35.8%) molecules and the shortest synthesis route of equivalent length for 260 (55.8%) molecules (**Fig. 5c**). This indicates that ACERetro can predict more optimized synthetic routes than Levin et al.'s tool.

To further study the difference in search space between ACERetro and Levin et al.'s tool, we compared the synthesis routes to (*S*)-verofylline (**1**), (3*S*)-3-hydroxy-β-ionone (**2**), and dimenoxadol (**3**) (**Fig. 5d-f**). In the synthesis of **1**, ACERetro predicted a three-step hybrid synthesis route including one enzymatic reaction, while the shortest synthesis route predicted by Levin et al.'s tool included four reactions in chemocatalysis (**Fig. 5d**). The route predicted by ACERetro first uses an enzymatic reaction to synthesize **5** from **4**. The recommended enzyme is 2,5-diamino-6-(5-phospho-D-ribitylamino)-pyrimidin-4(3H)-one deaminase (Rib2; EC number 5.4.99.28). **5** is subsequently alkylated with **6** containing a chiral center to form **7**. The final step constructs an imidazole ring using acetic acid with **7** to produce **1**. Note that the Levin et al.'s tool route uses the same strategy to introduce the chiral center and construct the imidazole ring to **1**, but the difference in starting materials makes the route longer. In the synthesis routes of **2**, ACERetro predicted a two-step enzymatic synthesis route, while Levin et al.'s tool predicted a four-step hybrid synthesis route (**Fig. 5e**). The former first uses a reductase to get the double bond starting with dihydro-beta-ionone (**8**) to form beta-ionone (**9**). Next, a hydroxylase is used to introduce the chiral hydroxyl group for **9** to form **2**. Recommended enzymes are 13,14-dehydro-15-oxoprostaglandin 13-reductase (PGR; EC number 1.3.1.48) and ent-isokaurene C2-hydroxylase (CYP71Z6; EC number 1.14.14.76) respectively. The latter uses a different starting material, beta-cyclocitral (**10**) to form **9**, and three steps to form **2** from **9**. In the synthesis routes of **3**, ACERetro predicted a two-step synthesis route including only chemical reactions, while Levin et al.'s tool predicted a four-step hybrid synthesis route (**Fig. 5f**). The former first constructs ether from benzilic acid (**11**) to form **12**, and then constructs ester to form **3**. The latter uses similar reaction to form the final product from **12**. However, it uses 1,1-diphenylethanol (**13**) as the starting material to synthesize **12** via a three-step hybrid synthesis route.

The routes of **1**, **2**, and **3** predicted by ACERetro cover three senarios: hybrid approach, purely synthetic approach, and purely enzymatic approach. The results show that ACERetro can often find shortcuts to synthesize compounds compared to Levin et al.'s tool, such as the synthesis of intermediate **7** in the synthesis route of **1**, the synthetic route from **9** to **2**, and the synthesis of **12** in the synthesis route of **3**. For the predicted enzyme reactions, although those enzymes have not been reported to use molecules in the predicted routes as substrates, the predicted reactions still provide effective guidance for future enzyme discovery and engineering. Among all routes predicted by ACERetro, the SPScore of each product except **5** is consistent with the corresponding catalytic method in the synthesis route. However, note that S_{Enzy} of **5** is higher than all other products in the route, and its $S_{Chem} - S_{Enzy}$ has the smallest value, which indicates **5** has higher potential to be synthesized by biocatalysis compared to other product molecules in the synthesis routes.

Case study for synthesis planning

Ethambutol is a drug used in the treatment of tuberculosis (TB). The (*S,S*)-enantiomer, ((*S,S*)-ethambutol; **(*S,S*)-14**), is the most active antimycobacterial agent compared with other three isomers[44–46]. Wilkinson et al. first reported the synthesis route for **(*S,S*)-14**, utilizing (2*S*)-2-aminobutan-1-ol (**15**) as the starting material[44]. Likewise, the synthesis routes of **(S,S)-14** developed by Butula et al.[47] and Stauffer et al.[48] also used starting materials containing chiral centers (**15** and **18**) directly. Trost et al. used palladium catalyzed stereoselective epoxide (**16**) opening on phthalimide (**17**) to construct the chiral center[49], while Kotkar et al. reported a synthesis route using proline-catalyzed α-aminooxylation on butyraldehyde (**19**)[50].

We conducted retrosynthesis planning on (*S,S*)-ethambutol by using ACERetro. The search parameters are the same as those used in the above-mentioned benchmarking study, except that the maximum search depth is set to 5 based on existing routes. The most promising predicted synthesis route connecting to buyable compounds is shown in **Fig. 6f**. The synthesis route first builds the chiral center through an enzymatic reaction of aminotransferase from cheap starting material 2-butanone (**20**) to form (*2R*)-butan-2-amine (**21**). **22** is synthesized by the acylation reaction of **23** and **21**, followed by reduction to form **24**. Two steps of symmetrical hydroxylation catalyzed by the same enzyme are used to complete the synthesis of **12**.

The predicted route effectively employs a single enzymatic reaction to construct the chiral portion of the molecule. Compared to chemical methods reported in the literature, the enzymatic reaction conditions are milder. The enzyme recommended for this step based on molecular similarity is L-glutamine: 2-deoxy-scyllo-inosose aminotransferase (G2DOIAT; EC number 2.6.1.100). The subsequent two symmetric hydroxylation reactions form a cascade, and a one-pot method can be employed to minimize the number of purifications. The CYP124 family of cytochrome P450 enzymes (CYP124; EC number 1.14.15.14) is recommended for this cascade. Moreover, introducing the hydroxyl group in the final step avoids side reactions during the acylation process and reduces the use of protecting groups. Although the predicted enzymatic reactions have not been experimentally verified for these substrates, the prediction still provides valuable guidance for future enzyme discovery and engineering.

Epidiolex is the brand name for the (−)-cannabidiol (**(-)-26**), which is used for the treatment of epilepsy disorders. Kobayashi et al. developed the synthesis route using olivetol dimethyl ether (**27**) and **30** as the starting materials[51]. The chirality is constructed through the nucleophilic addition of **28** and **31** to form **29**. Another synthesis route designed by Shultz et al. uses Ireland-Claisen rearrangements to build chirality starting from olivetol **32**[52]. Gong et al. used Friedel-Crafts reaction to build chirality starting with phloroglucinol (**35**) and cis-isolimonenol (**36**)[53]. The biosynthetic route of (−)-cannabidiol using hexanoyl-CoA as the starting material has also been reported[54]. The cannabidiolic acid synthase (CBDAS; EC number 1.21.3.8) uses cannabigerolic acid as substrate to close the ring and introduce stereochemistry.

The predicted route by ACERetro with a maximum search depth set to 4 and ignoring geometric isomerism in buyable molecule database is shown in **Fig. 7e**. The prediction provides a concise synthetic route, starting with the alkylation of olivetol **32** with geraniol **39** to form cannabigerol **40**. Then an enzymatic step is used to form the final product **(-)-26** with stereoisomerism. The first alkylation reaction has literature to support it[55], whereas the recommended enzyme for the second step, CBDAS, has not been proven to work using **40** as substrate. However, the high similarity between **40** and **38** points to the possibility of finding enzyme mutants that allow the reaction to occur.

Case study for optimizing synthesis routes

SPScore can also be used to optimize given synthesis routes by finding steps with opportunities for improvement that can be catalyzed by alternative catalytic methods in given routes. The steps with opportunities for improvement are selected based on the deviation between the SPScore predicted catalytic method and their catalytic method in the original route. ACERetro is then used to search alternative synthesis routes for the selected steps. The promising alternative synthesis route is appended to its original synthesis route to form a new optimized synthesis route for a given route. Herein, we conducted case studies on the synthesis route of rivastigmine (**41**) reported in the literature[56] and a synthesis route for (*R,R*)-formoterol (**42**) reported by a synthesis planning tool[33]. In the four-step organic synthesis route of the dementia drug rivastigmine **41**, the S_{Chem} of each molecule is greater than its S_{Bio}. Therefore, the intermediate **44** with the largest SPScore difference ("optimization score") was selected to search possible synthetic routes by the biocatalytic method. ACERetro with the same parameters was used for the search, and the maximum search depth was set to 1 because the intermediate **44** in the original synthesis route only takes one step to reach the commercially available molecule. The search results show that an enzymatic reaction can be found using the same starting material **43** (**Fig. 8a**). This enzymatic reaction has been validated in the literature[57], proving the effectiveness of SPScore in finding steps with opportunities for improvement and then optimizing the route.

The chemoenzymatic synthesis route for (*R,R*)-formoterol **42** was predicted by Levin et al.'s tool. Top 3 steps with opportunities for improvement (**46**, **48**, and **49**) were identified where their predicted SPScores are far away from their catalytic method in the original route. Specifically, the S_{Chem} of intermediates **46**, **48**, and **49** are larger than their corresponding S_{Bio}, yet enzymatic reactions were employed in the original route which causes a high optimization score. The new organic synthesis route for intermediate **46**, utilizing **45** as the starting compound, was predicted by ACERetro with a search depth capped at 1. Given that intermediates **49** and **48** are in the same branch, only the synthesis analysis for **48** was undertaken by ACERetro, with a maximum search depth of 2. The proposed route employs one-step chemical reaction to synthesize **48**, taking **50** and (+)-phenylethylamine as the precursor, which reduces the original three-step synthesis strategy to a single step (**Fig. 8b**). These predicted reactions for intermediates **46** and **48** have been corroborated by the literature[58,59].

Discussion

In this work, we have developed a synthetic potential guided asynchronous chemoenzymatic synthesis planning algorithm for designing chemoenzymatic synthesis routes for target molecules. When considering the evaluation of synthetic potential of molecules in each catalytic method for computer-assisted chemoenzymatic synthesis planning, our heuristic search algorithm can prioritize the exploration of the most promising catalytic method for a molecule. By leveraging the SPScore, we can also diagnose and then optimize existing synthesis routes through the identification of alternative bypasses. Consequently, the SPScore serves as a crucial link, constructing a bridge between step-by-step synthesis planning and synthesis route optimization in the design of chemoenzymatic synthesis routes. Performing asynchronous retrosynthetic searches in between the chemocatalysis and biocatalysis can significantly improve search efficiency and bolster the algorithm's robustness. This allows ACERetro to effectively address the challenge faced by the existing hybrid synthesis planners, which tend to be worse than single model planners in terms of efficiency and performance.

In addition, we capitalize on the characteristics of current chemical reaction and enzymatic reaction databases. A sufficiently large chemical reaction database can support the training of retrosynthesis tools based on language models, whereas a smaller-scale enzymatic reaction database is more suitable for rule-based reaction templates. Accordingly, we employ a template-free retrosynthesis tool, RXN4Chemistry, for chemocatalysis, and a template-based retrosynthesis tool, ASKCOS, for biocatalysis. Free from the limitations imposed by a template prioritization system, ACERetro guided by the synthetic potential possesses the capability to integrate seamlessly with any existing retrosynthesis tool.

By comparing the confidence of single-step retrosynthesis and single-step retrobiosynthesis of 11,003 molecules with the trend of SPScore distribution, it is shown that SPScore can effectively predict promising catalytic method for molecules. The performance of SPScore in multi-step retrosynthesis was further verified by catalytic method coverage, reaction retention rate, and route retention rate among predicted synthesis routes of 493 molecules. In the benchmarking study on 1,001 molecules, ACERetro outperformed the state-of-the-art method Levin et al.'s tool. Through a comparative analysis of the results obtained from FHSync, SPSync, and ACERetro, self-benchmarking reveals that the incorporation of the template-free model, the implementation of SPScore, and the adoption of asynchronous search methodologies each contribute to enhancing the performance of synthesis planning.

Examples of synthetic routes for (*S*)-verofylline, (*3S*)-3-hydroxy-β-ionone, and dimenoxadol reveal that our method can identify shortest synthesis routes with higher quality, and the predictions include not only hybrid synthesis routes, but also chemical reaction only synthesis routes and enzymatic reaction only synthesis routes. The case studies on synthesis planning of ethambutol and epidiolex demonstrate that our approach can effectively design hybrid synthesis routes for complex molecules and find potential enzyme candidates to perform the predicted enzymatic reactions. The complementarity of the two catalytic methods will further broaden the scope for designing efficient synthesis routes for molecules of interest. The case studies on synthesis route optimization for rivastigmine and (R,R)-formoterol illustrate that SPScore can be effectively applied to optimize existing synthesis routes. Existing synthesis tools are often inadequate for lengthy synthesis steps. Finding steps with opportunities for improvement that may be optimized in existing synthesis routes and then conducting retrosynthetic analysis can simplify the search process and make full use of existing parts of the synthesis routes that have been experimentally verified.

The concept underlying SPScore involves inferring the most promising catalytic method for a molecule based on existing catalysis data in a reaction database, employing a data-driven approach. This approach aims to differentiate the distinct reaction spaces of chemocatalysis and biocatalysis. Utilizing SPScore in chemoenzymatic synthesis planning can expedite the search process by avoiding less promising catalytic methods. However, there remains a risk that the model might overlook viable reactions in the catalytic methods it avoids. Consequently, a comprehensive and high-quality dataset encompassing various types of reactions is crucial to ensure optimal model performance. It is noteworthy that the reaction spaces of chemocatalysis and biocatalysis are dynamic. The unique reaction space of each may expand or contract with the discovery of new catalysts or enzymes. In ACERetro, the SPScore is firstly used to identify the promising catalytic method before conducting a retrosynthetic analysis. An alternative improvement strategy could be first conducting retrosynthetic analysis to identify all potential deconstruction sites and intermediates, then selecting the appropriate catalytic method for each step.

In summary, ACERetro demonstrates significant scalability and is not limited solely to template-based retrosynthesis tools. It represents a powerful strategy for designing efficient chemoenzymatic synthesis routes and identifying bypass opportunities to given routes. We believe that computer-aided chemoenzymatic synthesis planning will broaden the synthesis space by synergistically harnessing the unique properties of enzymes and chemocatalysis. Predicted synthesis routes are poised to accelerate the utilization of enzymes as eco-friendly catalysts in the synthesis of molecules, thereby facilitating the screening and engineering of enzymes for optimized performance.

Methods

Training the synthetic potential scoring model

The USPTO 480K database comprises 484,706 organic chemistry reactions from patents, and the ECREACT database comprises 62,222 enzymatic reactions from Rhea[60], BRENDA[61], PathBank[62], and MetaNetX[63]. After deduplication and excluding molecules with infeasible fingerprints (as detailed in Supplementary Information), we extracted 437,781 molecules from USPTO 480K (labeled with $y = 1$) and 37,939 molecules from ECREACT (labeled with $y = -1$). 515 overlapping molecules, found in both catalytic methods, are labeled with $y = 0$. An MLP model is trained to predict SPScores in chemocatalysis (S_{Chem}) and biocatalysis (S_{Bio}). The model input is molecular fingerprints. A sigmoid activation function is applied to the final layer, ensuring that the range of both scores lies between 0 and 1. The margin uses the relative value of S_{Chem} and S_{Enzy} to divide the output space to three areas corresponding to three scenarios of catalytic method as shown in **Fig. 1a**. A weighted margin ranking loss is applied to compute a criterion only when the prediction is out of the area of molecules' catalytic method. The weight is calculated based on reciprocal of the ratio among each label.

$$loss(S_{Chem}, S_{Enzy}, y) = \begin{cases} weight_i \cdot \max(0, -y(S_{Chem} - S_{Bio}) + margin) & \text{if } y = \pm 1 \\ weight_i \cdot \max(0, |S_{Chem} - S_{Bio}| - margin) & \text{if } y = 0 \end{cases} \quad (1)$$

The dataset was split into a training, validation, and test set (80%, 10% and 10%, respectively). We used a grid search to tune the hyperparameters including the type of the molecular fingerprints (ECFP4 and MAP4), the length of the molecular fingerprints (1024, 2048, or 4096), and the number of hidden layers (1, 3, or 5). The accuracy, F1, and recall are calculated on the validation set. To mitigate the risk of overfitting, the number of epochs is incorporated into the evaluation function to select the optimal models (see Supplementary Information). The optimal model, which utilizes ECFP4 embedding of 4096 length and comprises 3 hidden layers, trained for 10 epochs, was employed for all subsequent tasks.

Benchmarking the synthetic potential score

11,003 molecules were randomly selected from the "in-vitro" subset of ZINC15. SPScore was calculated for each molecule. RXN4Chemistry was employed for one-step retrosynthesis in chemocatalysis. Each predicted reaction is accompanied by a corresponding backward confidence score. Levin et al.'s enzymatic templates were employed for single-step precursor prediction in biocatalysis. Each predicted reaction is accompanied by a corresponding template score. The search parameters for RXN4Chemistry and Levin et al.'s enzymatic templates are listed in Supplementary Information. For molecules within different S_{Chem} intervals, we calculated the average confidence scores for top-5 predictions, and an analogous procedure was undertaken for S_{Bio} intervals and score difference ($S_{Chem} - S_{Bio}$) intervals.

Multi-step hybrid synthesis routes were derived from the retrosynthetic predictions for 493 molecules conducted by Levin et al.'s tool within a three-minute timeframe. Out of 493 target molecules, we enumerated 26,741 distinct product molecules in 397,040 synthetic routes. All the synthetic routes with the shortest length for each target molecule were collected, which contained 1,544 distinct product molecules. The catalytic method of a molecule (denoted as "Chem", "Bio", or "Both") is assigned based on whether the molecule has been synthesized by an organic chemical reaction or an enzymatic reaction. Catalytic method coverage out of all molecules counts the molecule whose SPScore predicted catalytic method includes the actual catalytic method out of all molecules. Saved searches out of "Chem" and "Bio" molecules count the molecule whose SPScore predicted catalytic method exactly matches the actual catalytic method for these molecules labeled with "Chem" or "Bio", so the search algorithm does not need to search the alternative catalytic method. Reaction retention rate counts the product molecule in a reaction whose SPScore predicted catalytic method includes the actual catalytic method of that reaction. The near shortest synthesis routes include synthesis routes whose lengths are less than or equal to the shortest synthesis route length plus two. Synthesis route retention rate counts the synthesis route whose reactions can be all retained (see Supplementary Information for the formulae).

Hybrid synthesis route search

1,001 compounds from the "boutique" subset of ZINC15 database are used in the benchmarking study. Three search algorithms used the identical search parameters of RXN4Chemistry and Levin et al.'s enzymatic templates as described in the previous section. Tree search architecture including iterative process of selection, expansion, and update is used for all three search algorithms. In the selection mode, the molecule that has the lowest score in the priority queue and is not in the buyable database will be selected. In the expansion mode of the fully hybrid search algorithm, RXN4Chemistry and Levin et al.'s enzymatic templates are used to predict single-step precursor for the selected molecule. The results from chemocatalysis and biocatalysis are combined, and all precursors which are not in the buyable database are scored based on the molecular complexity function (denoted as $f(P)$) and the depth with a depth exploration factor (denoted as d). In the SPScore guided synchronous hybrid search algorithm, only the promising syntheic field predicted by SPScore will be searched, precursors are scored by the same way. In the SPScore guided asynchronous hybrid search algorithm, the syntheic field corresponding to the score of the selected molecule will be searched. A catalytic method exploration factor (denoted as c) is used for the SPScore. A molecule will have two scores associated with two catalytic methods. In the update mode, all precursors with their associated scores will be appended to the priority queue and then reranked.

$$Score_i = (1 - c \cdot SPScore_i)Depth^d \cdot f(P) \quad i \in [Chem, Enzy] \tag{2}$$

The maximum search depth and the expansion time were 10 and 180s respectively. For a fair comparison, the above parameters together with commercially available compound database from the vendors eMolecules and Sigma-Aldrich are consistent with those used in Levin et al.'s tool (additional parameters in Supplementary Information). When the search reaches the time limit, all synthesis routes from buyable molecules to the target molecule are returned.

Case studies on synthesis planning

In the synthesis planning of (*S,S*)-ethambutol and epidiolex, ACERetro is used to search synthesis routes with a maximum search depth set to 5 and 4, respectively. Because the buyable compound database does not contain complete geometric isomerism information of molecules, when searching in the buyable database, geometric isomerism of molecules is ignored, and optical isomerism is retained. All other parameters of ACERetro are the same as those used in the benchmarking tools. In enzymatic reactions, enzymes are selected based on the similarity of products under the same reaction template.

Case studies on synthesis route optimization

For a given synthetic route, the SPScore is calculated for all molecules except the starting molecule. Steps with opportunities for improvement are determined by comparing the relative value of two SPScores and the actual catalytic method of the molecule in the route. The indexes of top-n steps with opportunities for improvement can be retrieved by Equation (3), where $y_i = -1$ if the molecule's catalytic method in the given synthesis route is chemocatalysis (labeled as "Chem"), and $y_i = 1$ if the molecule's catalytic method in the given synthesis route is biocatalysis "Bio". The equation aims to find top-n molecules with the largest SPScore difference away from the molecule's catalytic method ("optimization score"). ACERetro is used to search the synthesis route of steps with opportunities for improvement. The search depth to molecules is set to current length to starting molecules in the original synthesis route. For molecule **44** and **46,** the search depth is set as 1. The search depth to molecule **48** is set as 2. All other parameters of ACERetro are the same as those used in the case studies for synthesis planning.

$$I_n^{desc} = \mathrm{argsort}\big(y_i(SPScore_{Chem}^i - SPScore_{Bio}^i)\big)[-n:][::-1] \tag{3}$$

Data availability

The RXN4Chemistry models are the intellectual property of IBM, they are accessible through the IBM RXN for Chemistry website or at ref. 28. The ASKCOS model is available at ref. 33.

Code availability

The scripts for training SPScore, benchmarking, and synthesis planning in this manuscript are available at https://github.com/Zhao-Group/ACERetro[64].

Acknowledgements

This work was supported by the Molecule Maker Lab Institute: An AI Research Institutes program supported by the US National Science Foundation (NSF) under grant no. 2019897 and IBM-Illinois Accelerated Discovery Institute. We thank Teodoro Laino, Alain Claude Vaucher, and Amol Thakkar from IBM Research Europe for providing technical support on RXN4Chemistry, Dr. Haiyang Cui for help with analyzing case studies, and Aashutosh Boob for the constructive discussion on model design.

Author contributions

X.L. and H.Z. conceptualized the project. X.L. performed model development and analyses. H.L. supported model evaluation. X.L. and H.Z. wrote the manuscript with input from all authors.

Competing interests

The authors declare no competing interests.

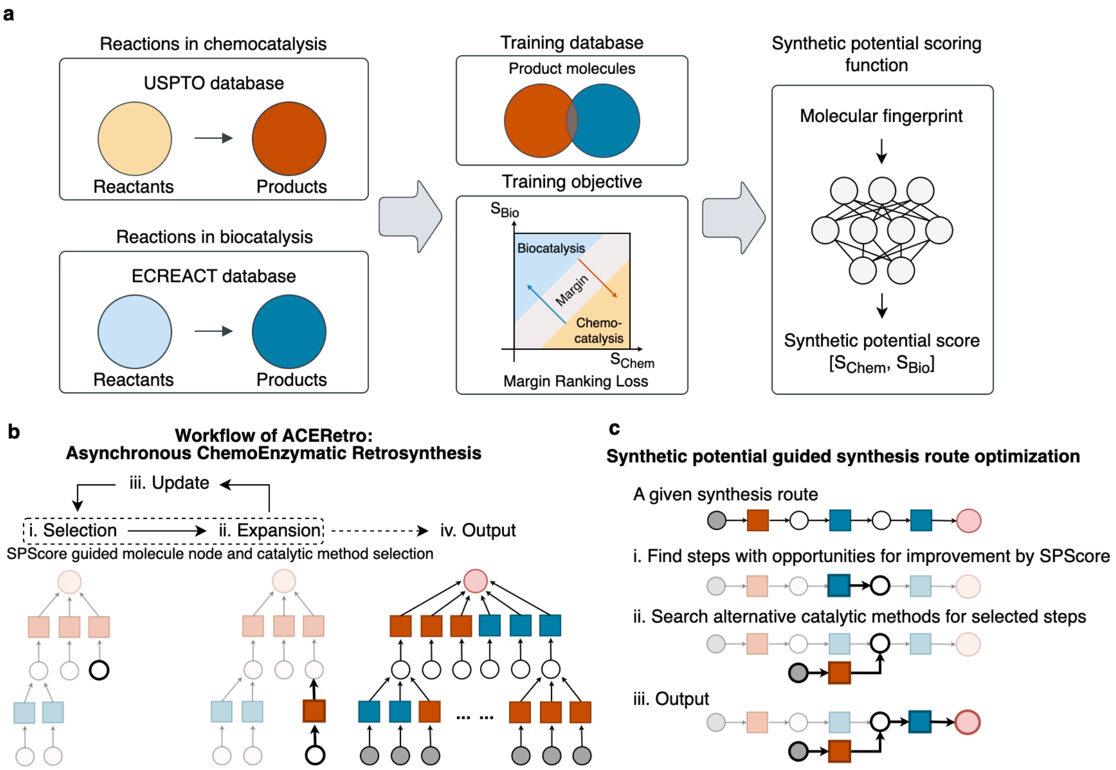
a
Reactions in chemocatalysis
USPTO database
Reactants
Products
Reactions in biocatalysis
ECREACT database
Reactants
Products
Training database
Product molecules
Training objective
S_Bio
Biocatalysis
Margin
Chemo-catalysis
S_Chem
Margin Ranking Loss
Synthetic potential scoring function
Molecular fingerprint
Synthetic potential score
[S_Chem, S_Bio]
b
Workflow of ACERetro:
Asynchronous ChemoEnzymatic Retrosynthesis
iii. Update
i. Selection
ii. Expansion
iv. Output
SPScore guided molecule node and catalytic method selection
... ...
c
Synthetic potential guided synthesis route optimization
A given synthesis route
i. Find steps with opportunities for improvement by SPScore
ii. Search alternative catalytic methods for selected steps
iii. Output

Fig. 1. Chemoenzymatic synthesis planning guided by synthetic potential score (SPScore). a. Development of the SPScore model. Reaction product molecules were extracted from USPTO (chemocatalysis) and ECREACT (biocatalysis) respectively. A neural network model is trained to infer the promising catalytic method for a given molecule through the predicted SPScore. **b.** Workflow of the synthetic potential guided chemoenzymatic synthesis planning process. The target molecule is labeled by a red circle, and reactions in chemocatalysis and biocatalysis are labeled by red squares and blue squares, respectively. (i) Selection: the molecule with the lowest score in the priority queue is selected. (ii) Expansion: the retrosynthesis tool using the catalytic method inferred by SPScore is used to predict reactions and precursors for the selected molecule. (iii) Update: the expansion results are added to the search tree. Precursors are scored and appended to the priority queue. (iv) Output: Step i, ii, and iii are executed recursively until a termination condition is met. When the search process is terminated, synthesis routes to the target molecule started with buyable molecules (gray circles) are returned. **c.** Workflow of the synthetic potential guided synthesis route optimization. (i) Identify steps with opportunities for improvement using SPScore. For a given synthetic route, the SPScore of each molecule is computed. The selected steps (bold) are determined based on their predicted SPScores that deviate significantly from the actual catalytic method used in the existing route. (ii) Search alternative catalytic methods for the selected steps. The synthesis planning algorithm in **b** is used to search synthesis routes with alternative catalytic methods for the selected steps. (iii) Output. The promising search results are appended to the original route, and the optimized route is returned.

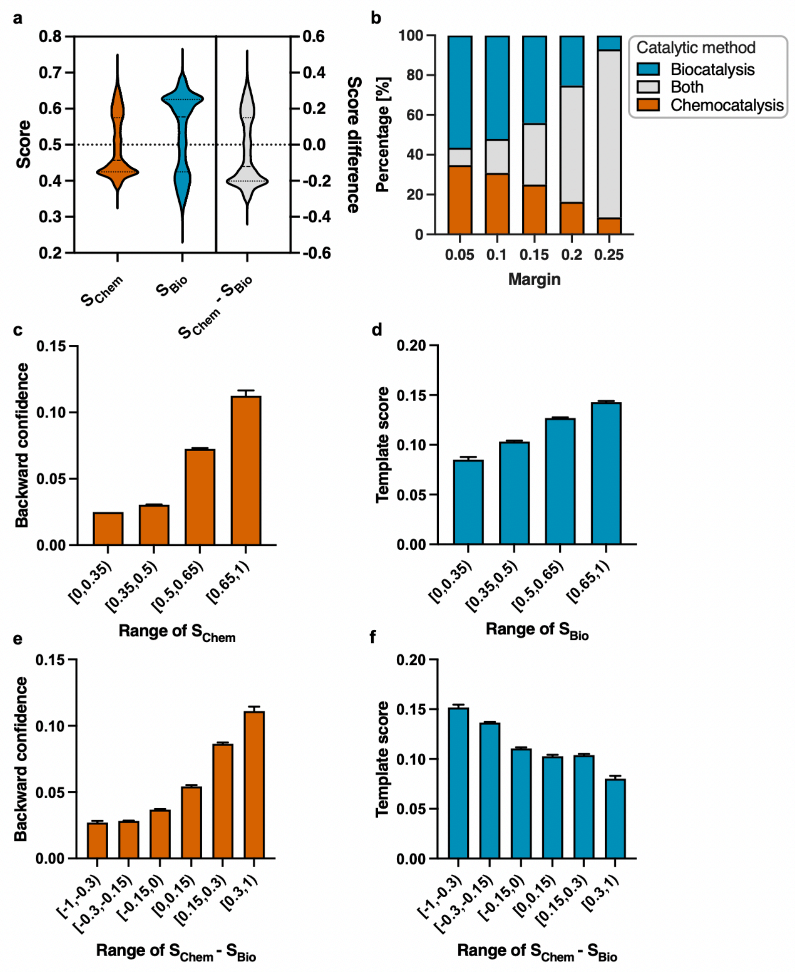

Fig. 2. Analysis of SPScore on molecules from ZINC "in-vitro" subset. a. The distribution of S_{Chem}, S_{Bio} and the score difference ($S_{Chem} - S_{Bio}$). **b.** The percentage of predicted catalytic method of molecules versus different margin settings. **c.** The mean backward confidence with SEM (the standard error of the mean) versus the range of S_{Chem}. In chemocatalysis, RXN4Chemistry is used to predict retrosynthetic reactions for the molecules. The average backward confidence of the top-5 predictions is sorted by the range of molecules' synthetic potential score in chemocatalysis. **d.** The mean template score with SEM versus the range of S_{Enzy}. In biocatalysis, Levin et al.'s enzymatic templates are used to predict enzymatic reactions for the molecules. The average template score of the top-5 predictions is sorted by the range of molecules' synthetic potential score in biocatalysis. **e.** The mean backward confidence with SEM versus the range of $S_{Chem} - S_{Bio}$. **f.** The mean template score with SEM versus the range of $S_{Chem} - S_{Bio}$.

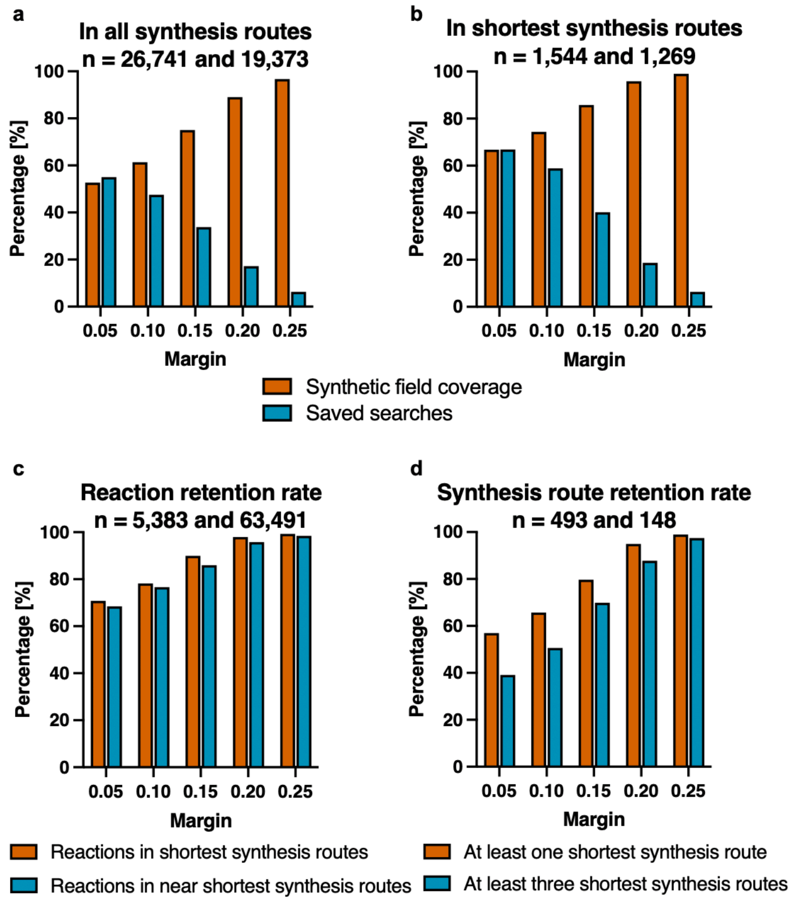

Fig. 3. Analysis of SPScore on synthesis routes. Reaction products are extracted from reactions in all synthesis routes and molecules' shortest synthesis routes predicted by Levin et al.'s tool on 493 molecules. Molecules are assigned with the ground truth label of catalytic method based on the extracted reaction set. The amount of "catalytic method coverage" is counted when the SPScore gives the correct prediction or predicts it as "both". The amount of "saved searches" is only counted when the SPScore gives the correct catalytic method prediction. The percentage of catalytic method coverage out of all molecules (red) and saved searches out of non-"both" molecules (blue) are calculated among different margin for product molecules **a** in all synthesis routes and **b** in shortest synthesis routes. **c.** The reaction retention rate for the shortest synthesis routes (red) and near shortest synthesis routes (blue) against different margin settings. The reactions from shortest synthesis routes and the near shortest synthesis routes (where the route length ≤ shortest route length + 2) are extracted. The amount of retained reactions is counted when the SPScore gives the correct catalytic method prediction for the product molecule or predicts it as "both". **d.** The shortest synthesis routes retention rate against different margin settings. The amount of retained synthesis routes is counted when all the reactions in the synthesis route can be retrained by the SPScore's prediction.

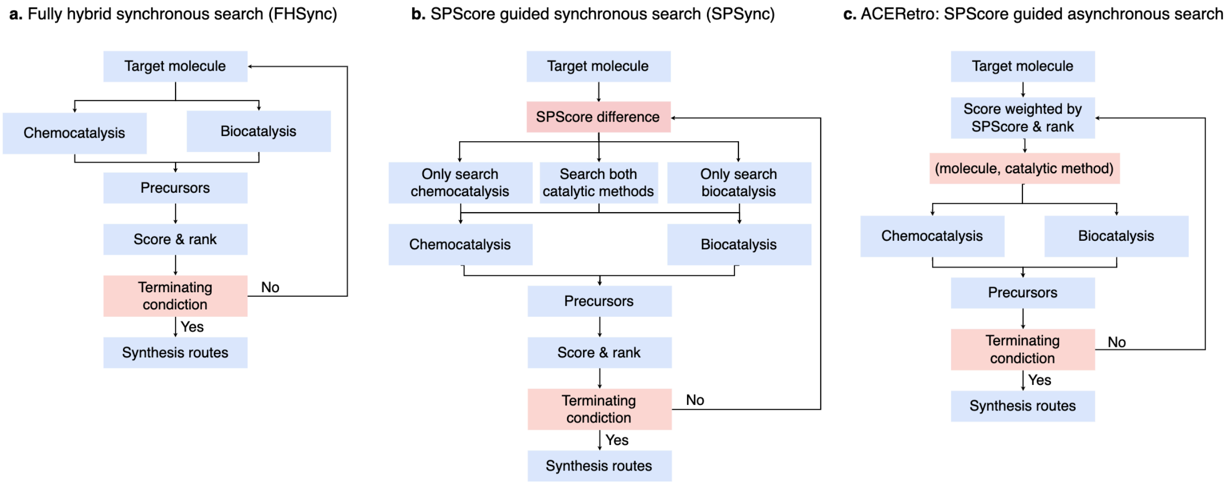

Fig. 4. Hybrid search algorithms for designing chemoenzymatic synthesis routes. a. Fully hybrid synchronous search algorithm (FHSync). **b.** SPScore guided synchronous search algorithm (SPSync). **c.** ACERetro: SPScore guided asynchronous search algorithm.

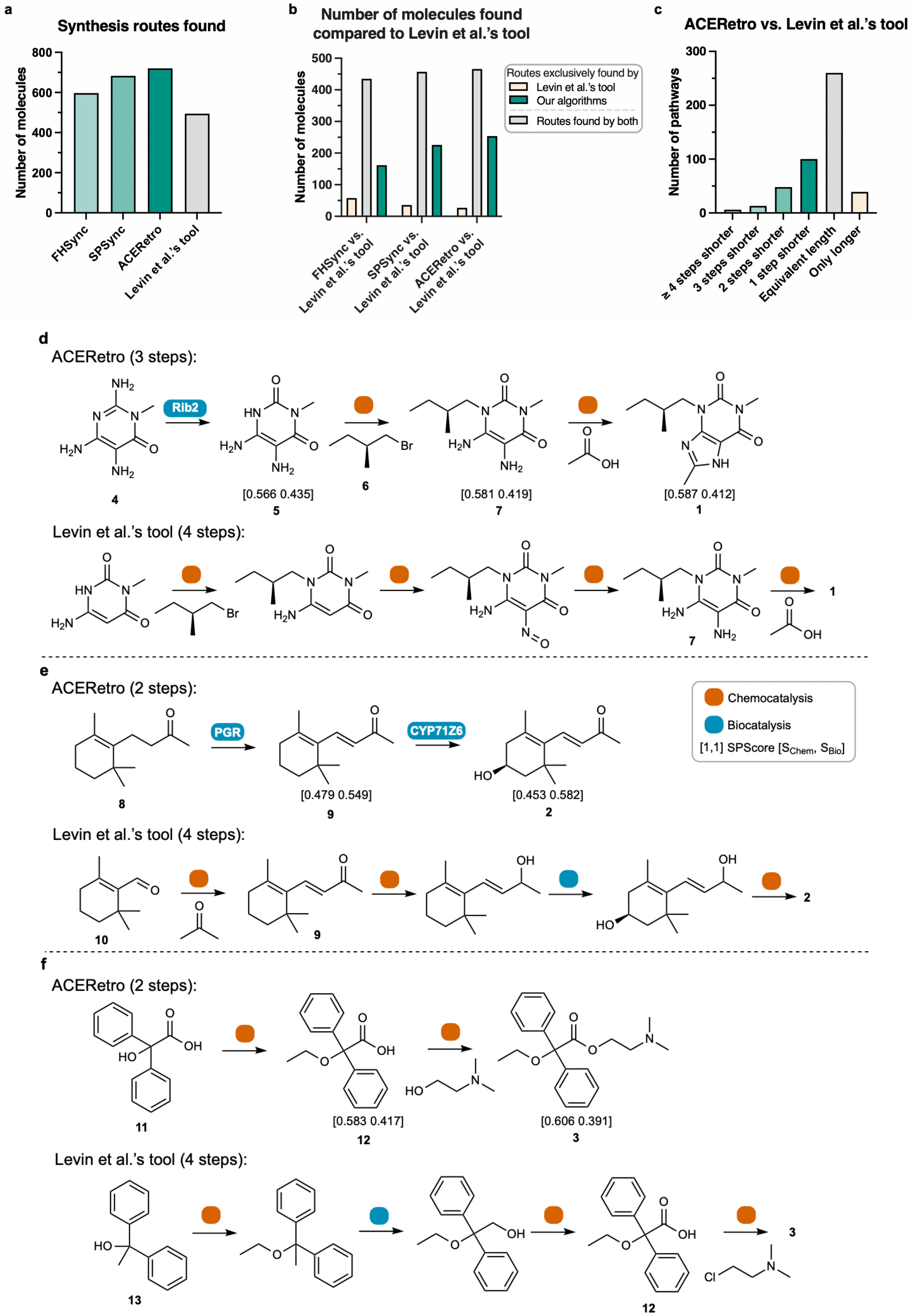
a
Synthesis routes found
Number of molecules
800
600
400
200
0
FHSync
SPSync
ACERetro
Levin et al.'s tool
b
Number of molecules found compared to Levin et al.'s tool
Number of molecules
500
400
300
200
100
0
FHSync vs. Levin et al.'s tool
SPSync vs. Levin et al.'s tool
ACERetro vs. Levin et al.'s tool
Routes exclusively found by
Levin et al.'s tool
Our algorithms
Routes found by both
c
ACERetro vs. Levin et al.'s tool
Number of pathways
300
200
100
0
≥ 4 steps shorter
3 steps shorter
2 steps shorter
1 step shorter
Equivalent length
Only longer
d
ACERetro (3 steps):
Rib2
4
[0.566 0.435]
5
6
[0.581 0.419]
7
[0.587 0.412]
1
Levin et al.'s tool (4 steps):
7
1
e
ACERetro (2 steps):
PGR
CYP71Z6
8
[0.479 0.549]
9
[0.453 0.582]
2
Chemocatalysis
Biocatalysis
[1,1] SPScore [S_Chem, S_Bio]
Levin et al.'s tool (4 steps):
10
9
2
f
ACERetro (2 steps):
11
[0.583 0.417]
12
[0.606 0.391]
3
Levin et al.'s tool (4 steps):
13
12
3

Fig. 5. Comparison of synthesis routes found by hybrid search algorithms. a. Number of molecules for which synthesis routes were found out of 1,001 molecules. **b.** Number of molecules that their synthesis routes can be found by Levin et al.'s tool (red) only, both (grey), and ours (blue). **c.** Comparison of the number of steps in the shortest synthesis route found by ACERetro compared to the shortest synthesis route found by Levin et al.'s tool for molecules for which synthesis routes were found by both (466 total) from the ZINC15 "boutique" subset. The example synthesis routes of (*S*)-verofylline (**1**), (*3S*)-3-Hydroxy-β-ionone (**2**), and dimenoxadol (**3**) are shown in **d-f**. All product molecules, except for **9**, do not appear in the training set of the SPScore. Cofactors and some non-primary reactants are ignored. Enzymes are selected based on the similarity of products under the same reaction template. Rib2: 2,5-diamino-6-(5-phospho-D-ribitylamino)-pyrimidin-4(3H)-one deaminase, PGR: 13,14-dehydro-15-oxoprostaglandin 13-reductase, CYP71Z6: ent-isokaurene C2-hydroxylase.

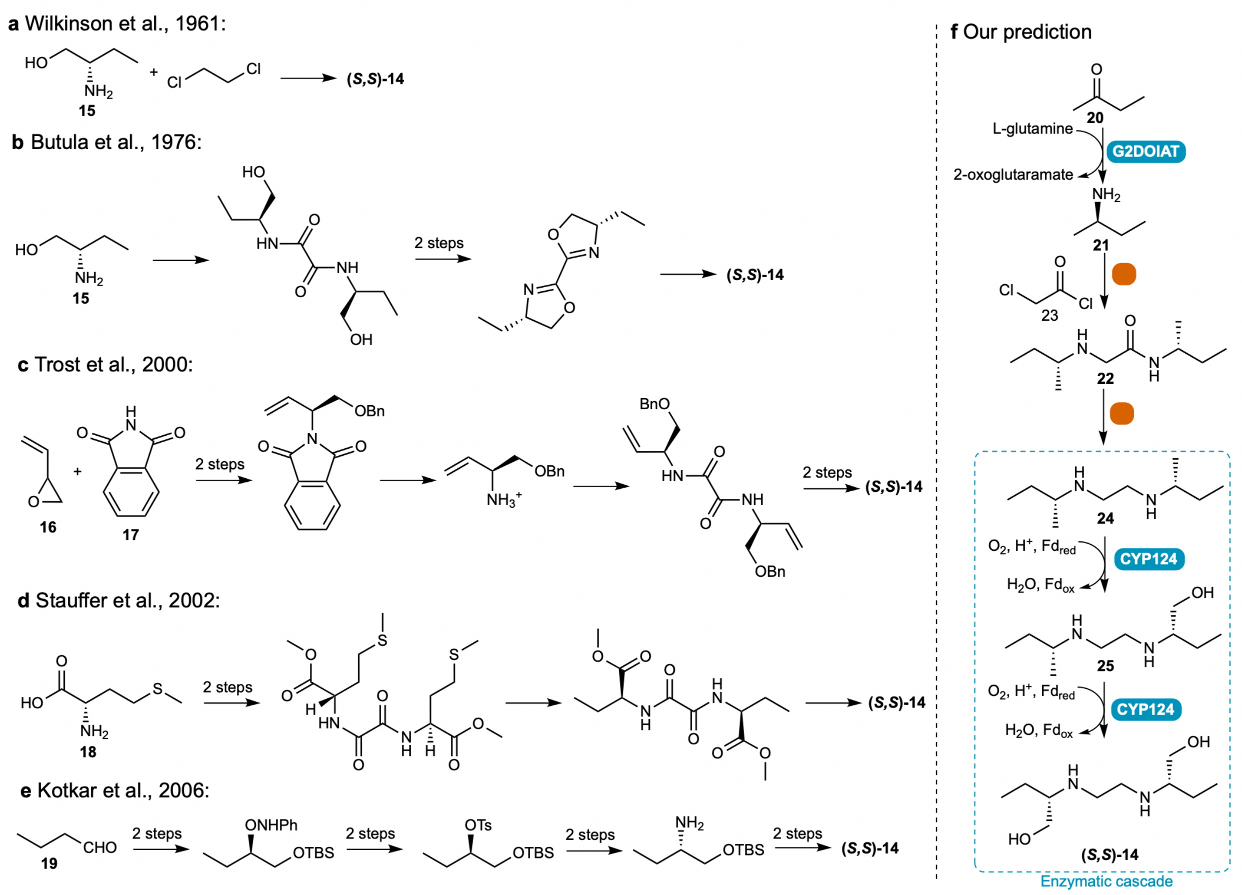

Fig. 6. Case study of ethambutol. a-e. Published synthesis routes of ethambutol. **f.** Predicted synthesis route of ethambutol. Ethambutol does not appear in the training set of the enzymatic model. Enzymes are selected based on the similarity of products under the same reaction template. G2DOIAT: L-glutamine:2-deoxy-scyllo-inosose aminotransferase, CYP124: CYP124 family of cytochrome P450 enzymes.

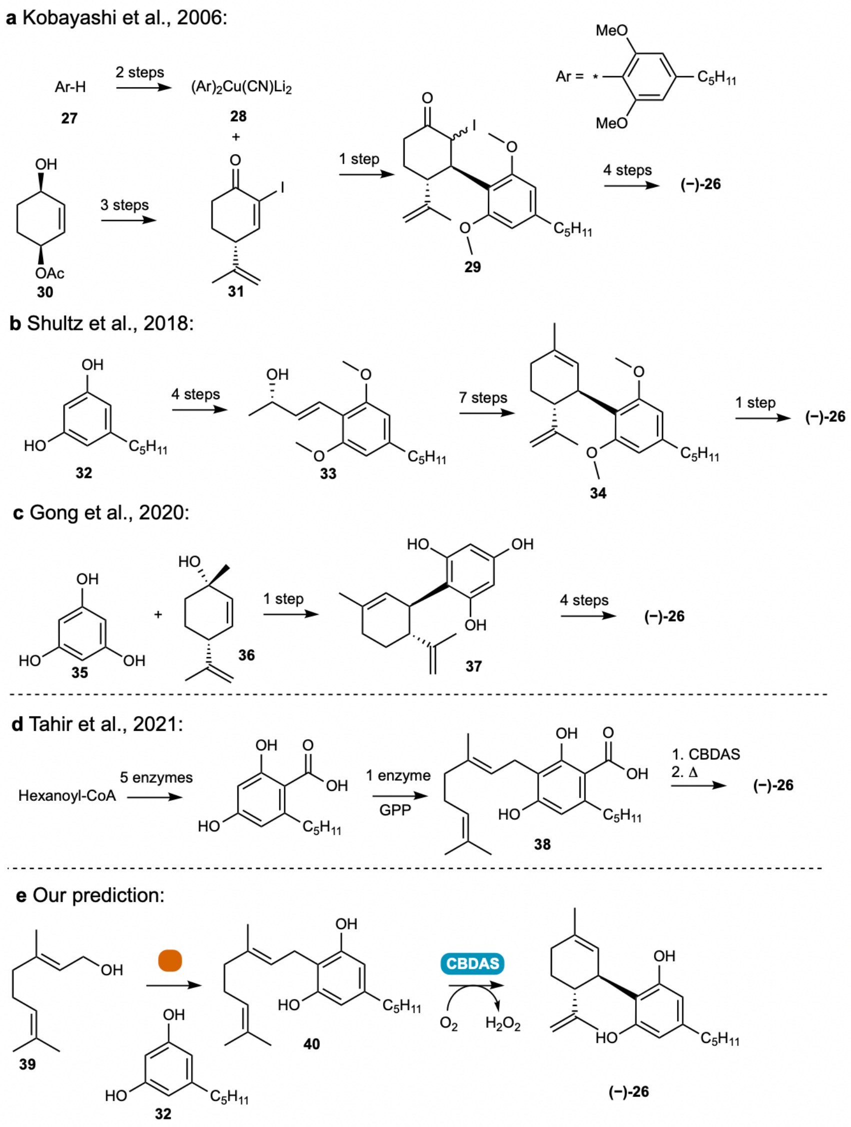

Fig. 7. Case study of epidiolex. a-d. Published synthesis routes of epidiolex. **e.** Predicted synthesis route of ethambutol. Epidiolex does not appear in the training set of the enzymatic model. Enzymes are selected based on the similarity of products under the same reaction template. CBDAS: Cannabidiolic acid synthase.

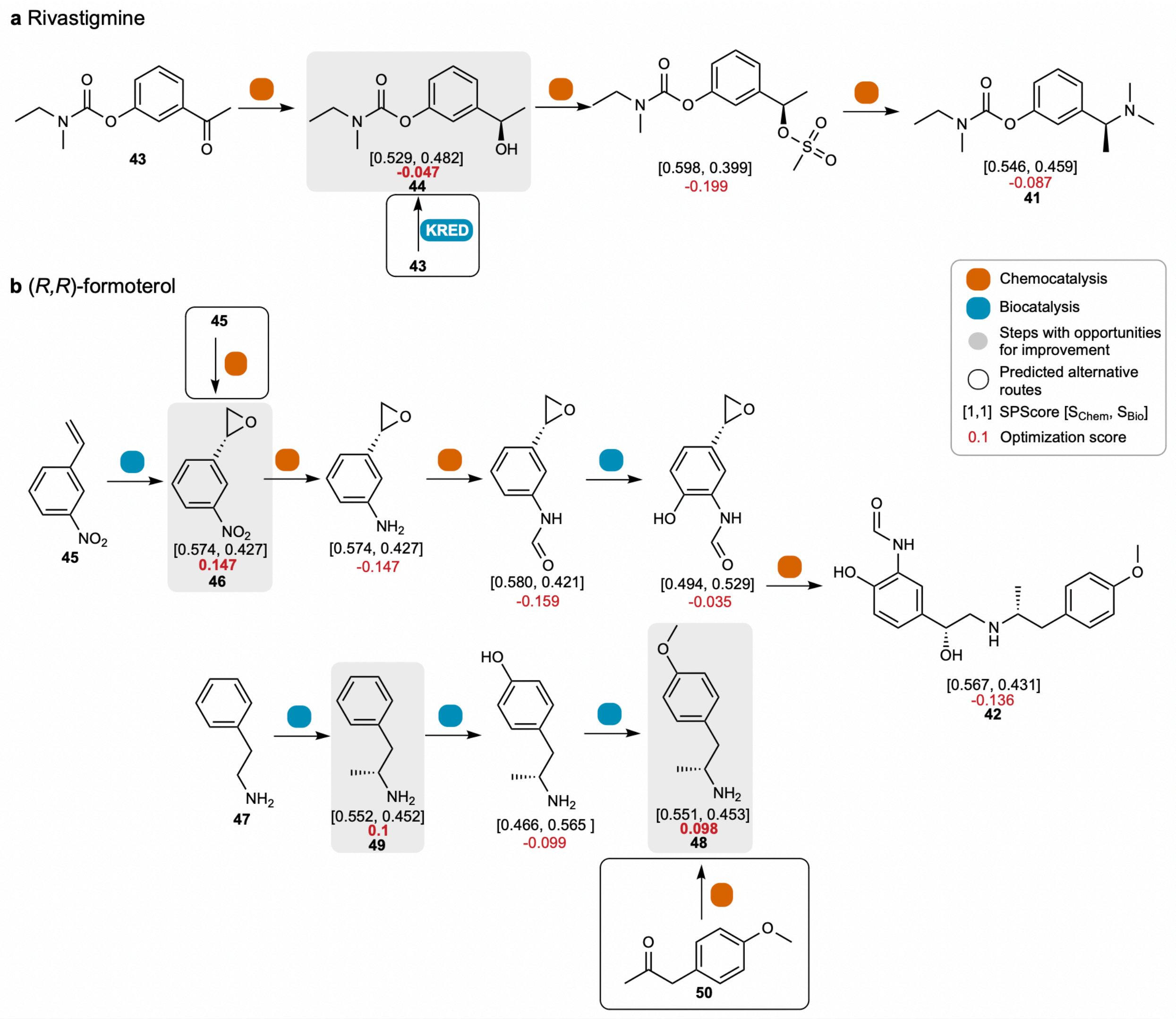

Fig. 8. SPScore guided synthesis route optimization. a. Synthesis route optimization of rivastigmine. **b.** Synthesis route optimization of (*R*,*R*)-formoterol. Steps with opportunities for improvement do not appear in the training set of the SPScore. (+)-Phenylethylamine, which is available in the buyable database, and other reagents are not shown in the diagram. The predicted bypasses have literature support. KRED: ketoreductase.

References

1. Simić, S. *et al.* Shortening synthetic routes to small molecule active pharmaceutical ingredients employing biocatalytic methods. *Chem. Rev.* **122**, 1052–1126 (2022).
2. Kaspar, F. & Schallmey, A. Chemo-enzymatic synthesis of natural products and their analogs. *Curr. Opin. Biotechnol.* **77**, 102759 (2022).
3. Chakrabarty, S., Romero, E. O., Pyser, J. B., Yazarians, J. A. & Narayan, A. R. H. Chemoenzymatic total synthesis of natural products. *Acc. Chem. Res.* **54**, 1374–1384 (2021).
4. Li, J., Amatuni, A. & Renata, H. Recent advances in the chemoenzymatic synthesis of bioactive natural products. *Curr. Opin. Chem. Biol.* **55**, 111–118 (2020).
5. Rudroff, F. *et al.* Opportunities and challenges for combining chemo- and biocatalysis. *Nat. Catal.* **1**, 12–22 (2018).
6. McIntosh, J. A. *et al.* Engineered ribosyl-1-kinase enables concise synthesis of molnupiravir, an antiviral for COVID-19. *ACS Cent. Sci.* **7**, 1980–1985 (2021).
7. Coley, C. W., Green, W. H. & Jensen, K. F. Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **51**, 1281–1289 (2018).
8. Thakkar, A. *et al.* Artificial intelligence and automation in computer aided synthesis planning. *React. Chem. Eng.* **6**, 27–51 (2021).
9. Sun, Y. & Sahinidis, N. V. Computer-aided retrosynthetic design: fundamentals, tools, and outlook. *Curr. Opin. Chem. Eng.* **35**, 100721 (2022).
10. Szymkuć, S. *et al.* Computer-assisted synthetic planning: the end of the beginning. *Angew. Chem. Int. Ed.* **55**, 5904–5937 (2016).
11. Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **555**, 604–610 (2018).
12. Coley, C. W. *et al.* A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **365**, eaax1566 (2019).
13. Genheden, S. *et al.* AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminformatics* **12**, 70 (2020).
14. Delépine, B., Duigou, T., Carbonell, P. & Faulon, J.-L. RetroPath2.0: A retrosynthesis workflow for metabolic engineers. *Metab. Eng.* **45**, 158–170 (2018).
15. Finnigan, W., Hepworth, L. J., Flitsch, S. L. & Turner, N. J. RetroBioCat as a computer-aided synthesis planning tool for biocatalytic reactions and cascades. *Nat. Catal.* **4**, 98–104 (2021).
16. Fortunato, M. E., Coley, C. W., Barnes, B. C. & Jensen, K. F. Data augmentation and pretraining for template-based retrosynthetic prediction in computer-aided synthesis planning. *J. Chem. Inf. Model.* **60**, 3398–3407 (2020).
17. Kumar, A., Wang, L., Ng, C. Y. & Maranas, C. D. Pathway design using de novo steps through uncharted biochemical spaces. *Nat. Commun.* **9**, 184 (2018).
18. Koch, M., Duigou, T. & Faulon, J.-L. Reinforcement learning for bioretrosynthesis. *ACS Synth. Biol.* **9**, 157–168 (2020).
19. Lin, K., Xu, Y., Pei, J. & Lai, L. Automatic retrosynthetic route planning using template-free models. *Chem. Sci.* **11**, 3355–3364 (2020).
20. Zheng, S., Rao, J., Zhang, Z., Xu, J. & Yang, Y. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *J. Chem. Inf. Model.* **60**, 47–55 (2020).
21. Wang, X. *et al.* RetroPrime: a diverse, plausible and transformer-based method for single-step retrosynthesis predictions. *Chem. Eng. J.* **420**, 129845 (2021).
22. Wan, Y., Hsieh, C.-Y., Liao, B. & Zhang, S. Retroformer: pushing the limits of end-to-end retrosynthesis transformer. in *Proceedings of the 39th International Conference on Machine Learning* 22475–22490 (PMLR, 2022).
23. Thakkar, A. *et al.* Unbiasing retrosynthesis language models with disconnection prompts. *ACS Cent. Sci.* **9**, 1488–1498 (2023).

24. Liu, S. *et al.* FusionRetro: molecule representation fusion via in-context learning for retrosynthetic planning. in *Proceedings of the 40th International Conference on Machine Learning* 22028–22041 (PMLR, 2023).
25. Zhang, K., Mann, V. & Venkatasubramanian, V. G-MATT: single-step retrosynthesis prediction using molecular grammar tree transformer. *AIChE J.* **70**, e18244 (2024).
26. Yu, T. *et al.* Machine learning-enabled retrobiosynthesis of molecules. *Nat. Catal.* **6**, 137–151 (2023).
27. Schwaller, P. *et al.* Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chem. Sci.* **11**, 3316–3325 (2020).
28. Toniato, A., Vaucher, A. C., Schwaller, P. & Laino, T. Enhancing diversity in language based models for single-step retrosynthesis. *Digit. Discov.* **2**, 489–501 (2023).
29. Ucak, U. V., Ashyrmamatov, I., Ko, J. & Lee, J. Retrosynthetic reaction pathway prediction through neural machine translation of atomic environments. *Nat. Commun.* **13**, 1186 (2022).
30. Probst, D. *et al.* Biocatalysed synthesis planning using data-driven learning. *Nat. Commun.* **13**, 964 (2022).
31. Zheng, S. *et al.* Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. *Nat. Commun.* **13**, 3342 (2022).
32. Zhang, C. & Lapkin, A. A. Reinforcement learning optimization of reaction routes on the basis of large, hybrid organic chemistry–synthetic biological, reaction network data. *React. Chem. Eng.* **8**, 2491–2504 (2023).
33. Levin, I., Liu, M., Voigt, C. A. & Coley, C. W. Merging enzymatic and synthetic chemistry with computational synthesis planning. *Nat. Commun.* **13**, 7747 (2022).
34. Zeng, T., Jin, Z., Zheng, S., Yu, T. & Wu, R. Developing BioNavi for hybrid retrosynthesis planning. *JACS Au* **4**, 2492–2502 (2024).
35. Sankaranarayanan, K. & F. Jensen, K. Computer-assisted multistep chemoenzymatic retrosynthesis using a chemical synthesis planner. *Chem. Sci.* **14**, 6467–6475 (2023).
36. Coley, C. W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).
37. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **50**, 742–754 (2010).
38. Capecchi, A., Probst, D. & Reymond, J.-L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminformatics* **12**, 43 (2020).
39. Coley, C. W., Rogers, L., Green, W. H. & Jensen, K. F. SCScore: synthetic complexity learned from a reaction corpus. *J. Chem. Inf. Model.* **58**, 252–261 (2018).
40. 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.* **11**, 154–168 (2019).
41. Sterling, T. & Irwin, J. J. ZINC 15 – ligand discovery for everyone. *J. Chem. Inf. Model.* **55**, 2324–2337 (2015).
42. Mo, Y. *et al.* Evaluating and clustering retrosynthesis pathways with learned strategy. *Chem. Sci.* **12**, 1469–1478 (2021).
43. Cornwall, P., Diorazio, L. J. & Monks, N. Route design, the foundation of successful chemical development. *Bioorg. Med. Chem.* **26**, 4336–4347 (2018).
44. Wilkinson, R. G., Shepherd, R. G., Thomas, J. P. & Baughn, C. Stereospecificity in a new type of synthetic antituberculous agent. *J. Am. Chem. Soc.* **83**, 2212–2213 (1961).
45. Shepherd, R. G. & Wilkinson, R. G. Antituberculous agents. II. N,N′-diisopropylethylenediamine and analogs. *J. Med. Pharm. Chem.* **5**, 823–835 (1962).
46. Wilkinson, R. G., Cantrall, M. B. & Shepherd, R. G. Antituberculous agents. III. (+)-2,2 -(ethylenediimino)-di-1-butanol and some analogs. *J. Med. Pharm. Chem.* **5**, 835–845 (1962).
47. Butula, I. & Karlović, G. Katalytische hydrierung von stickstoffhaltigen heterocyclen, IV1) hydrogenolyse von verbrückten oxazolinen und oxazolidinen. *Justus Liebigs Ann. Chem.* **1976**, 1455–1464 (1976).

48. Stauffer, C. S. & Datta, A. Efficient synthesis of (S,S)-ethambutol from l-methionine. *Tetrahedron* **58**, 9765–9767 (2002).

49. Trost, B. M., Bunt, R. C., Lemoine, R. C. & Calkins, T. L. Dynamic kinetic asymmetric transformation of diene monoepoxides: a practical asymmetric synthesis of vinylglycinol, vigabatrin, and ethambutol. *J. Am. Chem. Soc.* **122**, 5968–5976 (2000).

50. Kotkar, S. P. & Sudalai, A. Enantioselective synthesis of (S,S)-ethambutol using proline-catalyzed asymmetric α-aminooxylation and α-amination. *Tetrahedron Asymmetry* **17**, 1738–1742 (2006).

51. Kobayashi, Y., Takeuchi, A. & Wang, Y.-G. Synthesis of cannabidiols via alkenylation of cyclohexenyl monoacetate. *Org. Lett.* **8**, 2699–2702 (2006).

52. Shultz, Z. P., Lawrence, G. A., Jacobson, J. M., Cruz, E. J. & Leahy, J. W. Enantioselective total synthesis of cannabinoids - a route for analogue development. *Org. Lett.* **20**, 381–384 (2018).

53. Gong, X. *et al.* Synthesis of CBD and its derivatives bearing various C4'-side chains with a late-stage diversification method. *J. Org. Chem.* **85**, 2704–2715 (2020).

54. Stout, J. M., Boubakir, Z., Ambrose, S. J., Purves, R. W. & Page, J. E. The hexanoyl-CoA precursor for cannabinoid biosynthesis is formed by an acyl-activating enzyme in Cannabis sativa trichomes. *Plant J.* **71**, 353–365 (2012).

55. Seccamani, P. *et al.* Photochemistry of cannabidiol (CBD) revised. A combined preparative and spectrometric investigation. *J. Nat. Prod.* **84**, 2858–2865 (2021).

56. Yan, P.-C. *et al.* Industrial scale-up of enantioselective hydrogenation for the asymmetric synthesis of rivastigmine. *Org. Process Res. Dev.* **17**, 307–312 (2013).

57. Sethi, M. K. *et al.* Asymmetric synthesis of an enantiomerically pure rivastigmine intermediate using ketoreductase. *Tetrahedron Asymmetry* **24**, 374–379 (2013).

58. Agarwala, H. *et al.* Electronic structure and catalytic aspects of [Ru(tpm)(bqdi)(Cl/H2O)]n, tpm = tris(1-pyrazolyl)methane and bqdi = o-benzoquinonediimine. *Dalton Trans.* **42**, 3721–3734 (2013).

59. Laroche, B., Ishitani, H. & Kobayashi, S. Direct reductive amination of carbonyl compounds with H2 using heterogeneous catalysts in continuous fow as an alternative to N-alkylation with alkyl halides. *Adv. Synth. Catal.* **360**, 4699–4704 (2018).

60. Alcántara, R. *et al.* Rhea—a manually curated resource of biochemical reactions. *Nucleic Acids Res.* **40**, D754–D760 (2012).

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

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

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

64. Liu, X. ACERetro. Zenodo https://doi.org/10.5281/zenodo.10578664 (2024).