Digital Discovery

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: X. Liu, H. Zhao and H. Li, *Digital Discovery*, 2025, DOI: 10.1039/D5DD00008D.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/digitaldiscovery

View Article Online
DOI: 10.1039/D5DD00008D

ARTICLE

Chemoenzymatic synthesis planning guided by synthetic potential score

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

Xuan Liu, [a,b,c] Hongxiang Li [a,b,c,d] and Huimin Zhao *[a,b,c,d,e]

Computer-aided chemoenzymatic synthesis planning integrates the advantages of enzymatic and organic reactions to design efficient hybrid synthesis routes for a target molecule. Existing tools rely on either a step-by-step strategy or a bypass strategy. Here we introduce a synthetic potential score (SPScore) to unify these two strategies. This score is developed by training a multilayer perceptron on existing reaction databases to evaluate the potential of enzymatic or organic reaction for synthesis of a molecule. We systematically evaluate the effectiveness of SPScore in both single-step and multi-step hybrid retrosynthesis, demonstrating its strong ability to prioritize promising reaction types. In benchmarking various chemoenzymatic retrosynthesis algorithms guided by SPScore, we find an asynchronous search algorithm named ACERetro yields higher efficiency and robustness that can find hybrid synthesis routes to 46% more molecules compared with the state-of-the-art tool using a test dataset consisting of 1001 molecules. We then apply ACERetro to design efficient chemoenzymatic synthesis routes for 4 FDA-approved drugs. We anticipate that the application of SPScore will provide a new avenue for computer-aided chemoenzymatic synthesis planning, thereby advancing the synthesis of functional molecules.

Introduction

Enzymatic and organic reactions span distinct reaction spaces in terms of designing synthesis routes for molecules of interest due to their different characteristics (1). Enzymatic reactions typically exhibit excellent selectivity (stereo-, chemo-, or regio-), while organic reactions are advantageous due to their broad substrate scope, vast types of reactions, and numerous well-studied cases. Combining these two reaction types can capitalize on their unique advantages to build more efficient chemoenzymatic synthesis routes to many compounds (2–6). 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 (7).

Computer-aided synthesis planning (CASP) enables massive search and design of synthesis routes for a target molecule by integrating template-based (8,9) or template-free (10,11) single-step retrosynthesis predictors with a search algorithm (12,13). To achieve computer-aided chemoenzymatic synthesis planning, currently there are two distinct strategies to build a search algorithm (**Fig. 1A**) including (i) step-by-step (14,15) and (ii) bypass (16,17). The step-by-step strategy combines the results from single step enzymatic/organic reaction precursor predictors to build a hybrid synthesis route (14,15), whereas the bypass strategy identifies alternative reaction types in an existing or predicted synthesis route, i.e. identify enzymatic reactions as bypasses to chemical syntheses or vice versa (16,17). Levin et al.'s tool combines precursor prediction results from two reaction template prioritizers trained on separate reaction databases (14), but template prioritizers cannot heuristically identify the bypass without proper alignment as the two prioritizer models are trained separately. Similarly, Sankaranarayanan et al.'s tool employs an exhaustive search to identify biocatalytic opportunities for intermediates in predicted synthesis routes (16), but this tool is challenging to scale in an exponentially growing search space. More recently, Zeng et al.'s step-by-step strategy tool predicts reaction types (organic or enzymatic reaction) with a template-free precursor predictor (15), while Li et al.'s bypass strategy tool builds a reaction type score (RTscore) to distinguish synthesis reactions from decomposition reactions (17), highlighting the importance of heuristic methods for advancing computer-aided chemoenzymatic retrosynthesis. However, Zeng et al.'s method and Li et al.'s method can only be applied within their own step-by-step or bypass strategies. Finding a simple and effective way to unify the step-by-step strategy and the bypass strategy can help us further understand the principles behind computer-aided chemoenzymatic synthesis planning and promote the

[a] NSF Molecule Maker Lab Institute, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.
[b] Department of Chemical and Biomolecular Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.
[c] Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.
[d] Department of Chemistry, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.
[e] DOE Center for Advanced Bioenergy and Bioproducts Innovation, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.
* Corresponding to: zhao5@illinois.edu.
Supplementary Information available: [details of any supplementary information available should be included here]. See DOI: 10.1039/x0xx00000x

View Article Online
DOI: 10.1039/D5DD00008D

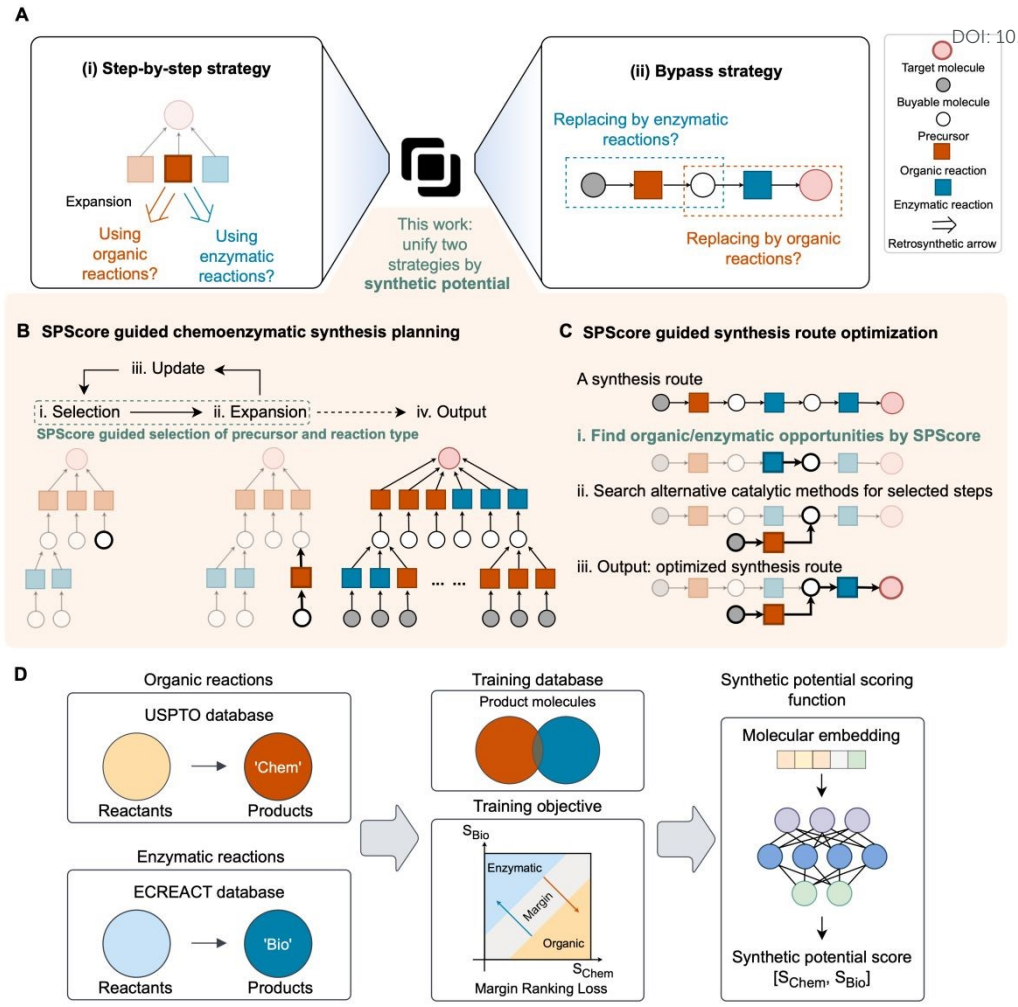

Fig. 1. Chemoenzymatic synthesis planning guided by a synthetic potential score. (A) Workflow of chemoenzymatic retrosynthesis strategies and representative work. **(B)** Workflow of the SPScore guided chemoenzymatic synthesis planning process. The target molecule is labeled by a red circle, and organic and enzymatic reactions 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 reaction type 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 SPScore guided synthesis route optimization. (i) Identify steps with opportunities for improvement using SPScore. For a given synthesis 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 reaction type used in the existing route. (ii) Search alternative reaction types for the selected steps. The synthesis planning algorithm in **(B)** is used to search synthesis routes with alternative reaction types for the selected steps. (iii) Output. The promising search results are appended to the original route, and the optimized route is returned. **(D)** Development of the synthetic potential scoring function. Reaction product molecules were extracted from USPTO (organic reactions) and ECREACT (enzymatic reactions), respectively. A neural network model is trained to infer the promising reaction type for a given molecule through the predicted SPScore.

development of efficient hybrid synthesis planning tools. One precedent is that the introduction of synthetic complexity in retrosynthesis can help synthesis planning tools efficiently find concise and feasible synthesis routes (18,19). The context of chemoenzymatic retrosynthesis motivates us to introduce a synthetic potential score (SPScore) — a hypothetical metric describing the suitability of enzymatic and organic reactions for synthesizing a molecule. This score not only bridges the gap between the step-by-step strategy and the bypass strategy but also enhances the step-by-step strategy algorithm, transforming it from a mere replica of traditional retrosynthesis algorithms designed for single reaction types into a more versatile and innovative approach.

In this work, we integrate the above-mentioned two strategies by using SPScore to prioritize reaction type (enzymatic or organic) in step-by-step chemoenzymatic synthesis planning (**Fig. 1B**) and identify alternative reaction types in given synthesis routes (**Fig. 1C**). This score was developed by training a multilayer perceptron on existing reaction corpora (**Fig. 1D**). We evaluated the performance of SPScore on both single-step and multi-step retrosynthesis and developed a SPScore guided

asynchronous search algorithm for chemoenzymatic synthesis planning. The resulting strategy named asynchronous chemoenzymatic retrosynthesis planning algorithm (ACERetro) can identify chemoenzymatic synthesis routes for 46% more molecules compared to the state-of-the-art tool when using 1,001 molecules as a test dataset. To demonstrate its utility, we applied ACERetro to design promising chemoenzymatic synthesis routes for two FDA-approved drugs, ethambutol and epidiolex. In addition, we applied ACERetro to optimize the synthesis routes of two additional FDA-approved drugs, rivastigmine and (*R,R*)-formoterol. A user-friendly web interface of ACERetro could be accessed at https://aceretro.platform.moleculemaker.org/search-routes.

Results

Development of the synthetic potential score

The synthetic potential of a molecule in the organic or enzymatic reactions is influenced by its structure and the current knowledge of synthetic chemistry and biochemistry. As a proof of concept, we demonstrated that synthetic potential can be learned from reaction data by using molecular fingerprints, which capture substructures, combined with a multilayer perceptron (MLP). The method is based on the premise that if a molecule has documented reactions for its synthesis, the reaction type of these reported reactions is the molecule's promising reaction type. A dataset comprising reaction products was extracted from two primary sources: USPTO 480K (20), which contains 484,706 organic reactions, and ECREACT (21), which contains 62,222 enzymatic reactions. After removing duplicates and molecules that could not be converted into valid molecular fingerprints for each respective reaction type, the resulting dataset comprised of 437,781 molecules in organic reactions and of 37,939 molecules in enzymatic reactions, while 515 molecules were present in both reaction types.

Molecules were represented by ECFP4 (22) (extended connectivity fingerprint, up to four bonds) and MAP4 (23) (MinHashed Atom Pair fingerprint, diameter *d* = 4) with three different lengths (*length* = 1024, 2048, and 4096) and used to train several MLP models. Rather than predicting a binary label indicating the preferred catalysis type, the MLP is trained to generate two continuous values: the synthetic potential score for organic reactions (S_{Chem}) and for enzymatic reactions (S_{Bio}). These two scores are directly output by the MLP and reflect how favourable each reaction type is for a given molecule. Because the reaction corpus may not cover all possible transformations, formulating this task as a binary classification is not ideal. Instead, we use Margin Ranking Loss as the training objective, which encourages the model to rank the more promising reaction type higher based on relative differences between S_{Chem} and S_{Bio} (see Methods). This loss function is well-suited for tasks where the goal is to learn a preference between two options. In our case, it allows the model to rank one catalytic option over the other, which better aligns with the decision-making nature of hybrid retrosynthetic planning. The SPScores range from 0 to 1, so they can act as the probability of a molecule being promisingly synthesized by each reaction type. When the difference between two SPScores of a molecule is within the margin, both reaction types are considered promising for the synthesis of that molecule. If a molecule's SPScore of one reaction type is greater than the other, and the difference is greater than the margin, the reaction type 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 (18). Since reactions sourced from the USPTO are confined to patents and differ in distribution from those documented in the literature (24), 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 (**Fig. S2**). 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 (25) 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 reaction types 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 reaction types (**Fig. 2B**).

Since there is no definitive ground truth for the synthetic potential of a given molecule, the evaluation of SPScore relies on indirect evidence on retrosynthesis results to demonstrate its practicality and robustness. To further explore whether SPScore effectively predicts a molecule's promising reaction type, we conducted one-step retrosynthesis in each reaction type by employing RXN4Chemistry (26) for organic reactions, and Levin et al.'s enzymatic templates (14) for enzymatic reactions. For a given target molecule, RXN4Chemistry predicts possible organic 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 organic reactions predicted by our scoring model (**Fig. 2C**). A similar

View Article Online
DOI: 10.1039/D5DD00098D

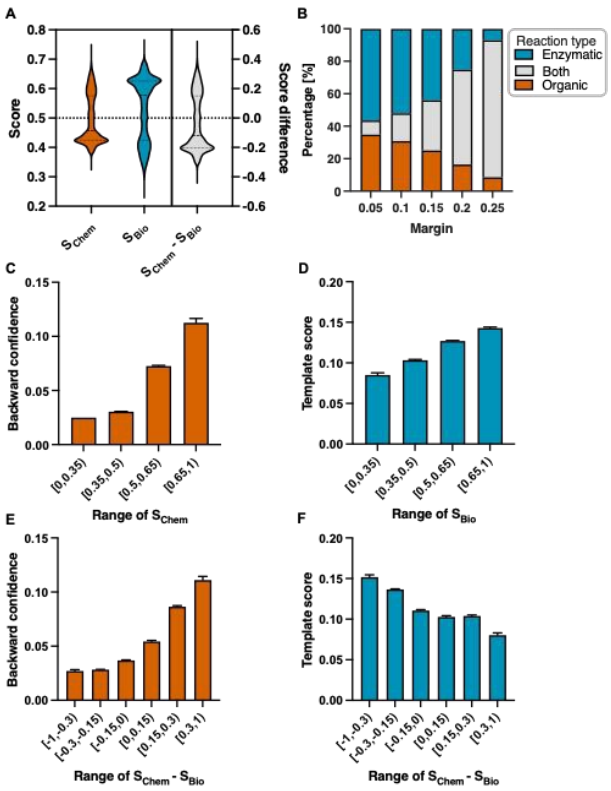

Fig. 2. Analysis of SPScore on molecules from ZINC "in-vitro" subset. (A) The distribution of S_{Chem}, S_{Bio} and the score difference ($S_{Chem} - S_{Bio}$). **(B)** The percentage of predicted type of catalysis of molecules versus different margin settings. **(C)** The mean backward confidence with SEM (the standard error of the mean) versus the range of S_{Chem}. In organic reactions, 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' SPScore in organic reactions. **(D)** The mean template score with SEM versus the range of S_{Enzy}. In enzymatic reactions, 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' SPScore in enzymatic reactions. **(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 $S_{Chem} - S_{Bio}$.

trend between the average template score and the mean synthetic potential score in enzymatic reactions is observed (**Fig. 2D**). This suggests that as the predicted probability of molecules being synthesized by a specific reaction type increases, the corresponding retrosynthesis tool's confidence in its predictions also tends to increase.

To assess whether the relative value of S_{Chem} and S_{Bio} can help identify the dominant reaction type, we analyzed both the average backward confidence and the average template score for top-5 predictions versus the mean SPScore difference ($S_{Chem} - S_{Bio}$) (see **Fig. 2, E and F)**. The result reveals that when S_{Bio} is larger than S_{Chem}, molecules tend to have a relatively high template score to be synthesized by enzymatic reactions and low backward confidence to be synthesized by organic reactions. Collectively, these trends in **Fig. 2** suggest that our scoring function exhibits a good ability to deduce the promising reaction type 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. Due to the limited availability of databases containing multi-step synthesis routes, particularly for hybrid synthesis, the synthesis routes of 493 target molecules identified by Levin et al.'s hybrid planner within three minutes were used for this in silico benchmarking study (14). The SPScore prediction of molecules is compared with the 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 reaction types were identified. Specifically, 9,162 (34.3%) molecules synthesized exclusively by organic reactions, 10,211 (38.2%) synthesized exclusively by enzymatic reactions, 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 reaction types were identified: 788 (51.0%) in

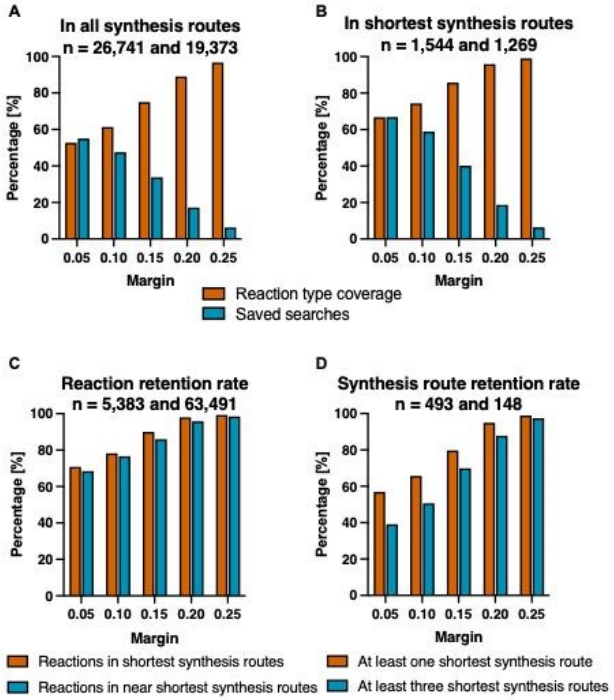

Fig. 3. Analysis of SPScore on synthesis routes. Reaction products are extracted from reactions in all synthesis routes and molecules' shortest synthesis routes predicted by Levin et al.'s tool on 493 molecules. Molecules are assigned with the ground truth label of reaction type based on the extracted reaction set. The amount of "reaction type coverage" is counted when the SPScore gives the correct prediction or predicts it as "both". The "saved searches" means SPScore gives the correct reaction type prediction. The percentage of reaction type 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 reaction type 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.

View Article Online
DOI: 10.1039/D5DD00008D

organic reactions, 481 (31.2%) in enzymatic reactions, and 275 (17.8%) spanning both fields.

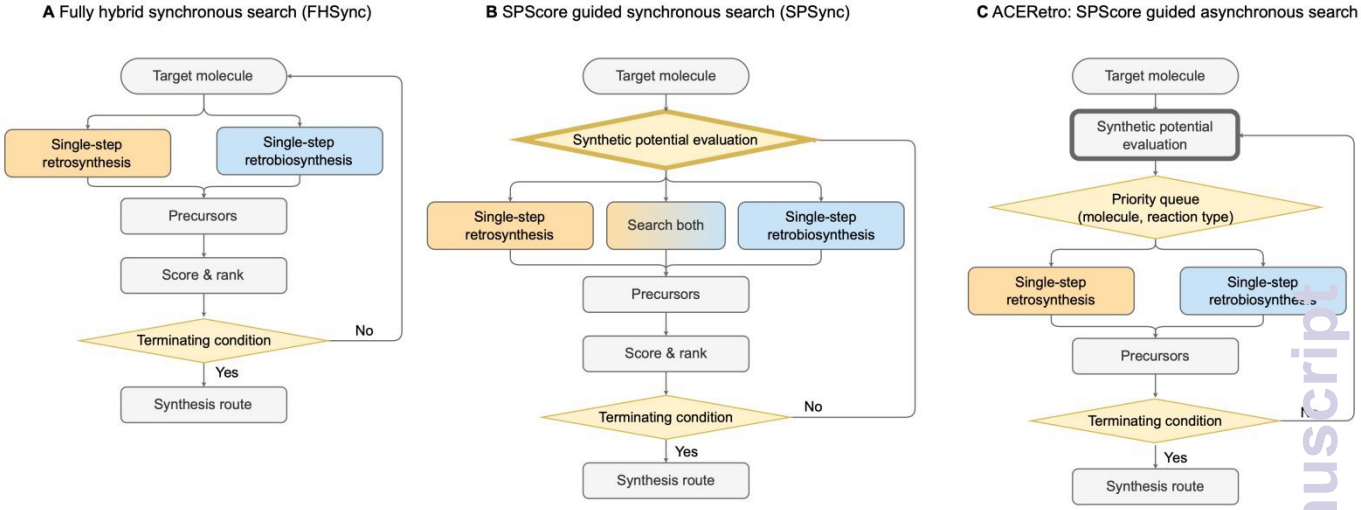

Fig. 4. Hybrid search algorithms for designing chemoenzymatic synthesis routes. (A) Fully hybrid synchronous search algorithm (FHSync). **(B)** SPScore guided synchronous search algorithm (SPSync). **(C)** ACERetro: SPScore guided asynchronous search algorithm. FHSync and SPSync are included as ablation studies to evaluate the components of ACERetro. FHSync represents a version without synthetic potential scoring and without asynchronous search, while SPSync includes synthetic potential scoring but does not use asynchronous search. Modules that incorporate SPScore are indicated with bold edges in the diagrams.

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 reaction type in the original synthesis routes (**Fig. 3, A and B**). The observed trend indicates that as the margin expands, the reaction type 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 defined as the proportion of reactions where the actual reaction type matches the predicted preferred reaction type for the product molecule, as determined by SPScore (see Methods and Supplementary Information). 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 reaction types 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 majority of favorable synthesis routes.

Searching for chemoenzymatic synthesis routes

Searching for chemoenzymatic synthesis routes requires predicting precursors in organic and enzymatic reactions. Synchronous methods, like Levin et al.'s tool (14), search precursors within in enzymatic and organic reactions simultaneously and then combine the search results. In this study, we report a SPScore guided asynchronous method, ACERetro, which prioritizes the search of the most promising reaction type for a given molecule. To evaluate the performance of ACERetro (**Fig. 4C**), we conducted a self-benchmark study using two simplified versions of the algorithm: a fully hybrid synchronous algorithm (FHSync, **Fig. 4A**) and a SPScore-guided synchronous algorithm (SPSync, **Fig. 4B**). These two variations serve as ablation studies to assess the contribution of each component within ACERetro. FHSync does not incorporate SPScore or asynchronous search, while SPSync includes SPScore-based guidance but uses a synchronous (non-asynchronous) search strategy. 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 organic reactions and enzymatic reactions. The fully hybrid search algorithm (FHSync) without SPScore directly searches both organic reactions and enzymatic reactions, and the results are combined in each step, while the SPScore guided synchronous hybrid search algorithm (SPSync) only searches for a promising reaction type predicted by the SPScore. In the SPScore guided asynchronous hybrid search algorithm (ACERetro), SPScore is used to guide separate search processes for each reaction type (See details in Methods). By running these searches asynchronously, ACERetro can prioritize the more promising reaction type while still exploring the alternative. This setup improves the robustness of

View Article Online
DOI: 10.1039/D5DD00085F

the planning process, allowing the algorithm to recover and switch paths if the initially favoured reaction type turns out to be less viable.

To explore search spaces of these three algorithms, we conducted a comparative study on a set of 1,001 molecules from ZINC, which Levin et al.'s tool had explored 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 molecules (**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

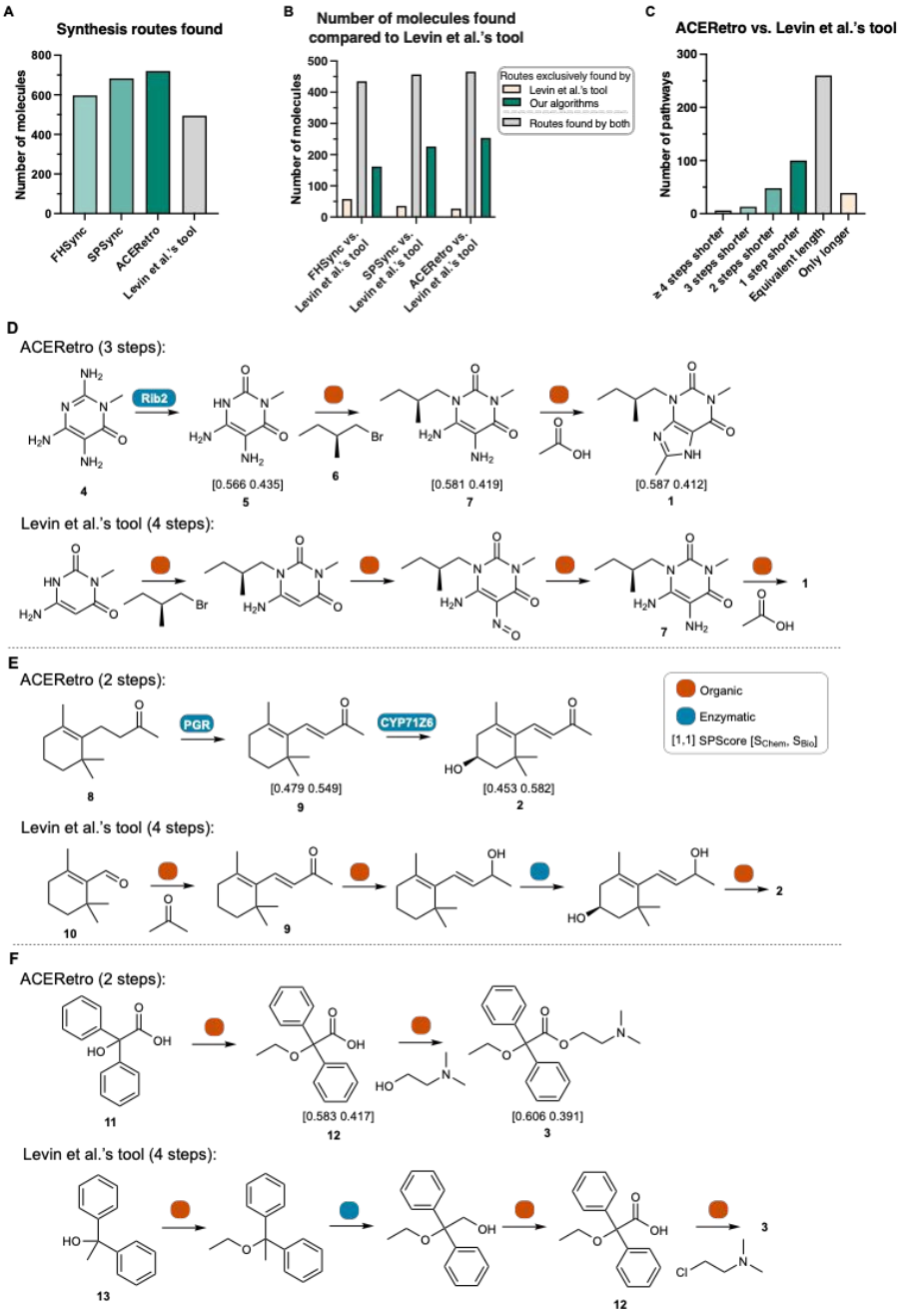

Fig. 5. Comparison of synthesis routes found by hybrid search algorithms. (A) Number of molecules for which synthesis routes were found out of 1,001 molecules. **(B)** Number of molecules that their synthesis routes can be found by Levin et al.'s tool (red) only, both (grey), and ours (blue). **(C)** 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. 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.

is mainly attributed to the incorporation of the template-free model, RXN4Chemistry, in organic reactions. 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 comparison between 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 drop the search for molecules' suboptimal reaction type, the asynchronous search keeps all suboptimal reaction type of molecules to the queue for later exploration. The algorithm will start to search suboptimal reaction type of molecules based on a comprehensive consideration including 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 (27). A smaller number of steps usually implies the use of fewer reagents and fewer purification steps (28). 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 synthesis 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**), dimenoxadol (**3**) (**Fig. 5, D-F**), and other 7 syntheses (**Fig. S8**). 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 organic reactions (**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 scenarios: hybrid approach, purely organic 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 synthesis 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 reaction type in the synthesis route. However, note that S_{Bio} of **5** is higher than all other products in the route, and its $S_{Chem} - S_{Bio}$ has the smallest value, which indicates **5** has higher potential to be synthesized by enzymatic reactions compared to other product molecules in the synthesis routes.

Applying ACERetro for synthesis planning

View Article Online
DOI: 10.1039/D5DD00008D

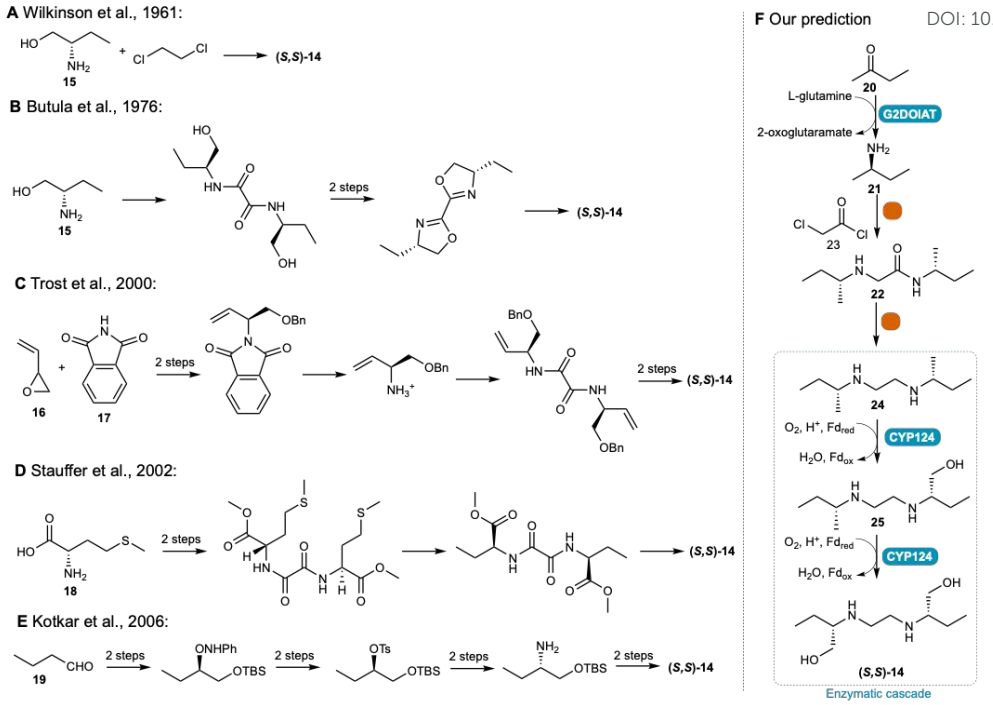

Fig. 6. Synthesis routes of ethambutol. (A-E) Published synthesis routes of ethambutol. **(F)** Predicted synthesis route of ethambutol. Ethambutol does not appear in the training set of the enzymatic model. G2DOIAT: L-glutamine:2-deoxy-scyllo-inosose aminotransferase, CYP124: CYP124 family of cytochrome P450 enzymes.

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 (29–31). Wilkinson et al. first reported the synthesis route for **(*S,S*)-14**, utilizing (*2S*)-2-aminobutan-1-ol (**15**) as the starting material (29). Likewise, the synthesis routes of **(S,S)-14** developed by Butula et al. (32) and Stauffer et al. (33) 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 (34), while Kotkar et al. reported a synthesis route using proline-catalyzed α-aminooxylation on butyraldehyde (**19**) (35) (see **Fig. 6, A-E**).

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 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 (36). 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** (37). Gong et al. used Friedel-Crafts reaction to build chirality starting with phloroglucinol (**35**) and cis-isolimonenol (**36**) (38). The biosynthetic route of (−)-cannabidiol using hexanoyl-CoA as the starting material has also been reported (39). The cannabidiolic acid synthase (CBDAS; EC number 1.21.3.8) uses cannabigerolic acid as substrate to close the ring and introduce stereochemistry (see **Fig. 7, A-D**).

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 synthesis 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 (40), 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

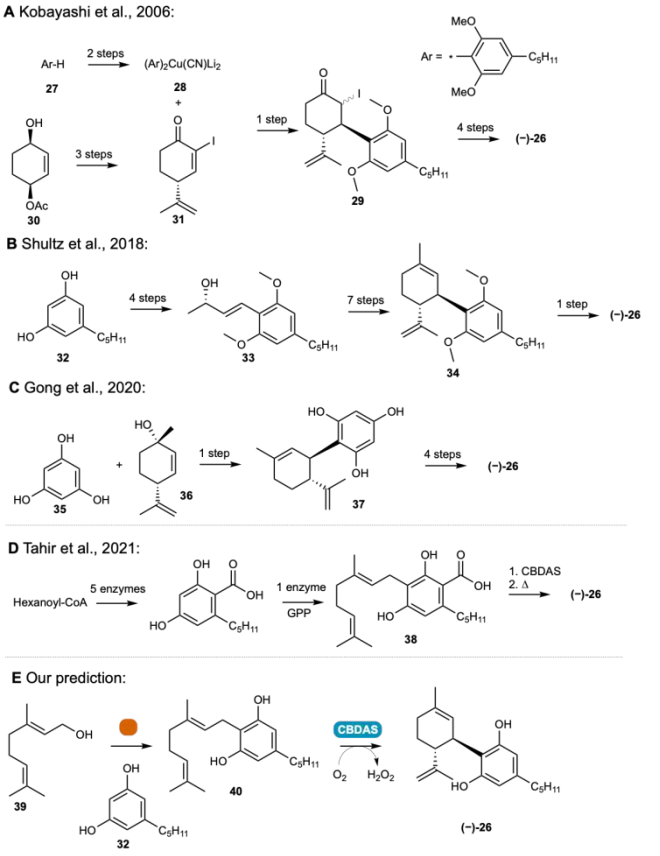

Fig. 7. Synthesis routes of epidiolex. (A-D) Published synthesis routes of epidiolex. **(E)** Predicted synthesis route of ethambutol. Epidiolex does not appear in the training set of the enzymatic model. CBDAS: Cannabidiolic acid synthase.

enzyme mutants that allow the reaction to occur.

Applying ACERetro 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 reaction types in given routes. The steps with opportunities for improvement are selected based on the deviation between the SPScore predicted reaction type and their reaction type 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. The utility of SPScore and ACERetro has been examined in the previous sections. To further showcase their combined ability to optimize synthesis routes, we present case studies on the synthesis route of rivastigmine (**41**) reported in the literature (41) and a synthesis route for (*R,R*)-formoterol (**42**) reported by a synthesis planning tool (14). 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

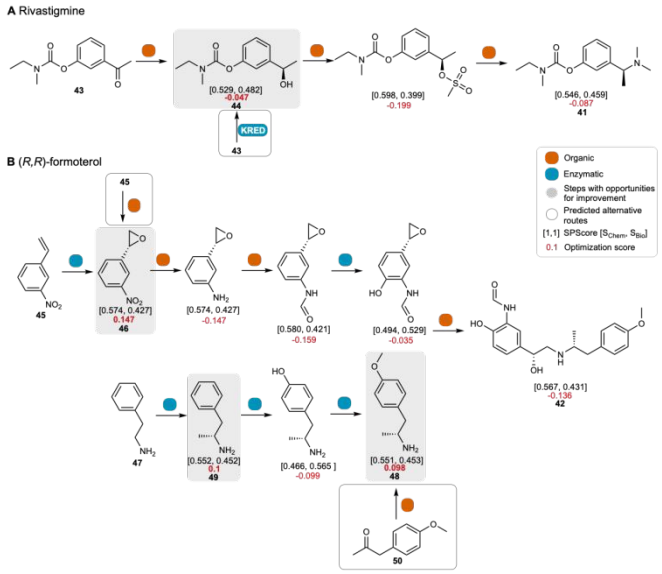

Fig. 8. SPScore guided synthesis route optimization. (A) Synthesis route optimization of rivastigmine. **(B)** Synthesis route optimization of (*R,R*)-formoterol. Steps with opportunities for improvement do not appear in the training set of the SPScore. (+)-Phenylethylamine, which is available in the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

("optimization score") was selected to search possible synthesis routes with enzymatic reactions. 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 (42), 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 reaction type 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 analysis for **48** is shown in the **Fig. 8B**, which 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

View Article Online
DOI: 10.1039/D5DD00008D

strategy to a single step. These predicted reactions for intermediates **46** and **48** have been corroborated by the literature (43,44).

Discussion

In this work, we identified the underlying principle between the step-by-step strategy and the bypass strategy by emphasizing the role of synthesis potential. We developed a SPScore guided asynchronous chemoenzymatic synthesis planning algorithm named ACERetro for designing chemoenzymatic synthesis routes for target molecules. When considering the evaluation of synthetic potential of molecules in each reaction type for computer-assisted chemoenzymatic synthesis planning, our heuristic search algorithm can prioritize the exploration of the most promising reaction type 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 evaluating synthetic potential of molecules constructs 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 organic reactions and enzymatic reactions 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 organic reaction and enzymatic reaction databases. A sufficiently large organic 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 organic reactions, and a template-based retrosynthesis tool, ASKCOS, for enzymatic reactions. Free from the limitations imposed by a template prioritization system, ACERetro guided by the SPScore 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 reaction type for molecules. The performance of SPScore in multi-step retrosynthesis was further verified by reaction type coverage, reaction retention rate, and route retention rate among predicted synthesis routes of 493 molecules. In the benchmarking study on 1,001 molecules, ACERetro incorporating two single-step precursor prediction tools outperformed Levin et al.'s tool, a state-of-the-art method. 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 synthesis 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 reaction types 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 reaction type 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 organic reactions and enzymatic reactions. In this work, we performed a simplified but fruitful verification of the synthetic potential with molecular fingerprint and MLP. Combining more complex models such as molecular graphs and reinforcement learning to predict the SPScore will be explored in a follow-up study. Utilizing SPScore in chemoenzymatic synthesis planning can expedite the search process by avoiding less promising reaction types. However, there remains a risk that the model might overlook viable reactions in the reaction types 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 organic reactions and enzymatic reactions 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 reaction type 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 reaction type for each step.

In summary, this study transforms chemoenzymatic synthesis planning from a fragmented process into a unified framework by combining the concept of synthetic potential with practical algorithmic design. ACERetro overcomes the limitations of existing methods by integrating both template-based and template-free strategies, enabling comprehensive synthesis

View Article Online
DOI: 10.1039/D5DD00008D

planning across diverse organic and enzymatic reaction databases. Its ability to design efficient synthesis routes and identify alternative pathways highlights its potential as a powerful tool in the field. We believe that computer-aided chemoenzymatic synthesis planning will expand the synthesis space by leveraging the complementary strengths of enzymatic reactions and organic reactions. This approach can accelerate the adoption of enzymes as eco-friendly catalysts, facilitating enzyme screening and engineering for improved catalytic performance.

Method

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(45), BRENDA(46), PathBank(47), and MetaNetX . 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 reaction types, are labeled with $y = 0$. An MLP model is trained to generate two continuous synthetic potential scores for each molecule: one for organic reactions (S_{Chem}) and one for enzymatic reactions (S_{Bio}). The model takes molecular fingerprints as input and outputs both scores through a final sigmoid activation layer, which bounds the values between 0 and 1. Importantly, the model is not trained to directly predict reaction type labels. Instead, it is optimized using a margin ranking loss based on ground truth labels $y \in \{-1, 0, 1\}$, which indicate molecule's reaction is enzymatic, organic, or compatible with both. This training approach encourages the model to assign a higher score to the more suitable reaction type for each molecule, allowing it to capture relative preferences rather than making hard classification decisions. The margin uses the relative value of S_{Chem} and S_{Bio} to divide the output space to three areas corresponding to three scenarios of reaction type as shown in **Fig. 1A**. To evaluate the MLP, we compute accuracy by checking whether the predicted scores satisfy the correct ranking implied by the ground truth label y∈{−1,0,1}. Specifically:

- If y=1 (i.e., the reaction type is organic), we consider the prediction correct if $S_{Chem} > S_{Bio} + margin$.
- If y=−1 (i.e., enzymatic), the prediction is correct if $S_{Bio} > S_{Chem} + margin$
- If y=0 (i.e., overlap), the prediction is considered correct if the absolute difference between the two scores is less than the margin, i.e., $|S_{Chem} - S_{Bio}| < margin$.

A weighted margin ranking loss, $loss(S_{Chem}, S_{Bio}, y) = weight_i \cdot max(0, -y(S_{Chem} - S_{Bio}) + margin)$ if $y = \pm 1$ and $loss(S_{Chem}, S_{Bio}, y) = weight_i \cdot max(0, |S_{Chem} - S_{Bio}| - margin)$ if $y = 0$, is applied to compute a criterion only when the prediction is out of the area of molecules' true labels. The weight is calculated based on reciprocal of the ratio among each label.

The dataset was randomly 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 organic reactions. 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 enzymatic reactions. 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 synthesis routes. All the synthesis routes with the shortest length for each target molecule were collected, which contained 1,544 distinct product molecules. The reaction type 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. Reaction type coverage out of all molecules counts the molecule whose SPScore predicted reaction type includes the actual reaction type out of all molecules. Saved searches out of "Chem" and "Bio" molecules count the molecule whose SPScore predicted reaction type exactly matches the actual reaction type for these molecules labeled with "Chem" or "Bio", so the search algorithm does not need to search the alternative reaction type. The reaction retention rate measures how often the actual reaction type used in a pathway agrees with the preferred reaction type predicted by SPScore for the product of that reaction. Specifically, for each reaction in a given dataset, we check whether the reaction type (e.g., chemical or enzymatic) matches one of the reaction types that SPScore

DOI: 10.1039/D5DD00098D

predicts as favorable for the reaction's product molecule. The retention rate is calculated as the percentage of reactions that meets this criterion across the entire dataset (see Supplementary Information for the formulae). 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).

Development and evaluation of ACERetro

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. All three search algorithms—FHSync, SPSync, and ACERetro—are built on a tree search framework that follows an iterative process of selection, expansion, and update. 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 step, the behavior differs across the three methods. In FHSync, both organic and enzymatic reaction models are applied to the selected molecule. RXN4Chemistry (26) and Levin et al.'s enzymatic templates (14) are used to predict single-step precursors for the selected molecule. The precursors generated from both reaction types are merged, 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 SPSync, the algorithm uses SPScore to determine which reaction type is more promising for the selected molecule. Only the predicted reaction type is used for precursor generation, and scoring is performed in the same way as FHSync.

In ACERetro, SPScore is used to guide an asynchronous search between the two reaction types. For each selected molecule, scores are calculated for both the organic and enzymatic pathways using the formula:

$$Score_i = (1 - c \cdot SPScore_i) Depth^d \cdot f(P) \quad i \in [Chem, Bio]$$

where c is reaction type exploration factor. A molecule will have two scores associated with two reaction types. This approach allows ACERetro to flexibly favour the more promising catalytic route based on SPScore, while still retaining the ability to switch paths if needed.

In the update step, all newly generated precursors, along with their associated scores, are added to the priority queue, which is then re-ranked before the next iteration begins. This shared architecture enables a fair comparison on how the integration of SPScore and asynchronous search affects planning performance.

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. After all pathways and precursors are predicted, the enzyme used in each biocatalytic reaction is selected based on the similarity of products under the same reaction template.

Case studies on synthesis route optimization

For a given synthesis route, SPScore is calculated for all intermediate molecules (i.e., all except the starting material). To identify steps with potential for optimization, we compare the reaction type used in the current route with the reaction type preferred by SPScore for each molecule. The top-n molecules with the highest potential for improvement are selected using the following expression:

$$I_n = \text{argsort}\left(y_i(SPScore^i_{Chem} - SPScore^i_{Bio})\right)$$

Here, i is the index of each molecule in the pathway, and n is the number of molecules we wish to identify as most optimizable. The term $y_i = -1$ if the molecule's reaction type in the given synthesis route is organic reaction (labeled as "Chem"), and $y_i = 1$ if the molecule's reaction type in the given synthesis route is enzymatic reaction ("Bio"). The equation aims to find top-n molecules with the largest SPScore difference away from the molecule's reaction type ("optimization score", $y_i(SPScore^i_{Chem} - SPScore^i_{Bio})$). 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.

Author contributions

Conceptualization: XL, HZ
Methodology: XL
Investigation: XL, HL
Supervision: HZ
Writing—original draft: XL
Writing—review & editing: XL, HL, HZ

Conflicts of interest

There are no conflicts to declare.

Data availability

The scripts for training scoring function, benchmarking, and synthesis planning in this manuscript are available at https://github.com/Zhao-Group/ACERetro. The corresponding data and code have been archived on Zenodo at https://doi.org/10.5281/zenodo.10578664. The RXN4Chemistry (26) is used for single-step precursor prediction of organic reactions, which is the intellectual property of IBM, they are also accessible through the IBM RXN for Chemistry website. Levin et al.'s enzymatic templates (14) are used for single-step precursor prediction of biocatalytic reactions.

Acknowledgements

This work was funded by 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.

Notes and references

1. Zhang C, Lapkin AA. Reinforcement learning optimization of reaction routes on the basis of large, hybrid organic chemistry–synthetic biological, reaction network data. React Chem Eng. 2023 Sep 26;8(10):2491–504.

2. Simić S, Zukić E, Schmermund L, Faber K, Winkler CK, Kroutil W. Shortening synthetic routes to small molecule active pharmaceutical ingredients employing biocatalytic methods. Chem Rev. 2022 Jan 12;122(1):1052–126.

3. Kaspar F, Schallmey A. Chemo-enzymatic synthesis of natural products and their analogs. Curr Opin Biotechnol. 2022 Oct 1;77:102759.

4. Chakrabarty S, Romero EO, Pyser JB, Yazarians JA, Narayan ARH. Chemoenzymatic total synthesis of natural products. Acc Chem Res. 2021 Mar 16;54(6):1374–84.

5. Li J, Amatuni A, Renata H. Recent advances in the chemoenzymatic synthesis of bioactive natural products. Curr Opin Chem Biol. 2020 Apr 1;55:111–8.

6. Rudroff F, Mihovilovic MD, Gröger H, Snajdrova R, Iding H, Bornscheuer UT. Opportunities and challenges for combining chemo- and biocatalysis. Nat Catal. 2018 Jan;1(1):12–22.

7. McIntosh JA, Benkovics T, Silverman SM, Huffman MA, Kong J, Maligres PE, et al. Engineered ribosyl-1-kinase enables concise synthesis of molnupiravir, an antiviral for COVID-19. ACS Cent Sci. 2021 Dec 22;7(12):1980–5.

8. Szymkuć S, Gajewska EP, Klucznik T, Molga K, Dittwald P, Startek M, et al. Computer-assisted synthetic planning: the end of the beginning. Angew Chem Int Ed. 2016 May 10;55(20):5904–37.

9. Finnigan W, Hepworth LJ, Flitsch SL, Turner NJ. RetroBioCat as a computer-aided synthesis planning tool for biocatalytic reactions and cascades. Nat Catal. 2021 Feb;4(2):98–104.

10. Wang X, Li Y, Qiu J, Chen G, Liu H, Liao B, et al. RetroPrime: a diverse, plausible and transformer-based method for single-step retrosynthesis predictions. Chem Eng J. 2021 Sep;420:129845.

11. Ucak UV, Ashyrmamatov I, Ko J, Lee J. RetroTRAE: retrosynthetic translation of atomic environments with Transformer [Internet]. Chemistry; 2021 Aug [cited 2022 Jan 14]. Available from: https://chemrxiv.org/engage/chemrxiv/article-details/6117ed347117505183e94d93

12. Chen B, Li C, Dai H, Song L. Retro*: Learning Retrosynthetic Planning with Neural Guided A* Search. In: Proceedings of the 37th International Conference on Machine Learning [Internet]. PMLR; 2020 [cited 2022 Aug 28]. p. 1608–16. Available from: https://proceedings.mlr.press/v119/chen20k.html

13. Hong S, Zhuo HH, Jin K, Shao G, Zhou Z. Retrosynthetic planning with experience-guided Monte Carlo tree search. Commun Chem. 2023 Jun 10;6(1):1–14.

14. Levin I, Liu M, Voigt CA, Coley CW. Merging enzymatic and synthetic chemistry with computational synthesis planning. Nat Commun. 2022 Dec 14;13(1):7747.

15. Zeng T, Jin Z, Zheng S, Yu T, Wu R. Developing BioNavi for hybrid retrosynthesis planning. JACS Au. 2024 Jul 22;4(7):2492–502.

16. Sankaranarayanan K, F. Jensen K. Computer-assisted multistep chemoenzymatic retrosynthesis using a chemical synthesis planner. Chem Sci. 2023;14(23):6467–75.

17. Li H, Liu X, Jiang G, Zhao H. Chemoenzymatic Synthesis Planning Guided by Reaction Type Score. J Chem Inf Model [Internet]. 2024 Dec 8 [cited 2024 Dec 15]; Available from: https://doi.org/10.1021/acs.jcim.4c01525

18. Coley CW, Rogers L, Green WH, Jensen KF. SCScore: synthetic complexity learned from a reaction corpus. J Chem Inf Model. 2018 Feb 26;58(2):252–61.

19. Schwaller P, Petraglia R, Zullo V, Nair VH, Haeuselmann RA, Pisoni R, et al. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. Chem Sci. 2020;11(12):3316–25.

20. Coley CW, Jin W, Rogers L, Jamison TF, Jaakkola TS, Green WH, et al. A graph-convolutional neural network model for the prediction of chemical reactivity. Chem Sci. 2019 Jan 2;10(2):370–7.

21. Probst D, Manica M, Nana Teukam YG, Castrogiovanni A, Paratore F, Laino T. Biocatalysed synthesis planning using data-driven learning. Nat Commun. 2022 Feb 18;13(1):964.

22. Rogers D, Hahn M. Extended-connectivity fingerprints. J Chem Inf Model. 2010 May 24;50(5):742–54.

23. Capecchi A, Probst D, Reymond JL. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Cheminformatics. 2020 Jun 12;12(1):43.

24. Thakkar A, Kogej T, Reymond JL, Engkvist O, Bjerrum EJ. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. Chem Sci. 2019 Dec 18;11(1):154–68.

25. Sterling T, Irwin JJ. ZINC 15 – ligand discovery for everyone. J Chem Inf Model. 2015 Nov 23;55(11):2324–37.

26. Toniato A, Vaucher AC, Schwaller P, Laino T. Enhancing diversity in language based models for single-step retrosynthesis. Digit Discov. 2023 Apr 11;2(2):489–501.

27. Mo Y, Guan Y, Verma P, Guo J, Fortunato ME, Lu Z, et al. Evaluating and clustering retrosynthesis pathways with learned strategy. Chem Sci. 2021;12(4):1469–78.

28. Cornwall P, Diorazio LJ, Monks N. Route design, the foundation of successful chemical development. Bioorg Med Chem. 2018 Aug 7;26(14):4336–47.

29. Wilkinson RG, Shepherd RG, Thomas JP, Baughn C. Stereospecificity in a new type of synthetic antituberculous agent. J Am Chem Soc. 1961 May;83(9):2212–3.

30. Shepherd RG, Wilkinson RG. Antituberculous agents. II. N,N'-diisopropylethylenediamine and analogs. J Med Pharm Chem. 1962 Jul 1;5(4):823–35.

31. Wilkinson RG, Cantrall MB, Shepherd RG. Antituberculous agents. III. (+)-2,2 -(ethylenediimino)-di-1-butanol and some analogs. J Med Pharm Chem. 1962 Jul;5(4):835–45.

32. Butula I, Karlović G. Katalytische hydrierung von stickstoffhaltigen heterocyclen, IV1) hydrogenolyse von verbrückten oxazolinen und oxazolidinen. Justus Liebigs Ann Chem. 1976;1976(7–8):1455–64.

33. Stauffer CS, Datta A. Efficient synthesis of (S,S)-ethambutol from l-methionine. Tetrahedron. 2002 Dec 2;58(49):9765–7.

34. Trost BM, Bunt RC, Lemoine RC, Calkins TL. Dynamic kinetic asymmetric transformation of diene monoepoxides: a practical asymmetric synthesis of vinylglycinol, vigabatrin, and ethambutol. J Am Chem Soc. 2000 Jun 1;122(25):5968–76.

35. Kotkar SP, Sudalai A. Enantioselective synthesis of (S,S)-ethambutol using proline-catalyzed asymmetric α-aminooxylation and α-amination. Tetrahedron Asymmetry. 2006 Jul 17;17(11):1738–42.

36. Kobayashi Y, Takeuchi A, Wang YG. Synthesis of cannabidiols via alkenylation of cyclohexenyl monoacetate. Org Lett. 2006 Jun 1;8(13):2699–702.

37. Shultz ZP, Lawrence GA, Jacobson JM, Cruz EJ, Leahy JW. Enantioselective total synthesis of cannabinoids - a route for analogue development. Org Lett. 2018 Jan 19;20(2):381–4.

38. Gong X, Sun C, Abame MA, Shi W, Xie Y, Xu W, et al. Synthesis of CBD and its derivatives bearing various C4'-side chains with a late-stage diversification method. J Org Chem. 2020 Feb 21;85(4):2704–15.

39. Stout JM, Boubakir Z, Ambrose SJ, Purves RW, Page JE. The hexanoyl-CoA precursor for cannabinoid biosynthesis is formed by an acyl-activating enzyme in Cannabis sativa trichomes. Plant J. 2012;71(3):353–65.

40. Seccamani P, Franco C, Protti S, Porta A, Profumo A, Caprioglio D, et al. Photochemistry of cannabidiol (CBD) revised. A combined preparative and spectrometric investigation. J Nat Prod. 2021 Nov 26;84(11):2858–65.

41. Yan PC, Zhu GL, Xie JH, Zhang XD, Zhou QL, Li YQ, et al. Industrial scale-up of enantioselective hydrogenation for the asymmetric synthesis of rivastigmine. Org Process Res Dev. 2013 Feb 15;17(2):307–12.

42. Sethi MK, Bhandya SR, Maddur N, Shukla R, Kumar A, Jayalakshmi Mittapalli VSN. Asymmetric synthesis of an enantiomerically pure rivastigmine intermediate using ketoreductase. Tetrahedron Asymmetry. 2013 Apr 15;24(7):374–9.

43. Agarwala H, Ehret F, Chowdhury AD, Maji S, Mobin SM, Kaim W, 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. 2013 Feb 12;42(10):3721–34.

44. 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. 2018;360(24):4699–704.

45. Alcántara R, Axelsen KB, Morgat A, Belda E, Coudert E, Bridge A, et al. Rhea—a manually curated resource of biochemical reactions. Nucleic Acids Res. 2012 Jan;40(Database issue):D754–60.

46. Schomburg I, Chang A, Schomburg D. BRENDA, enzyme data and metabolic information. Nucleic Acids Res. 2002 Jan 1;30(1):47–9.

47. Wishart DS, Li C, Marcu A, Badran H, Pon A, Budinski Z, et al. PathBank: a comprehensive pathway database for model organisms. Nucleic Acids Res. 2020 Jan 8;48(D1):D470–8.

48. 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 Mar 15;29(6):815–6.

View Article Online DOI: 10.1039/D5DD00008D

View Article Online
DOI: 10.1039/D5DD00008D

Data Availability Statement

The scripts for training scoring function, benchmarking, and synthesis planning in this manuscript are available at https://github.com/Zhao-Group/ACERetro. The corresponding data and code have been archived on Zenodo at https://doi.org/10.5281/zenodo.10578664. The RXN4Chemistry (1) is used for single-step precursor prediction of organic reactions, which is the intellectual property of IBM, they are also accessible through the IBM RXN for Chemistry website. Levin et al.'s enzymatic templates (2) are used for single-step precursor prediction of biocatalytic reactions.

1. Toniato A, Vaucher AC, Schwaller P, Laino T. Enhancing diversity in language based models for single-step retrosynthesis. Digit Discov. 2023 Apr 11;2(2):489–501.
2. Levin I, Liu M, Voigt CA, Coley CW. Merging enzymatic and synthetic chemistry with computational synthesis planning. Nat Commun. 2022 Dec 14;13(1):7747.