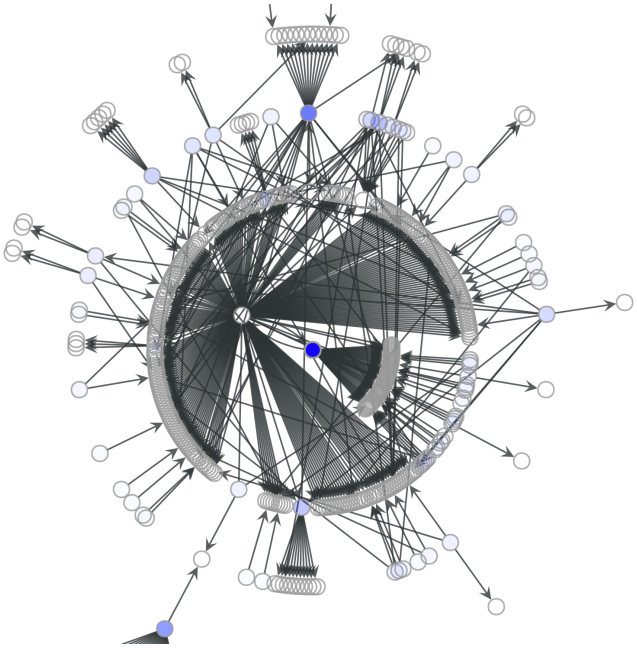
Data-driven automatic synthesis planning:

Synthesis routes of *S*-Zanubrutinib identified with CDI CASP



A white paper

2024

Chemical Data Intelligence (CDI) Ltd

Innovation Centre in Digital Molecular Technologies (*i*DMT)

Shionogi & Co. LTD

https://doi.org/10.26434/chemrxiv-2024-f0kcq **ORCID:** https://orcid.org/0000-0001-7621-0889

Data-driven automatic synthesis planning:
Synthesis routes of *S*-Zanubrutinib identified with CDI CASP

Zhen Guo,[1,2,4,*] Akihiro Takada[3] and Alexei A. Lapkin[1,2,4,*]

[1]*Chemical Data Intelligence (CDI) Pte Ltd, 9 Raffles Place #26-01, Republic Plaza Singapore, Singapore 048619*

[2]*Innovation Centre in Digital Molecular Technologies, Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge UK CB2 1EW*

[3]*Shionogi & Co. Ltd, 1, Futabacho, 3-chome, Toyonaka-shi, Osaka 561-0825, Japan*

[4]*Cambridge Centre for Advanced Research and Education in Singapore, CARES Ltd. 1 CREATE Way, CREATE Tower #05-05, 138602 Singapore*

Keywords: *computer-aided synthesis planning; object-oriented planning; expert-machine interactions; Active Pharmaceutical Ingredients (APIs); functional molecules*

* Zhen Guo: zhen.guo@cdi-sg.com
* Alexei Lapkin: a.lapkin@cdi-sg.com

Executive Summary

Advances in computer-assisted synthesis planning (CASP) are revolutionising how new functional molecules in many chemistry-using industries are being developed. CASP tools allow to assemble and analyse prior knowledge of a specified chemical system (a molecule, a reaction, a synthesis route), to generate hypotheses on experimental campaigns that could either be performed manually or using automated reaction systems. Advanced CASP tools are combining data science, chemoinformatics, machine learning and physical models-based predictive tools. Compared to expert-based synthesis planning, the power of CASP techniques allows for faster and more comprehensive planning, which could significantly improve the efficiencies of chemical process/product development.

This White Paper describes a recent collaboration project between Shionogi & Co. Ltd. and Chemical Data Intelligence (CDI) Ltd. The CASP system developed by CDI (CDI-CASP) was tested in developing a new synthesis of *S*-Zanubrutinib, a drug for lymphoma treatment. Three types of search in CDI-CASP - "search synthesis routes", "search analogue routes" and "search chiral reactions" - were iteratively applied for synthesis planning. Setting search criteria requires expert involvement. This 'human in the middle' interactive strategy leads to a shorter, greener, and more efficient synthesis route compared to the benchmark route filed in a patent.

Introduction

Organic synthesis plays an important role in modern industrial products, such as development of medicines, agrochemicals, functional additives to polymers and other products, and thus affects a wide range of daily consumer choices and activities. As a classical chemistry research topic, organic synthesis remains a dynamic field, because of many sophisticated challenges waiting to be resolved. In general, research process in developing new syntheses can be split in two parts: synthesis planning and validating plans in wet lab. Thanks to the developments in high-throughput experimentation techniques, robotics and lab automation, the validation part could be accelerated as demonstrated by several research groups based in both academia and industry.[1,2] For synthesis planning, significant progress had also been made by taking advantages of digitalising chemical data and having access to significant computing resources. The concept of computer-aided synthesis planning (CASP) has attracted significant interest over the past few decades. [3-5]

The idea of CASP can be traced back to 1960s pioneer works by Corey et. al.[6,7] In recent decades, this research area became highly active because of the dramatically improved computing power and the associated explosion of interest in new algorithms. Most CASP systems are developed based on two technologies: expert-coded rules[2] and Machine Learning (ML).[8]

Expert-coded reaction rules had been applied in automatic generation of reactions.[4] This approach is reliable yet limited to a specific set of rules (which should be continuously updated); as such, it lacks flexibility and is, potentially, biased by the process of how rules are generated. In contrast, millions of reaction templates, analogues to expert-coded rules but much more numerous, can be extracted from large reaction datasets using Machine Learning (ML) algorithms.[9] However, a new issue arises as there may be too many options for each individual reaction step. The number of possible synthesis routes increases exponentially with the number of synthetic steps in a route. This problem is termed “combinatorial explosion”.[3] Algorithms for ranking and selecting templates were developed to reduce the computing load, but it is non-trivial to balance computing cost and quality of results without, potentially risking losing important routes.

Template-free techniques have also been developed for CASP systems.[8,10,11] Different to the template-based methods, various ML models were trained to predict single-step reactions, then these models were applied recursively on a target molecule to be synthesized. These ML algorithms included, but not limited to, deep neural networks, Sequence to Sequence (seq2seq), transformer (both borrowed from Natural Language Processing), Graph Neural Networks (GNN) and Graph-to-Graph (G2G) methods.[12]

When approaching planning of multi-step syntheses, it is still critical to develop algorithms to prune reactions or sub-routes and, thus, avoid the excessive computational load; some approaches use Monte Carlo tree search (MCTS), for example.[8] Other challenges facing all ML-based synthesis prediction approaches include: potential model overfitting, locally optimal choices during route search, lack of reaction conditions information, poor quality of available data (bias and incomplete data), and lack of methods to evaluate the predicted results.[13] This explains why in recent years hybrid CASP systems, which incorporate expert-coded rules, ML and well-designed searching heuristics/scores, have become popular.[14,15]

CDI's CASP System

The CASP system of Chemical Data Intelligence (CDI) has been evolving through academic curiosity-driven research at the University of Cambridge in collaboration with the Reaxys™ team at Elsevier as part of their Research Network, and practical test projects with our end-user clients. Through the combination of the developed methods and interactions with industry expert CDI has created an iterative CASP system that supports creativity of synthetic chemists by enriching their data sources.

At present the CDI-CASP system is able to perform three types of data searches:

1. Search of synthetic routes

   Retrosynthesis of a target molecule using Reaxys™ as the main data source. Routes of up to 15 steps can be searched. Various heuristics are implemented to guide and customize the search to meet diverse search objectives / criteria (e.g., avoiding specific compounds or sub-structures in the routes, avoid patented reactions, using only bio-based feedstocks, and so on). This type of search helps users to explore new routes/sub-routes and also to identify potential new feedstocks or starting compounds for routes.

2. Search of "analogues" routes

   We define "analogues" routes as those where reactions are suggested by literature precedents of reactions with molecules containing sub-structures of the actual molecules of interest. These would typically include chemical transformations such as bond breaking, elimination/replacement of atoms/groups, hydrogenation, ring formation, etc. Analogue routes carry useful reaction information which could inspire design of new synthesis methodologies, but do not guarantee success.

3. Search of analogues chiral reactions

   This function helps find asymmetric synthesises affording chiral substructures defined by the user. By leveraging information from analogue chiral reactions, users can develop and experiment with new chiral reactions.

Based on these search types, we developed a web based CASP system with the following features:

1. Objective-oriented synthesis planning

   Users may have very different objectives for their synthesis tasks, including shorter synthesis routes, use of less hazardous reagents, higher yields, or greener feedstocks. These objectives are translated into heuristics (e.g., yield threshold, unfavoured solvents, preference of substructures etc.) to guide the search. This approach also allows close interaction between users and our system. A user iteratively refines their search to obtain the satisfactory set of results.

2. Conducting multistep retrosynthesis planning

   Algorithms have been developed to resolve the issues of "combinatorial explosion" and locally optimal routes.

3. Maximize utilization of molecular and reaction information

   For molecular information, CDI-CASP allows to search at molecular substructures/atom levels. Reaction information is evaluated during search. All resulting reactions can be validated experimentally since they are all reported reactions. Scores to evaluate reactions and routes were devised to complement the existing sparse reaction data.

4. Ability to design new routes

   New and innovative synthesis routes often arise from the application of unexpected routes/sub-routes or molecules based on the standard search algorithm and novel synthesis methods are inspired by findings from "analogous" search algorithms.

5. Ability to provide information for asymmetric syntheses

   "Search analogue chiral reactions" provides search of targeted chiral substructures, rather than a specific chiral molecule, resulting in more information than that resulting from a conventional literature research.

6. Easy analysis of results

   To facilitate analysis of the obtained results various visualization methods were developed, such as ranking routes using different metrics, selection and display of interesting specific molecules, generating network view, etc.

To demonstrate the functions and features of CDI-CASP, we conducted the search of synthesis routes of *S*-Zanubrutinib, a molecule for the treatment of lymphoma; commercial name Brukinsa. A synthesis route filed in a recent patent was chosen as the benchmark route, as shown in Fig. 1.[16] This synthesis route consists of 12-steps (three parallel steps from two feedstocks) and involves many hazardous molecules. The objectives of this case study are:

1. Reduce the number of synthesis steps.
2. Find routes with a higher overall yield.
3. Avoid hazardous reagents, solvents and intermediates.

4. Explore potential use of renewable feedstocks.

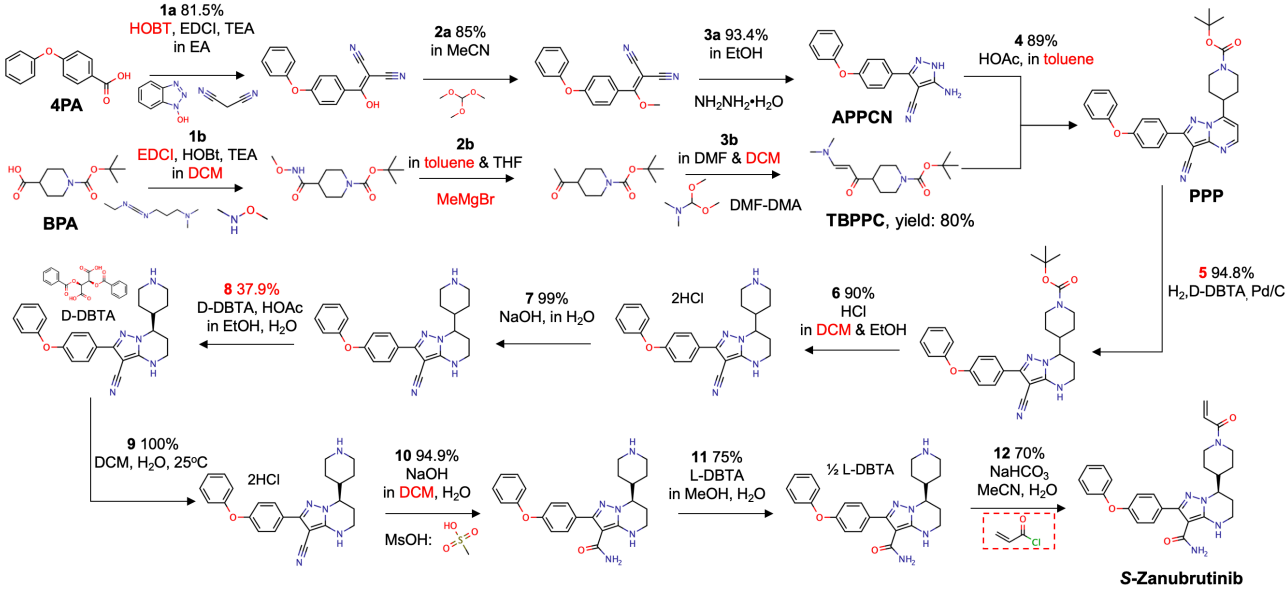


Fig. 1. A 12-steps synthesis of *S*-Zanubrutinib reported in the patent.[16] Problems of this benchmark routes are highlighted in red, including hazardous molecules and inefficient synthesis strategies (steps 5 and 8).

Overall methodology

As shown in Fig. 2, CDI-CASP offers three search functions that can be utilized either as standalone tools or in combination for interactive exploration of a synthesis challenge. The interactive mode is more common for most synthesis planning projects. Given a target molecule to synthesize, one can start from a multi-steps retrosynthesis search using the "Search synthesis routes" function to generate an initial pull of routes ranked according to the user-selected heuristics and thresholds. Based on these initial results, the user may re-run the retrosynthesis search or perform other search functions to modify and improve on the list of output routes. There are no rules on which type of search should be conducted first.

In synthesis planning, different search functions can be conducted iteratively, with expert judgement incorporated after each epoch, until the synthesis objective is achieved, or no more useful information can be extracted.

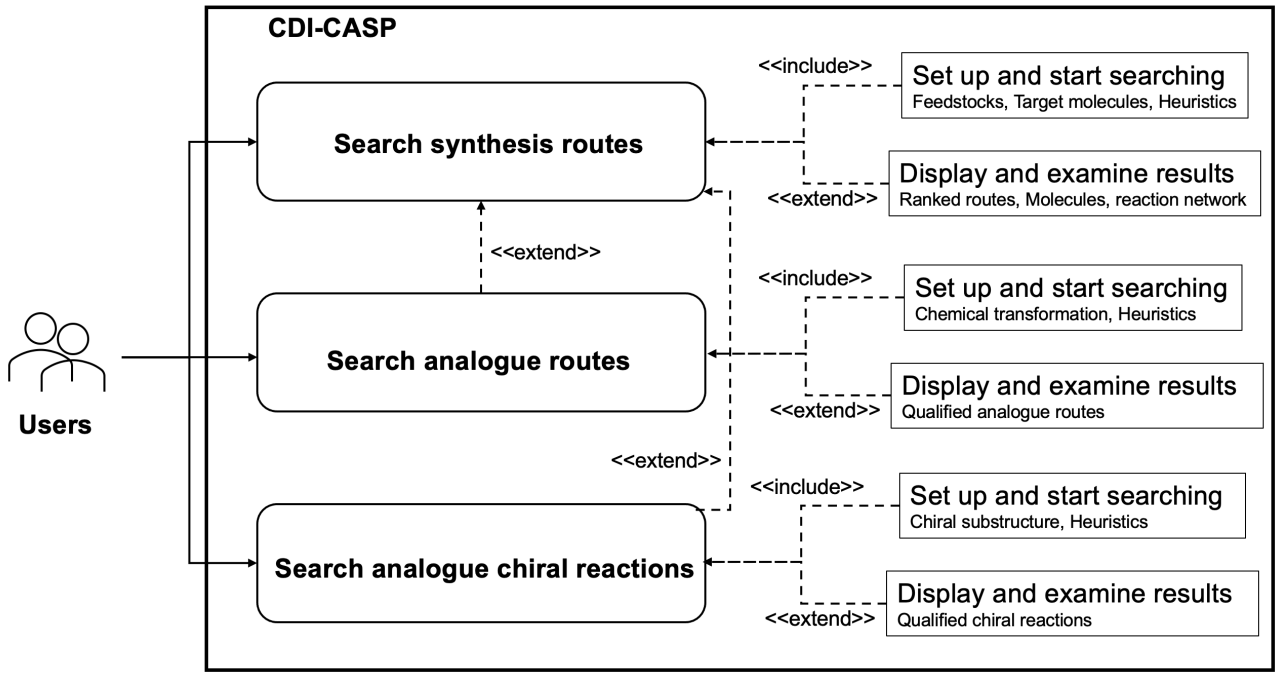


Fig. 2. A diagram illustrating user interaction with CDI-CASP.

Search synthesis routes

The “searching synthesis routes” function provides a retrosynthesis search, starting from one or multiple target molecules to be synthesized. In practice, multiple targets of similar molecules or isomers can be specified to gather as much relevant information as possible. Target molecules and feedstocks are specified by drawing molecular diagrams (Fig. 3.1). Users can also select feedstocks by types: “commercial”, “hubs” and “renewables”. Here “commercial” means commercially available molecules, “hubs” are the highly accessible molecules found by using ML and graph analysis,[17] and “renewables” are molecules derived from biomass.

After specifying targets and feedstocks, the number of steps to be searched backwards (up to 15) should be provided. Many heuristics have been devised to guide the search and to filter results, Table 1. Most are lumped in “Advanced Settings” with default configurations. New users may get better understanding of how to curate “Advanced Settings” after the first round of search. Preference of molecular fragments and yield threshold of reactions are picked in the first round of search, since fragment-based heuristic may be relevant to many search objectives, see Fig. 3d1.

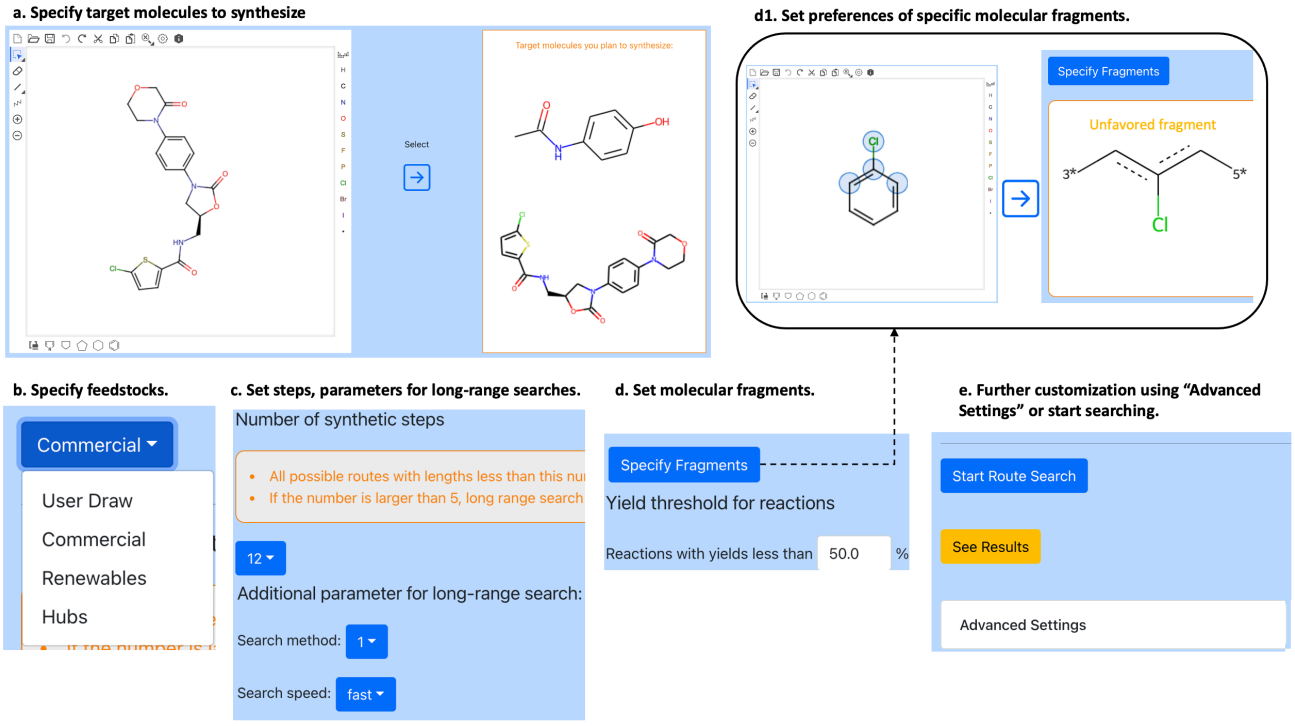


Fig. 3. Demonstration on setting up parameters and heuristics for “searching synthesis routes”.

Table 1. Parameters and heuristics for searching synthesis routes.

Inputs	Types	Definitions
Targets	Mandatory	Molecules to synthesize.
Feedstocks	Mandatory	Molecules as starting materials.
Synthesis steps	Mandatory	Total number of synthetic steps.
Parameters for long-range search	Mandatory	When synthesis steps larger than 5, need to choose one of two search methods and search speed (fast or slow).
Fragments	Optional, molecule heuristic	Unfavored molecular fragments to avoid or preferred fragments to keep along a route.
Yield threshold	Optional, reaction heuristic	Reactions with yield less than the threshold will not be included in results.
Preference on solvents	Optional, reaction heuristic	Avoid reactions containing unfavored solvents or only keep reactions with preferred solvents.
Preference on reagents	Optional, molecule reaction	Avoid reactions containing unfavored reagents or only keep reactions with preferred reagents.
Preference on reactants and products	Optional, molecule heuristic	Avoid reactions containing unfavored reactants/products or only keep reactions with preferred reactants/products.
Remove molecules and reactions in network	Optional, network heuristic	Remove unfavored molecules or reactions in reaction networks composed by synthesis routes.
Similarity threshold	Optional, molecule heuristic	Reactions with yield less than the threshold will not be included in results.
Yield-based heuristic for routes	Optional, route heuristic	Cumulative yields of target products less than this threshold will be removed from results.
Similarity-based heuristic for routes	Optional, route heuristic	Cumulative similarity-based score of routes less than this threshold will be removed from results.
Ranking methods	Optional, route heuristic	Routes will be ranked and filtered using either yield-based scores or similarity-based scores.

In many cases, search may end up with thousands of qualified routes. Users may consider five ways to handle the overwhelming number of results:

1. Rank routes using scores listed in Table 2. Different scores provide different perspectives of routes. Users may only need to examine *top-n* routes ranked by a specific type of score.
2. Group resulted routes by feedstocks. The number of available feedstocks can be limited. Probably only a few of feedstocks are of the interest. Routes grouped by feedstocks can also be ranked by other specific route scores.
3. Analyse reaction routes with the aid of network view. Note that the final results afford a small reaction network extending from targets. Each reaction route is a path from a feedstock to a target. Drawing of a network provides a straightforward view of reaction relationships. Hundreds of routes may be derived from a networks with only tens of molecules.

4. Remove reactions and/or molecules from the final reaction network. Removing edges (reactions) and/or nodes (molecules) may significantly reduce the number of routes, depending on centralities of the removed network components. Reactions and molecules to be removed normally fall into two classes: 1) they are undesired, such as inefficient reactions and hazardous molecules; 2) reactions and molecules that are shared by most synthesis routes from the network view of results. In this case, keep these useful and similar synthesis routes as a record, then re-run the ranking of routes by excluding these reactions and/or molecules. This operation allows users to focus on unconventional routes.
5. Use more strict heuristics to re-run the search, so that a smaller number of routes will be generated. For example, increase the threshold of similarity-based or yield-based scores to reduce the number of qualified reactions and routes. This method is recommended when search ends up with too many routes (e.g., over 10,000).

Table 2. Scores for ranking of routes and their definitions.

Scores	Definitions
Yield	Cumulative yields of the target molecule. Routes with the same number of steps with yield records are group and ranked together.
Similarity	A molecular similarity-based metric measure reliability and usefulness of routes. This score is designed to mitigate the issue caused by shortage of yield records and recommended as a default metric for most case studies.
Popularity	A higher popularity means more frequently reported intermediates/reactions in a route. Usually top-n routes ranked by these scores are reported routes or representing typical synthesis methods of the target molecules.
Novelty	Oppose to "Popularity", higher novelty of a route suggests lower number of popular intermediates/reactions, i.e., high novelty and low feasibility, but still qualified routes based on searching parameters.
Propitious	Routes with high propitious values are less "popular" but worthy for further study due to their significant reliability.

Search of analogues routes

The "search analogue routes" function provides "analogue reaction routes" leading to a target chemical transformation provided by a user. Two examples of analogue routes are shown in Fig. 4. Here, a chemical transformation is defined by changes from one substructure to a desired substructure (as highlighted in Fig. 4), rather than from one molecule to another molecule.

In general, "Search analogue routes" is used to explore alternative sub-routers in two scenarios:

1. Avoiding unfavoured molecules results in no qualified routes found when using the "Search reaction routes" function. The chances of finding qualified analogue routes is higher since it only

matches chemical transformations instead of specific molecules. As the example shown in Fig. 4a, reaction information of the analogue route can help to design a new route which may replace the unsatisfactory one involving hazardous molecules highlighted in red.

2. Finding analogue routes to support new synthesis ideas, as shown in Fig. 4b. In this example, we wanted to know if it was possible to link the *alpha* carbon of aldehyde group to the adjacent amine group. The single-step analogue route was found. Although it was not an intra-molecular reaction, the reaction information could be still useful, since the identified chemical transformation shared the same substructure with the target transformation.

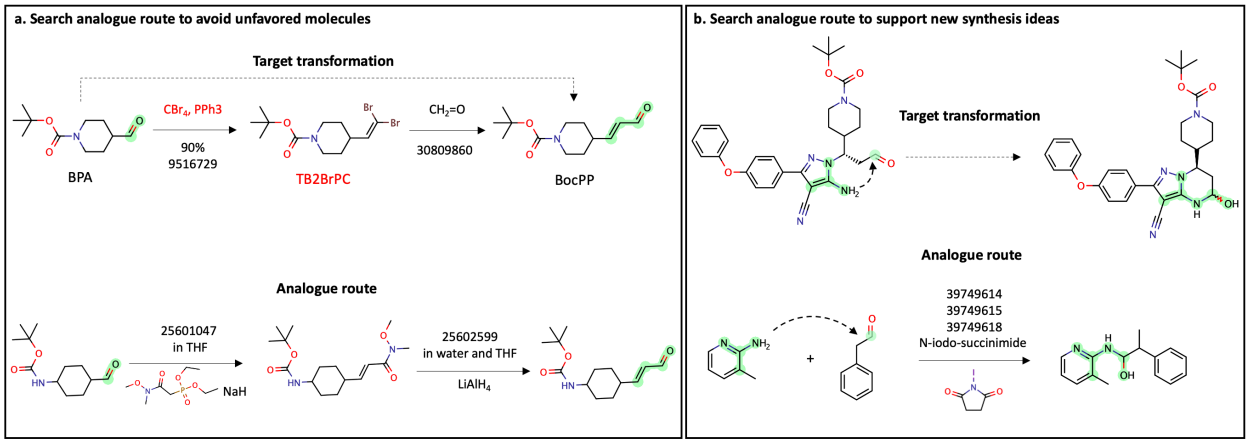


Fig. 4. Two examples of "Search analogue routes". a) Search analogue routes to avoid unfavoured molecules highlighted in red; b) Search analogue routes to support the new synthesis idea.

Searching analogue chiral reactions

Asymmetric synthesis plays an important role in accessing high value chemicals. This function in CDI-CASP is able to find reported asymmetric reactions that lead to a target chiral substructure, instead of a specific chiral molecule. This approximate search approach improves the exploration rate of reaction data, since the number of asymmetric synthesis records of a specific chiral molecule are limited in many cases. To initiate the search, a user needs to provide a target chiral substructure cut from a chiral molecule to be synthesized. As shown in Fig. 5, specification of a chiral substructure is conducted by selecting atoms expanding from a chiral center. Selected chiral substructures can be of any size, as long as they afford chiral centre (i.e., at least a chiral centre surrounded by four atoms). Users can include molecular segments that are considered important or unique for asymmetric synthesis. Like "search analogue routes", relevance and number of results are determined by the chiral substructure and heuristics listed in Table 3. Reaction conditions and asymmetric catalysts of analogue reactions are helpful for design of new asymmetric reactions for experimental validations.

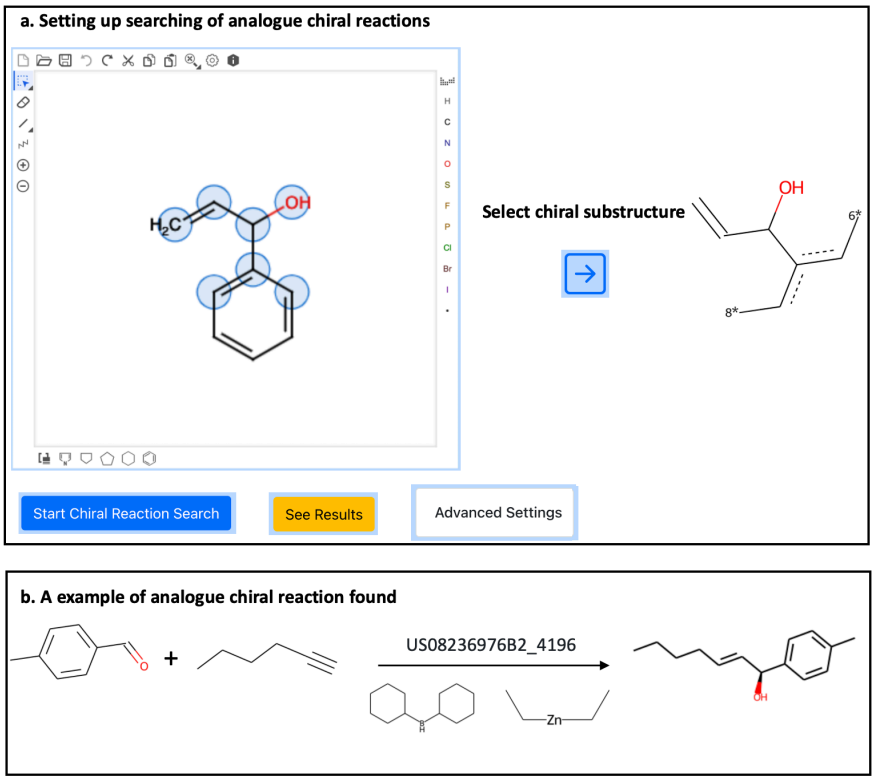


Fig. 5. Demonstration on setting up parameters and heuristics for: (a) "search analogue chiral reactions", as well as (b) an example of analogue chiral reaction found.

Table 3. Parameters and heuristics for searching analogue routes.

Inputs	Types	Definitions
Chiral substructure	Mandatory	The target chiral substructure will be shared by the chiral products of all resulted reactions.
Similarity threshold	Optional, molecule heuristic	A higher value means chiral molecules in results are more similar to the chiral molecule drawn by a user.
Explicit Hydrogen	Optional, molecule heuristic	Default value is true, meaning number of hydrogen atoms attached to heavy atoms in substructures will be considered. Try false, only when no results can be found.
Ring	Optional, molecule heuristic	Default value is false, meaning if ring property of atoms in substructures will be considered.
Aromaticity	Optional, molecule heuristic	Default value is false, meaning if aromaticity of atoms in substructures will be considered.

Searching synthesis routes of *S*-Zanubrutinib as a case study

Aiming to find synthetic routes to *S*-Zanubrutinib that would compare favourably to the patented route shown in Fig. 1, a long-range search (12-steps) was conducted first. Several interesting routes were found but not many feedstocks were in the class of 'renewable' molecules. Hence, a long-range search from renewable molecules to feedstocks found in the first round was performed. After examining these long-range routes, we wanted to shorten the synthesis length, improve yields, and avoid hazardous reagents,

solvents, and intermediates. Searchers of analogue routes and chiral reactions were conducted for this purpose. Inspired by analogue routes and chiral reactions, we proposed new routes and reactions together with new intermediates. A few more searches were run to support the new design. Finally, we obtained a shorter, safer, greener route composed of the reported and newly designed reactions. The proposed route was then verified in the *i*DMT lab.

Results and Discussion

Exploring synthetic routes through retrosynthetic search

Reaxys™ database contains *S*-Zanubrutinib (Reaxys IDs: 27490480) and its enantiomer (Reaxys IDs: 27490481) and racemic mixture (Reaxys ID: 27490479), as shown in Fig. 6. Searches of 12-steps routes from all of them were conducted, in case missing useful reaction information, especially sub-routes that were not relevant to the chiral centre could be identified. The long-range search was performed in a backwards manner staring from these three IDs. Main search parameters adopted include:

1. Feedstocks: renewable compounds derived from biomass.
2. Maximum synthetic steps: 12.
3. Similarity threshold: not less than 0.5.
4. Similarity-based heuristic for routes: overall similarity values of all routes should be not less than 0.6, only allowed 1 reaction step in a qualified route with a similarity value lower than 0.6.

The first round of search provided an automatic "literature review" of all reactions relevant to the target synthesis. For convenience, all routes were presented in a linear format without any branches from intermediates. Some examples can be found in Appendix, Fig. A1. It is interesting to see that top routes ranked by similarity-based scores are similar to the reported route starting from 4PA, while top routes ranked by the popularity metric are similar to the branch of the reported route starting from BPA, see Fig. 1. This could be due to the 4PA branch containing intermediates that are more similar to the target, compared to the branch from BPA.

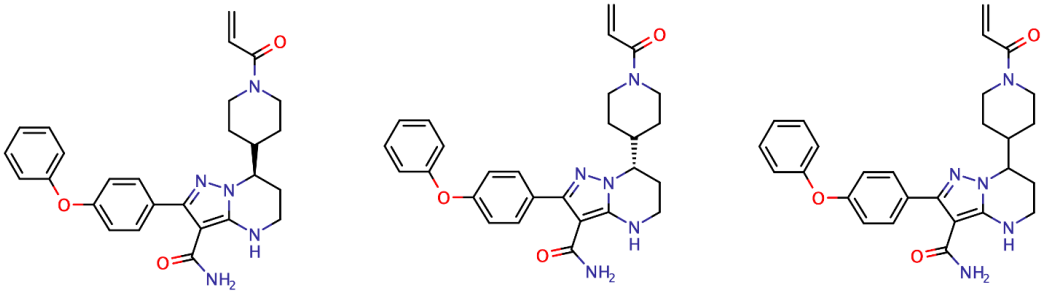


Fig. 6. A list of records related to *S*-Zanubrutinib in Reaxys™. All three molecules were considered as targets for a single epoch of long-range search.

One of the routes that was similar to the benchmark route is shown in Fig. 7. The exactly same route was not found because of the difference in definition of "single step reaction" among research groups. For example, steps 6 and 7 in the benchmark route, Fig. 1, were considered as a single step in step 8 of the route shown in Fig. 7. Most "single step" reactions are actually "one pot" reactions. Nevertheless, comparing the identified route, Fig. 7, and the reported route, Fig. 1, one can conclude that synthetic strategies of both were very close.

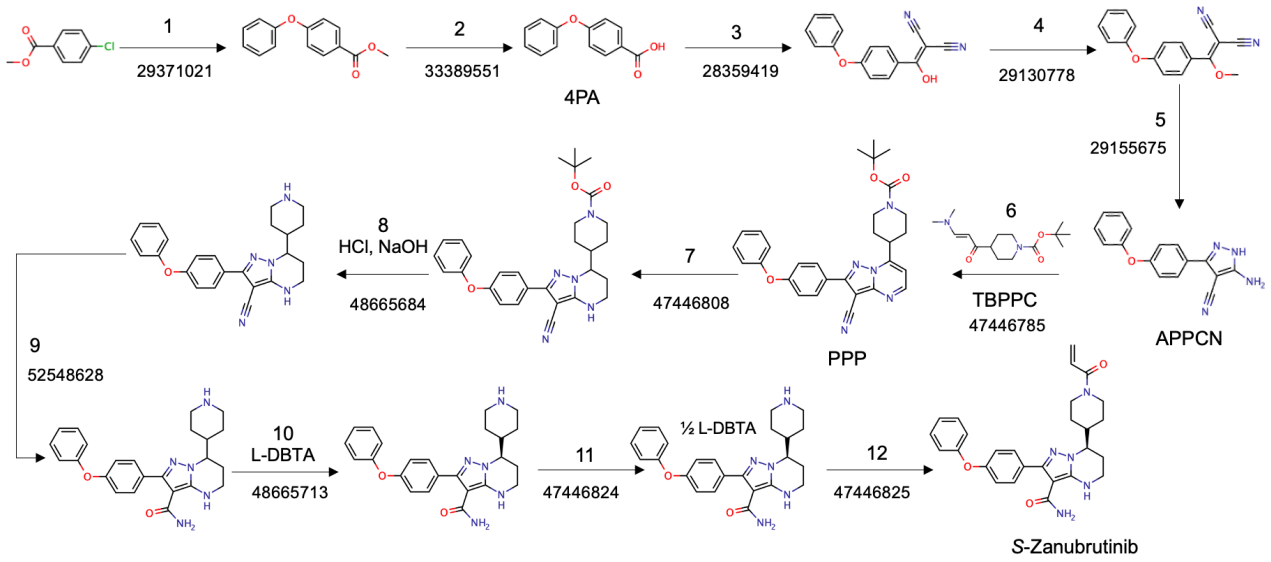


Fig. 7. An example of routes found through 12-steps search; this route is similar to the one reported in patent, see Fig. 1.

Both routes shared the same key intermediates (4PA, TBPPC, APPCN, PPP) and the same reaction for the formation of the main scaffold of *S*-Zanubrutinib (step 5 of the benchmark route and step 6 of the route in Fig. 7). The main difference was the priority of hydrolysis of nitrile and asymmetric resolution. In the benchmark route asymmetric resolution was performed before hydrolysis, while the order is reversed in the route found by CASP.

The 12-steps search not only yielded 12-steps routes but also all possible routes within the 12-steps range. Analysing these routes may unveil useful sub-routes. As shown in Fig. 8, a shorter sub-route from PPP to *S*-Zanubrutinib was found, which also avoided using organochlorine (acryloyl chloride) as a hazardous intermediate. Although development of chromatographic separation could be a challenge, this result may encourage researchers to devote efforts on this path. Other notable sub-routes for the synthesis of key intermediates (APPCN and TBPPC) are summarized in Figs. A2 and A3. These results may help to design shorter and safer routes. Many reactions identified in this search are not typically found in the routes to *S*-Zanubrutinib (e.g., a bio-synthesis 9010270 in Fig. A3).

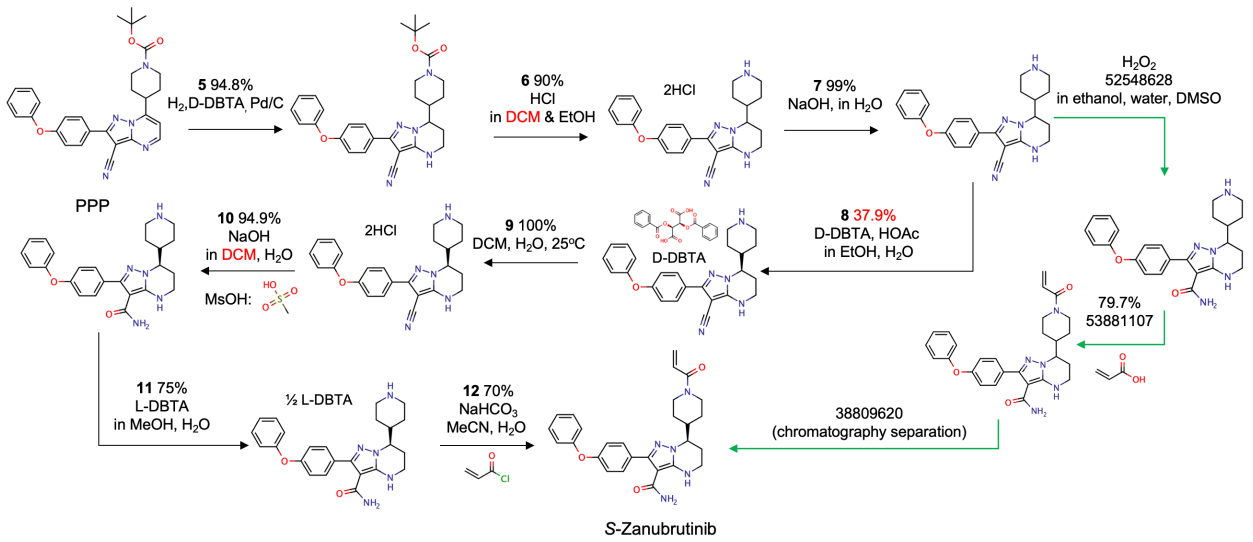


Fig. 8. A modified benchmark route by inserting an alternative sub-route (green arrows) from PPP to *S*-Zanubrutinib.

Although more than 5,000 routes were found in the first round of search, all routes started from a group of only 50 feedstocks, among which not many renewable feedstocks were identified, suggesting the searching was not deep-enough to reach to renewable feedstocks. Therefore, 4-step retrosynthesis searches were conducted from two key feedstocks, i.e., 4PA and BPA. Some encouraging routes are shown in Figs. 9 and 10. The small reaction network in Fig. 9 suggested that all carbons of 4PA can be derived from lignin and CO_2. For the synthesis of BPA, platform molecules from cellulose, e.g., *D*-glucose, *D*-sorbitol, furfural alcohol, may be employed as suitable building blocks, see Fig. 10. Many reactions were conducted in environmentally benign media such as water, ionic liquids, using bio-synthesis or solvent free. A promising route starting from a hub molecule (pyridine) was also found. This route contains many attractive features: incorporation of CO_2 via an electrochemical reaction, incorporation of salt of formic acid via heterogeneous catalysis, free of hazardous reagents, high yield of each reaction step.

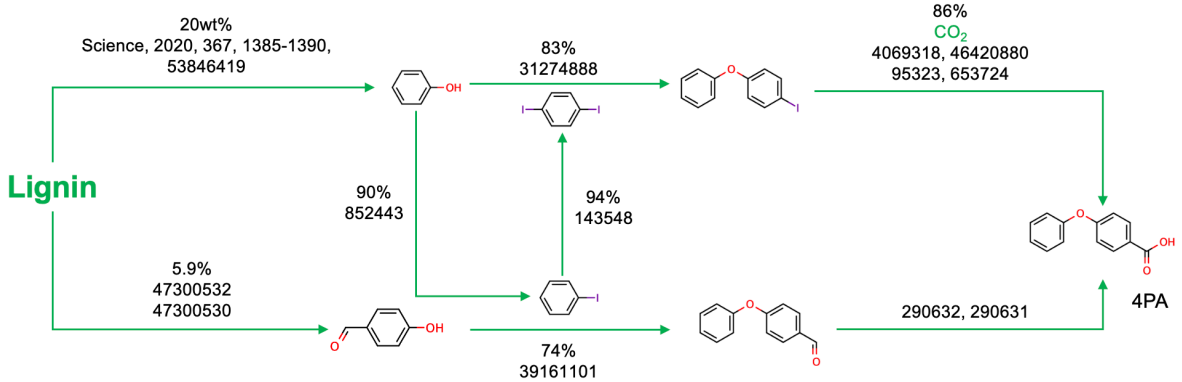


Fig. 9. Routes found from lignin derived molecules to the key feedstock 4PA.

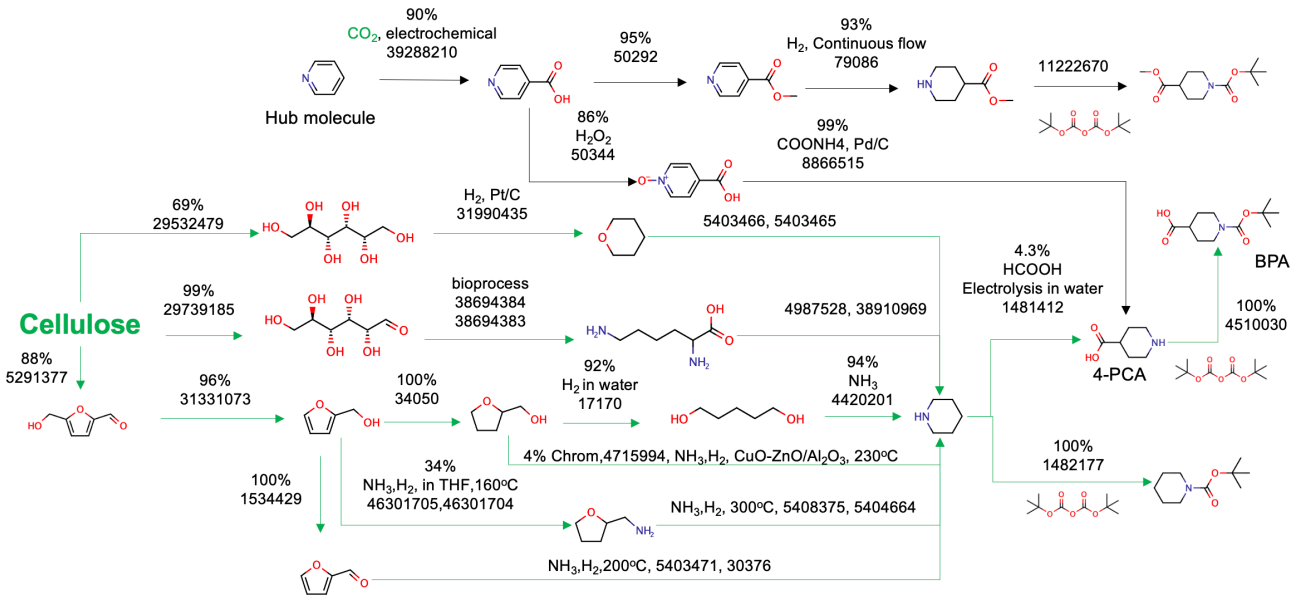


Fig. 10. Routes found from cellulose derived molecules to the key feedstock BPA, as well as other two precursors of the key intermediate TBPPC.

To sum up this section, we have achieved the following with the help of "search synthesis routes" routine of CDI-CASP:

1. Obtained fundamental understanding of the synthesis of *S*-Zanubrutinib based on routes found through long-range search. 4PA and BPA were found as key feedstocks, APPCN and TBPPC were key intermediates. The ring formation reaction between APPCN and TBPPC led to PPP bearing the main scaffold of *S*-Zanubrutinib.
2. Alternative sub-routes were identified, they may motivate design of shorter and greener sub-routes.
3. Exploration of renewable molecules as feedstocks yielded promising results.
4. New reactions were identified which may replace steps of the benchmark route involving hazardous molecules.

On the basis of these findings, modifications of the benchmark route are summarized in Fig. 11, from which further improvements of synthesis strategy were also identified:

1. All synthesis routes found in this stage involved asymmetric resolutions, which capped the maximum yield at 50% (step-8 in Fig. 11). The possibility to replace asymmetric resolution by asymmetric synthesis should be explored. As shown in Fig. 11, implementation of asymmetric synthesis at step-5 could also shorten synthetic length by overriding 4 steps (from step-6 to step-9).
2. A hazardous reagent (HOBT in step-1a) and a solvent (DCM in step-3b and step-6) remained in the route at this stage.

Therefore, subsequent searches were conducted aiming to solve these issues.

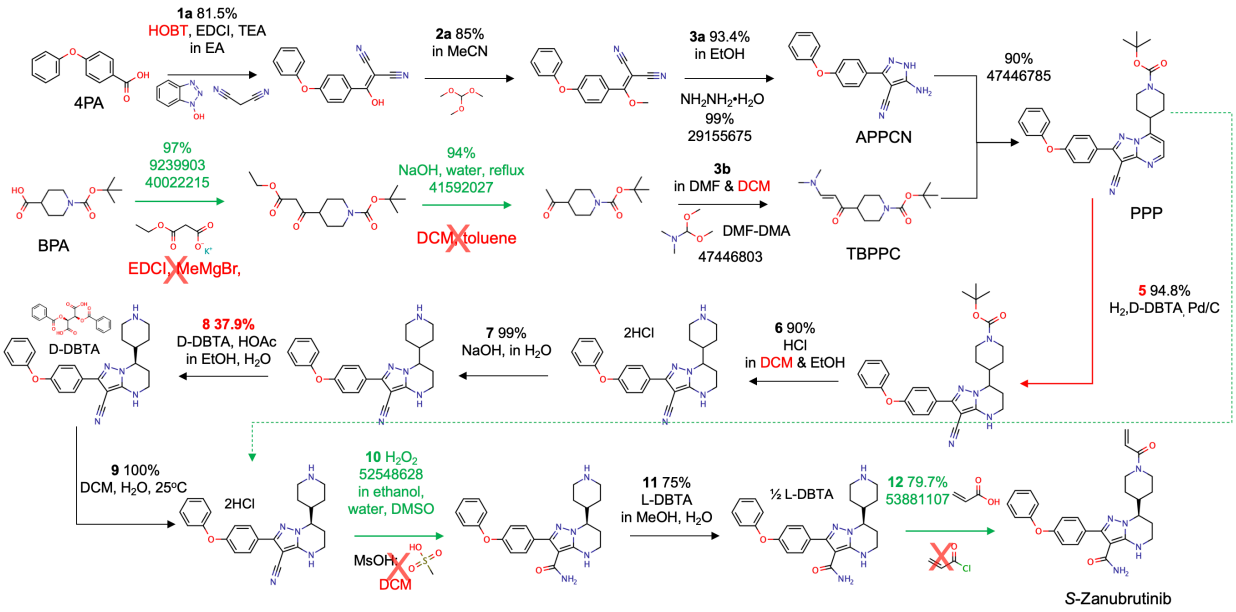


Fig. 11. Proposed routes based on results from retrosynthesis searching, as well as further improvements.

Design asymmetric synthesis with the aid of chiral search

The search of asymmetric synthesis started with specification of a substructure around chiral centres. As shown in Fig. 12, the target chiral substructure consisted of 7 atoms including two aromatic nitrogens and five aliphatic carbons, as highlighted by their atomic indices of *S*-Zanubrutinib. Two types of asymmetric transformations were of interest. The first type was catalytic asymmetric hydrogenation (see Fig. 12, type 1 chiral reaction). Decent enantiomeric excess (as high as 98%*ee*) was reported for the reaction leading to formation of the target chiral substructure. Therefore, one possible way could be the asymmetric hydrogenation of PPP, as shown in Fig. 11, by employing similar catalysts and reaction conditions. The other type of asymmetric transformation involved the reaction between an azole and an acetaldehyde, see Fig. 12, type 2 chiral reaction. Interestingly, all reactions were catalysed by metal-free catalysts. This type of reaction did not lead the formation of a ring, but it was unexpected and inspired us to design a new 2-step sub-route as shown at the bottom of Fig. 12. A new intermediate BocPP, bearing an acetaldehyde group, was proposed to react with the key intermediate APPCN producing an intermediate with the target chiral substructure, followed by a ring closing reaction between the *alpha*-carbon of the dangling aldehyde group and the amine group.

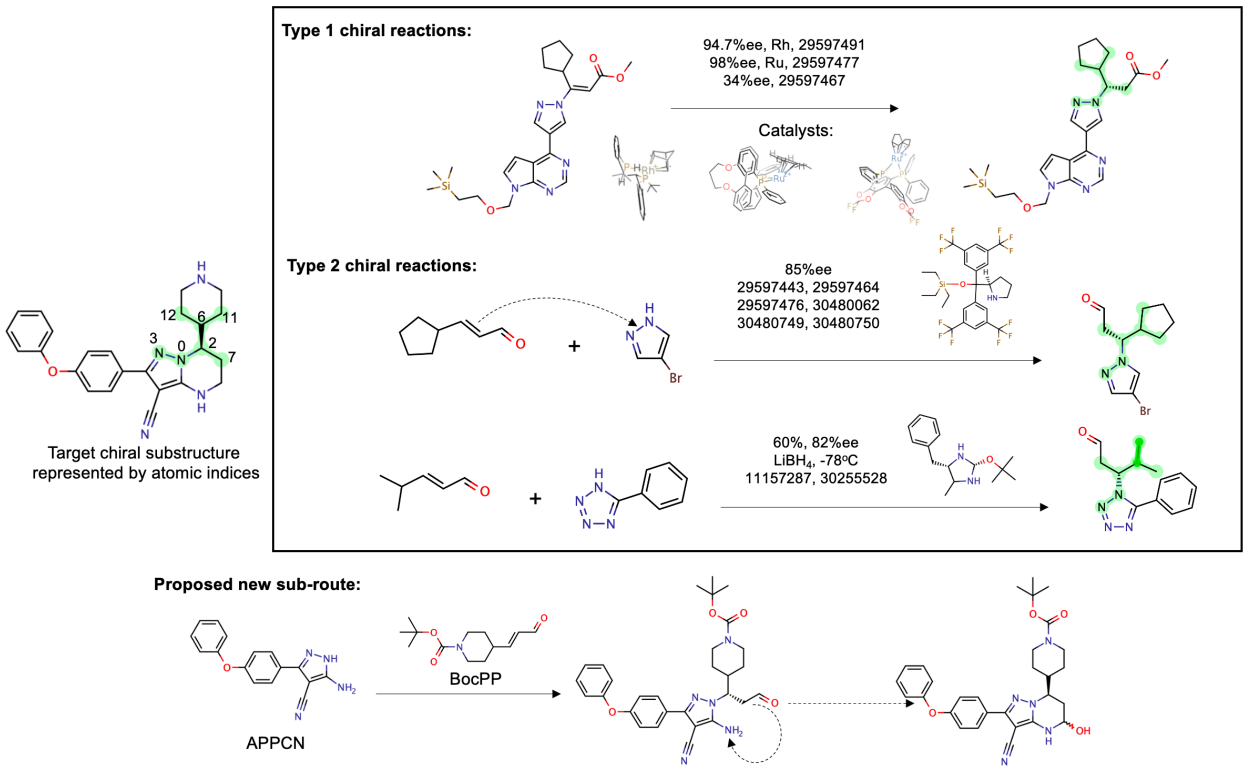


Fig. 12. Specification of the target chiral substructure and analogue chiral reactions.

In this section, we demonstrated that searching chiral reactions was not limited to certain types of chiral reactions, but all reactions that led to the target chiral substructure. A novel sub-route was proposed which also promoted new questions:

1. Is it possible to synthesize BocPP via a green synthesis route?

2. Can we find reported analogue reactions to support the ring closing reaction, as well as ways to remove the hydroxyl group yielding *S*-Zanubrutinib?

Bearing these questions and the second issue carried forward from previous section, we continued to the third round of search, mainly taking advantage of the "Searching analogue routes" function of CDI-CASP tool.

Design new routes based on analogue routes

A short-range retrosynthetic search was conducted first for the synthesis of BocPP, the intermediate proposed in the previous section. The search gave 3-step routes starting from a commercial molecule 4-FP, as shown in Fig. 13. The feedstock 4-FP can also be synthesized from 4-PCA (Reaxys ID: 37513873) which in turn can be prepared from platform molecules derived from cellulose, see Fig. 10. However, the sub-route from BPA to BocPP involved several hazardous molecules as highlighted in red in Fig. 13, and this was the only route found using "Search synthesis routes".

We then defined the target chemical transformation, highlighted in green in Fig. 13, to initiate the search of analogue routes. The searching led to three routes. Routes-1 and -2 were single step reactions, but both contained an unfavoured phenylphosphine derivative (PPh3A). The analogue route-3 was more appealing considering its solvents and reagents, as well as its high molecular similarity with respect to the starting and the target molecules. Thus, a new route from BPA to BocPP was proposed, see the bottom of Fig. 13.

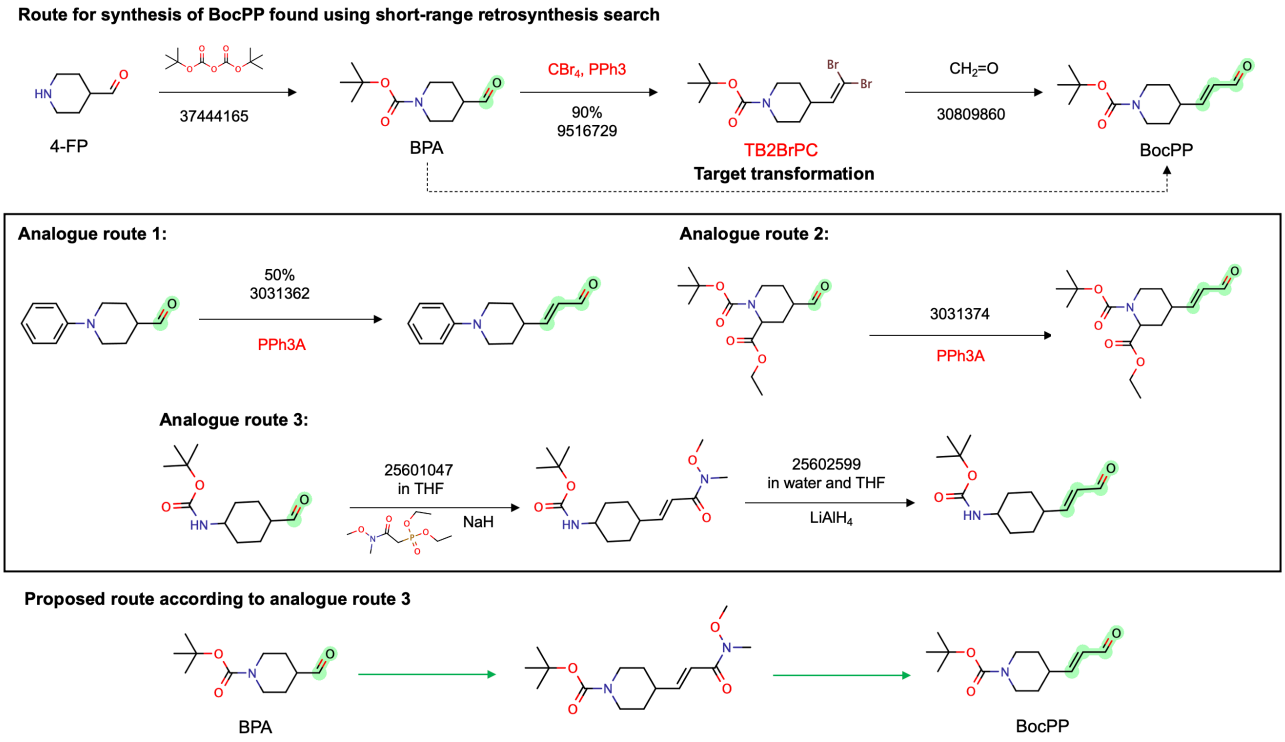


Fig. 13. Searching synthesis route of BocPP as a proposed intermediate.

To define a chemical transformation, selection of a substructure could be tricky, since both the number and quality of results are sensitive to atoms included. If the substructure selection included too many atoms, i.e., a very specific substructure, no qualified analogue routes can be found. Whereas a "loose" definition of chemical transformation with only a few atoms, (e.g., from -Cl to -OH), may lead to a very large number of routes, from which one can hardly find routes that suite the target chemical transformation. Therefore, multiple try-and-error exercises may be needed to find the balance point. The search of analogue routes for step-1a in the benchmark route, see Fig. 1, can serve as a good demonstration. As shown in Fig. 14, we purposefully excluded nitrile and hydroxyl groups of the product in step-1a, otherwise the only analogue route shown in the middle of Fig. 14 would be missed. Based on this analogue route, a new sub-route was designed to avoid the hazardous reagent HOBT. Note that the precursor of APPCN was bearing an amine group instead of a hydroxyl group as the one in the benchmark route. But the new route was still workable because of the reaction to APPCN found from the long-range search (please refer to the second route in Fig. A1a in Appendix).

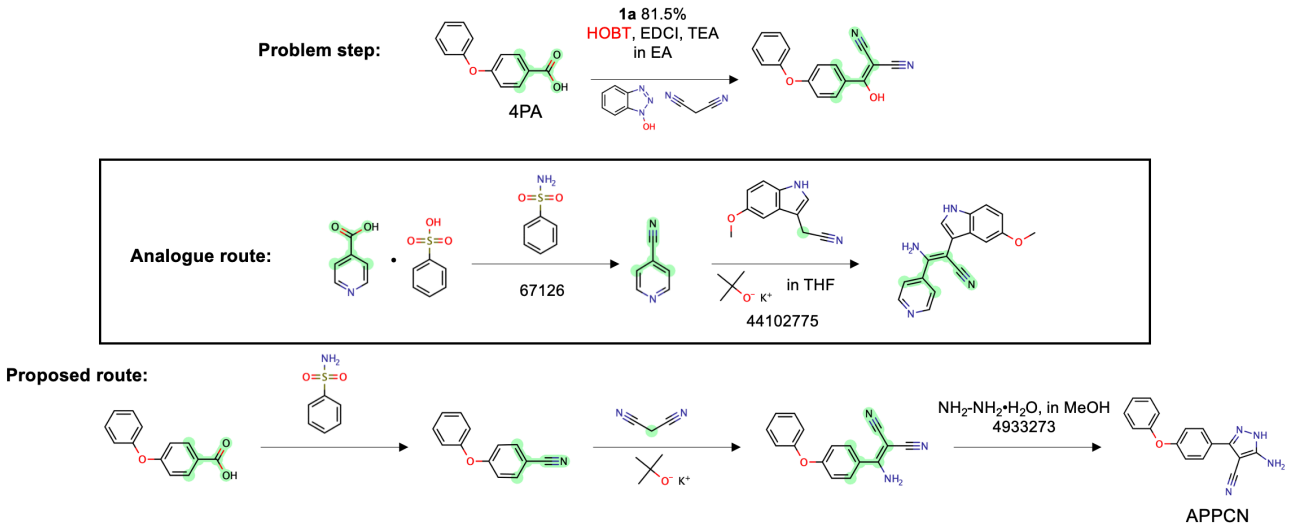


Fig. 14. Proposed new route to avoid using of HOBT based on a route found using analogue route search. The last reaction in the proposed route was found during long-range search, see the second route in Fig. A1a in Appendix.

In the new sub-route proposed in the previous section, a critical step for the formation of scaffold of *S*-Zanubrutinib was a ring-closing transformation. As shown in Fig. 15, the target transformation involved bond formation between the *alpha*-carbon of the aldehyde group and the amine group. After carefully tuning the selection of substructures (as highlighted in Fig. 15), three analogue reactions were found, encouraging further validation in the lab.

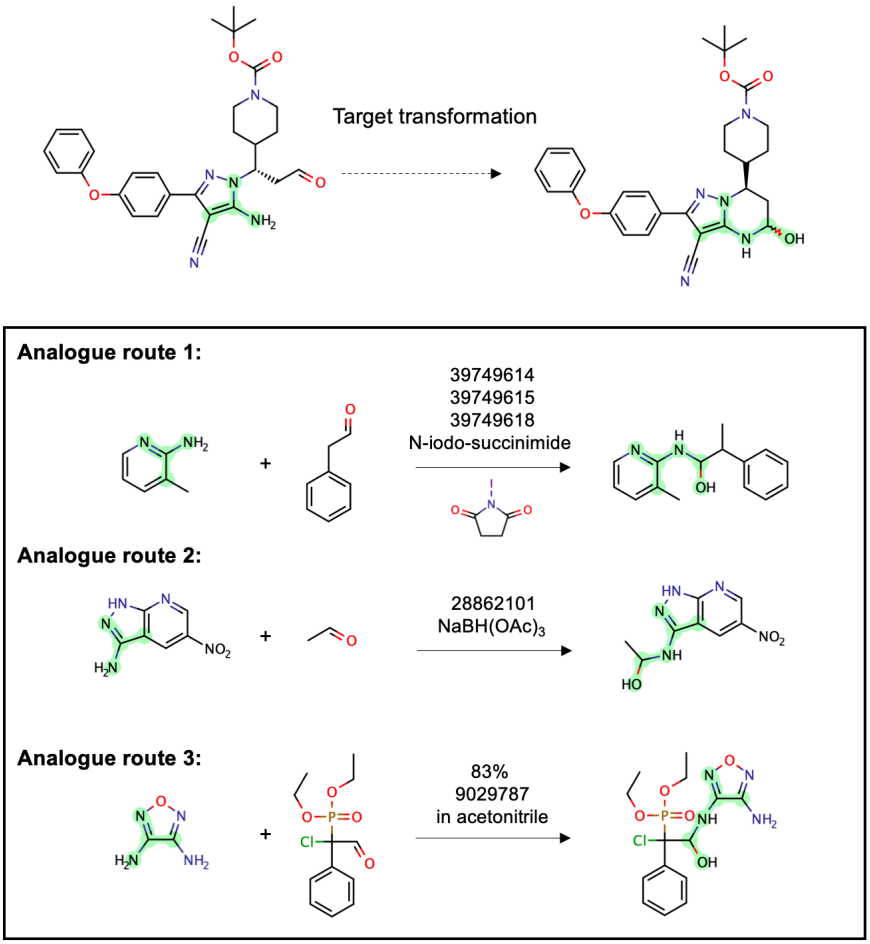


Fig. 15. Searching analogue reactions to support the ring-closing transformation proposed in previous section.

The last chemical transformation to explore was the removal of hydroxyl group on the bicyclic ring, giving rise to the final product. To our surprise, only one analogue route was found in which the chemical transformation sharing the same substructures as highlighted in Fig. 16. This sub-route was in line with general chemical knowledge, i.e., a dehydration step followed by a hydrogenation process. The hydrogenation part was relatively easier to carry out in practice compared to dehydration. Therefore, a single step dehydration reaction with substructure highlighted at the bottom of Fig. 16 was searched leading to another reaction as shown in Fig. 16. Both dehydration reactions happened at a high steric hindrance site. We hypothesised that solid acid catalysts could be applicable to this substrate, since there are no bulky groups around the hydroxyl group. Nevertheless, for both analogue dehydration reactions, the adjacent nitrogen was a tertiary amine instead of secondary amine, and the bonding pattern of carbon linking to the hydroxyl group was also different (i.e., tertiary vs quaternary). Therefore, feasibility of the proposed sub-route still needs to be verified.

In the last section, all remaining questions from previous searches are all resolved by searching analogue routes:

1. A synthesis route of the newly proposed intermediate, i.e., BocPP was designed. Based on the 2-step analogue route, it was possible to propose synthesis without incurring hazardous reagents and solvents.
2. A new route from the feedstock 4PA to APPCN was proposed to avoid using HOBT.
3. The chiral reaction proposed in the previous section led to two consecutive synthesis objectives: formation of six-membered ring and removal of hydroxyl group generated during ring-closing. Analogue routes found for both processes provided useful reaction information to complete the synthesis planning.

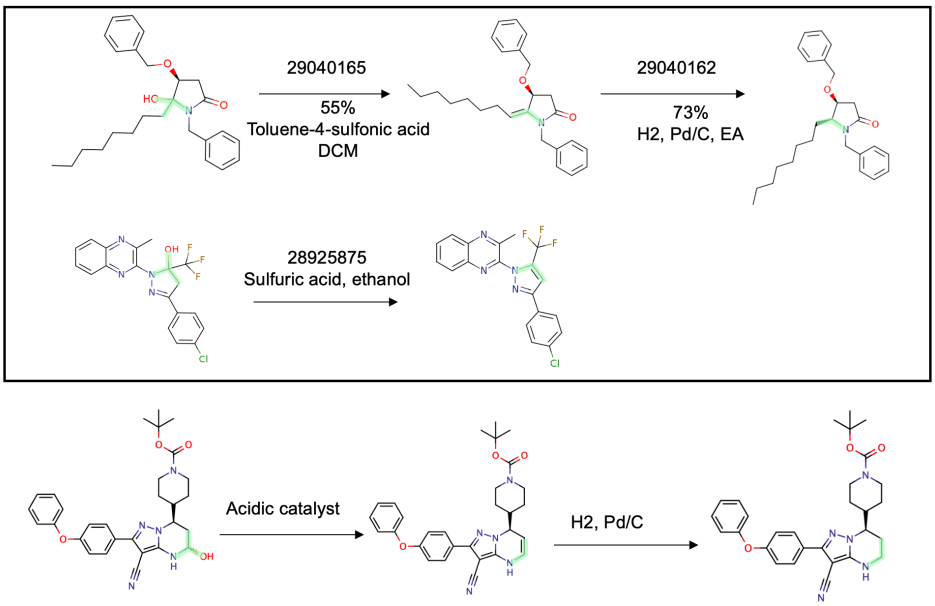


Fig. 16. Analogue routes found to remove hydroxyl group of an intermediate.

New route for synthesis of *S*-Zanubrutinib after in-depth search using CDI-CASP

By combining all search results, a new route to *S*-Zanubrutinib was proposed, see Fig. 17. Compared with the benchmark route in Fig. 1, only step-8 was similar to step-6 of the reported route, and we have achieved all synthesis objectives listed in the Introduction section:

1. The number of synthetic steps was reduced from 12 to 10, and one more step was saved for the synthesis branch of BocPP.
2. Yield of *S*-Zanubrutinib could be improved because chiral resolution was replaced by asymmetric synthesis. Most reactions are designed based on reactions found with high yields.
3. All hazardous reagents, solvents and intermediates in the benchmark route were absent in the new route.
4. The potential of using renewable molecules as feedstocks had been extensively explored. As shown in Figs. 9 and 10, both 4PA and BPA, can be prepared from renewable molecules.

Experimental verification

To confirm the feasibility of the CDI-suggested asymmetric synthesis route of *S*-Zanubrutinib, preparation of BocPP and APPCN were commenced (Fig. 17).

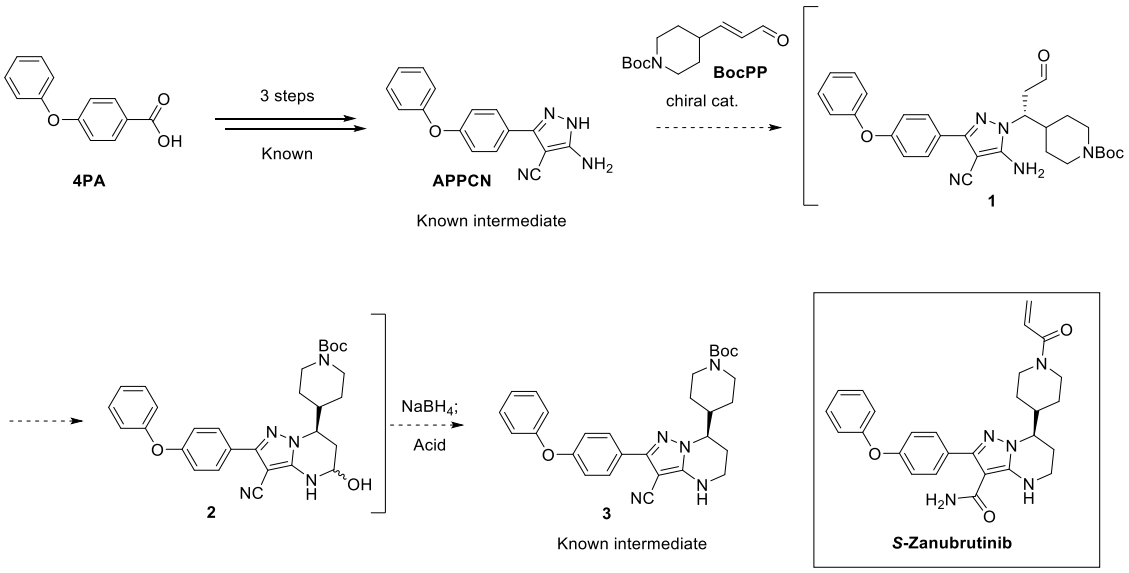


Fig. 17. Proposed asymmetric synthesis route of S-Zanubrutinib.

When aldehyde BPA was reacted with PPh_3CHCHO with reference to analogue routes 1 and 2 shown in Fig. 13, BocPP was obtained in low yield with impurities which were difficult to remove by column chromatography (Fig. 18a). On the other hand, the reported three-step synthesis[18] worked well to give desired BocPP in good yield without impurities.

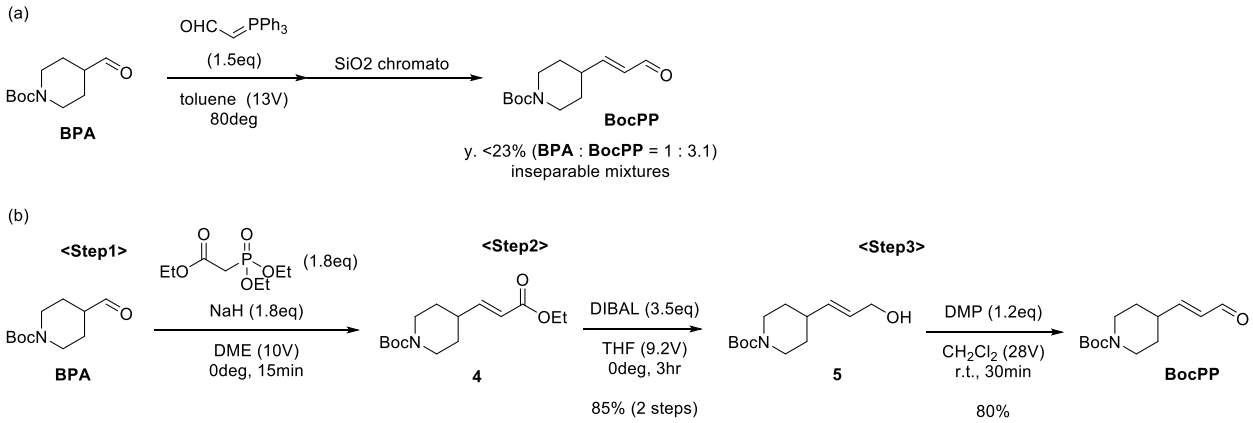


Fig. 18. Preparation of BocPP.

Synthesis of APPCN[19] was performed by numbering-up the flow reactor developed in the *i*DMT lab.[i] Condensation of 4PA and malononitrile with EDCI and DMAP gave **6**, which was treated with trimethylorthoformate to give **7**. Following cyclization with hydrazine provided APPCN, precursor of asymmetric Michael reaction (Fig. 19).

Having completed the synthesis of BocPP and APPCN, we next tried the asymmetric Michael reaction with reference to Type 2 chiral reaction[20] in Fig. 12. If the asymmetric Michael reaction proceeds, subsequent cyclization would then form hemiacetal, which could be reduced by $NaBH_4$ to give **3**. However, it was found that this reaction condition did not give the desired product but gave a complex mixture., Fig. 20. Since the reaction was messy at the first Michael addition step, it was considered that the side reaction occurred due to the high reactivity of the aldehyde moiety of BocPP.

Michael reaction was performed using unsaturated ester **4** instead of the unsaturated aldehyde to suppress reactivity, resulting in lactam rac-**8** that underwent cyclization after Michael addition, see Fig. 21. The obtained lactam rac-**8** was then reduced by $NaBH_4$ to give known intermediate rac-**3**. Due to competing nitrile reduction, rac-**3** was obtained in low yield, but the yield could be improved by changing the reaction conditions.

[i] This is described in the report available online from www.acceleratedmaterials.co.uk. Link: https://acceleratedmaterials.co.uk/wp-content/uploads/2022/09/Accelerated-Materials-Whitepaper_2022_Instant-scale-up-of-small-molecule-pharmaceuticals.pdf

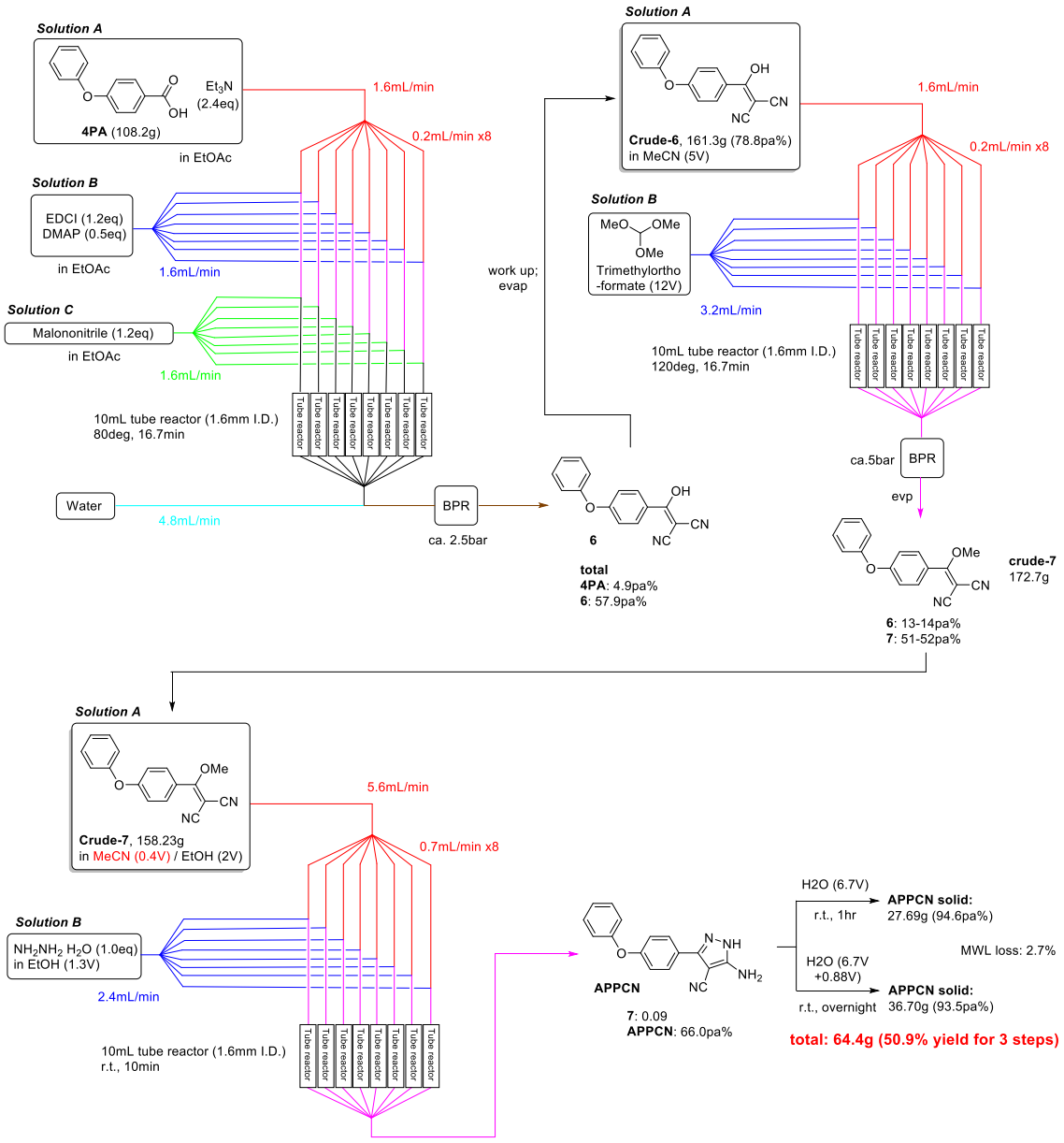


Fig. 19. Preparation of APPCN.

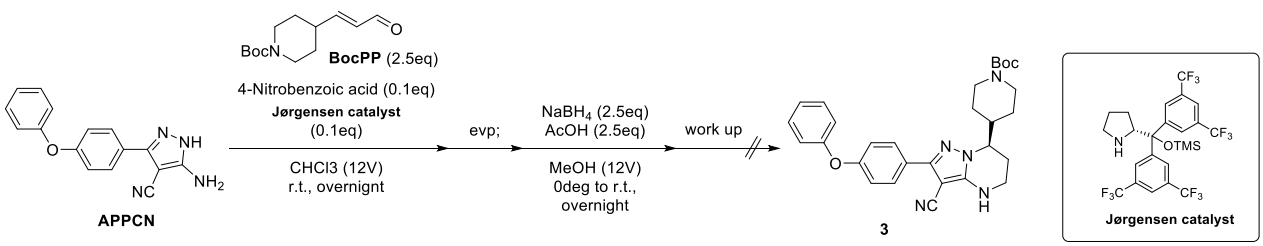


Fig. 20. Unsuccessful asymmetric Michael addition using Jørgensen catalyst.

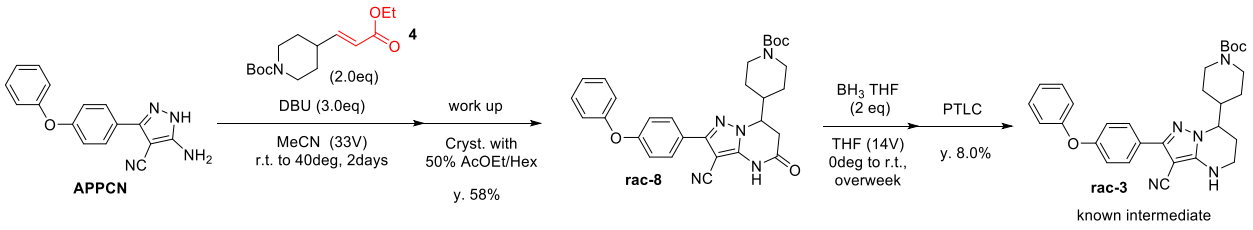


Fig. 21. Synthesis of rac-**3** by Michael reaction with unsaturated ester **4** and subsequent reduction of lactam ring.

Although we successfully synthesized known intermediate rac-**3**, asymmetric Michal reaction using organocatalyst would not be possible if unsaturated ester **4** is used. We therefore changed our plans and attempted diastereoselective Michael reaction using an ester having a chiral auxiliary.

Corey's (-)-8-phenylmenthol[21] was used as chiral auxiliary and chiral ester **11** was prepared in two steps.[22] Although enantiomeric excess has not been confirmed yet, Michael addition with chiral ester **11** and subsequent cyclization reaction was proceed to give lactam **8**.

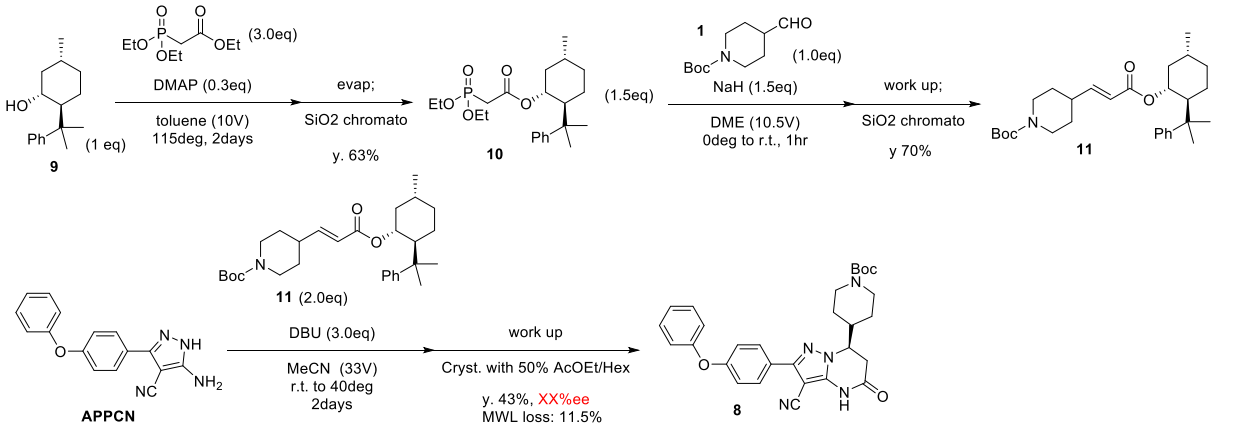


Fig. 22. An alternative approach using diastereoselective Michael reaction of chiral ester **11**.

Conclusions

In this paper, we presented the application of CDI-CASP tools for synthesis planning of complex molecules. Three search functions of CDI-CASP and their features were introduced. Synthesis of *S*-Zanubrutinib was adopted as a case study. We demonstrated that "object-oriented planning" system provides an expert-machine interaction platform equipped with functions for diverse tasks in design of organic synthetic reactions. The report highlights the successful experimental verification of the hypothesis for new chemical transformations generated by CDI-CASP tools.

Acknowledgements

This project was enabled by funding from European Regional Development Fund (ERDF) for Innovation Centre in Digital Molecular Technologies. This project was co-funded by Shionogi, AstraZeneca and the University of Cambridge. We acknowledge continuous support and collaboration with the Reaxys™ team at RELX Intellectual Properties SA. Copyright © 2020 Elsevier Limited except certain content provided by third parties. Reaxys is a trademark of Elsevier Limited.

References

(1) Gromski, P. S.; Granda, J. M.; Cronin, L. Universal Chemical Synthesis and Discovery with 'The Chemputer.' *Trends in Chemistry* **2020**, *2* (1), 4–12. https://doi.org/10.1016/j.trechm.2019.07.004.

(2) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, *365* (6453), eaax1566. https://doi.org/10.1126/science.aax1566.

(3) Molga, K.; Szymkuć, S.; Grzybowski, B. A. Chemist Ex Machina: Advanced Synthesis Planning by Computers. *Acc. Chem. Res.* **2021**, *54* (5), 1094–1106. https://doi.org/10.1021/acs.accounts.0c00714.

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

(5) Williams, C. M.; Dallaston, M. A.; Williams, C. M.; Dallaston, M. A. The Future of Retrosynthesis and Synthetic Planning: Algorithmic, Humanistic or the Interplay? *Aust. J. Chem.* **2021**, *74* (5), 291–326. https://doi.org/10.1071/CH20371.

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

(7) Corey, E. J.; Wipke, W. T.; Cramer, R. D. I.; Howe, W. J. Computer-Assisted Synthetic Analysis. Facile Man-Machine Communication of Chemical Structure by Interactive Computer Graphics. *J. Am. Chem. Soc.* **1972**, *94* (2), 421–430. https://doi.org/10.1021/ja00757a020.

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

(9) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51* (5), 1281–1289. https://doi.org/10.1021/acs.accounts.8b00087.

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

(11) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting Retrosynthetic Pathways Using Transformer-Based Models and a Hyper-Graph Exploration Strategy. *Chem. Sci.* 2020, *11* (12), 3316–3325. https://doi.org/10.1039/c9sc05704h.

(12) Shen, Y.; Borowski, J. E.; Hardy, M. A.; Sarpong, R.; Doyle, A. G.; Cernak, T. Automation and Computer-Assisted Planning for Chemical Synthesis. *Nat Rev Methods Primers* 2021, *1* (1), 1–23. https://doi.org/10.1038/s43586-021-00022-5.

(13) Schwaller, P.; Vaucher, A. C.; Laplaza, R.; Bunne, C.; Krause, A.; Corminboeuf, C.; Laino, T. Machine Intelligence for Chemical Reaction Space. *WIREs Computational Molecular Science* 2022, *12* (5), e1604. https://doi.org/10.1002/wcms.1604.

(14) Grzybowski, B. A.; Badowski, T.; Molga, K.; Szymkuć, S. Network Search Algorithms and Scoring Functions for Advanced-Level Computerized Synthesis Planning. *WIREs Computational Molecular Science* 2023, *13* (1), e1630. https://doi.org/10.1002/wcms.1630.

(15) Dong, J.; Zhao, M.; Liu, Y.; Su, Y.; Zeng, X. Deep Learning in Retrosynthesis Planning: Datasets, Models and Tools. *Briefings in Bioinformatics* 2022, *23* (1), bbab391. https://doi.org/10.1093/bib/bbab391.

(16) HILGER, J.; Zhang, X.; FENG, S.; Ro, S.; Huang, J. Treatment of Indolent or Aggressive B-Cell Lymphomas Using a Combination Comprising Btk Inhibitors. WO2019108795A1, June 6, 2019. https://patents.google.com/patent/WO2019108795A1/en?oq=WO2019%2f108795%2c+2019%2c+A1+Zanubrutinib (accessed 2023-02-15).

(17) Weber, J. M.; Lió, P.; Lapkin, A. A. Identification of Strategic Molecules for Future Circular Supply Chains Using Large Reaction Networks. *React. Chem. Eng.* 2019, *4* (11), 1969–1981. https://doi.org/10.1039/C9RE00213H.

(18) Speerschneider, A. C.; Yamashita, D. S.; Pitis, P. M.; Hawkins, M. J.; Liu, G.; Miskowski Daubert, T. A.; Yuan, C. C.K.; Borbo Kargbo, R.; Herr, R. J.; Romero, D.; Pacofsky, G. J. 6-Membered Aza-Heterocyclic Containing Delta-Opioid Receptor Modulating Compounds, Methods Of Using And Making The Same. WO 2017040545 A1, 2017.

(19) Guo, Y.; Yu, D.; Wang, Z. Crystalline Forms Of (S) -7- (1- (But-2-ynoyl) Piperidin-4-yl) -2- (4-Phenoxyphenyl) -4, 5, 6, 7-Tetrahydropyrazolo [1, 5-A] Pyrimidine-3-Carboxamide, Preparation, And Uses Thereof. WO 2018137681 A1, 2018.

(20) (a) Franzén, J.; Marigo, M.; Fielenbach, D.; Wabnitz, T. C.; Kjærsgaard, A.; Jørgensen, K. A. A General Organocatalyst for Direct α-Functionalization of Aldehydes: Stereoselective C-C, C-N, C-F, C-Br, and C-S Bond-Forming Reactions. Scope and Mechanistic Insights. J. Am. Chem. Soc., 2005, 127, 18296-

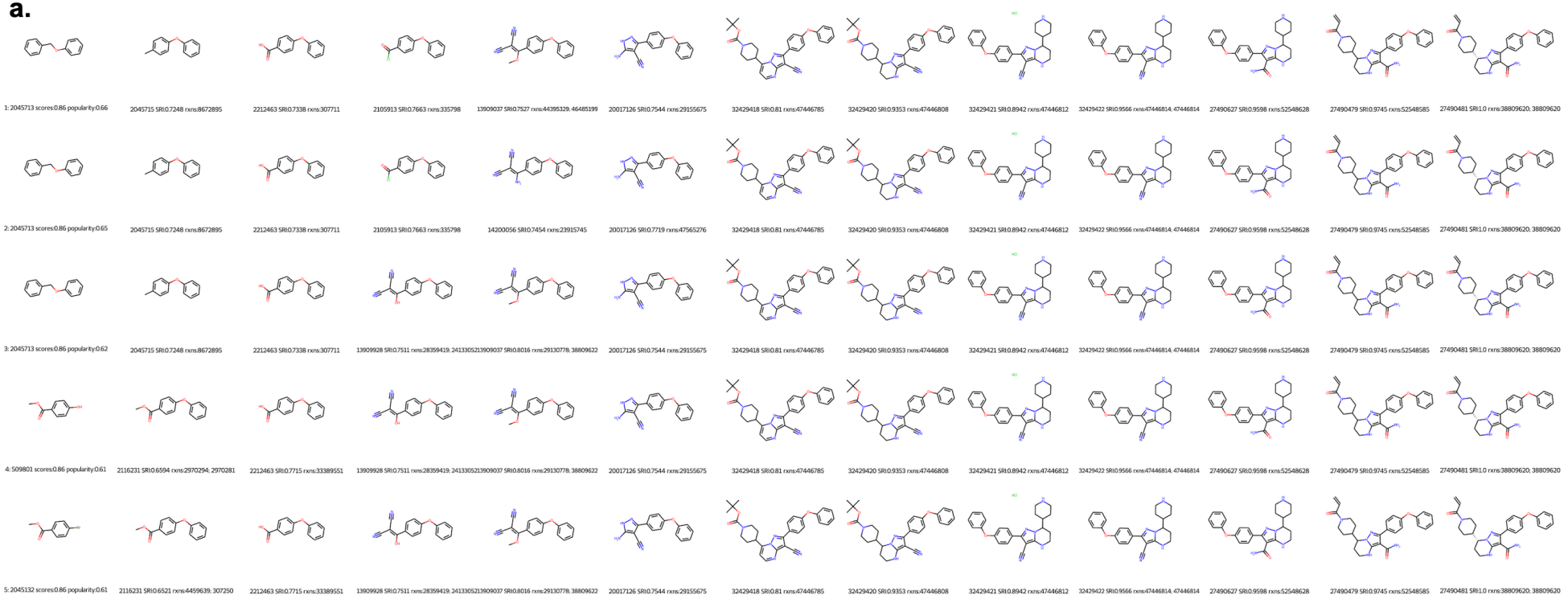
18304. (b) Babu, Y. S.; Kotian, P. L.; Kumar, V. S.; Wu, M.; Lin, T.-H. Heterocyclic Compounds As Janus Kinase Inhibitors. WO 2011031554 A2, 2011.

(21) Corey, E. J.; Ensley, H. E. Preparation of an optically active prostaglandin intermediate via asymmetric induction. J. Am. Chem. Soc., 1975, 97, 6908-6909.

(22) Zhu, J.-L.; Chen, P.-E. Huang, H.-W. Lewis acid mediated asymmetric Diels–Alder reactions of chiral 2-phosphonoacrylates. Tetrahedron Asymmetry, 2013, 24, 23-36.

Appendix

a.



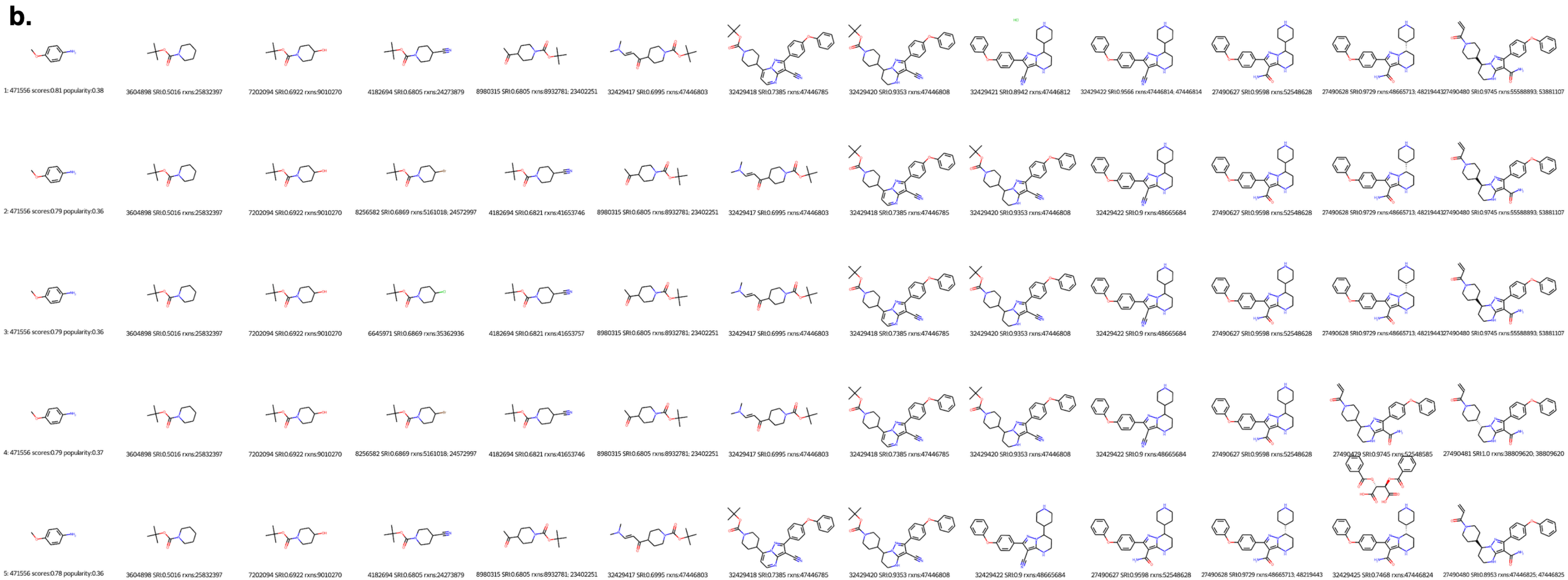


Fig. A1. Examples of 12-steps routes found by retrosynthetic search. a) top-5 routes ranked by similarity-based metric; b) top-5 routes ranked by popularity metric.

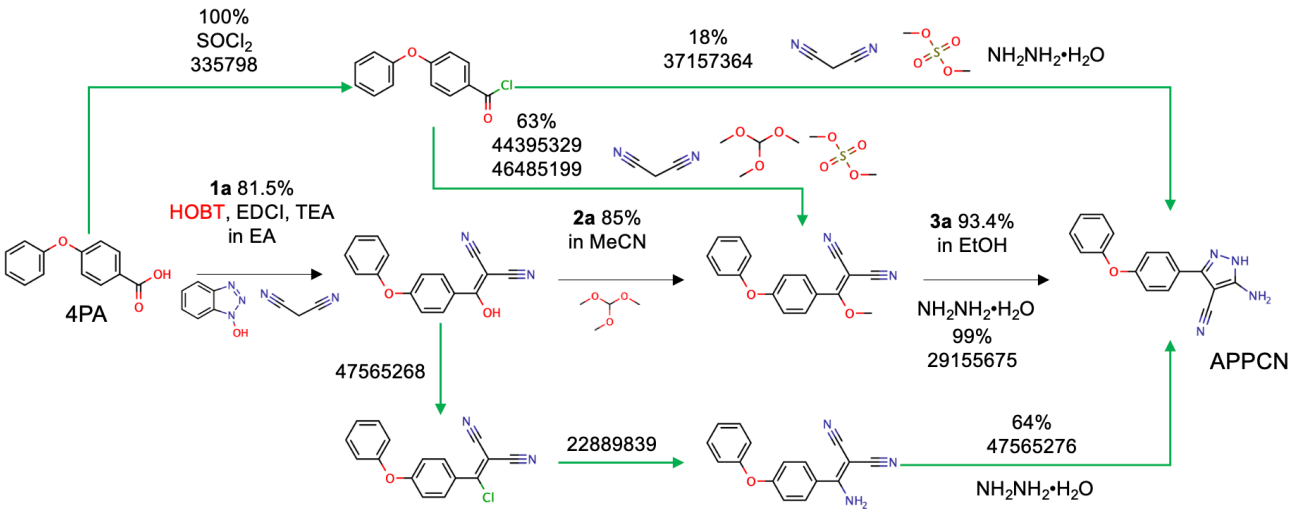


Fig. A2. Sub-routes to from 4PA to APPCN.

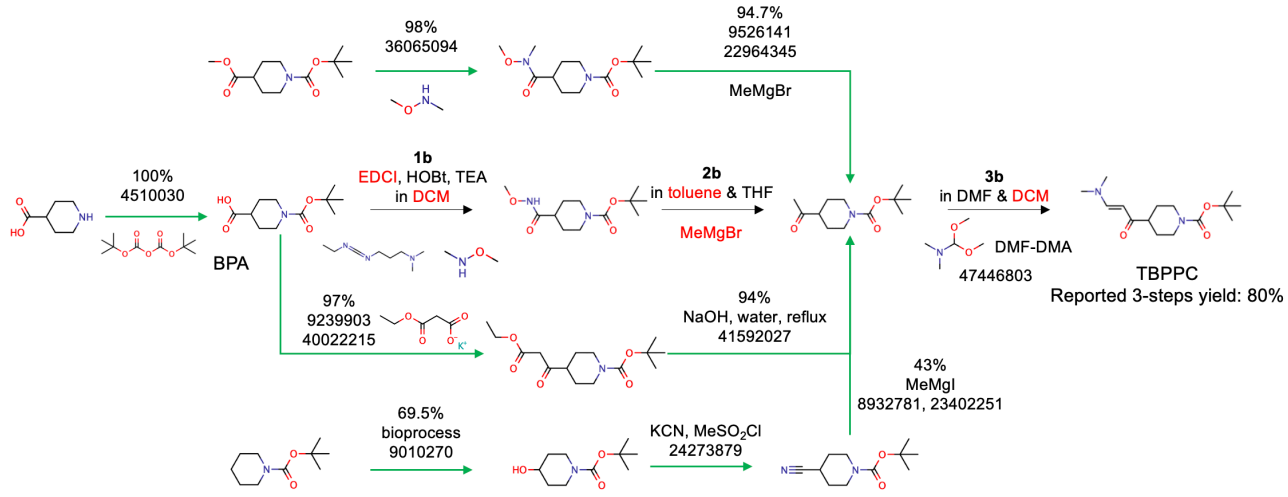


Fig. A3. Sub-routes to from BPA to TBPPC.