Multistep retrosynthesis combining a disconnection aware triple transformer loop with a route penalty score guided tree search

David Kreutter\(^a\) and Jean-Louis Reymond*\(^a\)

\(^a\) Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, Freiestrasse 3, 3012 Bern, Switzerland

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

Abstract

Computer-aided synthesis planning (CASP) aims to automatically learn organic reactivity from literature and perform retrosynthesis of unseen molecules. CASP systems must learn reactions sufficiently precisely to propose realistic disconnections while avoiding overfitting to leave room for diverse options, and explore possible routes such as to allow short synthetic sequences to emerge. Herein we report a CASP tool proposing original solutions to both challenges. First, we use a triple transformer loop (TTL) predicting starting materials (T1), reagents (T2), and products (T3) to explore diverse disconnections obtained by tagging potentially reacting atoms both systematically and using templates. Second, we integrate TTL into a multistep tree search algorithm (TTLA) prioritizing sequences using a route penalty score (RPScore) considering the number of steps, their confidence score, and the simplicity of all intermediates along the route. Our approach favours short synthetic routes to commercial starting materials, as exemplified by retrosynthetic analyses of recently approved drugs.
Introduction

Retrosynthetic analysis consists in drafting a synthetic sequence to produce a desired product from available starting materials. This analysis is one of the most useful but also difficult tasks in organic chemistry because it requires to integrate the large and complex set of rules that have emerged from millions of reactions reported in almost 200 years of organic synthesis. Computer-aided synthesis planning (CASP), initially conceived by E. J. Corey in the 1960’s, aims to harness the power of computers to automate retrosynthesis by exploiting data from experimental reactions collected in databases such as Reaxys or the open-access reaction dataset extracted from US patent office data. These databases list reactions of sets of starting materials (SM) and sets of reagents (R) to form one or several products (P).

While expert systems based on hand-written rules such as Chematica/Synthia perform quite well for synthesis planning, CASP ultimately aims to exploit artificial intelligence to automatically learn organic synthesis from reaction examples and propose synthetic routes for new molecules without human intervention. Template-based approaches extract reaction rules in form of substructure transformations and use machine learning to learn their applicability domain from the structure of P in the training data. On the other hand, transformer-based models use the linear SMILES notation of chemical reactions and learn to translate the character string of P into the character string of SM + R, or vice versa. The single-step predictions are then iterated to propose multistep retrosyntheses of target molecules from a selected set of building blocks (BB), which requires prioritizing possible routes using search algorithms such as Monte Carlo Tree Search, AND-OR trees, or a multistep graph exploration.

Any CASP system must overcome two critical challenges to propose realistic retrosyntheses. First, the system must learn the context of reactions sufficiently well to propose reactions that make sense, but without overfitting such as to propose diverse retrosynthetic
operations on previously unseen molecules. Second, the route-prioritizing algorithm must be designed to allow short sequences to emerge from the multitude of predicted possibilities. Herein we report a transformer-based retrosynthesis tool which proposes original solutions to both challenges. For single-step retrosynthesis, we use three different transformer models assembled as a triple transformer loop (TTL, **Figure 1a**). To broaden the scope of predicted disconnections on a given target molecule, the TTL explores multiple disconnections by using products with tagged reaction centers (P*) obtained by an exhaustive and template-based tagging procedure. Compared to a transformer model trained on predicting SM + R directly from P*, the TTL achieves better round-trip accuracy for single-step retrosynthesis. For multistep retrosynthesis predictions, we integrate the TTL into a multistep tree-search algorithm, here named TTLA, which selects reaction sequences using a new route penalty score (RPScore), which for a route of N steps, is the product of a step-penalty score P^N, the confidence scores of each single-step retrosynthesis (CS), and the simplicity scores of all SM along the route (**Figure 1b**). This selection scheme favours short sequences and is exemplified with the prediction of synthetic routes for recently approved drugs.
\[ \text{RPScore} = P^N \times \prod_{i=1}^{N} \left( CS_i \times \prod_{\forall \text{mol} \in SM_i} \text{Simplicity(mol)} \right) \]

**Figure 1.** Multistep retrosynthesis using TTLA. (a) Single-step retrosynthesis. At step \( i \), potentially reactive atoms in \( P_i \) are labelled systematically and using templates to produce multiple labelled \( P_i^* \). Transformer T1 is applied to each labelled \( P_i^* \) to predict \( SM_i \) (one or more starting materials), transformer T2 is applied to the top-scoring \( SM_i \rightarrow P_i \) to predict \( R_i \) (one or more reagents), and finally transformer T3 is applied to the top-scoring \( SM_i + R_i \) to produce \( P_{T3} \). The prediction is validated if \( P_{T3} = P_i \) with confidence score \( CS_i \) of T3. Each molecule in the \( SM_i \) set is then used as product \( P_{i+1} \) for the next iteration. The route branches out if \( SM_i \) contains multiple molecules. (b) TTLA sequence and route penalty scoring. All molecules in the \( SM_i \) set of each step are used in the RPScore calculation of a linear sequence. See text for details.
Methods

Dataset

The United States Patent and Trademark Office chemical reaction dataset (USPTO) of Lowe\textsuperscript{3,4} is used as a training dataset for all our models, however we only consider reactions with exactly one product (P) and between one and three starting materials (SM). The same dataset test, validation, and test split (90%/5%/5%) are used for all models.

Tagging reaction centers

Training the disconnection-aware retrosynthesis model requires a training dataset where all product SMILES have tagged atoms. To tag reacting atoms, we use the atom-mapping tool shared by Schwaller et al.\textsuperscript{28} to identify the atoms having an environmental change during the reaction. Reacting atoms are then re-labelled with the atom mapping label “1” while all other atom mapping labels are removed, as described by Byekwaso et al.\textsuperscript{29} In addition, we replace each mapped-tagged atom with its unmapped SMILES notation appended with another separated token (“!”) using RDkit.\textsuperscript{41} This modification allows to maintain an invariant SMILES token usage independent of the neighbouring hydrogen count or stereochemistry.

Single-step disconnection aware retrosynthesis (T1)

Being able to identify the reaction center of a given reaction, we apply our reaction tagging algorithm on USPTO to obtain a retrosynthesis-tagged training dataset. We remove reagents, catalysts, and solvents, which are identified as the unmapped species in atom-mapped reactions, and train the retrosynthesis model to predict the precursors given as input the tagged products. We use the Transformer architecture\textsuperscript{18} and train it using the OpenNMT\textsuperscript{42,43} library with standard previously-reported hyperparameters for this type of task.\textsuperscript{22}
**Automatic tagging of potentially reactive atoms**

We use two complementary methods to maximize the tagging possibilities while maintaining a reasonable number of predictions. First, we tag all possible combinations of neighbouring atoms, from 1 up to 3 atoms. Secondly, we tag reactive sets of atoms belonging to reactive substructures, which are themselves identified by analyzing the USPTO training set.

**Reagent Prediction (T2)**

Transformer T2 is trained from the untagged USPTO training set to identify reagents (R) from the combination of SM and P using the same hyperparameters as for T1. Note that R often include actual reagents and solvents.

**Forward Validation (T3)**

The third model of the triple-transformer loop is a forward reaction prediction model trained with untagged reactions (Molecular Transformer).\(^{22}\) We give this forward validation model the predicted SM\(_i\) (from T1) and the predicted R\(_i\) (from T2) as input separated by the “>” token. If T3 predicts the correct P\(_i\), those SM\(_i\) and R\(_i\) are stored for the tree search. The confidence score CS\(_i\) for the T3 prediction is used as confidence score for the reaction. T3 serves to filters down a large number of predictions to retain feasible reactions only.

**Route Penalty Score (RPScore)**

The RPScore is computed for each predicted linear retrosynthetic sequence of N steps leading from the final product P to starting material SM\(_N\) (Figure 1b). RPScore is the product of a route penalty P\(^N\) with 0 < P ≤ 1, the product of all confidence scores CS\(_i\) (from the T3 prediction) for each individual step and the Simplicity(mol) for all intermediates along the sequence of N steps. By default, the penalty value P is set to 0.8, but this could be adapted for every search in the configuration file of the multistep exploration. Simplicity(mol)\(^{24}\) ranges from 0 for complex to 1 for simple molecules and is derived from the molecular synthetic
complexity score (SCScore, ranging from 1 to 5) which describes molecular complexity taking synthetic accessibility into account. \(^{33}\) Here we assign a value of 1 if the molecule occurs in the BB set of commercial starting materials. In contrast to Schwaller et al., \(^{24}\) we exclude reagents \(R_i\) from the Simplicity calculation to avoid penalizing steps that use reagents with low calculated Simplicity, which is rarely a measure of their availability or ease of use.

\[
\text{RPScore} = P^N \times \prod_{i=1}^{N} \left( CS_i \times \prod_{\forall \text{ mol} \in SM_i} \text{Simplicity(mol)} \right)
\]

\[
\text{Simplicity(mol)} = 1 - \frac{\text{SCScore(mol)} - 1}{4}
\]

and \(\text{Simplicity(mol)} = 1\), if \(\text{Mol} \in \text{Commercial Database}\)

**Multistep Exploration Strategy**

We use a Tree Search Algorithm to iteratively explore retrosynthetic routes as previously reported for transformer-based retrosynthesis. \(^{24}\) Once predictions of an iteration are complete, the tree search updates and lists all possible routes, and computes the RPScore. Unsolved routes are sorted by decreasing RPScore. The top 20 unsolved routes are selected for expansion and the corresponding unexpanded and uncommercial SMs are predicted by the TTL. The resulting set of predicted single-step retrosynthesis is updated back to the tree wherever those SMs were present. The tree is updated for the next iteration. The process stops when a chosen minimum number of solved routes or a maximum number of iterations has been reached.

**Building block (BB) set**

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

**Triple transformer loop (TTL) for single-step retrosynthesis**

Initially, we use the atom-mapping transformer\textsuperscript{28} information to annotate reacting atoms in all products $P$ in the training data, which results in a training dataset containing labelled $P^*$. Our code is inspired by the recent report by Byekwaso et al.,\textsuperscript{29} however with a slightly simplified syntax for tagged atoms. Using the tagged $P^*$ data, we then train a transformer model $T1$ to predict SM from $P^*$. In contrast to Byekwaso et al. who use the tagged $P^*$ to predict SM+R,\textsuperscript{29} we train $T1$ to predict only SM from $P^*$.

To use $T1$ to predict possible SM$_i$ from a given product $P_i$ at step $i$, one must first tag potentially reacting atoms in $P_i$. We do this using two complementary methods. First, we tag all single atoms as well as pairs and triplets of adjacent atoms exhaustively in $P_i$. Second, we systematically apply templates extracted from tagged $P^*$ in the training dataset. These templates are substructures containing up to ten tagged atoms, with a peak at five atoms, which nicely complements the systematic single and adjacent atom tags (Figure 2a). In contrast to the tagging approach recently reported by Thakkar et al.\textsuperscript{30} where reacting atoms are identified using a tagging transformer trained to learn the detailed context from the tagged dataset, our tagging approach does not use any context and defines a multitude of potential disconnection sites.

To initiate the single-step retrosynthesis prediction for product $P_i$, we run $T1$ on all $P_i^*$ obtained by the systematic and template tagging procedure described above. The transformer outputs a series of possible SM$_i$, which are sorted in order of the $T1$ confidence score. For the top-$N$ SM$_i$ ($N = 1$ or more), we then apply a second transformer (T2) trained to predict R from SM→P. For each SM$_i$, T2 outputs a series of possible R$_i$, from which we retain the top-$N’$ ($N’ = 1$ or more). The TTL is completed with a forward validation\textsuperscript{31} transformer (T3) trained to
predict P from SM + R using the same training dataset used for T1 and T2. For each of the top-(N,N’) SMi + Rj predicted by T1 and T2, T3 predicts the most likely product PT3. The TTL prediction is validated if the top-1 predicted PT3 is identical to the input product Pi (Figure 1a). The T3 confidence scores CSi of the validated predictions SMi + Rj are used to select the best Rj if N’ > 1, and to calculate the route penalty score (RPScore, see below).

For single-step retrosynthesis, the TTL using only the top-1 predictions for T1 and T2 (N = N’ = 1) performs slightly better than the disconnection-aware retrosynthesis model of Thakkar et al.30 in terms of top-1 round-trip single-step accuracy. The performance increases further when considering the top-3 predictions of T2 (N = 1, N’ = 3). Similar to the observation by Thakkar et al.,30 we find that the prediction accuracy strongly decreases as a function of the number of tagged atoms (Figure 2b). Most importantly, when tested on unseen molecules, the TTL provides validated disconnections at several possible reactive sites. By contrast, the baseline transformer, trained as reported by Schwaller et al.24 to produce SM directly from P using the unannotated data for training, chooses fewer disconnection points, as exemplified here for the pro-nucleotide 1 (Figure 2c).32
Figure 2. Atom tagging and TTL. (a) Atom count distribution of extracted template-tagged substructures for automatic tagging. (b) Round-trip accuracies of the disconnection-aware (DA) TTL using the top-1 SM by T1 and the top-1 or top-3 R predicted by T2, compared to the disconnection-aware dual transformer baseline. (c) Highlighted disconnection sites of an antiviral compound using the baseline untagged retrosynthesis and forward validation models (top) and the TTL augmented by systematic tagging and template-based tagging after forward validation (bottom).
**Multistep retrosynthesis**

By integrating the single-step retrosynthesis TTL into a multistep tree search, we obtain a multistep retrosynthesis algorithm, here named TTLA. In each retrosynthesis iteration, TTLA runs the TTL exhaustively on all SM of the preceding iteration, newly defined as P, and ranks the routes to the newly predicted SM using a composite route penalty score RPScore (*Figure 1b*, see Methods for details).

When prioritizing multiple retrosynthesis options during the tree search, TTLA uses the RPScore to rank the different routes leading to the SM produced in the latest iteration of TTL, and only extends retrosynthesis on a small number (typically 20) of SM taken from the top RPSScoring routes. Because each additional step imposes a penalty (usually $P = 0.8$), lengthy routes and unproductive loops involving protection/deprotection cycles of the same functional group are rapidly falling down the RPScore priority list, which leads the algorithm to explore alternative routes, so that short synthetic sequences are eventually prioritized even if their first retrosynthetic steps were initially not top scoring.

As commonly observed with CASP tools as well as with transformer models in general, the top-scoring outputs of TTLA must be inspected to identify relevant predictions. While the RPScore is used in the tree-search, we find relevant routes by inspecting both the top-RPSScoring route and the top-CSScoring routes (CScore(route) = the product of CS, for all steps) in the TTLA output, as discussed below with examples.

TTLA is exemplified here for predicting the synthesis of two drug molecules approved in 2020, namely fostemsavir (*2, Figure 3*), a prodrug which upon phosphatase cleavage releases the antiretroviral agent temsavir as HIV entry inhibitor,\textsuperscript{34} and ozanimod (*9, Figure 4*), a sphingosine-1-phosphate receptor antagonist used as an immunomodulatory agent to treat multiple sclerosis.\textsuperscript{35} The commercial process for both drugs was recently reviewed.\textsuperscript{36} None of the synthetic steps involved in these two processes occurs in the USPTO dataset used for
training TTL, making them a good test case for TTLA. For these examples, we challenged TTLA to predict synthetic routes starting from a list of 534,058 commercially available BB.

The reported commercial process for the antiviral drug fostemsavir (2, Figure 3a) is a linear sequence involving the sequential C-acylation of pyrrolopyridine 3 with oxalyl monochloride 4a (step a) and benzyolpiperazone (5a, step b), followed by coupling of with triazole 6 (step c), N-alkylation of the pyrrole with the protected chloromethylphosphate 7a (step d), and finally deprotection of the tert-butyl ester protecting groups (step e).

When challenged with 2, TTLA proposes many possible routes from similar precursors as the commercial process but in a different order. The highest RPScoring route is a convergent sequence starting with N-acylation of oxalylpiperazone 4b with benzoylchloride (5b) on the one hand (step a’, Figure 3b), and alkylation of triazole 6 with pyrrolopyridine 3 on the other hand (step b’). The resulting intermediates are then coupled in a C-acylation (step c’). Finally, chloromethanol (9b) is combined with phosphate (8b, step d’) to form chloromethyl phosphate (7b), which alkylates the pyrrole nitrogen atom to form 2 (step e’). On the other hand, the highest CSscoring route uses a protected phosphate in the final steps similarly to the commercial route (Figure 3c). The TTLA route starts with the hydrolysis of the ethyl ester in 10 (step a’’) and condensation with benzyolpiperazone 5a (step b’’). In parallel, triazole 6 is alkylated with pyrrolopyridine 3 (step c’’). The resulting intermediates are then coupled (step d’’), and the product is N-alkylated with tert-butyl iodomethyl phosphate 7c (step f’’), itself obtained from tert-butyl phosphate 8c and diiodomethane (9c, step e’’). Deprotection of the tert-butyl esters finally gives product 2.
Figure 3. Summary of reported and TTLA predicted routes for fostemsavir 2. Bonds formed in each step are highlighted in colour. (a) commercial process. Reported reagents: a) AlCl₃, Bu₄NHSO₄, CH₂Cl₂, then KOH, then H₃PO₄; b) Ph₂POCl, NMM, NMP; c) KOH, CuI, then KOH, EtOH, LiI; d) Et₄NI, K₂CO₃, CH₃CN/H₂O; e) AcOH, H₂O. (b) Highest TTLA RPScoring route. Predicted reagents: a') HCl, NaOH, H₂O; b') K₂CO₃, DMF, CuI₂; c') HCl; d') HCl, PCl₃; e') Cs₂CO₃, DMF. (c) Highest TTLA CScoring route. Predicted reagents: a'') LiOH, THF, H₂O; b'') N-(3-Dimethylaminopropyl)-N'-ethyl-carbodiimide, Et₃N, CH₂Cl₂, HCl; c'') K₂CO₃, DMF, CuI₂; d'') MeOH, KOH; e'') N,N'-Dicyclohexylcarbodiimide, THF; f'') NaH, DMF; g'') HCl, EtOAc, dioxane.

In the second example, the drug ozanimod is synthesized commercially in a convergent sequence of 7 steps from ketone 10a (Figure 4a). After initial protection of the ketone as an acetal (step a) and reaction of the nitrile with hydroxylamine (step b), the resulting oxyamidine 11a is condensed with imidazolyl benzoate 13, obtained from the parent benzoic acid 12a (step c), to form the oxazole ring (step d). The ketone is then deprotected, condensed with
ethanolamine (14a) to the corresponding imine, which is reduced enantioselectively using a chiral ruthenium catalyst to form 9 (step e).

Many of the high-scoring routes identified with TTLA are extremely short sequences starting with a commercially available close analogs of the drug, and were removed from the list of top-scoring routes. However, TTLA also proposes routes starting from the chiral precursor aminobromoindane 10b, which avoids the enantioselective reaction used for the commercial process. For example, the third best CSoring route is a convergent synthesis starting with the Pd-catalyzed coupling of an organo-zinc obtained by transmetallation of aryl bromide 12b with dichlorooxadiazole 15 (step a’, Figure 4b). Exchange of the second chlorine in the formed intermediate with trimethylstannane then produces the aryl stannane 16 (step b’), which undergoes Stille coupling with the N-hydroxyethyleneamino-bromoindane obtained by alkylation of 10b with bromoethanol (14b) to form 9 (step c’ and d’). The 5th highest CSoring route employs the same alkylation of 10b as the first step (step a’’, Figure 4c), but then exchanges the bromo-substituent for a cyano group (step b’’) and converts it to the corresponding hydroxyamidine 11c (step c’’), which condenses with benzoid acid 12a in the last step to form 9 (step d’’).

The above syntheses proposed by TTLA involve questionable steps such as the use of the unprotected chloromethylphosphate intermediate 7b for fostemsavir and a selective mono-alkylation of the amino group of 12a, the absence of protecting groups on the resulting hydroxyethyleneamine, and the coupling chemistry of dichlorooxadiazole 15 for ozanimod. Nevertheless, all routes are short and follow realistic bond making strategies.
Figure 4. Summary of reported and TTLA predicted routes for ozanimod 9. Bonds formed in each step are highlighted in colour. (a) commercial process. Reported reagents: a) HC(O\text{Me})\textsubscript{3}, p-TsOH, PhCH\textsubscript{3}; b) NH\textsubscript{2}OH.HCl, Et\textsubscript{3}N; c) carbonyl diimidazole; d) NaOH; e) i) p-TsOH, acetone, ii) NH\textsubscript{2}CH\textsubscript{2}CH\textsubscript{2}OH, p-TsOH, PhCH\textsubscript{3}, iii) Chiral Ru-complex, Et\textsubscript{3}N/HCO\textsubscript{2}H. (b) 3\textsuperscript{rd} highest TTLA CScoring route. Predicted reagents: a') ZnCl\textsubscript{2}, Pd(PPh\textsubscript{3})\textsubscript{2}Cl\textsubscript{2}, BuLi, THF; b') XPhos Pd\textsubscript{2}(dba)\textsubscript{3}, dioxane; c') K\textsubscript{2}CO\textsubscript{3}, MeCN; d') Pd(PPh\textsubscript{3})\textsubscript{4}, toluene. (c) 5\textsuperscript{th} highest TTLA CScoring route. a'') K\textsubscript{2}CO\textsubscript{3}, MeCN; b'') N-methylpyrrolidone; c'') NaOAc, CH\textsubscript{2}Cl\textsubscript{2}, HCl, d'') 1-Ethyl-3-(3-dimethylaminopropyl) carbodiimide, HO\textsubscript{Bt}, DMF. The complete routes are shown in the supporting information.
Conclusion

In summary, our data shows that a triple transformer loop (TTL) operating on products with tagged reactive atoms achieves efficient single-step retrosynthesis predictions. TTL was integrated into a tree-exploration strategy using a route penalty scoring scheme to form the multistep retrosynthesis tool TTLA, which can predict short synthetic routes for drug molecules. Since our approach uses transformer models, it should be possible to specialize TTLA for specific reaction classes by transfer learning similar to transformer models for forward prediction.\(^{37}\) Furthermore, predicting SM from P and R from SM+P separately might be potentially adapted to reactions with more complex reagents such as enzymes\(^{38-40}\) and help expand the scope of CASP systems.

Availability of Data and Materials

The original USPTO dataset can be found at [https://doi.org/10.6084/m9.figshare.5104873.v1](https://doi.org/10.6084/m9.figshare.5104873.v1)

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work was supported financially by Novartis. Calculations were performed on UBELIX (http://www.id.unibe.ch/hpc), the HPC cluster at the University of Bern.
References


18


Supporting Information for Multistep retrosynthesis combining a disconnection aware triple transformer loop with a route penalty score guided tree search

Scheme S1. Literature reported retrosynthesis for fostemsavir. Orange-coloured compounds are commercially available. Reported reagents: a) AlCl₃, Bu₄NHSO₄, CH₂Cl₂, then KOH, then H₃PO₄; b) Ph₃POCl, NMM, NMP; c) KOH, CuI, then KOH, EtOH, LiI; d) Et₄NI, K₂CO₃, CH₃CN/H₂O; e) AcOH, H₂O.

Scheme S2. Best RPSscoring predicted retrosynthesis route for fostemsavir. Orange-coloured compounds are commercially available. Forward prediction confidence scores are shown under retrosynthesis arrows. Predicted reaction conditions: a’) HCl, NaOH, H₂O; b’) K₂CO₃, DMF, CuI; c’) HCl; d’) HCl, PCl₃; e’) Cs₂CO₃, DMF.
**Scheme S3.** Best overall confidence score predicted retrosynthesis route for fostemsavir. Orange-coloured compounds are commercially available. Blue-coloured compounds were present in the training set but led to a different retrosynthesis. Forward prediction confidence scores are shown under retrosynthesis arrows. Predicted reagents: a”) LiOH, THF, H2O; b”) N-(3-Dimethylaminopropyl)-N’-ethyl-carbodiimide, Et3N, CH2Cl2, HCl; c”) K2CO3, DMF, CuI2; d”) MeOH, KOH; e”) N,N’-Dicyclohexylcarbodiimide, THF; f”) NaH, DMF; g”) HCl, EtOAc, dioxane.

**Scheme S4.** Set of commercially available precursors of all solved routes for fostemsavir.
Scheme S5. Literature reported retrosynthesis for ozanimod. Orange-coloured compounds are commercially available. Reported reagents: a) HC(OMe)₃, p-TsOH, PhCH₃; b) NH₂OH.HCl, Et₃N; c) carbonyl diimidazole; d) NaOH; e) i) p-TsOH, acetone, ii) NH₂CH₂CH₂OH, p-TsOH, PhCH₃, iii) Chiral Ru-complex, Et₃N/HCO₂H.

Scheme S6. 3rd best confidence score predicted retrosynthesis route for ozanimod. Orange-coloured compounds are commercially available. Forward prediction confidence scores are shown under retrosynthesis arrows. Predicted reagents: a’) ZnCl₂, Pd(PPh₃)₂Cl₂, BuLi, THF; b’) XPhos Pd₂dba₃, dioxane; c’) K₂CO₃, MeCN; d’) Pd(PPh₃)₄, toluene.
Scheme S7. 5th best confidence score predicted retrosynthesis route for ozanimod. Orange-coloured compounds are commercially available. Forward prediction confidence scores are shown under retrosynthesis arrows. Predicted reagents: a”) K$_2$CO$_3$, MeCN; b”) N-methylpyrrolidone; c”) NaOAc, CH$_2$Cl$_2$, HCl, d”) 1-Ethyl-3-(3-dimethylaminopropyl) carbodiimide, HOBT, DMF.

Scheme S8. Set of commercially available precursors of all solved routes for ozanimod.