Finding relevant retrosynthetic disconnections for stereocontrolled reactions

Olaf Wiest,[1] Christoph Bauer,[2] Paul Helquist,[1] Per-Ola Norrby,[2] Samuel Genheden[3,*]

[1] Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana 46556, United States

[2]Data Science and Modelling, Pharmaceutical Sciences, R&D, AstraZeneca, Gothenburg, Pepparedsleden 1, SE-431 83 Mölndal, Sweden,

[3] Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Pepparedsleden 1, SE-431 83 Mölndal, Sweden

* Correspondence to: samuel.genheden@astrazeneca.com

Abstract:

Machine learning driven Computer Aided Synthesis Planning (CASP) tools have become important tools for idea generation in the design of complex molecule synthesis but do not adequately address stereochemical features of the target compounds. A novel approach to automated extraction of templates used in CASP that includes stereochemical information included in the USPTO and an internal AstraZeneca database containing reactions from Reaxys, Pistachio, and AstraZeneca electronic lab notebooks is implemented in the freely available AiZynthFinder software. 367 templates covering reagent- and substrate controlled as well as stereospecific reactions were extracted from the USPTO while 20,724 templates were from the AstraZeneca database. The performance of these templates in multi-step CASP are evaluated for 936 targets from the ChEMBL database and an in-house selection of 791 AZ compounds.

The potential and limitations are discussed for four case studies from the ChEMBL and examples of FDA-approved drugs.

Introduction

For the past few decades, enantioselective reactions have been at the forefront of the development of new methods in synthetic organic chemistry. A critical driving force has been the biomedical sector where the majority of the world's top-selling pharmaceuticals in recent years are chiral compounds, most of which are marketed as pure enantiomers.[1] Drug development has transitioned away from the "flatland" of therapeutic agents that were largely devoid of stereochemical features to modern drug discovery and manufacturing where specific enantiomers and diastereomers have to be synthesized in high stereochemical purity.[2, 3] It is for example well established that opposite enantiomers of chiral drugs may have greatly different biological activities whereby one enantiomer has desired therapeutic properties, but the other may have lower activity or even undesired toxic effects. The importance of enantioselective synthesis in the discovery and development of therapeutic agents[4] was recognized, for example, by the award of the 2001 Nobel Prize in Chemistry to Knowles, Noyori, and Sharpless who developed some the earliest and most important enantioselective methods while working in a combination of industrial and academic settings.[5]

In parallel with the growth of enantioselective methods has been the development of computer-aided synthesis planning (CASP) as a promising toolbox to assist chemists in the task of designing efficient and cost-effective synthetic routes for complex molecules. The first important steps in this direction were taken in the 1960s and 1970s with the efforts of Corey (LHASA),[6] Wipke (SECS),[7] Hendrickson (SYNGEN),[8] and Gelernter (SYNCHEM).[9] These programs were based upon algorithms for retrosynthetic analysis that were hand-coded by

human experts, interfaced with early forms of databases of organic reactions and the use of mathematical graph theory to treat chemical compounds as molecular graphs with atoms as joining points or nodes and bonds as the edges connecting them. These programs continued to be developed over the years. Some of them became commercially available,[10, 11] and have been used in industrial process development. Retrosynthetic elements are now features of widely available resources such as SciFinder[12] and Reaxys.[13]

More recently, CASP has experienced a paradigm shift with the incorporation of machine learning (ML) and deep neural networks.[14-18] ML models have proven to be invaluable in analyzing vast amounts of chemical data, extracting meaningful patterns, and generating predictive models. However, ML and data-driven CASP methods still face challenges with some classes of reactions. Despite the past decades of development, most of the computational retrosynthesis tools do not adequately address stereochemical features of compounds and the stereochemical outcome of reactions that are used to synthesize them. Stereochemistry has received only limited attention in recent studies in the CASP community. The reaction dataset derived from records of the US patent office (USPTO)[19] is often divided into a set where all stereochemical information is removed, and another set which contains stereoselective reactions. One-step retrosynthesis models, like the Molecular Transformer[16] or the Augmented Transformer,[20] often perform worse on the set containing stereoselective reactions, although it has been shown to be effective in predicting stereoselectivity in selected examples. Pesciullesi et al. used transfer learning to predict regio- and stereoselectivity in carbohydrate reactions using a transformer model.[21] For template-based models, stereoselective reactions could in principle be treated if the transformation can be encoded in a SMARTS pattern and this template can be applied.[22] The RDChiral package have enabled the latter criteria and increased the usefulness of data-driven extraction of templates. Although some successful examples of multistep retrosynthesis for chiral compounds have been demonstrated with the rule-based

Chematica[23] (commercialized as Synthia) and ASKCOS tools,[24] little general evaluation of the performance of the models has been carried out.[25] The sparsity of stereoselective transformations in datasets is also a potential issue, especially for template-based methods that often have issues in predictions for uncommon reaction classes.

As a result, most ML and data-driven CASP methods do not sufficiently consider the stereochemical information encoded in a molecule, limiting their use in synthesis planning of complex molecules such as drugs or natural products. Although some progress in addressing these shortcomings has been made,[26] our current industrial/academic collaborative team is aimed at continuing to fill this void. This work provides two important contributions towards better modelling of stereochemistry in retrosynthesis tools: (i) we have designed careful selection criteria for most common reactions that result in changes in stereochemistry, and (ii), we have trained a template-based retrosynthesis model for the selected stereoselective reactions and show its ability to suggest appropriate disconnections in a route prediction exercise within the AiZynthFinder[27, 28] workflow. These proposed routes can then, for example, be coupled with the Q2MM/CatVS method for the accurate prediction of the ratio of stereoisomers produced by asymmetric catalysis,[29-31] a topic beyond the scope of the present work. The long-term goal of this work is to develop a widely available toolbox for the generation of ideas for the synthesis of stereochemically complex molecules.

The large number of reactions providing stereochemically defined centers can be classified into a small number of different categories.[32] For the purposes of this work, we consider the following types:

- Stereospecific reactions: The reaction center is stereogenic in the reactant(s), as well as in the product. The stereochemical outcome depends on the enantiomeric purity in the starting material. Example: S_N2 reactions, where the stereochemistry at the reaction center is inverted.

- Stereoselective reactions: A new stereogenic center is formed in the product, with one stereoisomer in excess. Example: ketone reduction. Stereoselective reactions can be further divided into the following, possibly overlapping, classes:
    - Substrate controlled reactions: the new stereogenic center is influenced by other stereochemical information already present in the reactant(s).
    - Reagent controlled reactions: the stereochemistry of the product is influenced by reaction components other than the reactant(s). This class includes both reagent and catalyst controlled reactions.
    - Desymmetrizations: the new stereogenic center of the product is not at the same position as the reaction center. This class is not considered in the current work because we focus exclusively on reactions where stereochemical information is generated at the reaction center.

Methods

Reaction center identification. An essential part of the data processing is the identification of the atoms that form the reaction center. We therefore outline our novel algorithm to extract these atoms. From an atom-mapped reaction SMILES, we create an RDKit reaction object.[33] This object has some functionality to extract reactant atoms that form the reaction center. The functionality is based on finding reactant atoms where there is a change in atomic number, if there is a change in number of bonds to this reaction center, if it is bonded to an un-mapped atom, if the atom-mapping number of bonded atoms changes, or if any of the bond types change. We prune this list of reactant atoms using the following logic: for each reactant atom, we find the corresponding product atoms and we generate a list of the atom-mapping numbers for the neighboring atoms. If all the atom-mapping numbers agree between the reactant atom and the

product atom and the number of explicit hydrogens bonded to the two atoms is identical, we remove the reactant atom from the reaction center.

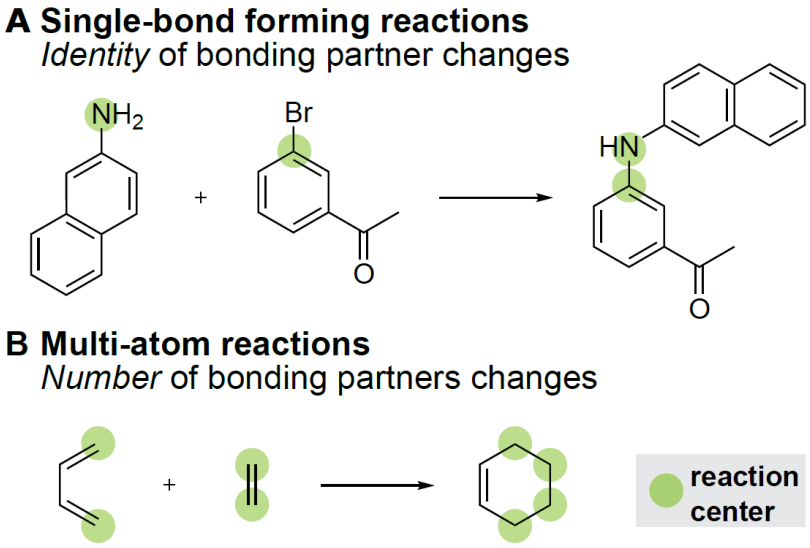

Figure 1: Determination of reaction centers

Data preparation. The extraction of training data is implemented as a pipeline in the AiZynthTrain package and is available free of charge on Github.[27] The algorithm for extracting templates for the modelling is summarized in Figure 2. We start from a clean set of atom-mapped reactions, which has been filtered according to the rules detailed previously. These reactions are then processed by a pipeline that serves to

1. remove all reaction SMILES lacking the @-character, marking a stereocenter anywhere in the reaction SMILES;
2. extract and flag any changes in stereochemical assignment between reactants and product;
3. flag if any of the reagent SMILES contain a stereocenter by identifying @-characters;

4. flag if there is a potential stereocenter in any of the reactants that are not marked in the SMILES string; this is based on RDKit routines to identify stereogenic centers;
5. flag if any of the reactants have a stereogenic center outside the reaction center;
6. flag if the product is a meso-compound based on the RDKit routines.

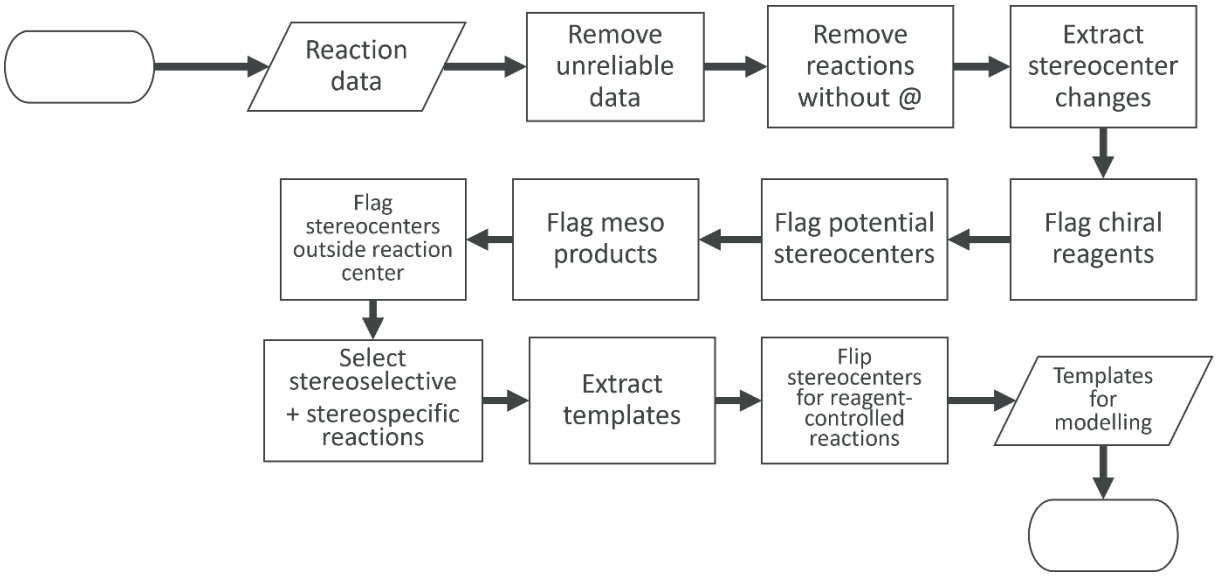

Figure 2. Flowchart summarizing the extraction of the templates from the reaction data. Oval boxes indicate start and end of the workflow.

From these calculations, we then identify three categories of stereoselective reactions on the reaction center(s) as outlined in Table 1. We only keep reactions that fall into any of these categories for the template extraction and one-step retrosynthesis modelling. The template extraction was then performed identically to the general one-step and RingBreaker models as detailed previously.[27] For the reagent-controlled reactions, we add additional templates from the extraction templates by flipping stereocenters in products with only one stereocenter, i.e. replacing @ with @@ and vice versa in the reaction template. For the modelling, we only keep templates that are supported by at least three reactions in the databases.

Table 1 Categories of stereoselective reactions treated in our model and the criteria used to identify them

Reaction category	Criteria
Reagent-controlled stereoselective	• A new stereocenter was created in the reaction • The reactants should not have any potential stereocenters not marked in the SMILES string • There should not be any stereocenters in the reactants outside the reaction center • The reagent should be chiral • The product should not be a meso-compound
Substrate-controlled stereoselective	• A new stereocenter was created in the reaction • The reactants should not have any potential stereocenters not marked in the SMILES string • There should at least one chiral atom in the reactants outside the reaction center • The reagent should not be chiral • The product should not be a meso-compound
Stereospecific	• A new stereocenter was not created in the reaction • A stereocenter was not destroyed in the reaction • The product should not be a meso-compound

One-step retrosynthesis model training. We train two retrosynthesis models for stereoselective disconnections: one based on the reactions extracted from the US Patent Office (USPTO),[19] and one based on our internal AstraZeneca reaction database, containing reactions from Reaxys,[34] Pistachio,[35] and AstraZeneca electronic lab notebooks (ELNs). The

retrosynthesis models were trained as previously detailed with the exception that the product atoms were featurized with an ECFP4 fingerprint containing chirality information.[36]

Model evaluation. The retrosynthesis model is used both in single-step and multi-step settings. For single-step evaluation, we constructed a set of 13,051 reactions from the test set of the AZ model by removing reactions also in the training and validation sets for the corresponding general retrosynthesis model previously detailed. Thus this set contains reactions from Reaxys, Pistachio, and AstraZeneca electronic lab notebooks and have not been featured in the training of any retrosynthesis model. For each of these reactions, we then extracted top-50 predictions from a one-step retrosynthesis model. We then computed *top-n* accuracies, i.e. the ability to find the recorded reactant set among the predictions. We also computed if any of the top-50 predictions changed the stereochemistry during the reactions; for the model trained only on stereocontrolled reactions, this is guaranteed if a top-50 template is applicable to the query compound, but for the general model it is not. Finally, we also record how many of the top-50 predicted templates could not be applied to the product, and therefore could not produce reactants.

In addition to the template-based models trained herein or in a previous publication, we also evaluated the performance of three contemporary one-step models: a template-free model, Chemformer,[37] a graph-based method LocalRetro,[38] and a template-based model trained for zero-shot learning, MHNReact.[39] For Chemformer, we downloaded the model weights trained on USPTO data, whereas for LocalRetro and MHNReact, we retrained those models on USPTO-50 data as explained on their Github pages.

Multistep route planning. We selected compounds from previous retrosynthetic analyses to evaluate the stereo model in multistep settings. We selected targets from the ChEMBL

database[40] and an internal AZ dataset (AZ designs). We selected compounds for which route predictions utilizing only the general retrosynthesis model failed to find any routes leading to commercial starting materials and where at least one of the starting materials (i.e. leaf compound of a synthetic route) had a stereogenic center. We then subjected these targets to multistep retrosynthesis analysis using the AiZynthFinder package.[28] The general retrosynthesis model and the new stereo model, both trained on AZ reaction data were put next to each other in the tree search. At each iteration, the top-50 suggestions from both the general and stereo models were added to the search tree, without altering the priors as given by the neural networks, hence at each expansion 100 potentially new children nodes were added. The list of available starting materials, i.e. the stock, used was an internal AstraZeneca stock or eMolecules for the AZ designs and ChEMBL, respectively. Default values were used for all other settings.

Table 2 – Statistics for extracted reactions from the USPTO and AstraZeneca sets

	USPTO		AZ set	
Number of reactions	Count	%	Count	%
Total	3285790		34741776	
With stereocenters	562933	17.13	5527989	15.91
Where stereocenter changes	46603	1.42	1840526	5.30
With chiral reagent	30712	0.93	461171	1.33
With potential stereocenter	29645	0.90	720224	2.07
With stereocenter outside reaction center	512378	15.59	4208117	12.11
Where product is meso-compound	31530	0.96	368612	1.06
Reagent-controlled reactions	1764	0.05	114873	0.33
Substrate-controlled reactions	10853	0.33	389210	1.12

Stereospecific reactions	7943	0.24	112851	0.32
Number of unique templates				
Reagent-controlled templates	84		8266	
Substrate-controlled templates	167		10169	
Stereospecific templates	116		2289	

Results

Dataset statistics. Table 2 shows statics on the two reaction datasets (USPTO and AZ sets) that we have analyzed. In both datasets, ~16-17% of the reactions have a stereocenter anywhere in the reaction SMILES, but the percentage of reactions where the stereochemistry changes during the reaction is much larger in the AZ set, ~5% compared to 1% in USPTO. The AZ set seems generally to be richer in reactions with stereochemistry, both the percentage of chiral reagents and potential stereocenters are enriched in this set. The most abundant type of stereochemistry is substrate-controlled, which amounts to 63% and 53% of the selected stereocontrolled reactions for the AZ set and USPTO, respectively. The USPTO set has only a low fraction of reagent-controlled reactions, only about 9% of the reactions fall into this category compared to about 19% for the AZ set. Finally, the stereospecific reactions make up 18% and 39% of the AZ and USPTO sets, respectively. We extract 20,724 unique templates from the AZ set, but only 367 from the USPTO set. Although the relative abundance among the different categories changes when extracting the templates, the order remains the same within each dataset after the template extraction.

Table 3 – Performance of template-based retrosynthesis model on test set of stereocontrolled reactions

	Exact match accuracy			Stereochemistry	Non-applicable
Model	top-1	top-5	top-50	change	templates
AZ stereo	0.43	0.90	0.93	0.98	30.5
AZ	0.04	0.07	0.08	0.21	37.1
USPTO stereo	0.01	0.02	0.02	0.96	48.9
USPTO	0.00	0.01	0.01	0.18	38.3

Single-step performance. The performance of one-step retrosynthesis models on the design test set of stereocontrolled reactions are shown in Table 3. The model based on the AZ stereo set is clearly the only model that is able to produce the ground truth reactants with a top-5 of 0.90 compared to 0.07 for the model trained on all reaction data and 0.01 to 0.02 for the USPTO models. The models trained on the AZ or USPTO sets can only suggest a disconnection leading to a stereochemical change for about 20% of the tested products, whereas the models trained on only stereocontrolled reactions can suggest those disconnections for 96-98% of the products. Of course, if more than the top-50 predictions were explored, this percentage would increase. For all models, the average number of non-applicable templates is high, with most of the top-50 ranked templates not being applicable to the query product. In Table S1 we show the corresponding performance of three other one-step models trained on USPTO data, LocalRetro, MHNreact, and Chemformer. None of those models have any predictive power when it comes to produce the ground truth, but are better than the template-based model trained on all USPTO data in suggesting disconnections leading to a stereochemical change. In fact, the LocalRetro

model approaches the performance of the template-based models trained on only stereocontrolled reactions, as it suggests those disconnections for 93% of the products.

Multi-step performance. We performed multi-step route planning for 936 targets from the ChEMBL database and an in-house selection of 791 AZ compounds. These compounds were previously used to benchmark the general retrosynthesis model,[27] but the multi-step retrosynthesis failed to provide routes that lead to purchasable starting materials when only using the general model. For these compounds, the set of leaf compounds for the top-ranked predicted route contained at least one compound with a chiral center. Table 4 shows that putting the new stereo model next to the general model results in successful predictions for 177 of the ChEMBL compounds and 242 of the AZ designs, i.e. a success rate of about 20%. The stereo model was used in slightly less than 10% of the disconnections in the search tree, and about 10% of the reactions in the top-10 ranked routes were disconnections suggested by the new stereo model. However, if we instead consider the compounds for which the prediction found a route leading to purchasable starting material, the percentage of disconnections coming from the new stereo model is greatly enriched. For the ChEMBL targets, close to a third of the disconnections in the top-10 ranked routes come from the stereo model, and for the AZ designs the proportion is close to a fifth.

Table 4 – Performance of multi-step retrosynthesis on two target sets

	ChEMBL	AZ designs
Number of targets	936	791
Number of solved targets	177	242
Usage of stereo model in search tree	9.2%	8.0%
Stereocontrolled reactions in top-10 routes	11.4%	9.0%
Stereocontrolled reactions in top-10 solved routes	27.7%	19.6%

Discussion

From the evaluation of the one-step retrosynthesis models on the diverse set of stereocontrolled reactions, it is clear that the USPTO dataset is not sufficient for training a widely applicable model for stereocontrolled reactions with high accuracy. First, we could only extract a few hundred unique reaction templates, indicating that the reaction diversity of the stereocontrolled reactions in the USPTO dataset is very low. The USPTO dataset has a reasonable fraction of reactions with stereocenters, in fact it is slightly higher than in the AZ dataset, but the fraction of those reactions that leads to a change in stereochemistry during the reaction is low. This could indicate that the information on stereocenters is simply missing from the reactions, but the fraction of reactions with a potential stereocenter in the reactant is comparable between the two datasets. Considering that USPTO is the only large reaction dataset in the public domain, it is worrisome that such an important class of reactions cannot be modelled with anything but large, diverse but proprietary datasets like Reaxys. For the general dissemination and development of better models for stereocontrolled reaction, this is far from ideal, and further confirms the need for better reaction data in the public domain.[41, 42]

All of the one-step retrosynthesis models, except the stereo model trained on AZ data, fail to reproduce the recorded reactant. For Chemformer,[37] LocalRetro,[38] and MHNReact,[39] this is expected considering that they were trained on USPTO data. However, all of these models have some predictive capability when considering reactants where the stereochemistry is different from the product, and therefore these models have some potential usefulness in an idea generation exercise. Especially LocalRetro[38] is very good and almost as good as the stereo model trained on AZ data in suggesting these kinds of disconnections, which could be an effect of the model's two-stage approach to retrosynthesis. LocalRetro first identifies a reaction center followed by predicting a suitable bond change. Considering that MHNReact[39] was trained for

zero-shot learning, it is disappointing that it does not perform better on a low-data regime like the stereocontrolled reactions.

In Figure 3, we highlight one example of a targeted product and the reactants generated with the different one-step retrosynthesis models. The AZ, AZ stereo and LocalRetro all generate the principal ground-truth reactant, although they are unable to suggest the second reactant. However, the generated reactants with these three models are probably sufficient to provide an understanding of how to synthesize the product. MHNReact on the other hand introduces a third stereocenter in the product, rather than remove one, and Chemformer introduces a Weinreb amide together with an epoxide precursor with an additional methyl group. It might be possible to find a metallated epoxide reactant that could react with the Weinreb amide to give the desired product, but the suggested chlorohydrin could not.

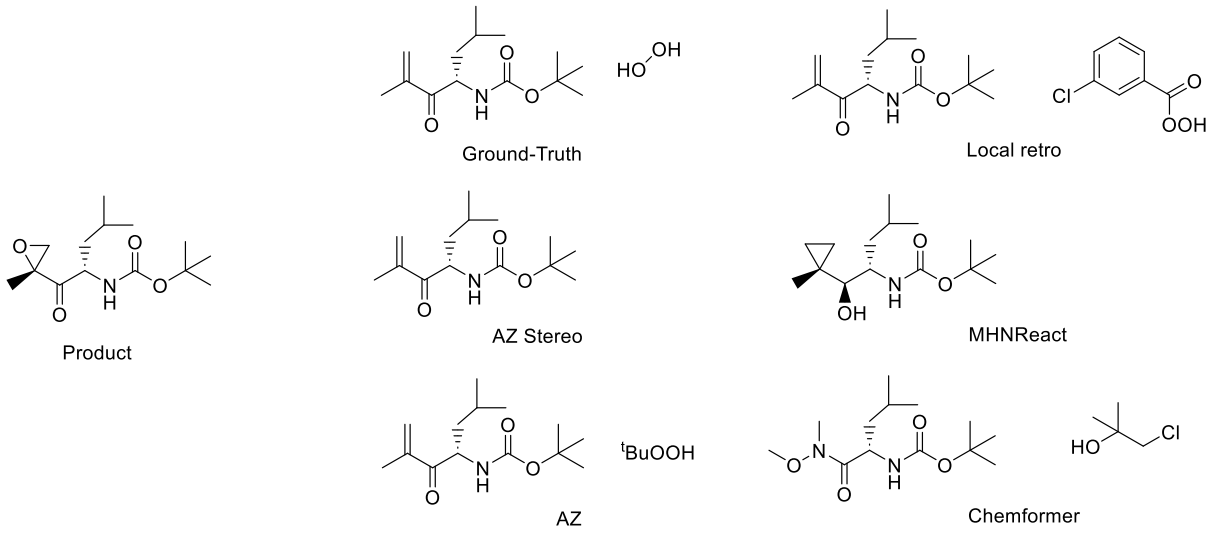

Figure 3 – Prediction of reactants for a product from the US20180298056A1 patent in the Pistacchio database using different one-step retrosynthesis models. The ground-truth is included as well as a reference. Three of the one-step models predicts the principal reactant correctly, although none of them predicts the minor reactant.

The evaluation of the product example in Figure 3 shows the limit of evaluating one-step retrosynthesis models in isolation with something like exact-match accuracy. As pointed

out previously, one-step retrosynthesis models need to be evaluated within the context of route prediction. To this end, we performed route planning for which AiZynthFinder and the general retrosynthesis model previously failed to break down starting material with at least one stereocenter. Encouragingly, incorporating the stereo model in the route planning algorithm shows an increased capability of breaking down compounds with stereocenters. For about 20% of the target compounds, the combined expansion protocol guides the planning to at least one route where all of the starting materials are in stock. However, in considering the entire dataset of 5,000 and 10,000 compounds for AZ designs and ChEMBL, respectively, the 177 ChEMBL and 242 AZ designs for which we now can identify a synthesis route, the increase in performance is rather modest. Hence, we can conclude that the combined expansion protocol is helpful in particular cases, but we are still a considerable way from being able to find synthesis routes for all molecules that may be selected as targets.

To demonstrate the performance of the model in the context of more complex synthesis and to highlight how it could be used in the synthesis of bioactive compounds, we examine the stereo-controlling steps in two example routes in Figure S1 (CHEMBL3112743) and Figure S2 (CHEMBL3559952). Figure S1 shows a seven-step synthesis for CHEMBL3112743, a compound with four stereogenic centers. One of these is an unspecified enolizable center, and one comes from a commercially available chiral amine. The other two stereocenters are derived from the stereocenter set in the first two steps of the proposed retrosynthesis (**Figure 4**). The first reaction is a ketone reduction, a classic example of reagent-controlled stereoselective reactions.[43] The second is a stereospecific inversion through an S_N2-type reaction, which could be realized in one pot by converting the free hydroxy to a sulfonate, or under Mitsunobu conditions.[44]

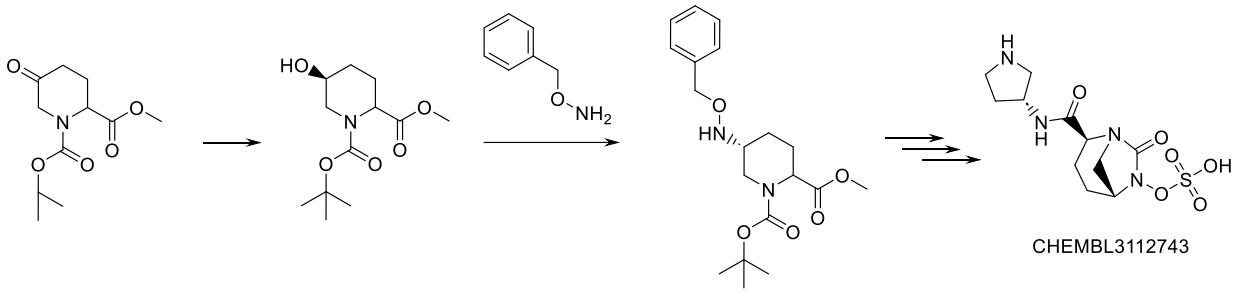

Figure 4. The first steps of the synthesis of CHEMBL3112743

CHEMBL3559952 has three stereogenic centers with the enolizable center again unspecified. In the proposed five-step route (Figure S2), the two stereogenic centers are created in the first two steps (**Figure 5**). The first step can be seen as a rearrangement of the alkyne to an allene, followed by addition of a carboxylic acid to the internal double bond. The reaction has precedent from rhodium-catalyzed transformation of terminal alkynes,[45] and does occur on model substrates with the desired regio- and stereo-selectivity, but will probably require protection of the α-hydroxy carboxylic acid. The second step is an epoxidation followed by an intramolecular 5-exo-trig ring closure of the free hydroxy group onto the epoxide. The epoxidation is proposed as substrate-controlled, but there are also ample opportunities to fine-tune the selectivity with well-known chiral epoxidation catalysts.[46]

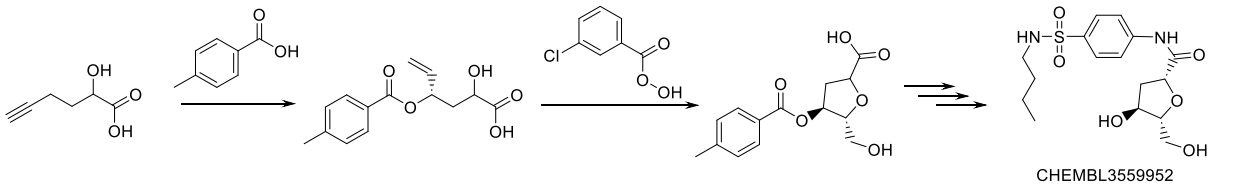

Figure 5. The first steps of the synthesis of CHEMBL3559952 .

We also examine two routes in Figure S3 and Figure S4, which highlight limitations inherent in the current approach of extracting and applying templates. Figure S3 shows a five-step synthesis for CHEMBL215018 with a single stereocontrolling disconnection based on use of a 3+2 cycloaddition suggested by the stereochemistry model (Figure 6). However, this template represents the substrate controlled category whereas the reactants in the predicted synthesis does not have any stereogenic centers. The stereogenic centers influencing the

transformation in the reaction precedents are outside the reaction center and thus are not included in the template. In the absence of a chiral controlling element, the ring formation is expected to be diastereoselective, but racemic. A solution would be the application of an enantioselective modification of the cycloaddition for which many variations are known using chiral catalysts or chiral auxiliaries[47, 48]

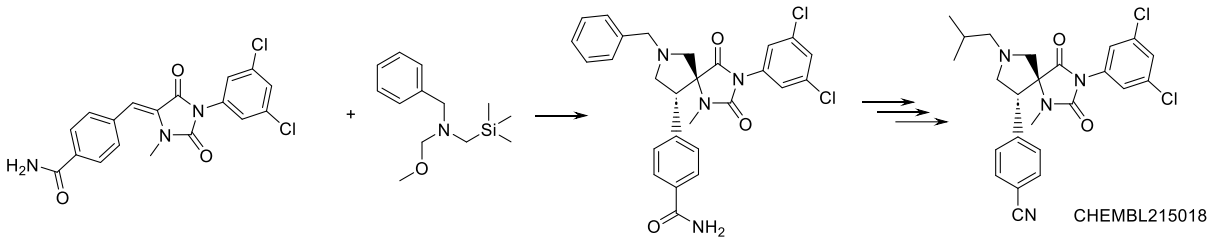

Figure 6. The stereocontrolling step in the synthesis of CHEMBL215018.

Figure S4 shows a predicted route for the drug sacubitril where we have forced the search to start with stereoselective disconnections shown in Figure 7. The first step is a methyl addition to an electron deficient double bond using copper catalysis (*i.e.*, reagent controlled stereochemistry according to our classification). In the second step, a reductive amination is proposed, in principle a good candidate for reagent-controlled stereoselectivity, but here the proposed reactant is an amide.[49] Precedents for such reactions are limited, and the templates do in fact come from much more common reductive aminations, but the limited radius of the template extraction does not allow a distinction between amine and amide reactants. A skilled chemist can still see that the synthesis could be accomplished, for example employing a chiral catalyst or a chiral auxiliary such as a phenethyl amine in the reductive amination, followed by benzylic hydrogenation and *N*-acylation of the resulting amine using succinic anhydride.[50] Thus, this type of proposal can still be useful for ideation, followed by fine-tuning to provide final routes.

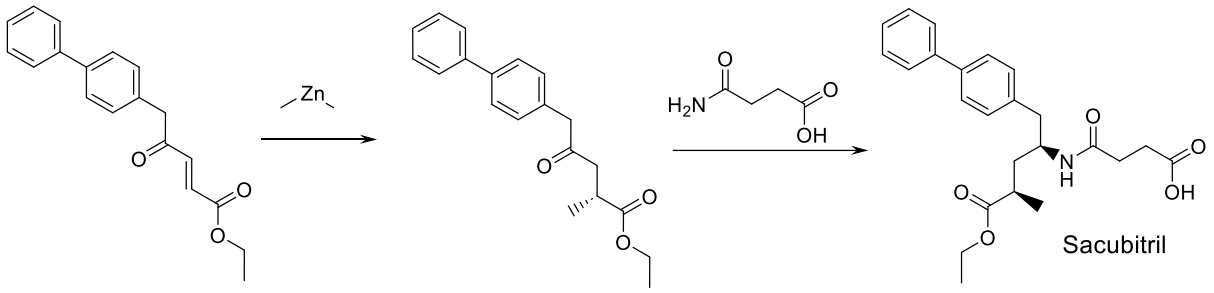

Figure 7. The stereocontrolling steps in the proposed route to sacubitril.

We have herein focused on three well-defined classes of stereocontrolled reactions for which we could design robust extraction rules. We have also taken the approach to be strict in identifying these reaction classes, and we have not attempted to correct any reaction in order to fit them into a category. In the future, it would be of interest to incorporate other types of stereocontrolled reactions.

Conclusions

We have devised a robust workflow to extract stereoselective and stereospecific reactions from historical reaction data and trained a template-based retrosynthesis model on these reactions. The one-step retrosynthesis model outperforms existing, more general, models but we have also identified room for improvements such as the approach we use to extract and apply templates. This becomes especially clear in the evaluation of the multi-step performance. We improve upon the general performance by mixing the general model with the new stereochemistry model, albeit at a modest rate. Detailed analyses of the stereocontrolling steps of a few case studies lead us to conclude that the predictions should be used primarily as an idea generation that should be carefully examined and elaborated on by chemists. The model is now implemented in the AiZynth workflow at AstraZeneca where it is used for this purpose in an industrial setting, and the USPTO-derived model, templates and workflows are available free of charge to the broader scientific community.

The results show that there is an urgent need for high-quality data for stereoselective and stereospecific reactions. The only large publicly available dataset, USPTO, contains only a small number of reactions from which only a few hundred templates can be extracted whereas proprietary datasets such as the in-house AZ dataset offer much richer stereochemical information. Thus the scarcity of data in the public domain limits the training and dissemination of a model for stereochemical reactions. This is unfortunate considering the importance of these reactions in modern drug development and the need for more robust computer-aided synthesis planning tools. The ongoing development of free, publicly available reaction databases such as the Open Reaction Database[41] provides an opportunity to address this issue early on by including stereocontrolled reactions and the appropriate information in the datasets.

In conclusion, the tools described in this work provide the framework to address a recognized weakness in CASP, the reliable identification of reactive centers and the inclusion of stereochemical information in automatically extracted templates. Even with the limited information in publicly available datasets, a significant improvement over existing methods was achieved with the goal of making CASP more useful as a hypothesis generator for the practicing organic chemist. The model can be further improved by applying the framework to stereochemical richer datasets that could include more focused parts of the reaction space or proprietary datasets without changes of the framework.

Associated Content

Additional performance metrics of three one-step retrosynthesis models and full proposed synthetic routes for case studies.

Data Availability Statement

All workflows and programs are part of the AiZynthTrain and AiZynthFinder packages which are available free of charge at the Github repository of the AstraZeneca Molecular AI group https://github.com/MolecularAI/. Models and other artifacts from the training of the USPTO model, as well as database IDs of the Pistachio and Reaxys reactions are available on Zenodo: https://zenodo.org/records/10548209

Author Information

Corresponding author: Samuel Genheden, samuel.genheden@astrazeneca.com

Authors:

Olaf Wiest: Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana 46556, United States

Christoph Bauer: Data Science and Modelling, Pharmaceutical Sciences, R&D, AstraZeneca, Gothenburg, Pepparedsleden 1, SE-431 83 Mölndal, Sweden

Paul Helquist; Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana 46556, United States

Per-Ola Norrby: Data Science and Modelling, Pharmaceutical Sciences, R&D, AstraZeneca, Gothenburg, Pepparedsleden 1, SE-431 83 Mölndal, Sweden

Samuel Genheden: Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Pepparedsleden 1, SE-431 83 Mölndal, Sweden

Author Contributions

S.G., P.-O. N. and O.W. conceptualized the work, S.G. and C.B. wrote the code and workflows, P.H., O.W. and P.-O. N. analyzed the proposed synthetic routes. All authors contributed to the design and execution of the study and manuscript writing.

Notes

The authors declare no competing financial interest.

Acknowledgments

This work was supported by the National Science Foundation (CHE–2202693) through the NSF Center for Computer Assisted Synthesis (C-CAS) and by AstraZeneca.

Dedication

This contribution is dedicated to our colleague, mentor, collaborator, and friend, Professor Björn Åkermark, on the occasion of his 90th birthday and in recognition of his ongoing accomplishments during seven decades of research.

References

1. Buntz, B., 50 of 2021's best-selling pharmaceuticals. *Drug Disc. Dev.* **2021,** *29*, www.drugdiscoverytrends.com/50-of-2021s-best-selling-pharmaceuticals/.

2. Lovering, F., Escape from Flatland 2: complexity and promiscuity. *MedChemComm* **2013,** *4*, 515-519.

3. Lovering, F.; Bikker, J.; Humblet, C., Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* **2009,** *52*, 6752-6756.

4. Yang, H.; Yu, H.; Stolarzewicz, I. A.; Tang, W., Enantioselective Transformations in the Synthesis of Therapeutic Agents. *Chem. Rev.* **2023,** *123*, 9397-9446.

5. Adam, D., Chemistry Nobel 2001. *Nature* **2001**, https://doi.org/10.1038/news011011-17

6. Pensak, D. A.; Corey, E. J., LHASA—logic and heuristics applied to synthetic analysis, ACS Symposium Series. In *Computer-Assisted Organic Synthesis*, ACS Publications: Washington DC, 1977; pp 1-31.

7. Wipke, W. T.; Ouchi, G. I.; Krishnan, S., Simulation and evaluation of chemical synthesis—SECS: An application of artificial intelligence techniques. *Artif. Int.* **1978,** *11*, 173-193.

8. Hendrickson, J. B.; Toczko, A. G., SYNGEN program for synthesis design: basic computing techniques. *J. Chem. Inf. Comp. Sci.* **1989,** *29*, 137-145.

9. Gelernter, H.; Sanders, A.; Larsen, D.; Agarwal, K.; Boivie, R.; Spritzer, G.; Searleman, J., Empirical Explorations of SYNCHEM: The methods of artificial intelligence are applied to the problem of organic synthesis route discovery. *Science* **1977,** *197*, 1041-1049.

10. Grzybowski, B. A.; Szymkuć, S.; Gajewska, E. P.; Molga, K.; Dittwald, P.; Wołos, A.; Klucznik, T., Chematica: a story of computer code that started to think like a chemist. *Chem* **2018,** *4*, 390-398.

11. Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A., Computer‐assisted synthetic planning: the end of the beginning. *Angew. Chem. Intl. Ed.* **2016,** *55*, 5904-5937.

12. Services, C. A., SciFinder-n. American Chemical Society: 2021.

13. Lawson, A. J.; Swienty-Busch, J.; Géoui, T.; Evans, D., The making of reaxys—towards unobstructed access to relevant chemistry information. In *The Future of the History of Chemical Information*, ACS Publications: 2014; pp 127-148.

14. Coley, C. W.; Green, W. H.; Jensen, K. F., Machine learning in computer-aided synthesis planning. *Acc. Chem. Res.* **2018,** *51*, 1281-1289.

15. Empel, C.; Koenigs, R. M., Artificial‐Intelligence‐Driven Organic Synthesis–En Route towards Autonomous Synthesis? *Angew. Chem. Intl. Ed.* **2019,** *58*, 17114-17116.

16. Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A., Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. *ACS Cent. Sci.* **2019,** *5*, 1572-1583.

17. Mervin, L.; Genheden, S.; Engkvist, O., AI for drug design: From explicit rules to deep learning. *Art. Intel. Life Sci.* **2022,** *2*, 100041.

18. Schwaller, P.; Vaucher, A. C.; Laplaza, R.; Bunne, C.; Krause, A.; Corminboeuf, C.; Laino, T., Machine intelligence for chemical reaction space. *WIRE: Comp. Mol. Sci.* **2022,** *12*, e1604.

19. Lowe, D., Chemical reactions from US patents, 1976–Sep 2016. *https://figshare.com/articles/* **2016**.

20. Tetko, I. V.; Karpov, P.; Van Deursen, R.; Godin, G., State-of-the-art augmented NLP transformer models for direct and single-step retrosynthesis. *Nature Comm.* **2020,** *11*, 5575.

21. Pesciullesi, G.; Schwaller, P.; Laino, T.; Reymond, J.-L., Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates. *Nature Comm.* **2020,** *11*, 4874.

22. Coley, C. W.; Green, W. H.; Jensen, K. F., RDChiral: An RDKit wrapper for handling stereochemistry in retrosynthetic template extraction and application. *J. Chem. Inf. Mod.* **2019,** *59*, 2529-2537.

23. Mikulak-Klucznik, B.; Gołębiowska, P.; Bayly, A. A.; Popik, O.; Klucznik, T.; Szymkuć, S.; Gajewska, E. P.; Dittwald, P.; Staszewska-Krajewska, O.; Beker, W., Computational planning of the synthesis of complex natural products. *Nature* **2020,** *588*, 83-88.

24. Coley, C. W.; Thomas III, D. A.; Lummiss, J. A.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H., A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019,** *365*, eaax1566.

25. Hardy, M. A.; Nan, B.; Wiest, O.; Sarpong, R., Strategic elements in computer-assisted retrosynthesis: A case study of the pupukeanane natural products. *Tetrahedron* **2022,** *104*, 132584.

26. Gallarati, S.; Fabregat, R.; Laplaza, R.; Bhattacharjee, S.; Wodrich, M. D.; Corminboeuf, C., Reaction-based machine learning representations for predicting the enantioselectivity of organocatalysts. *Chem. Sci.* **2021,** *12*, 6879-6889.

27. Genheden, S.; Norrby, P.-O.; Engkvist, O., AiZynthTrain: robust, reproducible, and extensible pipelines for training synthesis prediction models. *J. Chem. Inf. Mod.* **2023,** *63*, 1841-1846.

28. Genheden, S.; Thakkar, A.; Chadimová, V.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E., AiZynthFinder: a fast, robust and flexible open-source software for retrosynthetic planning. *J. Cheminf.* **2020,** *12*, 70.

29. Rosales, A. R.; Quinn, T. R.; Wahlers, J.; Tomberg, A.; Zhang, X.; Helquist, P.; Wiest, O.; Norrby, P.-O., Application of Q2MM to predictions in stereoselective synthesis. *Chem. Comm.* **2018,** *54*, 8294-8311.

30. Hansen, E.; Rosales, A. R.; Tutkowski, B.; Norrby, P.-O.; Wiest, O., Prediction of Stereochemistry using Q2MM. *Acc. Chem. Res.* **2016,** *49*, 996-1005.

31. Maloney, M. P.; Stenfors, B. A.; Helquist, P.; Norrby, P.-O.; Wiest, O., Interplay of Computation and Experiment in Enantioselective Catalysis: Rationalization, Prediction, and — Correction? *ACS Catalysis* **2023,** *13*, 14285-14299.

32. Eliel, E. L.; Wilen, S. H., *Stereochemistry of organic compounds*. John Wiley & Sons: 1994.

33. https://github.com/rdkit Accessed 2023-07-01.

34. https://www.reaxys.com/ Accessed 2022-09-01.

35. https://www.nextmovesoftware.com/pistachio.html Accessed 2022-09-01.

36. Rogers, D.; Hahn, M., Extended-connectivity fingerprints. *J. Chem. Inf. Mod.* **2010,** *50*, 742-754.

37. Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. J., Chemformer: a pre-trained transformer for computational chemistry. *Mach Learn: Sci. Tech.* **2022,** *3*, 015022.

38. Chen, S.; Jung, Y., Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au* **2021,** *1*, 1612-1620.

39. Seidl, P.; Renz, P.; Dyubankova, N.; Neves, P.; Verhoeven, J.; Wegner, J. K.; Segler, M.; Hochreiter, S.; Klambauer, G., Improving few-and zero-shot reaction template prediction using modern hopfield networks. *J. Chem. Inf. Mod.* **2022,** *62*, 2111-2120.

40. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B., ChEMBL: a large-scale bioactivity database for drug discovery. *Nucl. Acid. Res.* **2012,** *40*, D1100-D1107.

41. Kearnes, S. M.; Maser, M. R.; Wleklinski, M.; Kast, A.; Doyle, A. G.; Dreher, S. D.; Hawkins, J. M.; Jensen, K. F.; Coley, C. W., The open reaction database. *J. Am. Chem. Soc.* **2021,** *143*, 18820-18826.

42. Mercado, R.; Kearnes, S. M.; Coley, C. W., Data sharing in chemistry: lessons learned and a case for mandating structured reaction data. *J. Chem. Inf. Mod.* **2023,** *63*, 4253-4265.

43. Itsuno, S., Enantioselective reduction of ketones. *Org. React.* **2004,** *52*, 395-576.

44. Swamy, K. K.; Kumar, N. B.; Balaraman, E.; Kumar, K. P., Mitsunobu and related reactions: advances and applications. *Chem. Rev.* **2009,** *109*, 2551-2651.

45. Gellrich, U.; Meissner, A.; Steffani, A.; Kähny, M.; Drexler, H.-J.; Heller, D.; Plattner, D. A.; Breit, B., Mechanistic investigations of the rhodium catalyzed propargylic CH activation. *J. Am. Chem. Soc.* **2014,** *136*, 1097-1104.

46. Xia, Q.-H.; Ge, H.-Q.; Ye, C.-P.; Liu, Z.-M.; Su, K.-X., Advances in homogeneous and heterogeneous catalytic asymmetric epoxidation. *Chem. Rev.* **2005,** *105*, 1603-1662.

47. Jung, M. E.; Huang, A., Use of optically active cyclic N, N-dialkyl aminals in asymmetric induction. *Org. Lett.* **2000,** *2*, 2659-2661.

48. Yildirim, O.; Grigalunas, M.; Brieger, L.; Strohmann, C.; Antonchick, A. P.; Waldmann, H., Dynamic catalytic highly enantioselective 1, 3‐dipolar cycloadditions. *Angew. Chem. Intl. Ed.* **2021,** *133*, 20165-20173.

49. Kolesnikov, P. N.; Usanov, D. L.; Muratov, K. M.; Chusov, D., Dichotomy of Atom-Economical Hydrogen-Free Reductive Amidation vs Exhaustive Reductive Amination. *Org. Lett.* **2017,** *19*, 5657-5660.

50. Irrgang, T.; Kempe, R., Transition-metal-catalyzed reductive amination employing hydrogen. *Chem. Rev.* **2020,** *120*, 9583-9674.

Supporting Information:

Finding relevant retrosynthetic disconnections for stereocontrolled reactions

Olaf Wiest,[1] Christoph Bauer,[2] Paul Helquist,[1] Per-Ola Norrby,[2] Samuel Genheden[3,*]

1 Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana 46556, United States
[2]Data Science and Modelling, Pharmaceutical Sciences, R&D, AstraZeneca, Gothenburg, Pepparedsleden 1, SE-431 83 Mölndal, Sweden,
[3] Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, Pepparedsleden 1, SE-431 83 Mölndal, Sweden
* Correspondence to: samuel.genheden@astrazeneca.com

Table of Contents

Table S1 – Performance of three one-step retrosynthesis models on test set of stereocontrolled reactions

	Exact match accuracy			
Model	top-1	top-5	top-50	Stereochemistry change
LocalRetro	0.00	0.01	0.03	0.93
MhnReact	0.00	0.00	0.00	0.28
Chemformer	0.00	0.00	0.00	0.48

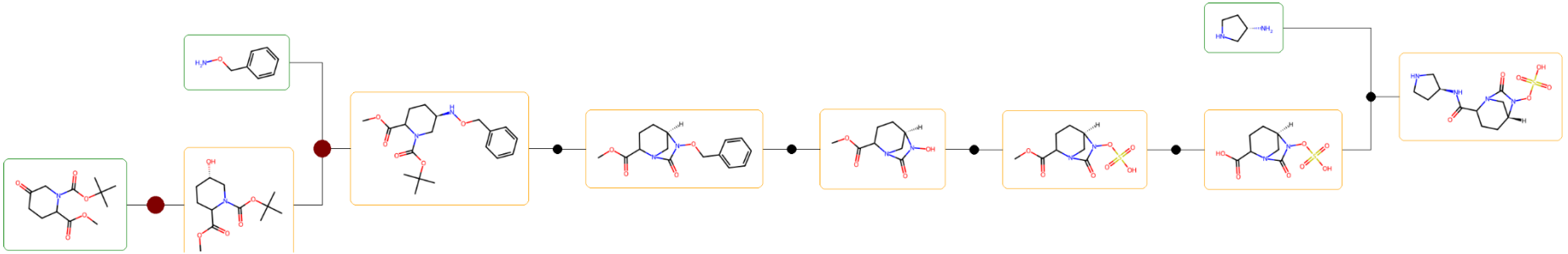

Figure S1 – Example route 1 showing a pathway for CHEMBL3112743. The disconnections are marked with a solid circle, and the disconnections suggested by the stereochemistry model are marked with a solid, red circle.

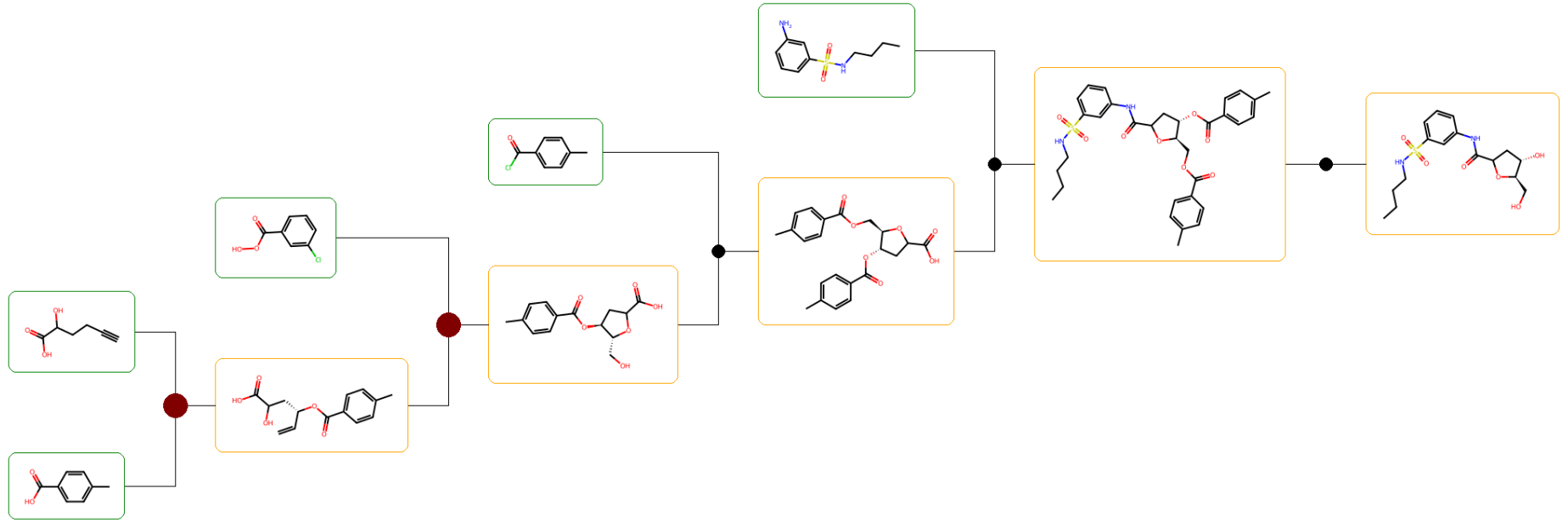

Figure S2 – Example route 2 showing a pathway for CHEMBL3559952. The disconnections are marked with a solid circle, and the disconnections suggested by the stereochemistry model are marked with a solid, red circle.

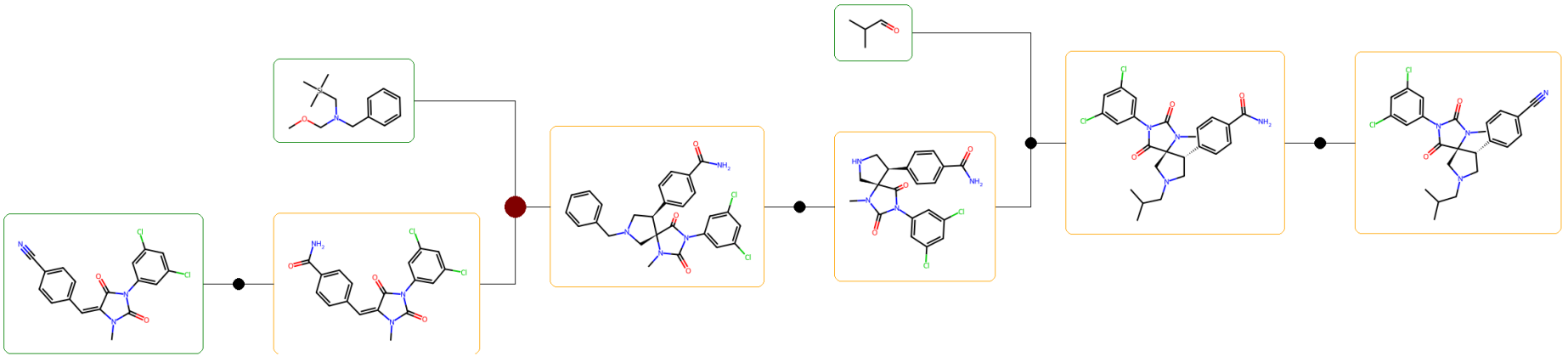

Figure S3 – Example route 3 showing a pathway for CHEMBL215018. The disconnections are marked with a solid circle, and the disconnections suggested by the stereochemistry model are marked with a solid, red circle.

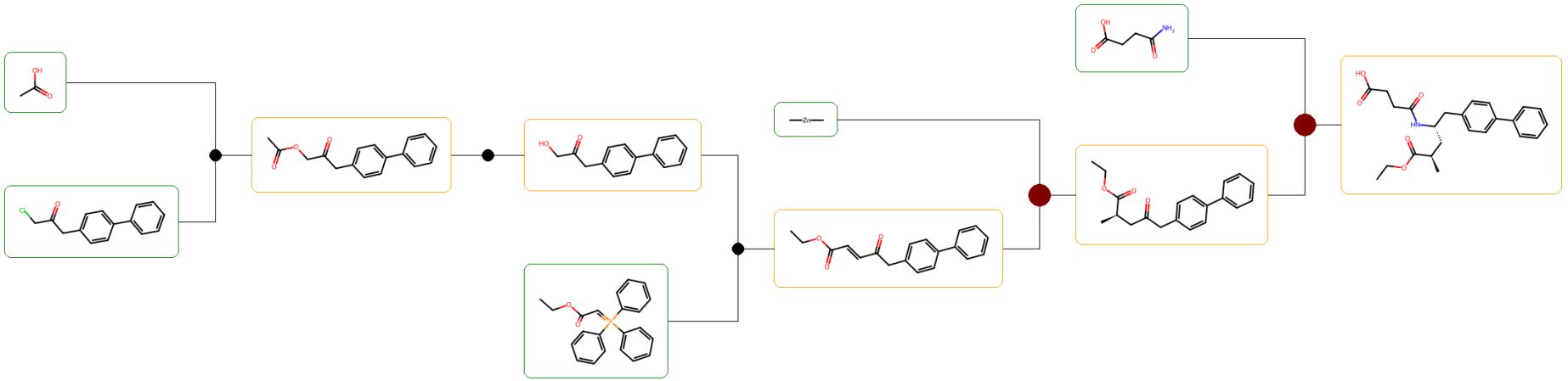

Figure S4 – Example route 4 showing a pathway to synthesize Sacubitril. The disconnections are marked with a solid circle, and the disconnections suggested by the stereochemistry model are marked with a solid, red circle.