JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

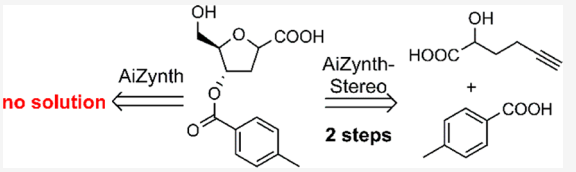
Finding Relevant Retrosynthetic Disconnections for Stereocontrolled Reactions

Olaf Wiest, Christoph Bauer, Paul Helquist, Per-Ola Norrby, and Samuel Genheden*

Cite This: J. Chem. Inf. Model. 2024, 64, 5796−5805 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

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 the stereochemical features of the target compounds. A novel approach to automated extraction of templates used in CASP that includes stereochemical information included in the US Patent and Trademark Office (USPTO) and an internal AstraZeneca database containing reactions from Reaxys, Pistachio, and AstraZeneca electronic lab notebooks is implemented in the freely available AiZynthFinder software. Three hundred sixty-seven 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 multistep CASP is evaluated for 936 targets from the ChEMBL database and an in-house selection of 791 AZ designs. The potential and limitations are discussed for four case studies from ChEMBL and examples of FDA-approved drugs.



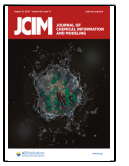
INTRODUCTION

For the past few decades, enantioselective reactions have been at the forefront of the development of new methods for 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.[6] The first important steps in this direction were taken in the 1960s and 1970s with the efforts of Corey (LHASA),[7] Wipke (SECS),[8] Hendrickson (SYNGEN),[9] and Gelernter (SYNCHEM).[10] 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,[11,12] and have been used in industrial process development. Retrosynthetic elements are now features of widely available resources such as SciFinder[13] and Reaxys.[14]

More recently, CASP has experienced a paradigm shift with the incorporation of machine learning (ML) and deep neural networks.[15−19] 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 computational retrosynthesis tools do not adequately address the 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 data set derived from records of the US Patent and Trademark Office (USPTO)[20] is often divided into a set where all stereochemical information is removed and another set that contains stereoselective reactions. One-step retrosynthesis

Received: March 4, 2024
Revised: June 27, 2024
Accepted: June 28, 2024
Published: July 12, 2024



https://doi.org/10.1021/acs.jcim.4c00370
J. Chem. Inf. Model. 2024, 64, 5796−5805

models, like the Molecular Transformer[17] or the Augmented Transformer,[21] 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.[22] 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.[23] The RDChiral package has 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[24] (commercialized as Synthia) and ASKCOS tools,[25] little general evaluation of the performance of the models has been carried out.[26] The sparsity of stereoselective transformations in data sets 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 the synthesis planning of complex molecules such as drugs or natural products. Although some progress in addressing these shortcomings has been made,[27] our current industrial/academic collaborative team is aimed at continuing to fill this void.

Over the past few years, we have created the AiZynth ecosystem, a freely available, open-source suite of retrosynthesis programs that provides algorithmic transparency that promotes reproducible results with sustainable software.[28,29] Figure 1 provides an overview of the AiZynth ecosystem. The

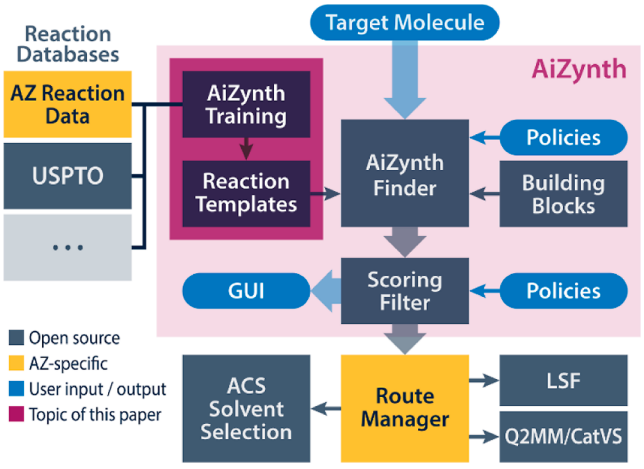


Figure 1. Overview of the AiZynth ecosystem.

two key components are AiZynthTrain,[30] which generates templates from reaction databases such as the publicly available USPTO or the proprietary AZ Reaction data. These templates are then used in the second key component, AiZynthFinder,[28] to generate possible routes to user-supplied target molecules starting from a library of available building blocks (e.g., commercially available molecules) using a Monte Carlo tree search.[31,32] User-defined policies (e.g., search depth) and a recommendation engine based on a neural network trained on template libraries[33] are used to rank possible routes generated by AiZynthFinder. It should be noted that, like all retrosynthesis programs, AiZynth provides suggestions about possible routes but not quantitative information about yields, conditions, or stereoselectivity. After further selection by scoring filters that exclude known incompatibilities and additional user-defined policies, the results can be visualized in a graphical user interface (GUI). Within AstraZeneca, they are forwarded to an in-house synthesis planning tool that incorporates quantitative prediction tools for late-stage functionalization (LSF) or the Q2MM/CatVS[34−37] workflow for the accurate prediction of the ratio of stereoisomers produced by asymmetric catalysis. The Aizynth components, including associated tools developed by other groups (e.g., the ACS solvent selection tool),[38] and the Q2MM/CatVS tools described above, are available as open-source programs free of charge on GitHub.

In the present study, we present two important contributions toward better modeling of stereochemistry in retrosynthesis tools: (i) we have designed careful selection criteria for the 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. These templates are incorporated in the AiZynth ecosystem, specifically using the AiZynthTrain workflow to generate reaction templates that consider the stereochemistry of reactions (red box in Figure 1) that are then used within the AiZynthFinder[28,30] workflow. The long-term goal of this work is to develop a freely available, transparent toolbox for the generation of ideas for the synthesis of stereochemically complex molecules, thus expanding the capabilities of AiZynth.

The large number of reactions providing stereochemically defined centers can be classified into a small number of different categories.[39] 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 of 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.

https://doi.org/10.1021/acs.jcim.4c00370
J. Chem. Inf. Model. 2024, 64, 5796−5805

METHODS

Data Sets. For model training, we used two sets of reaction data. First, we use reactions from the USPTO that have been processed with the rxnutils and AiZynthTrain packages as detailed previously.[30] In particular, the reactions have been assigned atom-mapping using the rxnmapper tool.[40] The size of the USPTO data set is 1,198,554. Second, we use reactions from the internal database, which is a combination of data from Reaxys, Pistachio, and AstraZeneca electronic lab notebooks (ELNs) that have been described previously.[30] We will refer to this database as AZ, and models derived from it will have the prefix AZ. In particular, the reactions have been assigned atom-mapping with NameRXN software whenever possible and otherwise with Biovia software. The size of this data set is 18,697,432.

For retrosynthesis experiments, we use targets from mainly two sources: 10,000 random ChEMBL compounds and an internal AstraZeneca data set of 5000 compounds that we will refer to as AZ designs. In addition to these data sets, we searched marketed drugs containing stereogenic centers and selected one, sacubitril, to highlight the developed methodology.

Reaction Center Identification. An essential part of data processing is the identification of the atoms that form the reaction center. We, therefore, outline our novel algorithm to extract these atoms (see Figure 2). From an atom-mapped reaction SMILES, we create an RDKit reaction object.[41] 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 the number of bonds to this reaction center if it is bonded to an unmapped 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.

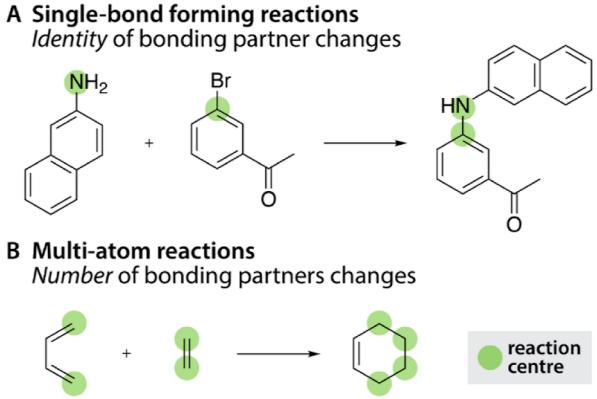


Figure 2. Determination of reaction centers.

Data Preparation. The extraction of training data is implemented as a pipeline in the AiZynthTrain package and is available free of charge on Github.[30] The algorithm for extracting templates for the modeling is summarized in Figure 3. We start from a clean set of atom-mapped reactions, which have 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 reagents SMILES contains 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 that can provide an exhaustive list of stereogenic centers, not only the ones provided by the reaction SMILES
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.

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 template extraction and one-step retrosynthesis modeling. The template extraction was then performed identically to the general one-step and RingBreaker models, as detailed previously.[30] 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

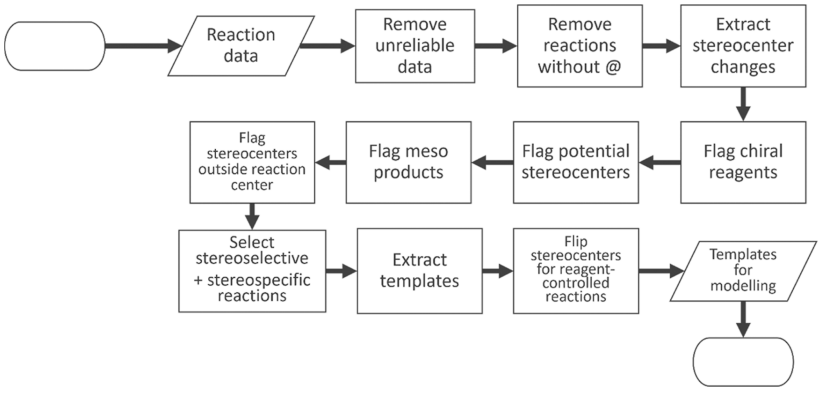


Figure 3. Flowchart summarizing the extraction of the templates from the reaction data. Oval boxes indicate the start and end of the workflow.

https://doi.org/10.1021/acs.jcim.4c00370
J. Chem. Inf. Model. 2024, 64, 5796−5805

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 be 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

template. For the modeling, we only keep templates that are supported by at least three reactions in the databases. The data sets with only the stereocontrolled reactions will be referred to as USPTO-stereo and AZ-stereo if they stem from the USPTO or AZ set, respectively.

One-Step Retrosynthesis Model Training. We train two template-based retrosynthesis models for stereoselective disconnections: one based on the reactions extracted from USPTO,[20] and one based on the AZ set described above. These models will be referred to as USPTO-stereo and AZ-stereo, respectively. The retrosynthesis models were trained as previously detailed, with the exception that the product atoms were featured with an ECFP4 fingerprint containing chirality information.[42] These models are multiclass classifiers that consist of a single-layer feed-forward network with dropout. The input size is 2048 (size of fingerprint), and the output size is equal to the unique number templates. The task of the model is to rank the templates based on the input product molecules. For further details on the models, we refer to previous studies. We opted not to update the model architecture or change any hyperparameters from what has been previously described in the literature[28,30] so that the new models can be compatible with the previously trained models and the AiZynthFinder software. The previously trained retrosynthesis models on all AZ data and on all USPTO will be referred to as the AZ and USPTO models, respectively.

Model Evaluation. The retrosynthesis model is used both in single-step and multistep settings. For single-step evaluation, we constructed an evaluation set of 13,051 reactions from the test set of the AZ-stereo model by removing reactions also in the training and validation sets for the AZ model. Thus, this set contains reactions from Reaxys, Pistachio, and AstraZeneca electronic lab notebooks and has not been featured in the training of any retrosynthesis model. Figure S1 shows the chemical space spanned by the evaluation set as well as USPTO-50K and training and validation sets for USPTO-stereo and AZ-stereo. For each of the reactions in the evaluation, 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,[43] a graph-based method, LocalRetro,[44] and a template-based model trained for zero-shot learning, MHNReact.[45] 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. We did not attempt to retrain or optimize any of the models on the stereosubset of the AZ data because the aim of this paper is not to benchmark different one-step retrosynthesis architectures.

Multistep Route Planning. For multistep route planning, we selected targets from the ChEMBL and AZ designs. We selected compounds for which route predictions utilizing the AZ 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 expansion policy used in the tree search was a concatenation of the AZ and AZ-stereo models. 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.

RESULTS

Data Set Statistics. Table 2 shows statics on the two reaction data sets (USPTO and AZ sets) that we have analyzed. In both data sets, ∼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 small 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 extracted 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 data set after the template extraction.

Single-Step Performance. The performance of one-step retrosynthesis models on the design test set of stereocontrolled

https://doi.org/10.1021/acs.jcim.4c00370
J. Chem. Inf. Model. 2024, 64, 5796−5805

Table 2. Statistics for Extracted Reactions from the USPTO and AZ Sets

	USPTO		AZ set	
number of reactions	count	%	count	%
total	3,285,790		34,741,776	
with stereocenters	562,933	17.13	5,527,989	15.91
where stereocenter changes	46,603	1.42	1,840,526	5.30
with chiral reagent	30,712	0.93	461,171	1.33
with potential stereocenter	29,645	0.90	720,224	2.07
with stereocenter outside reaction center	512,378	15.59	4,208,117	12.11
where product is meso-compound	31,530	0.96	368,612	1.06
reagent-controlled reactions	1764	0.05	114,873	0.33
substrate-controlled reactions	10,853	0.33	389,210	1.12
stereospecific reactions	7943	0.24	112,851	0.32
number of unique templates				
reagent-controlled templates	84		8266	
substrate-controlled templates	167		10,169	
stereospecific templates	116		2289	

reactions is shown in Table 3 (model convergence is shown in Figure S2). 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 suggest only 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 nonapplicable 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 LocalRetro, MHNreact, and Chemformer. None of those models have any predictive power when it comes to producing the ground truth, but it is better than the template-based model trained on all USPTO data (USPTO model) 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.

Multistep Performance. We performed multistep route planning for 936 targets from the ChEMBL database and an in-house selection of 791 AZ designs. These compounds were previously used to benchmark the AZ model,[30] but the multistep retrosynthesis failed to provide routes that lead to purchasable starting materials. However, for these compounds, the set of nonpurchasable leaf compounds for the top-ranked predicted route contained at least one compound with a chiral center, indicating that retrosynthesis with the AZ model failed to break down a chiral compound further, and hence, there is a possibility that a special stereo model could find a better synthetic route. Table 4 shows that putting the AZ-stereo model next to the AZ 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 AZ-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 AZ-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. For completeness, we add statistics for the ChEMBL targets and AZ designs when using USPTO models (USPTO and USPTO-stereo) in Table S2. We observe that the USPTO models can only find synthesis plans for 61 ChEMBL targets and 80 AZ designs. For the ChEMBL targets, we see a marginal drop in the usage of the stereo model in the search but a more pronounced drop in stereocontrolled reactions in the extracted routes. Conversely, for AZ designs, we see a significant drop in the usage of the stereo model but only a marginal decrease in stereocontrolled reactions in the extracted routes.

Table 4. Performance of Multistep 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[a]	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%

[a]The percentage of template-applications in the search tree that comes from the AZ-stereo model.

DISCUSSION

From the evaluation of the one-step retrosynthesis models on the diverse set of stereocontrolled reactions, it is clear that the USPTO data set is not sufficient for training a widely

Table 3. Performance of Template-Based Retrosynthesis Model on the Test Set of Stereocontrolled Reactions

model	exact match accuracy			stereochemistry change[a]	nonapplicable templates[b]
	top-1	top-5	top-50		
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

[a]The fraction of reactions where the model predicted a change in stereochemistry, not necessarily the one in the ground-truth data set. [b]The average number of top-50 templates that could not be applied to the product.

https://doi.org/10.1021/acs.jcim.4c00370
J. Chem. Inf. Model. 2024, 64, 5796−5805

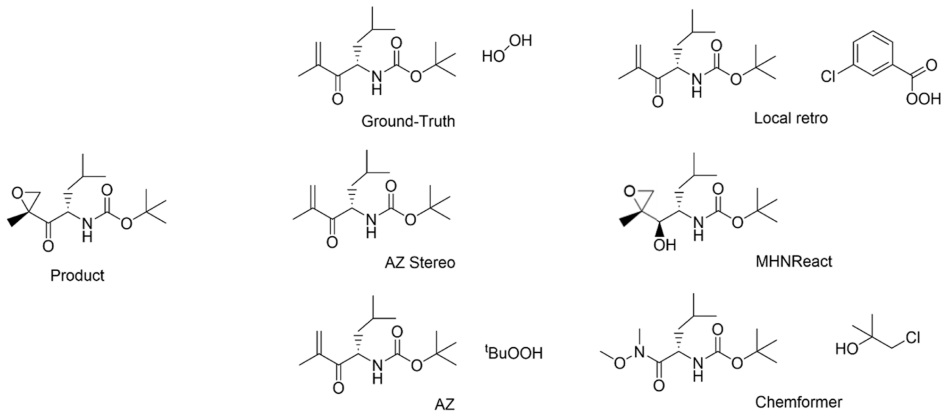


Figure 4. Prediction of reactants for a product from the US20180298056A1 patent in the Pistacchio database using different one-step retrosynthesis models. The ground-truth is included as well as a reference. Three of the one-step models predict the principal reactant correctly, although none of them predicts the minor reactant. The template for the ground-truth reaction is shown in Figure S3.

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 data set is very low. The USPTO data set has a reasonable fraction of reactions with stereocenters; in fact, it is slightly higher than in the AZ data set, but the fraction of those reactions that lead 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 data sets. Considering that USPTO is the only large reaction data set in the public domain, it is worrisome that such an important class of reactions cannot be modeled with anything but large, diverse, but proprietary data sets 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.[46,47]

All of the one-step retrosynthesis models, except the stereo model trained on AZ data, fail to reproduce the recorded reactant. For Chemformer,[43] LocalRetro,[44] and MHNReact,[45] 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[44] 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[45] was trained for zero-shot learning, it is disappointing that it does not perform better on a low-data regime like the stereocontrolled reactions. It is left to future work to retrain or finetune Chemformer, LocalRetro, and MHNReact on a data set like AZ-stereo or USPTO-stereo to see whether they have any advantages compared to the AZ-stereo model. Because we set out to devise a model that fits within our current framework (Figure 1), we did not perform a rigorous benchmark exercise, something that could be valuable considering the flaws of the models we have highlighted in this study.

In Figure 4, 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 models all generate the principal ground-truth reactant, although they are unable to suggest the second reactant. However, the reactants generated 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 metalated epoxide reactant that could react with the Weinreb amide to give the desired product, but the suggested chlorohydrin could not.

The evaluation of the product example in Figure 4 shows the limitation of evaluating one-step retrosynthesis models in isolation with something like exact-match accuracy. As pointed out previously, one-step retrosynthesis models must 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 materials 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 planning to at least one route where all of the starting materials are in stock. However, in considering the entire data set of 5000 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 a more complex synthesis and to highlight how it could be used in the synthesis of bioactive compounds, we examine the stereocontrolling steps in two example routes in Figure S4 (CHEMBL3112743) and Figure S5 (CHEMBL3559952). Figure S4 shows a seven-step synthesis of CHEMBL3112743, a compound with four stereogenic centers. One of these is an unspecified enolizable center, and the other comes from a commercially available chiral amine. The other two stereo-

https://doi.org/10.1021/acs.jcim.4c00370
J. Chem. Inf. Model. 2024, 64, 5796−5805

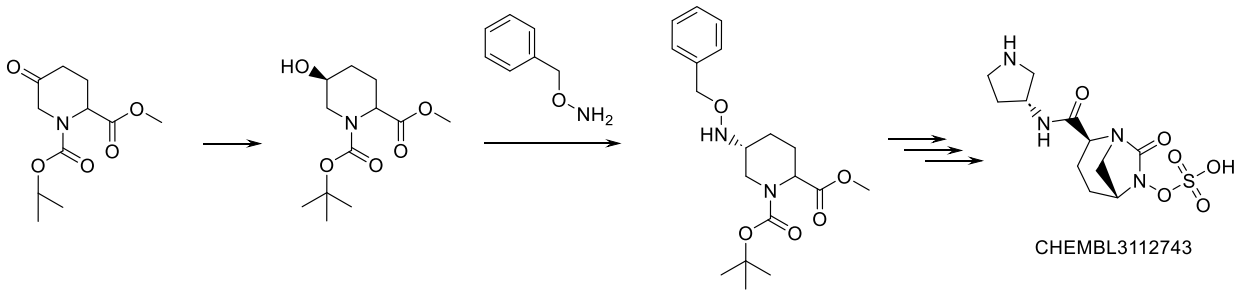


Figure 5. First steps of the synthesis of CHEMBL3112743.

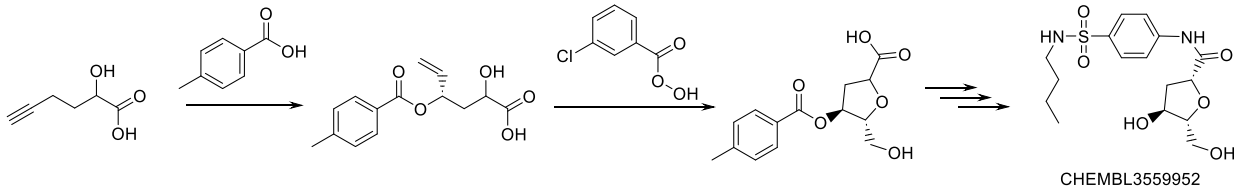


Figure 6. First steps of the synthesis of CHEMBL3559952.

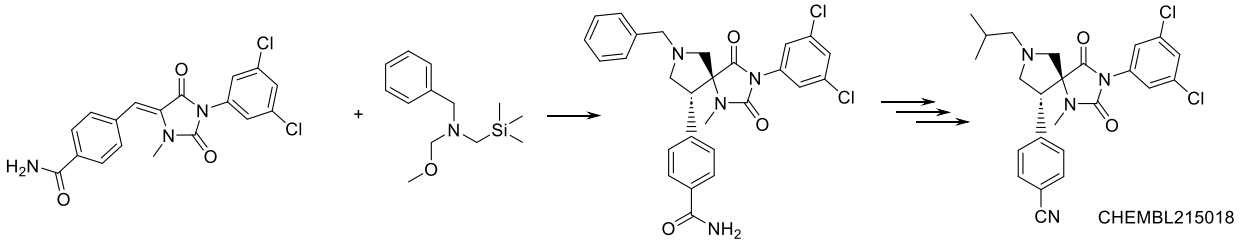


Figure 7. Stereocontrolling step in the synthesis of CHEMBL215018.

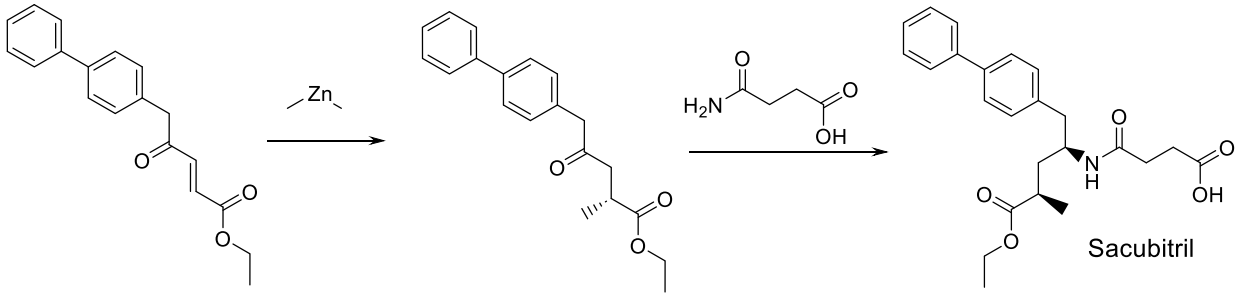


Figure 8. Stereocontrolling steps in the proposed route to sacubitril.

centers are derived from the stereocenter set in the first two steps of the proposed retrosynthesis (Figure 5). The first reaction is a ketone reduction, a classic example of reagent-controlled stereoselective reactions.[48] 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.[49]

CHEMBL3559952 has three stereogenic centers, with the enolizable center again unspecified. In the proposed five-step route (Figure S5), the two stereogenic centers are created in the first two steps (Figure 6). The first step can be seen as a rearrangement of the alkyne to an allene, followed by the addition of a carboxylic acid to the internal double bond. The reaction has precedent from a rhodium-catalyzed transformation of terminal alkynes,[50] and does occur on model substrates with the desired regio- and stereoselectivity but will probably require protection of the α-hydroxy carboxylic acid. The second step is 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.[51]

We also examine two routes in Figures S6 and S7, which highlight limitations inherent in the current approach of extracting and applying templates. Figure S6 shows a five-step synthesis for CHEMBL215018 with a single stereocontrolling disconnection based on the use of a 3 + 2 cycloaddition suggested by the stereochemistry model (Figure 7). However, this template represents the substrate-controlled category, whereas the reactants in the predicted synthesis do 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

https://doi.org/10.1021/acs.jcim.4c00370
J. Chem. Inf. Model. 2024, 64, 5796−5805

the cycloaddition for which many variations are known using chiral catalysts or chiral auxiliaries[52,53]

Figure S7 shows a predicted route for the drug sacubitril, where we have forced the search to start with stereoselective disconnections, as shown in Figure 8. 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.[54] Precedents for such reactions are limited; 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, by employing a chiral catalyst or a chiral auxiliary such as a phenethyl amine in the reductive amination, followed by benzylic hydrogenolysis and *N*-acylation of the resulting amine using succinic anhydride.[55] Thus, this type of proposal can still be useful for ideation, followed by fine-tuning to provide final routes.

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. Furthermore, we reiterate that we have developed a model that fits well within our current framework (see Figure 1) and, therefore, there is likely plenty of room for model optimization with regard to, e.g., architecture, training data selection, and hyperparameters. However, in our experience, it is very difficult to translate this sort of model optimization to performance in multistep route planning.[56] Therefore, we see an advantage in keeping cross-compatibility with already existing tools and models.

CONCLUSIONS

We devised a robust workflow to extract stereoselective and stereospecific reactions from historical reaction data and trained a template-based retrosynthesis model for these reactions. The one-step retrosynthesis model outperforms existing, more general models, but we have also identified room for improvement. For instance, there might be room for model optimization in terms of the architecture or hyperparameters. Furthermore, the approach we use to extract and apply templates can be improved, which becomes especially clear in the evaluation of the multistep performance. We may 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 tool that should be carefully examined and elaborated upon 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 data set, USPTO, contains only a small number of reactions from which only a few hundred templates can be extracted, whereas proprietary data sets, such as the in-house AZ data set 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 limitation 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[46] provides an opportunity to address this issue early on by including stereocontrolled reactions and the appropriate information in the data sets.

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 data sets, a significant improvement over existing methods was achieved with the goal of making CASP more useful as an idea generator for the practicing organic chemist. The model can be further improved by applying the framework to stereochemically richer data sets that could include more focused parts of the reaction space or proprietary data sets without changes of the framework.

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

ASSOCIATED CONTENT

Data Availability Statement

All workflows and programs are part of the AiZynthFinder package, which is available free of charge at the GitHub repository of the AstraZeneca Molecular AI group https://github.com/MolecularAI/

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.4c00370.

Chemical space of the evaluation set for the one-step models as well as some sets used for training other models; training statistics for AZ-stereo and USPTO-stereo; performance of three one-step retrosynthesis models on the test set of stereoselective reactions; performance of multistep retrosynthesis on ChEMBL target sets with the USPTO model; reaction from the US20180298056A1 patent with atom-mapping assigned; example route 1 showing a pathway for CHEMBL3112743; example route 2 showing a pathway for CHEMBL3559952; example route 3 showing a pathway for CHEMBL215018; and example route 4 showing a pathway to synthesize sacubitril (PDF)

AUTHOR INFORMATION

Corresponding Author

Samuel Genheden − *Molecular AI, Discovery Sciences, R&D, AstraZeneca, Gothenburg, SE-431 83 Mölndal, Sweden;* orcid.org/0000-0002-7624-7363; Email: samuel.genheden@astrazeneca.com

Authors

Olaf Wiest − *Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana 46556, United States;* orcid.org/0000-0001-9316-7720

https://doi.org/10.1021/acs.jcim.4c00370
J. Chem. Inf. Model. 2024, 64, 5796−5805

Christoph Bauer − *Data Science and Modelling, Pharmaceutical Sciences, R&D, AstraZeneca, Gothenburg, SE-431 83 Mölndal, Sweden;* orcid.org/0000-0002-6035-6944

Paul Helquist − *Department of Chemistry and Biochemistry, University of Notre Dame, Notre Dame, Indiana 46556, United States;* orcid.org/0000-0003-4380-9566

Per-Ola Norrby − *Data Science and Modelling, Pharmaceutical Sciences, R&D, AstraZeneca, Gothenburg, SE-431 83 Mölndal, Sweden;* orcid.org/0000-0002-2419-0705

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.4c00370

Author Contributions

S.G., P.-O.N., and O.W. conceptualized the work, S.G. and C.B. wrote the code and workflows, and 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 Discovery and Development*, Vol. *29*, 2021. 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**.

(6) Strieth-Kalthoff, F.; Szymkuć, S.; Molga, K.; Aspuru-Guzik, A.; Glorius, F.; Grzybowski, B. A. Artificial Intelligence for Retrosynthetic Planning Needs Both Data and Expert Knowledge. *J. Am. Chem. Soc.* **2024**, *146*, 11005−11017.

(7) 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.

(8) Wipke, W. T.; Ouchi, G. I.; Krishnan, S. Simulation and evaluation of chemical synthesis—SECS: An application of artificial intelligence techniques. *Artif. Intell.* **1978**, *11*, 173−193.

(9) Hendrickson, J. B.; Toczko, A. G. SYNGEN program for synthesis design: basic computing techniques. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 137−145.

(10) 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.

(11) 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.

(12) 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., Int. Ed.* **2016**, *55*, 5904−5937.

(13) Chemical Abstracts Services *SciFinder-n*; American Chemical Society, 2021.

(14) 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.

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

(16) Empel, C.; Koenigs, R. M. Artificial-Intelligence-Driven Organic Synthesis—En Route towards Autonomous Synthesis? *Angew. Chem., Int. Ed.* **2019**, *58*, 17114−17116.

(17) 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.

(18) Mervin, L.; Genheden, S.; Engkvist, O. AI for drug design: From explicit rules to deep learning. *Artif. Intell. Life Sci.* **2022**, *2*, 100041.

(19) Schwaller, P.; Vaucher, A. C.; Laplaza, R.; Bunne, C.; Krause, A.; Corminboeuf, C.; Laino, T. Machine intelligence for chemical reaction space. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2022**, *12*, No. e1604.

(20) Lowe, D., Chemical reactions from US patents. 1976−Sep 2016, https://figshare.com/articles/2016.

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

(22) Pesciullesi, G.; Schwaller, P.; Laino, T.; Reymond, J.-L. Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates. *Nat. Commun.* **2020**, *11*, 4874.

(23) 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. Model.* **2019**, *59*, 2529−2537.

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

(25) Coley, C. W.; Thomas, D. A., III; Lummiss, J. A.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; et al. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019**, *365*, No. eaax1566.

(26) 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.

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

(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) Shields, J. D.; Howells, R.; Lamont, G.; Leilei, Y.; Madin, A.; Reimann, C. E.; Rezaei, H.; Reuillon, T.; Smith, B.; Thomson, C.; et al. AiZynth impact on medicinal chemistry practice at AstraZeneca. *RSC Med. Chem.* **2024**, *15*, 1085−1095.

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

(31) Genheden, S.; Engkvist, O.; Bjerrum, E. Clustering of synthetic routes using tree edit distance. *J. Chem. Inf. Model.* **2021**, *61*, 3899−3907.
(32) Genheden, S.; Engkvist, O.; Bjerrum, E. Fast prediction of distances between synthetic routes with deep learning. *Mach. Learn. Sci. Tech* **2022**, *3*, 015018.
(33) Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.* **2020**, *11*, 154−168.
(34) Rosales, A. R.; Wahlers, J.; Limé, E.; Meadows, R. E.; Leslie, K. W.; Savin, R.; Bell, F.; Hansen, E.; Helquist, P.; Munday, R. H.; et al. Rapid virtual screening of enantioselective catalysts using CatVS. *Nat. Catal.* **2019**, *2*, 41−45.
(35) 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 Catal.* **2023**, *13*, 14285−14299.
(36) 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. Commun.* **2018**, *54*, 8294−8311.
(37) Hansen, E.; Rosales, A. R.; Tutkowski, B.; Norrby, P.-O.; Wiest, O. Prediction of Stereochemistry using Q2MM. *Acc. Chem. Res.* **2016**, *49*, 996−1005.
(38) Diorazio, L. J.; Hose, D. R.; Adlington, N. K. Toward a more holistic framework for solvent selection. *Org. Process Res. Dev.* **2016**, *20*, 760−773.
(39) Eliel, E. L.; Wilen, S. H. *Stereochemistry of Organic Compounds*; John Wiley & Sons, 1994.
(40) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Unsupervised attention-guided atom-mapping. *ChemRxiv* **2020**.
(41) https://github.com/rdkit (accessed 2023-07-01).
(42) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(43) Irwin, R.; Dimitriadis, S.; He, J.; Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Mach Learn: Sci. Technol.* **2022**, *3*, 015022.
(44) Chen, S.; Jung, Y. Deep retrosynthetic reaction prediction using local reactivity and global attention. *JACS Au* **2021**, *1*, 1612−1620.
(45) 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. Model.* **2022**, *62*, 2111−2120.
(46) 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.
(47) 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. Model.* **2023**, *63*, 4253−4265.
(48) Itsuno, S. Enantioselective reduction of ketones. *Org. React.* **1998**, *52*, 395−576.
(49) Swamy, K. K.; Kumar, N. B.; Balaraman, E.; Kumar, K. P. Mitsunobu and related reactions: advances and applications. *Chem. Rev.* **2009**, *109*, 2551−2651.
(50) 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.
(51) 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.
(52) Jung, M. E.; Huang, A. Use of optically active cyclic N, N-dialkyl aminals in asymmetric induction. *Org. Lett.* **2000**, *2*, 2659−2661.
(53) Yildirim, O.; Grigalunas, M.; Brieger, L.; Strohmann, C.; Antonchick, A. P.; Waldmann, H. Dynamic catalytic highly enantioselective 1, 3-dipolar cycloadditions. *Angew. Chem.* **2021**, *133*, 20165−20173.
(54) 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.
(55) Irrgang, T.; Kempe, R. Transition-metal-catalyzed reductive amination employing hydrogen. *Chem. Rev.* **2020**, *120*, 9583−9674.
(56) Torren-Peraire, P.; Hassen, A. K.; Genheden, S.; Verhoeven, J.; Clevert, D.-A.; Preuss, M.; Tetko, I. V. Models Matter: the impact of single-step retrosynthesis on synthesis planning. *Digital Discovery* **2024**, *3*, 558−572.

https://doi.org/10.1021/acs.jcim.4c00370
J. Chem. Inf. Model. 2024, 64, 5796−5805