

MIT Open Access Articles

Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: Struble, Thomas J. et al. "Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis" *Journal of Medicinal Chemistry*, "Artificial Intelligence in Drug Discovery" Special issue, 2020, © 2020 The Author(s)

As Published: 10.1021/acs.jmedchem.9b02120

Publisher: American Chemical Society (ACS)

Persistent URL: <https://hdl.handle.net/1721.1/125681>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of use: Creative Commons Attribution 4.0 International license



Current and Future Roles of Artificial Intelligence in Medicinal Chemistry Synthesis

Thomas J. Struble, Juan C. Alvarez, Scott P. Brown, Milan Chytil, Justin Cisar, Renee L. DesJarlais, Ola Engkvist, Scott A. Frank, Daniel R. Greve, Daniel J. Griffin, Xinjun Hou, Jeffrey W. Johannes, Constantine Kreatsoulas, Brian Lahue, Miriam Mathea, Georg Mogk, Christos A. Nicolaou, Andrew D. Palmer, Daniel J. Price, Richard I. Robinson, Sebastian Salentin, Li Xing, Tommi Jaakkola, William H. Green, Regina Barzilay, Connor W. Coley,* and Klavs F. Jensen*

Cite This: <https://dx.doi.org/10.1021/acs.jmedchem.9b02120>

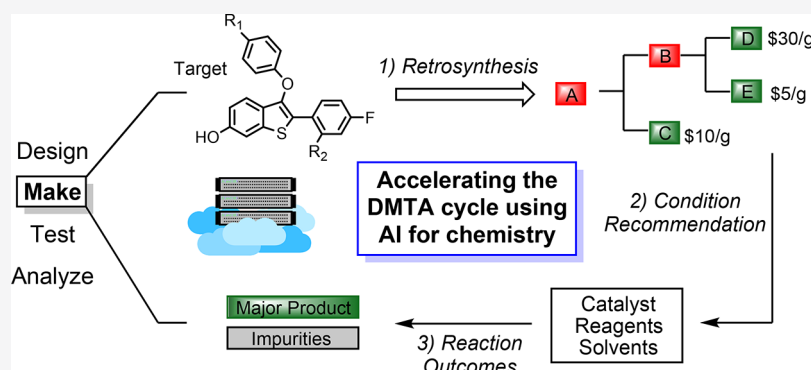
Read Online

ACCESS |

Metrics & More

Article Recommendations

Supporting Information



ABSTRACT: Artificial intelligence and machine learning have demonstrated their potential role in predictive chemistry and synthetic planning of small molecules; there are at least a few reports of companies employing *in silico* synthetic planning into their overall approach to accessing target molecules. A data-driven synthesis planning program is one component being developed and evaluated by the Machine Learning for Pharmaceutical Discovery and Synthesis (MLPDS) consortium, comprising MIT and 13 chemical and pharmaceutical company members. Together, we wrote this perspective to share how we think predictive models can be integrated into medicinal chemistry synthesis workflows, how they are currently used within MLPDS member companies, and the outlook for this field.

SECTION 1: WHERE CAN COMPUTER AIDED SYNTHESIS PLANNING AID MEDICINAL CHEMISTRY DISCOVERY?

Introduction. Current estimates of the cost to bring a single drug to market are in excess of 2–3 billion dollars.^{1–3} A significant portion of this cost may be attributed to two factors: the historically high attrition of candidate molecules going through clinical trials (an attrition rate of over 85%⁴) and the complexity of the preceding discovery phase, which requires considerable investments in time and resources. A stronger pipeline of preclinical candidates will have beneficial downstream effects in terms of total approvals. Advances in both computer hardware and *in silico* methods aim to expedite as well as improve various aspects of the medicinal chemistry's quintessential design-make-test-analyze (DMTA) drug discovery cycle (Figure 1). One area of increasing interest is the use of data-driven synthetic prediction tools for the make phase to accelerate and reduce failures in the synthesis of new molecular entities.

Computer aided synthesis planning (CASP) has a rich history that dates back to the 1960s when the Corey group first disclosed LHASA,⁵ a rule-based approach to retrosynthetic planning. This seminal publication was key in defining the heuristics of chemical synthesis that might be necessary for a synthesis planning software. Many groups disclosed advances in computer-assisted synthesis planning between the 1960s and 1990s but were largely limited by computational resources and primarily relied on human-encoded reaction rules.^{6–9}

These early progenitors serve as the inspiration for some of the commercial software packages, such as Synthia (formerly

Special Issue: Artificial Intelligence in Drug Discovery

Received: December 19, 2019

Published: April 3, 2020

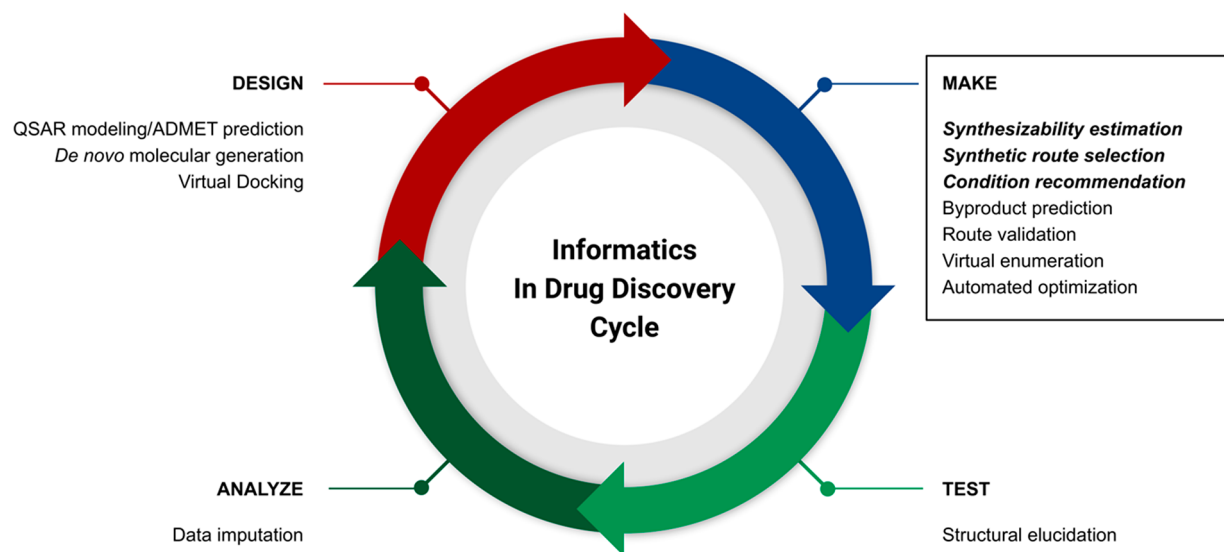


Figure 1. Some opportunities for informatics and AI techniques in the design-make-test-analyze drug discovery cycle.

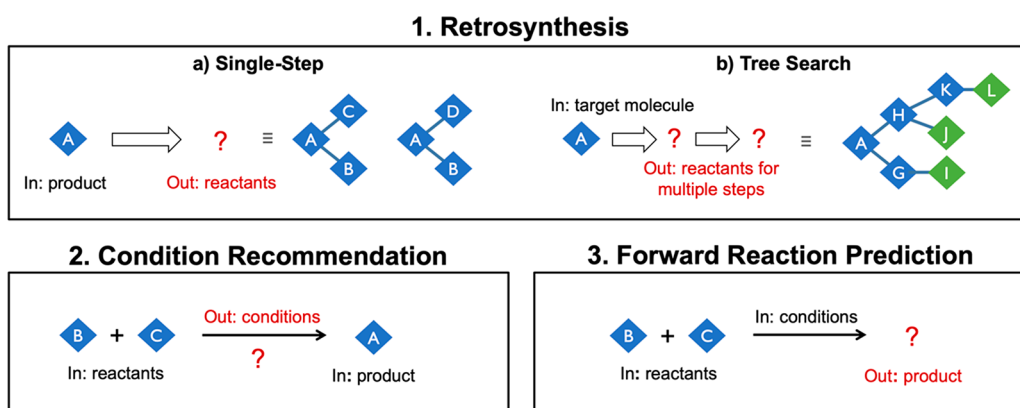


Figure 2. Three primary tasks for computer-aided synthesis planning. (1) Retrosynthesis can be broken into subproblems of (a) the generation of retrosynthetic suggestions one step at a time and (b) the recursive use of the single step suggestions to identify full, multistep routes. (2) Reaction conditions that will lead to a successful forward reaction must be recommended in order for suggestions to be actionable. (3) Reaction prediction, predicting the possible products from a set of starting materials and conditions, is used to validate the proposed synthetic steps.

Chemica) and ICSynth, where hand-coded reaction rules are used in conjunction with guiding heuristics to navigate synthetic pathways.^{10,11} Only in the past two decades have more automated methods for synthesis planning, such as those that use a subset of AI methods called machine learning (ML) to infer patterns of reactivity from published reaction data, emerged as viable alternatives to “expert” rule-based algorithms.¹² Both expert encoded rules and ML methods can be considered AI approaches: the former as an example of the so-called “first wave of AI” using crafted knowledge and the latter as an example of the “second wave” using statistical learning. Each brings its own distinct advantages to synthesis planning software. Expert-encoded rules have the opportunity to excel in low-data regimes where only 1–4 reactions might be recorded for a particular transformation. Although there is active research on using machine learning for low data, this has not yet been successfully applied to synthesis planning. However, machine learning methods can be easily extended to incorporate new reactions, as they are published due to the automation of extraction/training pipelines, which reduces the burden on experts (Ph.D.-level chemists). As more reactions are run within a company, the automated pipeline allows for

predictions to be more robust. Ref 13 provides detailed descriptions of and comparisons between available software tools.¹³

Both machine learning and rule-based approaches have demonstrated successes in planning synthetic routes that have been executed in the lab or evaluated by chemists as worth attempting. For example, Synthia has been used to find routes to medically relevant compounds and has even improved the overall yields compared to expert-developed routes;¹⁴ Segler et al. found that chemists expressed no preference for literature-validated routes over their algorithmically proposed routes in a double-blind evaluation.¹⁵ Automated platforms have been coupled to synthesis planning tools, with albeit varying levels of human intervention.^{16,17} Although the field is still in the early stages of using CASP for fully automated synthesis planning, these initial successes demonstrate the utility of the tools in a DMTA cycle.

Starting in May 2018, a team of researchers from MIT has been working closely with 13 pharmaceutical and chemical companies within the context of the Machine Learning for Pharmaceutical Discovery and Synthesis (MLPDS) Consortium.^{18,19} Among the many goals of the consortium is the

development of machine-learning-based algorithms and tools to accelerate the make phase of the DMTA cycle (Figure 1). This article will encompass the topic of machine learning for synthesis planning through the lens of the activities of the consortium and the open-source tools^{18,20} being developed at MIT and adopted by member companies.

More specifically, this perspective will describe the many roles of artificial intelligence in medicinal chemistry synthesis, including those that (1) can be integrated within a medicinal chemistry workflow, (2) are already integrated in certain pharmaceutical companies, and (3) require further development to accomplish even more ambitious tasks. We focus on the three primary tasks of computer-aided synthesis planning (CASP) in Figure 2: retrosynthetic planning, condition recommendation, and forward-reaction prediction.

Use of ML-Based CASP in Retrosynthetic Planning. Identification of Synthesizable Targets and Route Planning.

The traditional approach for synthesizing new small molecules in DMTA iterations involves manual planning and manual execution. Expert chemists are tasked with assessing the synthesizability of proposed targets, resulting in a slowdown when evaluating hundreds or possibly thousands of molecules. A particular series of lead compounds might be preferred over another due to synthetic accessibility (SA), as financial resources and time constraints limit the number of compounds that can be pursued or designed in parallel. Retrosynthesis software mitigates the bottleneck of manual synthesis assessment by generating hypothetical synthetic routes that can be used to rapidly prioritize compounds by ease of synthesis, thereby providing chemists a more focused set of compounds as a starting point for expert route planning. Finally, the use of a retrosynthetic planning platform can be beneficial for team members that do not have years of training in synthetic chemistry by providing these individuals with synthetic suggestions that may not have come to mind.

Two categories of methods for scoring compounds by synthesizability are the use of simplified structure-based heuristics^{21–23} or full retrosynthetic tree expansions. Heuristics aim to capture broad trends in SA from molecular structures and have traditionally been using expert-defined functions of molecular attributes. Nonlinear regression (e.g., using machine-learning techniques) can instead recapitulate subjective scores assigned by expert chemists²⁴ or be used in a semisupervised setting to learn from examples of chemical reactions.²⁵ In reality, however, the ability to synthesize a target is highly dependent on the availability of specific buyable building blocks and is not a smooth function of molecular structure. Since the availability of building blocks depends on the setting (e.g., organization, budget, and discovery versus process development), a more generalizable method for assessing synthesizability is to use retrosynthetic expansion with a custom database of buyable compounds that is tailored to the application. The benefit of an explicit retrosynthetic expansion is the knowledge that transformations to access a target of interest *do exist* and suitable starting materials *are available*; however, it comes at a higher computational cost. With access to a retrosynthetic planning tool and enough time and training, however, neural network models can begin to approximate this highly nonlinear function.²⁶

The two major categories of retrosynthetic planning software are those that use expert-encoded rules or heuristics to generate recommendations and those that learn (or infer) how to generate recommendations. Many retrosynthetic

methods rely on the use of reaction templates, which are reaction rules that can be stored in a SMARTS or SMIRKS format. A general procedure for the algorithmic extraction of templates from a reaction data set is to (1) identify the reaction center or changing atoms, (2) identify atoms adjacent to the reaction center, and (3) add generalized functional groups that are involved in the reaction.^{27,28} This approach captures the local reaction environments but, in most algorithmic implementations, does not capture the global features of molecules that contribute to reactivity. Expert-encoded methods¹¹ may better describe functional group requirements but cannot be tailored to an individual organization's capabilities. Automated pipelines for extracting reaction templates allow for facile (re)training on propriety data sets but are also inconsistent with the expert approach.

For the actual application of reaction templates to generate reactant molecules from an input product molecule, several machine-learning-based approaches have focused on learning which templates provide the most strategic disconnections, with varying degrees of sophistication.^{29,30} An alternative approach is the use of sequence-to-sequence models that treat the one-step retrosynthetic task as a translation between products and reactants.^{31–33} A single-step retrosynthetic recommender is sufficient for a chemist to construct routes manually, one step at a time.³⁴

The single-step retrosynthetic capabilities can be extended to full route design by using a tree search. Each step can produce thousands of precursors, which requires a guided search strategy to prevent combinatorial explosion. Candidate precursors can be filtered by SA heuristics¹¹ or by a learned expansion policy¹⁵ to have a more tractable list of chemicals to transform in the next cycle. Full pathways can be constructed by recursively suggesting single-step retrosynthetic precursors that become progressively simpler until a stop criterion is met. Different implementations of tree searches have been investigated including depth-first, best-first, and proof-number search and Monte Carlo tree search algorithms;^{15,35} a direct comparison of methods is difficult because quantitative scoring remains a challenge. Generally, a retrosynthetic search is terminated once precursors are found that can be purchased. This complicates benchmarking retrosynthesis algorithms because a larger, more diverse database of buyable chemicals will have a higher probability of termination and naturally appear more successful. Other stop criteria such as the number of occurrences in the literature or chemical logic (defining the allowable number of carbon, nitrogen, and oxygen atoms) can be used, the latter of which can provide greater standardization but is less relevant to actual applications. Moreover, the ability to identify a pathway does not guarantee its chemical feasibility. Since there are multiple routes that can synthesize the same target, the best method of validation would be to perform the chemistry in the lab; this is obviously prohibitively expensive to undertake for every route generated, time-intensive, and not a scalable approach to validating new methods in retrosynthesis planning.

Recommendation and Evaluation of Reaction Conditions. Planning a retrosynthetic route is arguably only one aspect of a full CASP system. To be an actionable suggestion that a chemist can take into the laboratory, we must propose a set of reaction conditions that are able to achieve the desired transformation. Finding the optimal or acceptable set of conditions for a reaction can require time-consuming empirical screening to determine what works best; often, a chemist will

employ “typical” conditions for that family of reactions without tailoring their choice to the particular substrate of interest. Biases for choosing reaction conditions can result from individual experience or immediate availability of reagents. In principle, machine-learning models for condition recommendation can more objectively infer suitable conditions if appropriately trained on historical condition data.

In practice, such models are difficult to develop due to a lack of high-quality data. The main data issues that hamper progress are inadequate disclosure of the (1) amounts, volumes, or concentrations; (2) reaction times or kinetics; and (3) order of the addition of reagents and catalysts. Despite these issues, data-driven approaches have demonstrated the ability to suggest conditions for specific reaction classes^{36,37} and for more diverse reaction sets.^{38,39} These models provide a strong basis for empirical optimization of reaction conditions but still lack the full details necessary for execution. Condition recommendation models would likely be developed to suit the needs of a particular area of chemistry such as medicinal chemistry or process chemistry. The objective of the reaction is different in many cases, such as the importance of yields and side-product formation. One objective might be the prediction of the “best” conditions for a set of reactions that we would like to run in parallel in a single well-plate. More specific predictions may be required for finding the optimal conditions for a single reaction where a new combination of conditions or new catalysts or reagents are designed.

Even though it is hard to escape empirical optimization of reaction conditions, particularly for complex substrates or in tandem catalysis, opportunities for techniques in artificial intelligence to accelerate this process also exist. Reaction optimization is a well-established field,^{40–42} and there exist many statistical techniques for selecting experimental conditions to iteratively improve performance (e.g., in terms of yield, turnover number, or throughput) exist. In machine learning parlance, these are active learning frameworks. The most popular methods are model-based techniques, which construct a surrogate model of reaction performance as a function of reaction conditions. Various search strategies (e.g., Bayesian optimization) can be layered on top of these models to help select the next set of conditions to try and refine the model. While these concepts are not new, machine-learning-based models have the potential to provide better estimates of performance and uncertainty to accelerate the search.^{43–45}

Forward-Reaction Prediction. The third key task of CASP is to ensure that recommendations obtained through algorithmic synthesis design are robust and actionable by anticipating, at least qualitatively, reaction products. A chemist might assess the feasibility of a reaction by searching for similar transformations, reading the literature, and determining if the synthetic method will generalize to the substrates of interest. Data-driven techniques can learn to perform the same generalization when trained on a broad set of reactions. Machine learning methods for reaction prediction include attempting to infer reaction rules from a predefined list of rules or templates,^{29,46,47} graph convolutional neural networks that predict atom and bond changes from starting materials to products,^{48,49} and sequence-to-sequence models which predict product SMILES.^{50–52} Compared to the evaluation of retrosynthesis models, forward synthesis models are more straightforward to evaluate quantitatively, as there is, in principle, only one true answer. In practice, however, the

absence of precise concentration, time, and temperature data make reaction prediction an ill-posed problem.

These forward-reaction predictors can also be used for side-product prediction. Knowledge of the most probable products helps identify reactions that would produce potentially harmful or difficult-to-separate intermediates. Many reactions can lead to multiple regio- or stereoisomeric compounds. Information about a reaction’s selectivity and possible side products is a crucial aspect of prioritizing syntheses, and can potentially assist with structure assignment. Once these models are able to make *quantitative* predictions, they will be indispensable for the consideration and design of purification strategies.

In addition to use in CASP, there are other applications for reaction prediction. Many make-on-demand virtual libraries are enumerated based on expert-defined reaction templates that focus on a limited set of chemistries intended to be as robust as possible.⁵³ Reportedly, successful delivery of compounds in make-on-demand libraries within 4 weeks is around 85% and within 6 weeks is 93%.⁵⁴ This high success rate demonstrates the robustness of rule-based methods using well-established chemistry. Either using heuristically extracted templates or template-free methods, new reaction space (e.g., novel synthetic methods described in new publications) can be included in an automated pipeline in real-time. If targets have been identified and retrosynthetic plans are in place, a search can be performed for all the combinations of available starting materials that could be substituted. For example, if the first reaction is a Suzuki coupling reaction, all the combinations of available boronic acids and aryl halides can be enumerated. A forward predictor is then useful for scoring which combinations will likely lead to successful reactions. By further ranking this set in terms of compounds’ properties of interest, a rapid assessment of the accessible chemical space surrounding the target can be made, e.g., for hit expansion in drug discovery. This capability is closely related to integrating the goals of diversity-oriented synthesis into CASP.

SECTION 2: HOW IS CASP CURRENTLY USED IN THE PHARMACEUTICAL AND CHEMICAL INDUSTRY?

There are many ways to envision integration of AI-based CASP tools into the medicinal chemistry workflow, and adoption appears to be on the rise. The discussion below will largely focus on the use of tools within the open source ASKCOS software suite.²⁰ However, the applications are general. We will break the use cases into multistep route planning, forward-reaction prediction, and condition recommendation, as outlined in the introduction. Finally, we will briefly discuss how the incorporation of programmatic interfaces can aid DMTA workflow and the general feedback from MLPDS member companies on ASKCOS functionality and its adoption in their organizations.

Multistep Route Planning. Many of the available commercial and academic synthetic route planning software provide a stand-alone graphical user interface (GUI) or web-based interface where users can interact with the suggested routes and predictions. The target users of the software range from nonchemists, without much knowledge of chemical reactivity, to highly trained expert chemists who want to streamline their synthesis workflow. The MLPDS consortium member companies report that the primary users of the software are expert, Ph.D.-level chemists, and adoption is reported to vary from indifference up to enthusiastic and

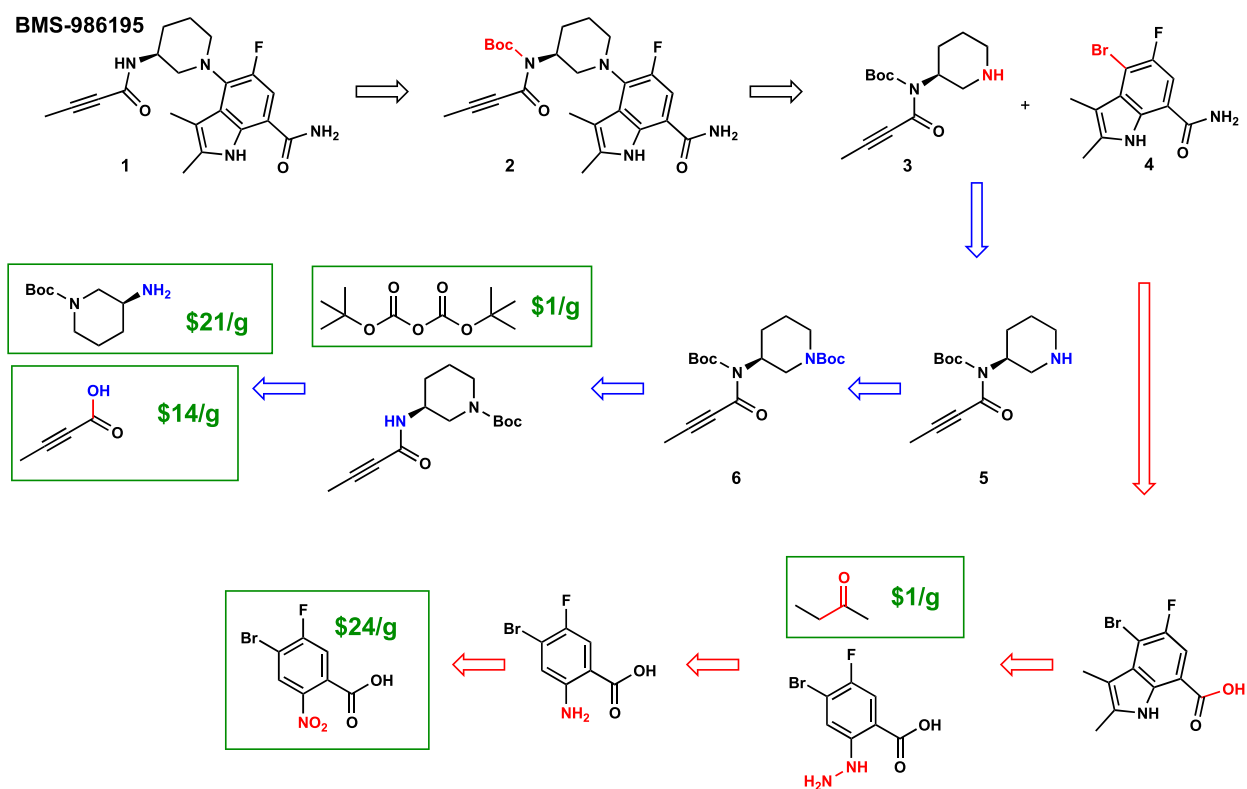


Figure 3. Retrosynthetic analysis of branebrutinib performed by ASKCOS. The route is similar to that in ref 56, with the difference that the authors found a nitrile analogue of 4 to be optimal for the C–N-bond-coupling step.

everyday use. At Janssen, many chemists use synthesis planning tools in parallel with traditional database lookups of known reactions to generate ideas more quickly. Other users are computational chemists and chemical engineers who may not have as much practice at retrosynthetic planning but are involved with molecular design or process development. Most companies pilot small rollouts to select expert chemists, who are in the strongest positions to evaluate the capabilities of machine learning CASP tools and identify key limitations. At BASF, experts from different stages in product development (e.g., early phase and process development) provided understanding of the different expectations across business areas. These small rollouts are necessary to understand the obstacles to wider adoption and further integration into synthesis pipelines. A close contact is necessary between the company's beta testers and the developers of retrosynthesis algorithms since the true assessment of performance must be carried out by trained experts who can validate the model(s) suggestions.

The proof of principle for full pathway planning has been established, but further refinement will require the input of chemists who can objectively evaluate retrosynthetic predictions. Input from the MLPDS member companies has identified some general trends in which the machine learning algorithms perform well and poorly. Generally, target molecules that are in a similar chemical space to product molecules found in Reaxys or USPTO tend to perform well using the ASKCOS suite of tools. These target molecules can be accessed using well-established chemistries and the models can perform adequately within their domains of applicability.

For example, input of the structure of branebrutinib (BMS-986195, 1, Figure 3) resulted in an ASKCOS proposed synthesis. While the synthetic route was first reported in 2016,^{55,56} the training data for ASKCOS stops before the initial

disclosure of this molecule, which demonstrates the ability of the machine-learning models to generalize to new target compounds. The ASKCOS-proposed synthesis begins from commercially available starting materials that are similar to those in the reported route and uses several types of reactions (C–N cross-coupling, heterocycle formation, diazotization/reduction) to arrive at the final product. While the overall synthesis appears plausible, the selective Boc deprotection of 6 to 5 is likely problematic and could be easily substituted with an orthogonal protecting group when deciding on the final route. As previously noted, one area for improvement is better prediction of regioselectivity of reactions where there are two pendant reactive groups. For the proposed CN-bond-forming reaction, ASKCOS suggests Boc protection of the N–H of the alkynamide intermediate (3) while the second coupling partner (4) contains a free carboxamide. In the literature synthesis of 1, the authors note that a carboxamide unexpectedly prevents the Buchwald (C–N) coupling from proceeding. Thus, this team performs the reaction with the nitrile substituted for the carboxamide of 4. While the exact ASKCOS C–N coupling has never been tried, the fact that chemists attempted the C–N coupling with a free carboxamide demonstrates that the ASKCOS prediction is reasonable (i.e., worth trying), but the carboxamide would likely need to be converted to a nitrile. This example is just one small piece of evidence that ASKCOS can reasonably disconnect modern drug-like molecules and how nuanced experimental results (including failures), not captured in large reaction databases, can negatively impact algorithm performance.

Many different aspects play into the “success” of machine-learning-based path-planning tools. One of the simplest factors in whether these programs are able to find pathways is the coverage of the database of compounds considered to be

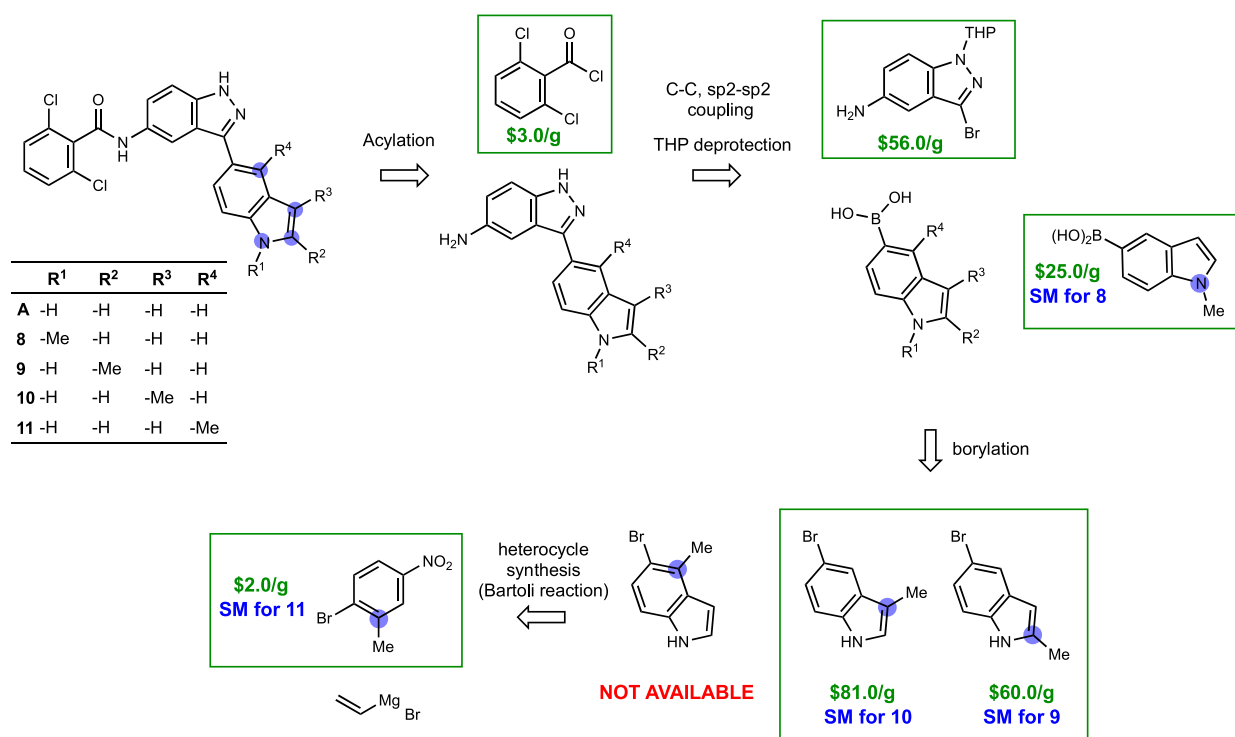


Figure 4. Retrosynthetic analysis of methylated analogues of compound **A**. ASKCOS proposes different length of routes which is dependent on the availability of starting materials, where the tree search stop criteria is commercial availability (<\$100/g). Compound **8** can be accessed from commercially available starting materials in 3 steps, compounds **9** and **10** require one extra step, and compound **11** requires 2 extra steps.

commercially available; simply put, a larger starting material database increases the odds that a search will terminate successfully. In an effort to better understand how the database of buyable chemicals affects tree search outcomes, GlaxoSmithKline compared the stock ASKCOS database of buyable compounds (138k) and a larger set that was augmented with their internal compounds/vendors (8M). On an internal set of 69 target molecules, using the most liberal path-planner settings, a route was found by ASKCOS for 54% of compounds with the stock database and 67% of compounds with their internal database. These results highlight the dependency of path-planning algorithms on the database used for a stop criterion. The dependence on a buyable database, however, complicates the comparison of CASP tools since every software package uses a different (typically undisclosed) buyable database. This problem may be alleviated by the implementation of straightforward utilities to load and use custom building-block sets in every CASP tool. This requirement is generally useful since all MLPDS corporate members maintain internally large collections of building blocks.

How the availability of starting materials affects the SA can be seen by enumerating methylated variants of compound **A**⁵⁷ followed by an evaluation using ASKCOS's retrosynthesis tree search. While an expert chemist can easily infer that compound **A** is disconnected by amide formation and C–C cross-coupling, the general knowledge of the commercial availability of the methylated starting materials is less likely. The input of methyl analogues of **A** results in the expected bond disconnections (Figure 4, representative ASKCOS results shown). Since the stop criteria for the tree search is commercial availability, the algorithm will assess at each disconnection whether the suggested starting materials are

purchasable. In this example, the starting material for compound **8** can be purchased after only a few retrosynthetic steps. Compounds **9** and **10** require an extra step compared to **8**. Finally, access to indole **11** would necessitate further steps to synthesize, which draws the step count to almost double that of the synthesis of compound **8**. Notably, the information is obtained using ASKCOS with one search per compound. This assessment now provides chemists with the information on which analogues are most synthetically accessible and can factor into the decision-making process for prioritization of target molecules.

An expected feature of machine-learning methods for predictive chemistry is that retraining models on proprietary data ought to allow companies to achieve better predictive ability on chemistries that are used in-house.⁵⁸ These in-house chemistries may not be well represented in public or published data sets, which most of the CASP systems are trained on. Researchers from AstraZeneca and the University of Bern applied a workflow for retrosynthetic template extraction²⁸ and training/application²⁹ to several public and proprietary data sets and compared the performance of the different models.⁵⁹ They found that Reaxys has the most unique reaction templates, of which 2% are shared between all the data sets used in the study, and only 0.6% are shared between Reaxys and a subset of their proprietary ELN data. Eli Lilly identified a subset of 6k target compounds from approved, experimental, and investigational drugs to represent the chemical space of interest to the company. Using the Lilly database of building blocks and ChemoPrint, an in-house synthetic planning platform, retrosynthetic expansion was performed using a template set from (1) only Lilly eLN data (13297 templates) and (2) Lilly eLN data plus patent data (13297 + 50275 templates). Routes could be found for 40.1% of the 6k

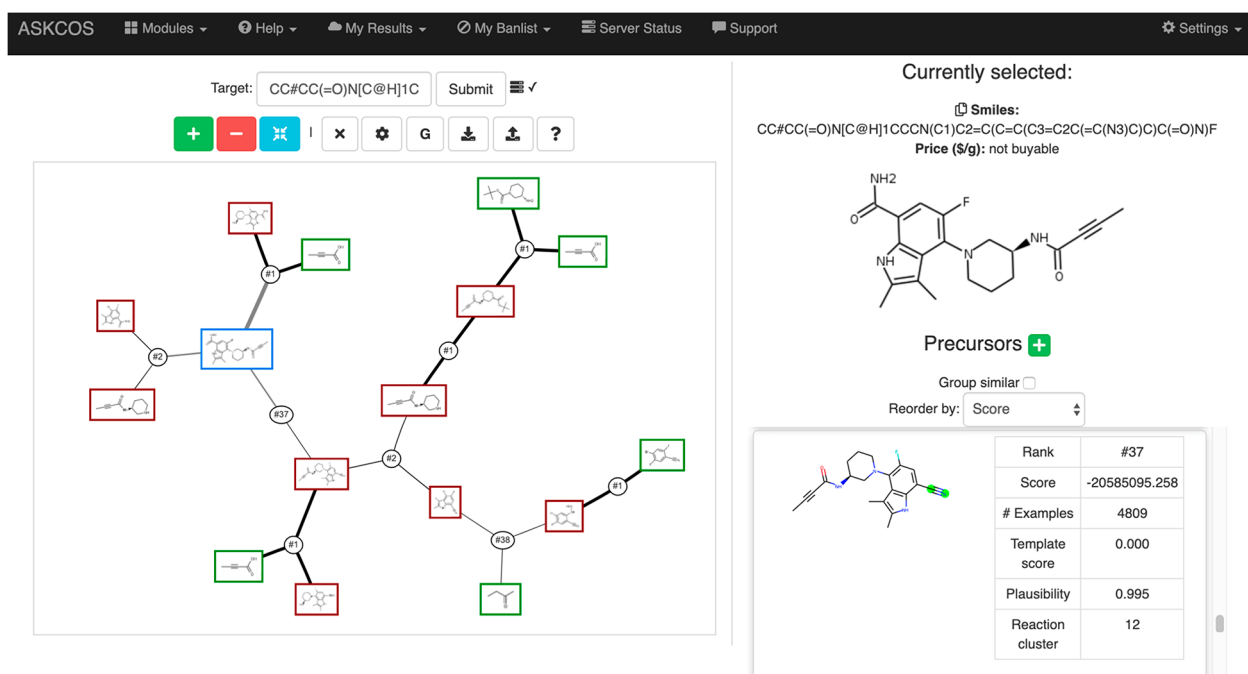


Figure 5. Screenshot of the ASKCOS interactive path planner. (Left) Visualization of a full synthetic graph constructed by the user. Boxes are color-coded green if they are purchasable and blue for the root target compound (branebrutinib). (Right) The selected molecule is displayed on top for which a single-step retrosynthesis prediction has been performed and on the bottom are all of the predicted precursors.

compounds with the first template set. Supplementing the template set with additional patent templates only provided a 5.8% increase in the ability to successfully furnish a route, corresponding to a success rate of 46.9%.⁶⁰ For full-pathway planning, these results demonstrate the necessity of further testing on internal and proprietary data sets and the influence that company data may have on multistep path planning.

There are still many molecular structures for which retrosynthesis planning fails to find any route. The MLPDS consortium members have identified lack of coverage for several company-specific target molecules or reactions in full-path planning. Commonly identified substructures that are not successful in full-path planning are small, densely functionalized carbon cores with or without many contiguous stereocenters, caged structures where 3D geometry is crucial for selectivity, newly discovered heterocycles, and complicated polycyclics. Some of these substructures, such as densely functionalized carbon cores, require chemistry that is specific to each core's substructural environment (perhaps with <5 precedents in the literature). Using the conventional template extraction procedures, the model will not be able to generalize due to the high specificity of the template. Conversely, path planning of some target molecules will find numerous pathways but include many poor retrosynthetic suggestions where regio- or stereoselectivity may not be predicted appropriately. To correct the issues of selectivity, further filtering using an accurate forward-prediction model will provide richer route suggestions. Another set of failures are due to the limitation of the search methods for navigating a synthetic tree. Since recursive retrosynthetic expansion has to restrict the search to avoid combinatorial explosion, most implementations cannot yet navigate a search path deeper than 15 synthetic steps. If chemists are using CASP tools for the ideation of routes and pathway planning cannot successfully

navigate a synthetic graph to produce a route, another solution is necessary.

When full-path planning fails, a chemist may resort to using single-step retrosynthetic predictions to manually construct a route. Figure 5 is an example where a path to branebrutinib is manually explored. Interestingly, the suggestion of using a nitrile, which was found to be ideal in practice, is in the precursors lists but is ranked #37, so a chemist would have to sort through many higher-ranked suggestions. Manually constructing a route from tens to thousands of disconnections is a time-consuming task. A synthesis planning feature that was born from discussions between MLPDS member companies and MIT was the implementation of an interactive path planner using single-step retrosynthetic predictions. The interactive planner addresses the issue of displaying diverse suggestions and having more control over a synthetic plan. When chemists are initially developing a route, the precise choice of leaving groups matters less, and as the routes are refined, specific leaving groups are chosen based on the desired reactivity. Machine-learning models for retrosynthesis generally handle all possible reactants as distinct options. For example, the chloride, bromide, and iodide form of a halogenated precursor are not normally lumped into a single category. It is inconvenient for chemists to sort through numerous suggestions that have the same fundamental disconnection but different leaving groups. Thus, a clustering algorithm was developed to group similar suggestions (based on a k-means clustering using a reaction fingerprint⁶¹) and expedite the exploration of distinct disconnections. Several routes are displayed using one visualization, which can be download and shared. Although none of the underlying machine-learning models were changed, expert users are much happier with exploring pathways interactively when an automated path-planning job fails. This success demonstrates the that tight collaboration between end users and the developers of

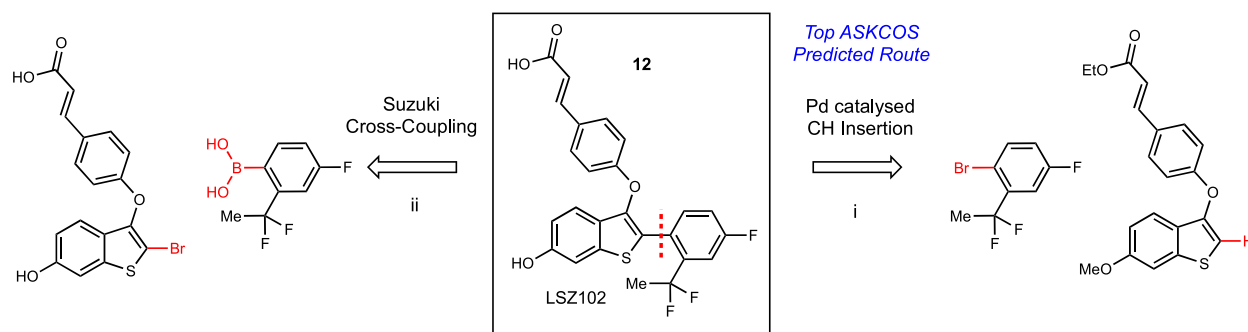


Figure 6. Single-step retrosynthetic predictions for LSZ102. (i) Palladium catalyzed C–H insertion and (ii) Suzuki cross-coupling. Conditions are listed for the top-ranked disconnection, and the optimized conditions are listed along with the range of yields for a variety of substrates.

synthetic planning software is helpful for adoption, particularly when it comes to the user interface.

An advantage of many synthesis planning packages is that reaction templates, or rules, are associated with a specific set of literature precedents. MLPDS member companies report that CASP tools are used more often when literature examples, upon which predictions are based, are easily accessible. For example, ASKCOS provides a mechanism to use reaction IDs tied to reaction examples in the training data and can direct users to literature lookups or in-house reaction entries.

Forward-Reaction Prediction. The purpose of a machine-learning-based forward prediction is to validate routes that are furnished from full-path planning. In our implementation, forward prediction is not carried out automatically during the tree search via the GUI but can be performed on reactions after expansion. In practice, forward-reaction prediction tools are currently used mainly to identify potential side products and impurities rather than for confirming routes. Similar to retrosynthetic planning, the use of company data should improve the quality of predictions for in-house use by aligning the types of chemicals/reactions used for training and prospective prediction. A recent study between Pfizer and the University of Cambridge demonstrates that retraining a sequence-to-sequence forward prediction model on proprietary data does boost accuracy for company-specific chemistry.⁵⁸

Condition Recommendation. Of all the MLPDS modules deployed at the member companies, condition recommendation is used least often and receives the least feedback. Previous research has reported recommendations of very specific conditions limited to a single reaction class.^{37,62,63} These focused models do not approach the global intuition of reactivity that expert chemists have but may be useful when very specific conditions are necessary. A general model for condition recommendation, such as the one included in ASKCOS,³⁹ that can provide a good starting point for reaction execution would be preferable for medicinal chemistry workflows. However, these generalized models encounter limitations subject to the training sets' domain of applicability. Chemists currently can use ASKCOS to get a good starting point for planning a reaction, but many reasons may contribute to the lower adoption of condition recommendation. One is that the model suggestions are not specific enough (concentrations, time, order of addition, etc. are missing) to give conditions that are actionable. The conditions the model provides can be obtained through a literature search of similar transformations, which is the mechanism still preferred by practicing chemists. We find that chemists often use the model to confirm some set of the conditions they have already

proposed or simply to evaluate the suggestions and give feedback to the model developers. Long-term, there is an opportunity to impact automated experimentation once it is possible to make quantitative recommendations, but currently, the utility of condition recommendation is limited.

A retrospective analysis was performed of reactions used in the SAR discovery phase of Novartis's LSZ102 (compound 12) and its derivatives. ASKCOS path planning identified routes where LSZ102 can be assembled via two classes of palladium catalyzed C–C coupling (C–H activation and Suzuki–Miyaura), as shown in Figure 6. Indeed both coupling strategies were widely utilized during the SAR discovery phase toward LSZ102.⁶⁴

Further retrospective analysis of the top-rated disconnection, a Pd-catalyzed C–H activation, identified the requirement for both high temperature and polar aprotic solvents (DMF/DMA) in the top 3 conditions proposed. In reality, a screening optimization of the ligand and base was necessary to maximize the performance for the specific scaffold. The optimized conditions were applied to a diverse range of substrate starting materials with yields in the range of 39–97%. The initial temperature and solvent conditions proposed by ASKCOS were not far from those actually employed and would have provided a good starting point for either scouting or screening efforts.

One appealing application for context recommendation models is helping chemists and chemical engineers identify opportunities to leverage specific technologies at the onset of synthesis design. In doing so, more efficient and sustainable conditions aligned with green chemistry principles could be readily identified. One such example of this would be in the application of surfactant-based technology at Novartis, which seeks to replace undesirable solvents with a greener micellar-water surfactant system. Indeed, in the above example of LSZ102, the Suzuki–Miyaura retrosynthetic analysis identified by ASKCOS could be realized under such surfactant conditions.^{65–68} In comparison with the same reaction performed in a standard organic solvent, the generation of the desbromo side product was significantly reduced from 8 to 0.7%.⁶⁵ By training ASKCOS with relevant internal data, it is envisaged that context recommendation models will be able to identify and propose more favorable conditions versus the historical conditions more prevalent in the existing literature using flexible user-provided definitions of “favorable”.

Programmatic Interfaces for Incorporation into Company Platforms. Although a graphical user interface is the primary method of use by chemists, computational tools can be straightforwardly integrated with other computational

pipelines. A closer integration with in-house tools for molecular design represents an additional value proposition for CASP and could lead to greater adoption. For example, a programmatic interface for sending requests to path-planning software from in house design modules allows the automation of running retrosynthetic expansions and the accumulation of the necessary data for prioritization of target molecules.

Eli Lilly has designed an in-house workflow named Kernel where target compounds from chemists, or screening hits, are submitted and prioritized in an automated fashion. Once Kernel identifies the prioritized compounds, full retrosynthetic path planning is performed on all of the molecules utilizing the ChemoPrint API⁶⁹ and the Lilly building block collection, which is then added to the compound listing; team members are informed of the results by e-mail.⁷⁰ This frees users of the design software from having to open standalone CASP tools. In addition, the use of a centralized data repository to store ideated compounds and their routes facilitates sharing and collaborative prioritization.

As do many companies, Merck & Co., Inc., Kenilworth, NJ, US uses Spotfire to organize designed molecules and their measured or predicted properties. They have initiated incorporation of the programmatic interface of the ASKCOS single-step retrosynthesis into a workflow to triage hits from virtual screening libraries. The results can be presented in many ways, but they analyze hits by availability and price of precursors broken down by each step. This rapidly informs chemists which target molecules may be able to be synthesized in parallel and, with incorporation of forward scoring, allow for consideration of which reactions may be most successful.

BASF has developed an integrated platform for linking literature references and internal electronic lab notebooks to the synthesis reaction template suggestions and integrated in-house compound stock databases into the recursive path planning to optimize usage of internal resources. Molecules used in a proposed synthetic pathway are connected to an in-house suite of tools for the prediction of physical and toxicological properties enabling an *in silico* assessment of reaction feasibility and safety before undertaking laboratory work.

One could envision that programmatic interfacing could be useful for *de novo* molecular generation as well. A common complaint from chemists about *de novo* methods is that the molecules are not synthetically accessible. Calculated SA scores have the benefit of speed, but imposing a bias to generate molecules using full recursive path planning would ensure that routes do exist to the generated molecules. Of course, this limits the chemical space in which the generative model will operate, but the improvement in synthesizability may be worth the trade-off.

Automated Synthesis Platforms. Synthesis planning is a crucial component of a fully automated reaction platform. Research toward automated synthesis platforms has been restricted to a relatively small set of reactions and largely remains in the proof-of-concept phase in both academia and industry. Current automated platforms still require a significant amount of human setup and planning but the process may become more streamlined with the integration of predictive chemistry tools. One opportunity was demonstrated using ASKCOS synthesis planning software that was coupled to a robotic flow synthesis platform.¹⁶ This is proof-of-principle that machine-learning CASP tools can be useful for recommending routes and conditions for automated execution;

however, the route and conditions suggestions still needed to be refined (e.g., to specify concentrations and reaction times) and were optimized offline (e.g., to be amenable for flow chemistry) before being executed on the robotic platform. In this case, the requirement for manual intervention is partially attributable to the dearth of training data for flow chemistry compared to the prevalence of batch chemistry results but could have been circumvented by using more traditional batch methods or parallel plate-based methods. Other options for automated systems include a closed-loop DMTA cycle using cyclofluidics,^{71–73} automated laboratories,⁷⁴ and ultrahigh throughput experimentation.^{75–77}

Integration of retrosynthesis planning software into closed-loop automation is currently underway at some pharmaceutical companies. At Eli Lilly, ChemoPrint has been successfully integrated into an automated platform for chemical synthesis.⁷⁴ Lilly has previously demonstrated as a proof-of-concept that the whole DMTA cycle can be automated and executed with minimal intervention from expert chemists.¹⁷ At present, these examples are limited to single-step synthetic plans and in initial literature reports⁷⁰ and did not have a large impact on driving the project. As a proof-of-concept, this experiment demonstrates the feasibility of coupling CASP and automation to drive a DMTA cycle. Although closed-loop lead optimization has not been fully implemented for multistep synthesis, rapid progress is being made by both academic and industrial researchers.

Adoption by Users. In 2017, a small group of chemists surveyed at three pharmaceutical companies were asked to define the most important features of a synthesis planning platform to encourage adoption.¹³ The top 6 important features to respondents were (1) an easy use and intuitive interface for interaction with the routes, (2) a method to explore the literature precedents associated with route suggestions, (3) the user can define bonds they desire to be broken to guide the search, (4) routes are terminated in purchasable starting materials, (5) functional group incompatibilities and unstable compounds are identified with protecting group strategies proposed to bypass these complications, and (6) a scoring system is implemented for ranking routes (discussed further in **Establishment of Success Metrics**). In our experience, these desired data are shared by end users at most organizations. Out of all of the important features, many are implemented, to varying extents, in the ASKCOS software package and many of the companies' in-house tools.

At AstraZeneca, historically, CASP tools have mainly been used in pharmaceutical development to identify alternative routes.¹⁰ However, during 2019, a customized adaption of the ASKCOS interface has been rolled out to all medicinal chemists. The interface includes both the models developed by the MLPDS consortium and internally developed models based on integrating internal ELN data with licensed and public data.^{78,79} So far, the uptake of the ASKCOS tool at AstraZeneca has been positive.

As stated previously, the users can range from nonexpert chemists to practicing chemists. Many of the early evaluators at companies are computational chemists and informaticians who are deciding the correct method for integration into workflows. Evaluation by chemists is tricky without first defining what success and failure look like. The natural tendency of expert synthetic chemists is to input a favorite target compound (often a very complicated natural product) into the full

pathway search and look for familiar routes. Users may be dissuaded from future use of the tools if known/published routes are not displayed or ranked near the top suggestions. Adoption is anecdotally higher when basic training is provided for chemists to introduce the theory behind the software and examples of how to effectively use the different modules in each package. Importantly, this training should convey that one goal of data-driven programs such as ASKCOS is to go beyond lookups of known routes; the proposed routes are predictions based on generalizing from known reaction data. Instructions on how the models work, what the goal of the methods are, the model limitations, and how to change the inputs to obtain useful information have been noted to greatly increase the engagement of chemists. At many companies, a 1–2 h tutorial and interactive session has been seen as sufficient to train new users and enable them to effectively use CASP tools.

SECTION 3: HOW CAN WE, AS A RESEARCH COMMUNITY, MAKE CASP BETTER?

Integration of CASP into medicinal chemistry workflows is a work in progress but there remain many challenges to developing and deploying machine-learning CASP tools in practice. Adoption of synthesis planning software is gaining momentum and is beginning to have an impact on the DMTA cycle by facilitating the “make” portion. Although more chemists are employing CASP tools, progress and reproducibility is hampered by a number of groups who publish advances in synthesis planning without making their code open-source or available upon request. In addition, standardization of metrics should be agreed upon with publicly available data sets since proprietary data are usually not, or cannot be, shared.

Establishment of Success Metrics. The most common metric for assessing single-step retrosynthetic model performance is top- k accuracy. This metric is evaluated using a test set of known single-step reactions and is calculated based on an exact match of the true disconnection being in the top- k of the predictions. Although top-1 accuracy is informative for the model development, it is a poor metric since there are always multiple retrosynthetic disconnections that *could* be successfully executed in the lab. Model evaluation using top- k accuracy for small k (1–3) implies that the published method is one of a few “correct answers”, when in reality retrosynthesis is a fuzzier prediction. While multiple answers are not recorded in the database, there may be many correct ground truths, so metrics like top-10 accuracy (or larger k) are more appropriate but can also inflate the accuracy, which might not correctly reflect the model performance. A simple example is if a program chose bromine and chlorine as possible leaving groups for a simple substitution; either of these might be successful in an experiment, depending on the attempted reaction.

Another important but less commonly reported metric is the diversity of predictions. From most chemists’ point of view, the top- k accuracy may not always be the most important factor for choosing a retrosynthesis tool. For route planning, a key disconnection that has not been thought of yet is just as important as the feasibility of the suggestions. There is a trade-off when developing models for retrosynthesis, where the suggestions have to be feasible, useful, and sometimes nonobvious for idea generation. An example of a highly feasible, not useful, but obvious suggestion would simple functional group interconversions, where the complexity is not

being built (but is often seen in the historical reaction data). Conversely, a nonfeasible, very useful, and nonobvious suggestion would be one that suggests breaking a bond where no known chemistry has been developed to actually carry out the reaction. One can envision where this would be very useful for idea generation but not practical when planning routes where timelines are of the essence. The holy grail would be to suggest a feasible, useful, and nonobvious disconnection, which is a difficult task without true metrics for each of these categories. Using top- k accuracy to score single-step predictions allows us to achieve *feasible* suggestions while employing heuristics guides the models toward *useful* disconnections. The trade-off can only be assessed by expert chemists who can sort through many suggestions; however, these chemists’ scores are subjective and often biased to the chemistries they are familiar with. The difficulty in defining “ideal” metrics for community-wide adoption is to balance the development of accurate models and ones that provide diverse suggestions.

Similar to defining metrics for single-step retrosynthesis predictions, the main obstacle in developing *full* path-planning algorithms is the difficulty in assessing the predicted pathways. Each individual retrosynthetic step can be evaluated as above, with the efficiency of the path search as an additional criterion. Simple metrics such as the existence of a route, the length of the longest linear sequence, or the price of the starting materials are sometimes used for assessment of route planning software but cannot fully capture the complexity of the many desires from different types of chemists. A question that one might ask is whether the models are able to suggest a route that has been previously published. One would not want only published routes to be suggested because a lookup would be sufficient. Since there is a combinatorial space of options for many disconnections (e.g., chloride or iodide instead of bromide), it is not desirable to penalize route planning based on only suggestions appearing in the literature. Another question for path planning is whether the models are able to suggest routes that are chemically feasible. Since scoring the feasibility has yet to be addressed quantitatively and would have its own errors and limitations, the evaluation of different CASP packages, *based on synthetic feasibility*, is difficult.

A major requirement for path planning is the diversity of pathways, but diversity, in terms of full pathways, remains undefined. Route diversity depends not only on the single-step suggestions but also on the method for navigating the full synthetic trees. Diversity could mean the suggestion of many routes, some of which are very similar, but among all the pathways, there are some that employ completely different disconnections. Similar to single-step suggestions, there is no point to diverse route recommendations if they are not feasible, and currently, the only true validation available is to carry out the suggested syntheses.^{14,70} The final important factor for path planning is speed. Speed is dependent on what the stop criteria used for the search is and what one considers to be buyable chemicals (in addition to what transformations are allowed, of course). A trade-off between the speed and quality of routes is often observed, but this could be tuned to the needs of the user.

The relative importance of each of these desires is subjective and highly dependent on the area of chemistry in which one is operating in. Medicinal chemists might like to see routes that share common intermediates that can be elaborated. This delivers a common route that leads to the highest diversity of

target molecules but might not be the optimal route to any individual compound. On the other hand, process chemists might want to see highly convergent routes but would like to visualize many diverse route suggestions for ideation, as they will have more complex considerations that are not captured by computational models. Finally, results need to be available quickly to provide the most value over traditional database searching and manual path planning. Balancing the scoring of retrosynthetic algorithms that encompass the accuracy to assess model performance, the diversity to satisfy the application to chemistry, and the convergency to a difficult multiobjective optimization problem.

In their seminal publication, Segler et al. used double blind A/B testing to evaluate whether chemists preferred published routes over computer generated routes.¹⁵ If the participants of A/B testing span many organizations and departments (i.e., process chemistry, medicinal chemistry, etc.), personal biases may average out. This evaluation, however, does not represent a reproducible or scalable approach to obtaining a quantitative score.

Data, Determining Common Benchmarking, and Evaluation Methods. Of course, machine-learning models are thought to benefit from a higher quantity and richer data. The mechanism that companies and universities use to capture and report data is vital for the further advancement of data-driven methods in synthetic planning. An example of data that are not often recorded in databases is alternate reactions or conditions that have been tested in route to a new compound or natural product. A discussion about the evolution of a route is recorded in the literature reports but does not always get captured when being translated to database entries. This information is highly useful for chemists to determine a strategy in planning routes but is not captured when building models. Another consideration is that databases that record literature reports typically only include positive data with higher yields. Most reaction predictors are trained on successful reactions (USPTO and Reaxys data sets), and as a result, they cannot predict whether a reaction will have a low conversion. Additionally, the full characterization of side or byproducts in reaction mixtures is not often disclosed, due to the high time and costs associated with identifying *all* chemical species. This limits our ability to construct predictive models for reactivity. Finally, there are data that are being captured but not reported such as unpublished catalyst screens or HTE campaigns. However, data capture is increasingly becoming a topic of interest at many companies, and their reporting will hopefully make its way into the public data sets. CASP approaches using expert-encoded rules are less sensitive to data availability than ones using statistical learning because humans can facilitate the generalization of small numbers of reaction precedents to broader rules. Nevertheless, these methods would still benefit from the availability of richer data (e.g., more complete substrate scope tables and byproduct identification) because the expert that encodes the rules would have a better understanding of the reaction.

Retraining machine-learning models on company data has not been thoroughly investigated by all MLPDS member companies. As discussed previously, Eli Lilly only found a modest benefit to including USPTO with their own internal data when training a retrosynthesis model. These results indicate that internal reaction data sets may contain enough representative examples of the main “workhorse” reactions⁸⁰ most often used in medicinal chemistry programs. This brings

up the question: will retraining models on company data simply give suggestions that will reinforce⁸¹ the chemistries that are most popular? The answer depends on the chemistry setting where a CASP tool may be employed.

In a medicinal chemistry program, where accessing a chemical space dissimilar to current molecules is desired, new reaction types might be essential to synthesize nontraditional, increasingly complex target molecules. However, if targets *can* be synthesized through robust chemistries, it is appropriate for CASP to recommend a route with well-established chemistry rather than a creative route with many unknowns. Identifying routes with common chemistry also allows medicinal chemist to more accurately predict timelines to targets by allowing for prioritization of syntheses that can be outsourced versus ones that will need to be executed in-house. Nevertheless, users of CASP tools describe wanting to see more creative recommendations, particularly those working in process chemistry divisions. This is often because there are more complex or subjective considerations of pathway optimality, as discussed above, than the tools can handle. A skilled process chemist may have already considered the obvious disconnections from a step-efficiency perspective and may be looking for a more process-friendly and safer alternative.

Furthermore, if CASP tools are being used in conjunction with automated synthesis platforms, the question of reinforcing reactions⁸¹ may not be as important. For automated synthesis, if a CASP system recommends simple and robust chemistries that have been used frequently within companies, a large burden could be taken off the chemists from having to plan or execute simple chemistry. Even if only a small portion of reactions in a medicinal chemistry program could be automated, a large impact could be made on the timelines for accessing new target molecules. This allows chemists more time to focus on rare chemistries that are key complexity building steps, which facilitates the expansion into new chemical space. As new chemistries are developed, they may be used to further train and refine the CASP models.

Different approaches to synthesis planning exist, and comparisons between different models are not currently standardized. Although current metrics for scoring full synthetic pathways are imperfect, an open access shared benchmarking platform or data set(s) need to be developed so researchers can compare retrosynthetic software and algorithms. Providing a common test set for investigators to benchmark their systems is a nontrivial task. As the field progresses, molecules provided for a test set will have to evolve, because over time, they will be included in databases for training. The underlying distribution of training data evolves with time as well, so it would be likely that the metrics for a common test set on newer models look better due to a higher representation of new reactions and structures in the training data. Additionally, it would be even better to also provide an open data set for training such that the training and test set will be common for every data-driven model that is published. Benchmarking retrosynthetic software is also complicated by the fact that some systems incorporate expert-encoded rules in their algorithms. This means that there is likely an overlap between these coded rules with a test set, whereas pure data-driven methods have a clear separation between the training and test data.⁸² Even with better or more standardized metrics, chemists ultimately will use a program that fits their definition of useful within the scope of chemistry

that they are operating and proves its utility by suggesting routes that are successful in the lab.

Additional Opportunities for CASP. Although discussion has focused on the difficulty of evaluating and benchmarking current synthesis planning tools, advances in many other predictive tasks may have a large impact on medicinal chemistry synthesis workflows. For instance, condition recommendation systems generally focus on predicting known reagent(s) in known or possibly unique combinations. Data for catalytic reactions are very limited in that there are often a lower number of distinct catalysts/ligands proven to be successful. Improvement in machine-learning models will be necessary to handle the many low-data environments that exist in drug discovery. One can envision principles similar to *de novo* molecular design being applied to the generation of unique catalysts/ligands structures, which might raise the probability of success for a given transformation. Advances for the design of large catalytic systems in materials science^{83–85} have already been reported but less have been disclosed in catalysis for the synthesis of small molecule organic compounds. The chemical space of catalytic reactions is often very constrained (in terms of starting materials and reagents), which poses a problem for the generation of data sets that are useful for molecular generation. As with many prediction tasks in medicinal chemistry, the further development of models that can learn from small, constrained data sets is crucial and will likely require new input representations to capture a richer description of the molecular structure. Additional opportunities exist for the prediction of ligands in stereoselective reactions but will require the development of new 3D representations. The final complicating factor to *de novo* design of catalysts/ligands is that often the synthesis and characterization of new catalysts/ligands can consume a significant amount of time. The addition of a multistep synthesis simply for a catalyst/ligand would be prohibitive in most medicinal chemistry programs but may be of interest to academic chemists or process chemists who are highly focused on optimizing each step of the reaction sequence.

A time-consuming step for all synthetic organic research is the characterization of products and side products and the unequivocal assignment of final target molecule structure. Incorrect assignment of target structures can lead to incorrect data for further structure/activity optimization and can even result in patent disputes.⁸⁶ A naïve approach for structural assignment would be to use forward-prediction models to identify possible side products in a reaction. These predictions could be verified with mass spectrometry (MS) or IR to confirm side products in a reaction mixture. However, this method would not be able to distinguish between constitutional isomeric or diastereomeric compounds, which limits its use to reactions that give well-defined isomeric products. Standardized data that are required to elucidate small molecule organic structures and are required for publication include MS, ¹H NMR, and ¹³C NMR and can often require further experiments using 2-dimensional NMR experiments or other NMR active nuclei. When all these data are combined, the structural features can be determined, and it is feasible to train models to predict structures from their spectra. Learning complex nonlinear patterns between the different pieces of data is a perfect application for machine learning. However, data sets containing *all* these experiments are rare. The less common analytical methods (e.g., 2D NMR) tend to be more useful for structural determination. Models could conceivably

be trained on data sets where each compound does not have every data point necessary, and the model could learn to use the higher quality data when needed.

Finally, this discussion of outlooks, standardization of data and models, and dissemination of code can have a large impact on the full pipeline working toward fully autonomous synthesis. A recent review points out many areas where improvement is needed to realize autonomous chemical synthesis both in terms of data/software and hardware.⁸⁷ Among these, the development of data-efficient and interpretable models are discussed. Interpretability of models is important to many users because they want to understand *why* a machine-learning model makes certain predictions. With the multitude of data that could be generated by automated experimentation, the ability to use that data to build predictive models with a low computational overhead and a short time to produce results will enable constructing rounds of experiments that most efficiently reach the goal. Another consideration that is important for machine learning in synthesis planning and automated experimentation is the improvement of uncertainty estimation, particularly in the low-data regime. Improved uncertainty estimates in active learning will produce richer experiments that will reduce time and cost. Finally, evaluation metrics that are specific to the goals of automated synthesis need to be established and standardized, which can focus on molecular targets that test the ability of the models and hardware to reach a new chemical space.

CONCLUSION

The integration of machine-learning models for predictive chemistry into the DMTA cycle is currently underway at companies within and beyond the MLPDS consortium. Companies have begun integration of ASKCOS (and independently developed CASP platforms) in workflows, and computational developers are working closely with synthetic chemists to find the emerging areas where new research will have the greatest impact. For the pace of machine-learning-based CASP research to accelerate, standardized metrics and shared data sets need to be established with a common benchmarking scheme. Fundamental advances to representation, robustness in low-data scenarios, and generalizability will be important for more robust machine-learning-based synthetic tools. Further research into hybrid machine learning and expert-encoded CASP⁸² tools may be able to leverage the most useful aspects of each approach. The impact of machine-learning-based predictive chemistry is already being observed at some companies, and adoption by chemists is on the rise.

Many of the current CASP tools are developed for planning routes using robust reproducible chemistry. The aim of these tools is *not* to suggest only transformations that an experienced chemist could not identify. Rather, particularly with the current machine-learning-based CASP tools, their aim is to enable chemists to lighten the cognitive burden of synthesis planning. If chemists within an organization can each offload planning “easy” routes to computers, even for a relatively small fraction (e.g., 10%) of targets, there is the potential for significant total time savings. With the continued development of machine-learning models for synthesis planning and an increase in chemists’ acceptance of using CASP to lighten their workload, the tools will be improved to fit the needs of different fields of chemistry and to handle synthetic challenges of increasing complexity.

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jmedchem.9b02120>.

Author biographies (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Connor W. Coley – Department of Chemical Engineering, MIT, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0002-8271-8723; Email: ccoley@mit.edu

Klavs F. Jensen – Department of Chemical Engineering, MIT, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0001-7192-580X; Email: kfjensen@mit.edu

Authors

Thomas J. Struble – Department of Chemical Engineering, MIT, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0003-1695-2367

Juan C. Alvarez – Computational and Structural Chemistry, Merck & Co. Inc., Kenilworth, New Jersey 07033, United States; orcid.org/0000-0003-2031-2135

Scott P. Brown – Sunovion Pharmaceuticals Inc., Marlborough, Massachusetts 01752, United States; orcid.org/0000-0003-2797-5409

Milan Chytil – Sunovion Pharmaceuticals Inc., Marlborough, Massachusetts 01752, United States; orcid.org/0000-0001-5280-6741

Justin Cisar – Janssen Research & Development LLC, Spring House, Pennsylvania 19477, United States; orcid.org/0000-0002-8693-9670

Renee L. Desjarlais – Janssen Research & Development LLC, Spring House, Pennsylvania 19477, United States; orcid.org/0000-0002-6752-2976

Ola Engkvist – Hit Discovery, Discovery Sciences, R&D, AstraZeneca 431 83, Mölndal, Sweden; orcid.org/0000-0003-4970-6461

Scott A. Frank – Eli Lilly and Company, Indianapolis, Indiana 46285, United States; orcid.org/0000-0001-6680-9695

Daniel R. Greve – LEO Pharma A/S, DK-2750 Ballerup, Denmark; orcid.org/0000-0001-5790-1131

Daniel J. Griffin – Amgen Inc., Cambridge, Massachusetts 02141, United States; orcid.org/0000-0002-5051-2069

Xinjun Hou – Pfizer Inc., Cambridge, Massachusetts 02139, United States; orcid.org/0000-0002-3473-8455

Jeffrey W. Johannes – Medicinal Chemistry, Early Oncology, Oncology R&D, AstraZeneca, Boston, Massachusetts 02451, United States; orcid.org/0000-0001-7242-6682

Constantine Kreatsoulas – GlaxoSmithKline, Collegeville, Pennsylvania 19426, United States; orcid.org/0000-0002-3654-8209

Brian Lahue – Computational and Structural Chemistry, Merck & Co. Inc., Kenilworth, New Jersey 07033, United States; orcid.org/0000-0002-5399-3603

Miriam Mathea – BASF SE, 67056 Ludwigshafen am Rhein, Germany; orcid.org/0000-0002-3214-1487

Georg Mogk – Bayer GmbH, D-51368 Leverkusen, Germany; orcid.org/0000-0002-3553-5625

Christos A. Nicolaou – Eli Lilly and Company, Indianapolis, Indiana 46285, United States; orcid.org/0000-0002-1466-6992

Andrew D. Palmer – BASF SE, 67056 Ludwigshafen am Rhein, Germany; orcid.org/0000-0003-3863-8491

Daniel J. Price – GlaxoSmithKline, Collegeville, Pennsylvania 19426, United States; orcid.org/0000-0001-7436-0883

Richard I. Robinson – Novartis Institutes for BioMedical Research, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0002-7209-3424

Sebastian Salentin – BASF SE, 67056 Ludwigshafen am Rhein, Germany; orcid.org/0000-0003-0662-2209

Li Xing – WuXi AppTec, Cambridge, Massachusetts 02142, United States; orcid.org/0000-0002-9056-7996

Tommi Jaakkola – Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0002-2199-0379

William H. Green – Department of Chemical Engineering, MIT, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0003-2603-9694

Regina Barzilay – Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, Massachusetts 02139, United States; orcid.org/0000-0002-2921-8201

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.jmedchem.9b02120>

Funding

This work was supported by the Machine Learning for Pharmaceutical Discovery and Synthesis Consortium and the DARPA Make-It program under contract ARO W911NF-16-2-0023.

Notes

The authors declare no competing financial interest.

■ ABBREVIATIONS USED

AI, artificial intelligence; DMTA, design-make-test-analyze; ML, machine learning; CASP, computer aided synthesis planning; MLPDS, Machine Learning for Pharmaceutical Discovery and Synthesis Consortium; ASKCOS, Automated System for Knowledge-Based Continuous Organic Synthesis; SMILES, simplified molecular-input line-entry system; SMARTS, SMILES arbitrary target specification; SMIRKS, a restricted version of reaction SMARTS involving changes in atom-bond patterns; SA, synthetic accessibility; GUI, graphical user interface; MS, mass spectrometry; IR, infrared spectrometry; NMR, nuclear magnetic resonance; USPTO, United States Patent and Trademark Office; UDM, unified data model

■ REFERENCES

- (1) Paul, S. M.; Mytelka, D. S.; Dunwiddie, C. T.; Persinger, C. C.; Munos, B. H.; Lindborg, S. R.; Schacht, A. L. How to improve R&D productivity: the Pharmaceutical Industry's Grand Challenge. *Nat. Rev. Drug Discovery* **2010**, *9* (3), 203–214.
- (2) Murcko, M. A. Envisioning the Future: Medicine in the Year 2050. *Disruptive Sci. Technol.* **2012**, *1* (2), 89–99.
- (3) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *J. Health Econ.* **2016**, *47*, 20–33.
- (4) Wong, C. H.; Siah, K. W.; Lo, A. W. Estimation of Clinical Trial Success Rates and Related Parameters. *Biostatistics* **2019**, *20* (2), 273–286.
- (5) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science* **1969**, *166* (3902), 178–192.
- (6) Hirst, J.; Lim, C.; Jordan, K. D.; Thiel, W.; Judson, P. *Knowledge-Based Expert Systems in Chemistry: Not Counting on Computers; Theoretical and Computational Chemistry Series*; The Royal Society of Chemistry: Cambridge, UK, 2009.

- (7) Cook, A.; Johnson, A. P.; Law, J.; Mirzazadeh, M.; Ravitz, O.; Simon, A. Computer-Aided Synthesis Design: 40 Years On. *Wiley Interdiscip. Rev.: Comput. Mol. Sci.* **2012**, *2* (1), 79–107.
- (8) Ihlenfeldt, W.-D.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angew. Chem., Int. Ed. Engl.* **1996**, *34* (23–24), 2613–2633.
- (9) Todd, M. H. Computer-Aided Organic Synthesis. *Chem. Soc. Rev.* **2005**, *34* (3), 247–266.
- (10) Bøgevig, A.; Federsel, H.-J.; Huerta, F.; Hutchings, M. G.; Kraut, H.; Langer, T.; Low, P.; Oppawsky, C.; Rein, T.; Saller, H. Route Design in the 21st Century: The ICSYNTH Software Tool as an Idea Generator for Synthesis Prediction. *Org. Process Res. Dev.* **2015**, *19* (2), 357–368.
- (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., Int. Ed.* **2016**, *55* (20), 5904–5937.
- (12) Coley, C. W.; Green, W. H.; Jensen, K. F. Machine Learning in Computer-Aided Synthesis Planning. *Acc. Chem. Res.* **2018**, *51* (5), 1281–1289.
- (13) Engkvist, O.; Norrby, P. O.; Selmi, N.; Lam, Y. H.; Peng, Z.; Sherer, E. C.; Amberg, W.; Erhard, T.; Smyth, L. A. Computational Prediction of Chemical Reactions: Current Status and Outlook. *Drug Discovery Today* **2018**, *23* (6), 1203–1218.
- (14) Klucznik, T.; Mikulak-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; Touthkine, A.; Dittwald, P.; Startek, M. P.; Kirkovits, G. J.; Roszak, R.; Adamski, A.; Sieredzińska, B.; Mrksich, M.; Trice, S. L. J.; Grzybowski, B. A. Efficient Syntheses of Diverse, Medically Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem.* **2018**, *4* (3), 522–532.
- (15) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning Chemical Syntheses with Deep Neural Networks and Symbolic AI. *Nature* **2018**, *555* (7698), 604–610.
- (16) Coley, C. W.; Thomas, D. A.; Lummiss, J. A. M.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H.; Hicklin, R. W.; Plehiers, P. P.; Byington, J.; Piotti, J. S.; Green, W. H.; Hart, A. J.; Jamison, T. F.; Jensen, K. F. A Robotic Platform for Flow Synthesis of Organic Compounds Informed by AI Planning. *Science* **2019**, *365* (6453), No. eaax1566.
- (17) Nicolaou, C. A.; Humblet, C.; Hu, H.; Martin, E. M.; Dorsey, F. C.; Castle, T. M.; Burton, K. I.; Hu, H.; Hendle, J.; Hickey, M. J.; Duerksen, J.; Wang, J.; Erickson, J. A. Idea2Data: Toward a New Paradigm for Drug Discovery. *ACS Med. Chem. Lett.* **2019**, *10* (3), 278–286.
- (18) <http://askcos.mit.edu> (accessed Dec 17, 2019).
- (19) <http://mlpds.mit.edu> (accessed Dec 17, 2019).
- (20) <https://github.com/connorcoley/ASKCOS> (accessed Dec 17, 2019).
- (21) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminf.* **2009**, *1* (1), 8.
- (22) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4* (2), 90–98.
- (23) Baber, J. C.; Feher, M. Predicting Synthetic Accessibility: Application in Drug Discovery and Development. *Mini-Rev. Med. Chem.* **2004**, *4* (6), 681–692.
- (24) Sheridan, R. P.; Zorn, N.; Sherer, E. C.; Campeau, L.-C.; Chang, C. Z.; Cumming, J.; Maddess, M. L.; Nantermet, P. G.; Sinz, C. J.; O'Shea, P. D. Modeling a Crowdsourced Definition of Molecular Complexity. *J. Chem. Inf. Model.* **2014**, *54* (6), 1604–1616.
- (25) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58* (2), 252–261.
- (26) Schreck, J. S.; Coley, C. W.; Bishop, K. J. M. Learning Retrosynthetic Planning through Simulated Experience. *ACS Cent. Sci.* **2019**, *5* (6), 970–981.
- (27) Law, J.; Zsoldos, Z.; Simon, A.; Reid, D.; Liu, Y.; Khew, S. Y.; Johnson, A. P.; Major, S.; Wade, R. A.; Ando, H. Y. Route Designer: A Retrosynthetic Analysis Tool Utilizing Automated Retrosynthetic Rule Generation. *J. Chem. Inf. Model.* **2009**, *49* (3), 593–602.
- (28) 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* (6), 2529–2537.
- (29) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **2017**, *23* (25), 5966–5971.
- (30) Dai, H.; Li, C.; Coley, C.; Dai, B.; Song, L. Retrosynthesis Prediction with Conditional Graph Logic Network. In *NIPS; Curran Associates, Inc.*: Red Hook, NY, 2019; pp 8870–8880.
- (31) Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. *ACS Cent. Sci.* **2017**, *3* (10), 1103–1113.
- (32) Karpov, P.; Godin, G.; Tetko, I. V. A Transformer Model for Retrosynthesis. In *Artificial Neural Networks and Machine Learning - ICANN 2019; Workshop and Special Sessions; Springer International Publishing*; pp 817–830.
- (33) Zheng, S.; Rao, J.; Zhang, Z.; Xu, J.; Yang, Y. Predicting Retrosynthetic Reactions Using Self-Corrected Transformer Neural Networks. *J. Chem. Inf. Model.* **2020**, *60* (1), 47–55.
- (34) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Cent. Sci.* **2017**, *3* (12), 1237–1245.
- (35) Kishimoto, A.; Buesser, B.; Chen, B.; Botea, A. Depth-First Proof-Number Search with Heuristic Edge Cost and Application to Chemical Synthesis Planning. In *NIPS; Curran Associates, Inc.*: Red Hook, NY, 2019; pp 7224–7234.
- (36) Marcou, G.; Aires de Sousa, J.; Latino, D. A.; de Luca, A.; Horvath, D.; Rietsch, V.; Varnek, A. Expert System for Predicting Reaction Conditions: The Michael Reaction Case. *J. Chem. Inf. Model.* **2015**, *55* (2), 239–250.
- (37) Nielsen, M. K.; Ahneman, D. T.; Riera, O.; Doyle, A. G. Deoxyfluorination with Sulfonyl Fluorides: Navigating Reaction Space with Machine Learning. *J. Am. Chem. Soc.* **2018**, *140* (15), 5004–5008.
- (38) Segler, M. H. S.; Waller, M. P. Modelling Chemical Reasoning to Predict and Invent Reactions. *Chem. - Eur. J.* **2017**, *23* (25), 6118–6128.
- (39) Gao, H.; Struble, T. J.; Coley, C. W.; Wang, Y.; Green, W. H.; Jensen, K. F. Using Machine Learning To Predict Suitable Conditions for Organic Reactions. *ACS Cent. Sci.* **2018**, *4* (11), 1465–1476.
- (40) Ley, S. V.; Fitzpatrick, D. E.; Ingham, R. J.; Myers, R. M. Organic Synthesis: March of the Machines. *Angew. Chem., Int. Ed.* **2015**, *54* (11), 3449–3464.
- (41) Sans, V.; Cronin, L. Towards Dial-a-Molecule by Integrating Continuous Flow, Analytics and Self-Optimisation. *Chem. Soc. Rev.* **2016**, *45* (8), 2032–2043.
- (42) Gromski, P. S.; Henson, A. B.; Granda, J. M.; Cronin, L. How to Explore Chemical Space Using Algorithms and Automation. *Nat. Rev. Chem.* **2019**, *3* (2), 119–128.
- (43) Zhou, Z.; Li, X.; Zare, R. N. Optimizing Chemical Reactions with Deep Reinforcement Learning. *ACS Cent. Sci.* **2017**, *3* (12), 1337–1344.
- (44) Simm, G. N.; Reiher, M. Error-Controlled Exploration of Chemical Reaction Networks with Gaussian Processes. *J. Chem. Theory Comput.* **2018**, *14* (10), 5238–5248.
- (45) Hase, F.; Roch, L. M.; Kreisbeck, C.; Aspuru-Guzik, A. Phoenix: A Bayesian Optimizer for Chemistry. *ACS Cent. Sci.* **2018**, *4* (9), 1134–1145.
- (46) Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of Organic Reaction Outcomes Using Machine Learning. *ACS Cent. Sci.* **2017**, *3* (5), 434–443.

- (47) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2* (10), 725–732.
- (48) Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. In *NIPS*; Curran Associates, Inc.: Red Hook, NY, 2017; pp 2607–2616.
- (49) Coley, C. W.; Jin, W.; Rogers, L.; Jamison, T. F.; Jaakkola, T. S.; Green, W. H.; Barzilay, R.; Jensen, K. F. A Graph-Convolutional Neural Network Model for the Prediction of Chemical Reactivity. *Chem. Sci.* **2019**, *10* (2), 370–377.
- (50) Nam, J.; Kim, J. Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions. *ArXiv*. **2016** ArXiv161209529.
- (51) Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. Found in Translation: Predicting Outcomes of Complex Organic Chemistry Reactions Using Neural Sequence-to-Sequence Models. *Chem. Sci.* **2018**, *9* (28), 6091–6098.
- (52) 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* (9), 1572–1583.
- (53) Walters, W. P. Virtual Chemical Libraries. *J. Med. Chem.* **2019**, *62* (3), 1116–1124.
- (54) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Alga, E.; Tolmacheva, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566* (7743), 224–229.
- (55) Liu, Q.; Watterson, S. H.; Ahmad, S. Indole Carboxamide Compounds. U.S. Patent 9,688,629, July 27, 2017.
- (56) Watterson, S. H.; Liu, Q.; Beaudoin Bertrand, M.; Batt, D. G.; Li, L.; Pattoli, M. A.; Skala, S.; Cheng, L.; Obermeier, M. T.; Moore, R.; Yang, Z.; Vickery, R.; Elzinga, P. A.; Discenza, L.; D'Arienzo, C.; Gillooly, K. M.; Taylor, T. L.; Pulicchio, C.; Zhang, Y.; Heimrich, E.; McIntyre, K. W.; Ruan, Q.; Westhouse, R. A.; Catlett, I. M.; Zheng, N.; Chaudhry, C.; Dai, J.; Galella, M. A.; Tebben, A. J.; Pokross, M.; Li, J.; Zhao, R.; Smith, D.; Rampulla, R.; Allentoff, A.; Wallace, M. A.; Mathur, A.; Salter-Cid, L.; Macor, J. E.; Carter, P. H.; Fura, A.; Burke, J. R.; Tino, J. A. Discovery of Branebrutinib (BMS-986195): A Strategy for Identifying a Highly Potent and Selective Covalent Inhibitor Providing Rapid in Vivo Inactivation of Bruton's Tyrosine Kinase (BTK). *J. Med. Chem.* **2019**, *62* (7), 3228–3250.
- (57) Liu, Z.; Chen, L.; Yu, P.; Zhang, Y.; Fang, B.; Wu, C.; Luo, W.; Chen, X.; Li, C.; Liang, G. Discovery of 3-(Indol-5-Yl)-Indazole Derivatives as Novel Myeloid Differentiation Protein 2/Toll-like Receptor 4 Antagonists for Treatment of Acute Lung Injury. *J. Med. Chem.* **2019**, *62* (11), 5453–5469.
- (58) Lee, A. A.; Yang, Q.; Sresht, V.; Bolgar, P.; Hou, X.; Klug-McLeod, J. L.; Butler, C. R. Molecular Transformer Unifies Reaction Prediction and Retrosynthesis across Pharma Chemical Space. *Chem. Commun.* **2019**, *55* (81), 12152–12155.
- (59) 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* (1), 154–168.
- (60) Nicolaou, C. A.; Watson, I. A.; LeMasters, M.; Masquelin, T.; Wang, J. Context Aware Data-Driven Retrosynthetic Analysis. Unpublished Results.
- (61) Schneider, N.; Lowe, D. M.; Sayle, R. A.; Landrum, G. A. Development of a Novel Fingerprint for Chemical Reactions and Its Application to Large-Scale Reaction Classification and Similarity. *J. Chem. Inf. Model.* **2015**, *55* (1), 39–53.
- (62) Li, J.; Eastgate, M. D. Making Better Decisions during Synthetic Route Design: Leveraging Prediction to Achieve Greenness-by-Design. *React. Chem. Eng.* **2019**, *4* (9), 1595–1607.
- (63) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting Reaction Performance in C-N Cross-Coupling Using Machine Learning. *Science* **2018**, *360* (6385), 186–190.
- (64) Tria, G. S.; Abrams, T.; Baird, J.; Burks, H. E.; Firestone, B.; Gaither, L. A.; Hamann, L. G.; He, G.; Kirby, C. A.; Kim, S.; Lombardo, F.; Macchi, K. J.; McDonnell, D. P.; Mishina, Y.; Norris, J. D.; Nunez, J.; Springer, C.; Sun, Y.; Thomsen, N. M.; Wang, C.; Wang, J.; Yu, B.; Tiong-Yip, C.-L.; Peukert, S. Discovery of LSZ102, a Potent, Orally Bioavailable Selective Estrogen Receptor Degradator (SERD) for the Treatment of Estrogen Receptor Positive Breast Cancer. *J. Med. Chem.* **2018**, *61* (7), 2837–2864.
- (65) Gallou, F.; Parmentier, M.; Zhou, J.; Guo, P. Reaction Medium Containing Water-Surfactant Mixture. WO/2017/168303, Oct 5, 2017.
- (66) Gallou, F.; Isley, N. A.; Ganic, A.; Onken, U.; Parmentier, M. Surfactant Technology Applied toward an Active Pharmaceutical Ingredient: More than a Simple Green Chemistry Advance. *Green Chem.* **2016**, *18* (1), 14–19.
- (67) Parmentier, M.; Wagner, M. K.; Magra, K.; Gallou, F. Selective Amidation of Unprotected Amino Alcohols Using Surfactant-in-Water Technology: A Highly Desirable Alternative to Reprotoxic Polar Aprotic Solvents. *Org. Process Res. Dev.* **2016**, *20* (6), 1104–1107.
- (68) Gallou, F.; Guo, P.; Parmentier, M.; Zhou, J. A General and Practical Alternative to Polar Aprotic Solvents Exemplified on an Amide Bond Formation. *Org. Process Res. Dev.* **2016**, *20* (7), 1388–1391.
- (69) Watson, I. A.; Wang, J.; Nicolaou, C. A. A Retrosynthetic Analysis Algorithm Implementation. *J. Cheminf.* **2019**, *11* (1), 1.
- (70) Vidler, L. R.; Baumgartner, M. P. Creating a Virtual Assistant for Medicinal Chemistry. *ACS Med. Chem. Lett.* **2019**, *10* (7), 1051–1055.
- (71) Parry, D. M. Closing the Loop: Developing an Integrated Design, Make, and Test Platform for Discovery. *ACS Med. Chem. Lett.* **2019**, *10* (6), 848–856.
- (72) Pant, S. M.; Mukonoweshuro, A.; Desai, B.; Ramjee, M. K.; Selway, C. N.; Tarver, G. J.; Wright, A. G.; Birchall, K.; Chapman, T. M.; Tervonen, T. A.; Klefström, J. Design, Synthesis, and Testing of Potent, Selective Hepsin Inhibitors via Application of an Automated Closed-Loop Optimization Platform. *J. Med. Chem.* **2018**, *61* (10), 4335–4347.
- (73) Desai, B.; Dixon, K.; Farrant, E.; Feng, Q.; Gibson, K. R.; van Hoorn, W. P.; Mills, J.; Morgan, T.; Parry, D. M.; Ramjee, M. K.; Selway, C. N.; Tarver, G. J.; Whitlock, G.; Wright, A. G. Rapid Discovery of a Novel Series of Abl Kinase Inhibitors by Application of an Integrated Microfluidic Synthesis and Screening Platform. *J. Med. Chem.* **2013**, *56* (7), 3033–3047.
- (74) Godfrey, A. G.; Masquelin, T.; Hemmerle, H. A Remote-Controlled Adaptive Medchem Lab: An Innovative Approach to Enable Drug Discovery in the 21st Century. *Drug Discovery Today* **2013**, *18* (17), 795–802.
- (75) Buitrago Santanilla, A.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L.-C.; Schneeweis, J.; Berritt, S.; Shi, Z.-C.; Nantermet, P.; Liu, Y.; Helmy, R.; Welch, C. J.; Vachal, P.; Davies, I. W.; Cernak, T.; Dreher, S. D. Nanomole-Scale High-Throughput Chemistry for the Synthesis of Complex Molecules. *Science* **2015**, *347* (6217), 49.
- (76) Perera, D.; Tucker, J. W.; Brahmabhatt, S.; Helal, C. J.; Chong, A.; Farrell, W.; Richardson, P.; Sach, N. W. A Platform for Automated Nanomole-Scale Reaction Screening and Micromole-Scale Synthesis in Flow. *Science* **2018**, *359* (6374), 429.
- (77) Tu, N. P.; Dombrowski, A. W.; Goshu, G. M.; Vasudevan, A.; Djuric, S. W.; Wang, Y. High-Throughput Reaction Screening with Nanomoles of Solid Reagents Coated on Glass Beads. *Angew. Chem., Int. Ed.* **2019**, *58* (24), 7987–7991.
- (78) Thakkar, A.; Bjerrum, E. J.; Engkvist, O.; Reymond, J.-L. Neural Network Guided Tree-Search Policies for Synthesis Planning. In *Artificial Neural Networks and Machine Learning - ICANN 2019: Workshop and Special Sessions*; Tetko, I. V., Kůrková, V., Karpov, P., Theis, F., Eds.; Lecture Notes in Computer Science; Springer International Publishing: Cham, Switzerland, 2019; pp 721–724.

- (79) Thakkar, A.; Selmi, N.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. 'Ring Breaker': Assessing Synthetic Accessibility of the Ring System Chemical Space. *ChemRxiv Prepr.* **2019**, No. 2.
- (80) Roughley, S. D.; Jordan, A. M. The Medicinal Chemist's Toolbox: An Analysis of Reactions Used in the Pursuit of Drug Candidates. *J. Med. Chem.* **2011**, *54* (10), 3451–3479.
- (81) Jordan, A. M. Artificial Intelligence in Drug Design: The Storm Before the Calm? *ACS Med. Chem. Lett.* **2018**, *9* (12), 1150–1152.
- (82) Badowski, T.; Gajewska, E. P.; Molga, K.; Grzybowski, B. A. Synergy Between Expert and Machine-Learning Approaches Allows for Improved Retrosynthetic Planning. *Angew. Chem., Int. Ed.* **2020**, *59* (2), 725–730.
- (83) Freeze, J. G.; Kelly, H. R.; Batista, V. S. Search for Catalysts by Inverse Design: Artificial Intelligence, Mountain Climbers, and Alchemists. *Chem. Rev.* **2019**, *119* (11), 6595–6612.
- (84) Goldsmith, B. R.; Esterhuizen, J.; Liu, J.-X.; Bartel, C. J.; Sutton, C. Machine Learning for Heterogeneous Catalyst Design and Discovery. *AIChE J.* **2018**, *64* (7), 2311–2323.
- (85) Li, Z.; Wang, S.; Xin, H. Toward Artificial Intelligence in Catalysis. *Nat. Catal.* **2018**, *1* (9), 641–642.
- (86) Borman, S. Dispute over the Legal Rights to an Anticancer Agent Continues. *Chemical & Engineering News.* **2017**, *95* (7), 7.
- (87) Coley, C. W.; Eyke, N. S.; Jensen, K. F. Autonomous Discovery in the Chemical Sciences Part II: Outlook. *Angew. Chem., Int. Ed.* **2019**. (accessed 2019-12-17).