ACCOUNTS of chemical research

Article

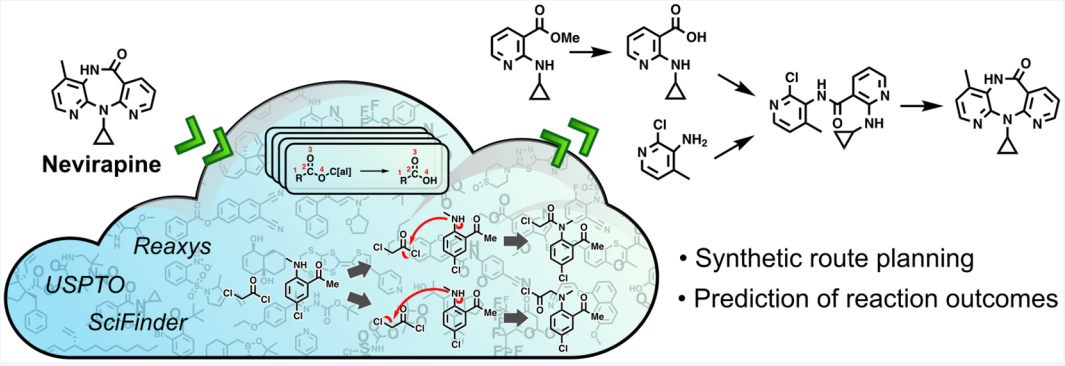
Cite This: Acc. Chem. Res. 2018, 51, 1281−1289

pubs.acs.org/accounts

Machine Learning in Computer-Aided Synthesis Planning

Connor W. Coley, William H. Green,* and Klavs F. Jensen*

Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts 02139, United States



CONSPECTUS: Computer-aided synthesis planning (CASP) is focused on the goal of accelerating the process by which chemists decide how to synthesize small molecule compounds. The ideal CASP program would take a molecular structure as input and output a sorted list of detailed reaction schemes that each connect that target to purchasable starting materials via a series of chemically feasible reaction steps. Early work in this field relied on expert-crafted reaction rules and heuristics to describe possible retrosynthetic disconnections and selectivity rules but suffered from incompleteness, infeasible suggestions, and human bias. With the relatively recent availability of large reaction corpora (such as the United States Patent and Trademark Office (USPTO), Reaxys, and SciFinder databases), consisting of millions of tabulated reaction examples, it is now possible to construct and validate purely data-driven approaches to synthesis planning. As a result, synthesis planning has been opened to machine learning techniques, and the field is advancing rapidly.

In this Account, we focus on two critical aspects of CASP and recent machine learning approaches to both challenges. First, we discuss the problem of retrosynthetic planning, which requires a recommender system to propose synthetic disconnections starting from a target molecule. We describe how the search strategy, necessary to overcome the exponential growth of the search space with increasing number of reaction steps, can be assisted through a learned synthetic complexity metric. We also describe how the recursive expansion can be performed by a straightforward nearest neighbor model that makes clever use of reaction data to generate high quality retrosynthetic disconnections. Second, we discuss the problem of anticipating the products of chemical reactions, which can be used to validate proposed reactions in a computer-generated synthesis plan (i.e., reduce false positives) to increase the likelihood of experimental success. While we introduce this task in the context of reaction validation, its utility extends to the prediction of side products and impurities, among other applications. We describe neural network-based approaches that we and others have developed for this forward prediction task that can be trained on previously published experimental data.

Machine learning and artificial intelligence have revolutionized a number of disciplines, not limited to image recognition, dictation, translation, content recommendation, advertising, and autonomous driving. While there is a rich history of using machine learning for structure−activity models in chemistry, it is only now that it is being successfully applied more broadly to organic synthesis and synthesis design. As reported in this Account, machine learning is rapidly transforming CASP, but there are several remaining challenges and opportunities, many pertaining to the availability and standardization of both data and evaluation metrics, which must be addressed by the community at large.

1. INTRODUCTION

1.1. What Is Synthesis Planning?

Synthesis planning is the process of determining *how* to synthesize a chemical compound from available starting materials through a series of chemically feasible reaction steps. *Retrosynthetic analysis*, formalized in the 1960s by E. J. Corey,[1] approaches this problem in reverse: beginning with the product and choosing suitable disconnections recursively. Retrosynthesis is particularly conducive to computer assistance because it requires searching a large space of possible "moves" starting from a particular state. Computational approaches date back to Corey's work on the development of LHASA,[2] with the goal of suggesting synthetic pathways to chemist users.

Received: February 28, 2018
Published: May 1, 2018

DOI: 10.1021/acs.accounts.8b00087
Acc. Chem. Res. 2018, 51, 1281−1289

1.2. Anatomy of a CASP Tool

Computer-assisted synthesis planning (CASP) tools generally consist of five major components:

i. A template library containing the rules by which disconnections are proposed.
ii. A recursive template application engine that generates candidate reactants for target product molecules.
iii. A database containing compounds that do not need to be expanded retrosynthetically (e.g., are commercially available).
iv. A strategy to guide the retrosynthetic search toward chemicals in that database.
v. A method for single-step or pathway-level scoring, for example, a preference for fewer synthetic steps.

An example workflow is shown in Figure 1. While much of this framework was established decades ago,[1,2] significant improvements have been realized in recent years by leveraging large reaction databases and advancements in data science and machine learning.

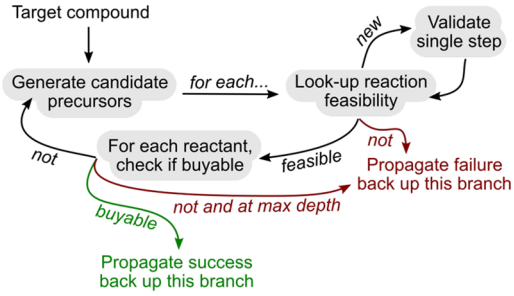


Figure 1. Example recursive workflow for synthesis planning.

1.3. Role of Machine Learning

The power of machine learning techniques is in their ability to approximate complex functions where the exact relationship between input and output variables is not easily codified. Problems previously tackled using expert systems have been revolutionized by machine learning, including image recognition, content recommendation, and machine translation. This is generally attributed to a combination of improvements in hardware capability and in data availability.

1.4. Focus

In this Account, we focus on two major domains within organic synthesis where machine learning can be and has been applied:

i. Retrosynthesis, synthetic route planning
ii. Forward synthesis, prediction of reaction outcomes

Machine learning is more broadly useful for inferring nonobvious relationships within high-dimensional data, which is increasingly relevant to chemistry as the field's ability to rapidly generate high-fidelity experimental data improves.[3] Some of the topics discussed below have been mentioned in previous perspectives and reviews on machine learning in synthesis planning.[4−6]

Due to space constraints, we will *not* describe in detail the quintessential cheminformatics problem of developing quantitative structure−activity/property relationship (QSAR/QSPR) models,[7] to which machine learning is routinely applied as a regression tool[8,9] and, more recently, to learn representations directly from molecular structures.[10−13] Several important aspects of QSAR/QSPR are described in ref 14.

2. ROUTE PLANNING

2.1. Introduction

The history of computer-assisted synthesis planning is well-reviewed.[15−20] Here, rather than provide an exhaustive recapitulation of historical systems, we summarize contemporary strategies grouped into three categories: template library-based, template-free, and focused template application (Figure 2b−d).

2.2. Template Library-Based

Template library based retrosynthesis involves matching generalized reaction rules to target molecules to yield one or more candidate precursors. Early programs required chemists to manually codify these rules using user-unfriendly syntax,[2,21] resulting in often-incomplete template databases able to predict a limited set of chemistries. With sufficient time investment, however, this manual approach can be made to cover much of known chemistry as in the commercial program Chematica.[22]

Contemporary approaches use algorithmic template extraction from atom-mapped reaction examples (Figure 2a).[23−25] An extracted rule must contain atoms that change connectivity, but the degree to which auxiliary atoms are included is flexible. There is an inevitable trade-off between specificity and speed: including too many neighboring atoms leads to large, poorly generalized template libraries, which are computationally expensive to apply; too few neglects the necessary context and leads to unfeasible disconnections.

We find that using simple heuristics provides an appropriate balance. Atoms adjacent to the reaction center are included if they are terminal, are required for unambiguous specification of chirality, or belong to substructures known to influence reactivity (e.g., adjacency to a carbonyl).[26] From 12.5 million single-step reactions from Reaxys, we extract 2.5 million templates, although just 100 000 are observed 10+ times. In practice, rare templates are excluded to reduce computational expense.

Application of the full template library to a target produces candidate precursors, often numbering in the 100s or 1000s. If no precursors are commercially available, they must themselves be expanded; this continues recursively until a fully buyable path is found or some maximum depth is reached (cf. Figure 1 without validation). Careful handling of stereochemistry is required for faithful preservation, inversion, or destruction of tetrahedral centers and cis/trans chirality, for example, using the open-source RDChiral[27] wrapper for RDKit.[28]

To avoid combinatorial explosion, recursive expansion must focus on only the most promising disconnections that yield easily synthesizable compounds. Numerous metrics attempt to quantify the complexity of molecular structures.[29,30] One very crude metric is the length of a molecule's SMILES representation[31] raised to the three-halves power ($\mathrm{SMILES}^{3/2}$), which favors dividing molecules into the smallest possible components as in a convergent synthesis. Chematica uses user-definable Chemical Scoring Functions (CSFs): algebraic functions of molecular parameters (e.g., $2N_{\mathrm{atoms}}^{3/2} + N_{\mathrm{ringbonds}}^{3/2} + 2N_{\mathrm{chiral}}^{2}$).[22] The SA_Score[32] is a more sophisticated fragment-contribution method that scores rare structural motifs as complex. An alternative to these state-agnostic metrics is the use of proof-number search.[33]

Unfortunately, these heuristics are not suited to retrosynthesis due to the nonequivalence of *structural* complexity and *synthetic* complexity. A large multifunctional molecule may appear hard to synthesize yet be easily produced from available

DOI: 10.1021/acs.accounts.8b00087
Acc. Chem. Res. 2018, 51, 1281−1289

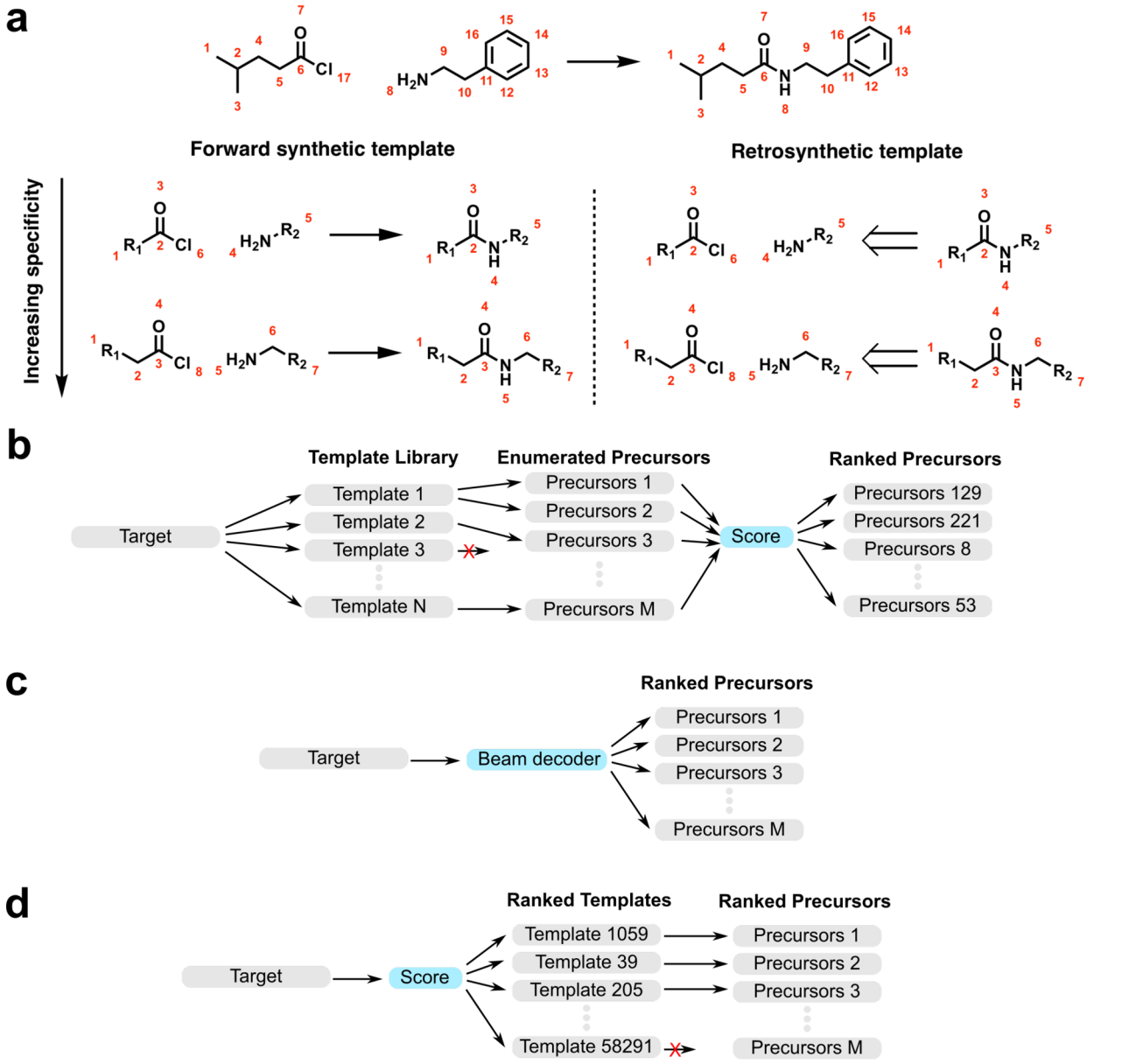


Figure 2. (a) Algorithmic extraction of a reaction template; retrosynthetic expansion using (b) template library-based enumeration, (c) template-free enumeration, and (d) focused template application. Not all templates yield precursors, but some yield multiple.

building blocks. Moreover, retro-deprotections are always considered "uphill" moves, because the molecule becomes larger, when they may actually be "downhill", that is, productive.

To account for this, we developed the SCScore[34] as a data-driven metric designed to describe real syntheses, modeled after the premise that the products of published reactions should, on average, be more *synthetically* complex than each of their reactants. Roughly 12 million single-product reactions from Reaxys were divided into 22 million (reactant, product) pairs. A feedforward neural network model was constructed to compute a synthetic complexity score from an extended-connectivity fingerprint (ECFP[35]) and trained using a hinge loss function to promote separation between scores assigned to the molecules in each (reactant, product) pair. Through this pairwise ranking objective, the model implicitly learns what structures and motifs are more prevalent as reactants.

The advantage of the SCScore is best illustrated through examination of linear syntheses. In the comparison shown in Figure 3 using four different metrics (SMILES$^{3/2}$ score, Chematica-like CSF, SA_Score,[32] and SCScore), only the SCScore perceives the desired monotonic increase in complexity; for 40 additional comparisons, not all as favorable, see ref 34.

2.3. Template-Free

While the dominant paradigm of retrosynthetic enumeration is based on templates, template-free alternatives are attractive for several reasons. First, calculating subgraph isomorphism is computationally expensive, especially for large libraries. Second, the degree of template generality/specificity can lead to either low-quality or incomplete recommendations. And third, template-based methods cannot propose fundamentally novel disconnections.

Retrosynthesis requires the prediction of reactant molecules. Such structured predictions where the output is not natively

DOI: 10.1021/acs.accounts.8b00087
Acc. Chem. Res. 2018, 51, 1281−1289

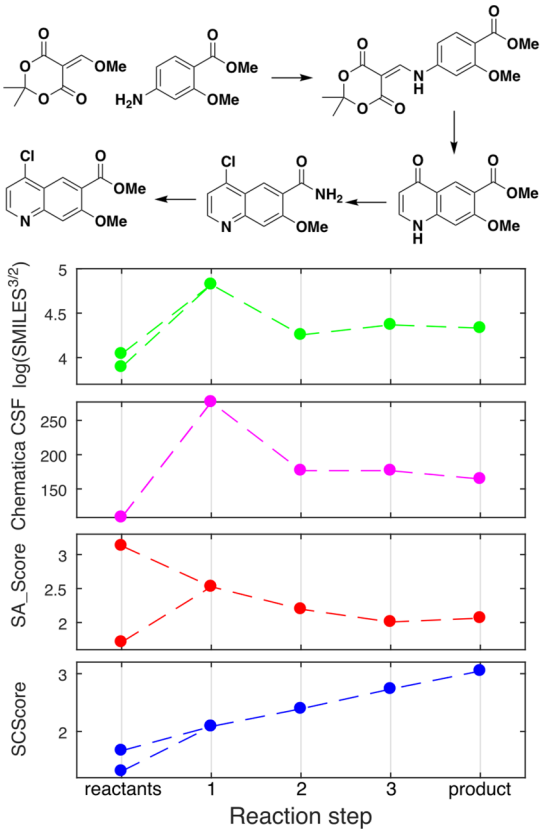


Figure 3. Use of different metrics to analyze the synthesis of a precursor to lenvatinib.[36] Only the SCScore correctly perceives a monotonic increase in complexity.

numerical can require specialized neural network architectures, but molecule prediction can be recast as sequence prediction via SMILES representations; generating sequences using recurrent networks is commonplace. Liu et al. describe a sequence-to-sequence (seq2seq) model that converts a product SMILES to reactant(s) SMILES.[37] While no significant benefit in accuracy, the ability to propose the true, recorded reactants as high-ranked suggestions, over template library application was reported (57.0% versus 59.1% top-5 accuracy), neural translation *has* been successfully applied to the inverse problem of reaction prediction,[38] suggesting that straightforward model improvements might improve its performance.

2.4. Focused Template Application

Rather than forego templates entirely, focused template methods select *relevant* templates to apply, mitigating the computational expense of full library application. Segler and Waller employed a neural network to score template relevance based on molecular fingerprints.[39] This focused expansion policy was later used in a Monte Carlo Tree Search (MCTS) framework for full pathway design, which enables recommendations to be generated with impressive speed.[40]

This does not overcome the question of generalization during template extraction, however, nor the need to exclude rare templates (as the data are too sparse to train over the full set). To overcome these challenges, we devised a strategy for retrosynthetic expansion based on the concept of molecular similarity reminiscent of a nearest neighbor model.[41]

Consider how a chemist might approach the synthesis of a new molecule: they might search Reaxys or SciFinder to find syntheses of similar compounds and determine if those strategies are applicable to their target. To automate this, we first calculate the structural similarity (using ECFP fingerprints and the Tanimoto distance) between the target and all known products to recall relevant precedents from a reaction corpus. A highly generalized template is extracted from the precedent, with no attempt to incorporate surrounding context, and applied to the target. For any resulting precursor sets, their structural similarity to the reaction precedent's reactants is calculated and multiplied by the product similarity to provide an overall score. This overall similarity score quantifies how strongly the precedent reaction supports the proposed reaction and is used for ranking suggestions.

Similarity-based scoring implicitly considers potential functional group conflicts or missing activating groups. Further, recalling individual precedents makes use of *all* known reactions, not just those corresponding to popular templates. This approach outperforms the template-free seq2seq model[37] (81.2% versus 57.0% top-5 accuracy) and extends to full route planning when applied recursively.

3. PREDICTION OF REACTION OUTCOMES

3.1. Introduction

Advancements in data-driven retrosynthetic planning have helped avoid manual curation of template libraries and improve computational speed. An additional benefit is increased confidence in template applicability without explicitly encoding reactivity conflicts. An important goal of recommender systems is this reduction of false positives: proposed reactions incorrectly thought to be chemically feasible.

To validate retrosynthetic suggestions, one can solve the inverse problem of forward synthesis: *given specific experimental parameters (reactants, reagents, catalysts, solvent, concentrations, temperature, time, etc.), what is the product distribution?* The absence of detailed concentration information in reaction databases necessitates simplification to only reactants and reagents, perhaps also including catalysts, solvents, and temperature. Note that these simplified problems are underdetermined, although one can assume reactions are run under implicitly defined "standard conditions". Another limitation, albeit understandable, is the absence of side-product information; prediction of the full product distribution must be recast as prediction of the major (recorded with >50% yield) product.

The following subsections categorize applications of machine learning to organic reaction prediction by their problem formulations. We do not describe early approaches that relied heavily on expert heuristics to define possible mechanistic reactions[42−45] nor models targeting one specific reaction class,[46,47] which are structurally equivalent to regression models.

3.2. Classifying Reaction Feasibility

One approach is to estimate reaction feasibility without explicit enumeration or consideration of side reactions. Segler et al. describe such a neural network model that classifies reactions as true or false based on their fingerprint representations, trained using true experimental data augmented by synthetic negative data.[40]

3.3. Predicting Mechanistic Steps

The Baldi group has approached reaction prediction from a mechanistic perspective.[48−50] Their ReactionPredictor identifies electron sources and sinks, enumerates possible inter-

DOI: 10.1021/acs.accounts.8b00087
Acc. Chem. Res. 2018, 51, 1281−1289

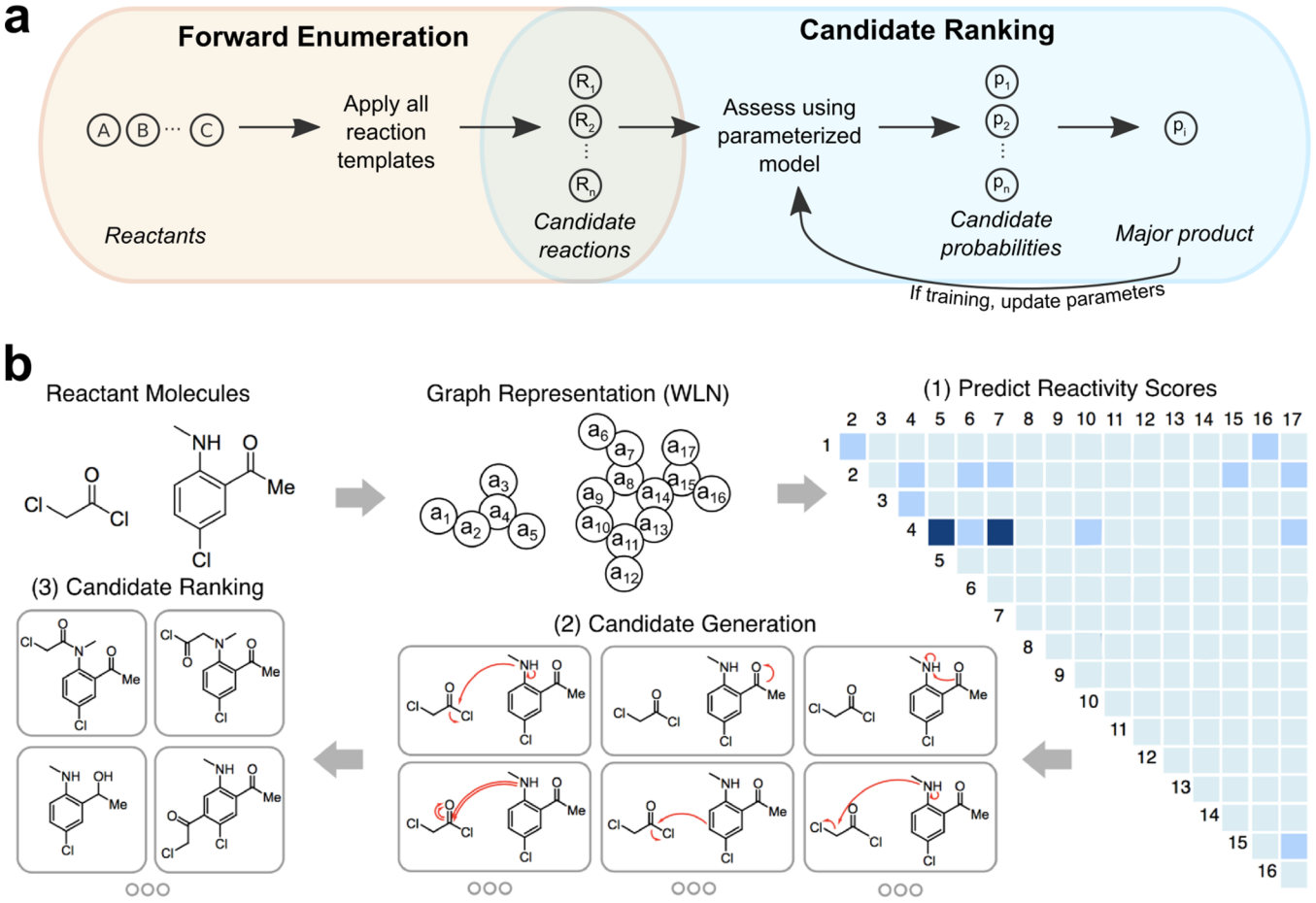


Figure 4. Two approaches to predicting outcomes of organic reactions: (a) combined template-based enumeration and neural network-based candidate ranking; (b) learned pairwise atom reactivities for focused enumeration prior to ranking. Reproduced with permission from ref 52, Copyright 2017 American Chemical Society, and ref 53 Courtesy of Wengong Jin, MIT.

actions, and ranks those interactions using graph-based representations of molecules with pseudomolecular orbital considerations. An openly acknowledged limitation is the need to manual encode mechanistic rules for artificial data generation.

3.4. Ranking Templates

Applying a library of synthetic templates generates many products; provided that the major product is among those, only one template is needed to propose it.

Wei et al. established a proof-of-concept for using machine learning to predict template applicability.[51] Their model, given reactants and reagents, predicts which of 16 rules is most relevant, using simulated data generated by these rules. Segler and Waller extend this approach to experimental data from Reaxys; each reactant fingerprint yields a probability distribution over a library of ≤8720 algorithmically extracted templates.[39] Strictly speaking, neither approach directly predicts the major product, as application of the highest-ranked template may yield several nonequivalent product species, for example, when site selectivity is ambiguous as in a halogenation where multiple aromatic C−H bonds could be activated.

3.5. Ranking Products

Our work on forward prediction began with a hybrid system that directly predicts product(s) of chemical reactions, rather than templates.[52] The model framework combines template-based forward enumeration and machine learning-based candidate ranking. Importantly, this two-step approach (Figure 4a) overcomes the literature bias toward positive reaction examples. For reaction A + B → C recorded with above-50% yield, we know that D, E, etc. were produced to a lesser extent than C. The alternate outcomes that most help the model learn chemical reactivity are those that seem chemically plausible. In this workflow, applying reaction templates defines the *scope* of chemically valid outcomes; they are overgeneralized to increase product coverage at the expense of specificity.

Scoring candidate reactions is complicated by the lack of an established "best" numerical representation. Conceptually, the likelihood should be a function of the corresponding structural changes. Due to limited performance using standard fingerprints (perhaps due to feature obfuscation through hashing), we devised an "edit-based" representation to enable richer, more explicit encoding of prior chemical knowledge. Reactions are represented by atom- and bond-level features of atoms gaining/losing hydrogen atoms and pairs of atoms gaining/losing bonds. Features include structural information (e.g., atomic number, aromaticity, degree) as well as rapidly calculable geometric and electronic features (e.g., estimated partial charge, surface area contribution). These feature vectors constitute the inputs to a feedforward neural network model, which embeds each edit separately before sum-pooling their latent representations and further transforming that combined representation into one aggregated score. Scores of all candidate outcomes are converted to probabilities via a softmax

DOI: 10.1021/acs.accounts.8b00087
Acc. Chem. Res. 2018, 51, 1281−1289

activation. The model was trained to maximize the log-probability assigned to the true (recorded) major product. In the initial demonstration on 15000 United States Patent and Trademark Office (USPTO) reactions augmented by synthetic negative examples obtained via template application, this combined two-step approach predicted the recorded product with 72% accuracy (87% top-3 accuracy, 91% top-5) in a 5-fold cross-validation.

This study was the first large-scale demonstration of product prediction using experimental data, but predictions cannot be made outside the template library scope, and using templates for data augmentation limits its scalability.

3.6. Generating Products

To overcome the drawbacks of the hybrid template/neural network approach, our next study forwent templates and used a trainable model for candidate enumeration in addition to ranking (Figure 4b).[53] Molecules are represented as attributed graphs, with atom- and bond-level features as before. Atom features are iteratively embedded using a Weisfeller–Lehman Network (WLN),[54] a graph convolutional neural network, which incorporates information from neighboring atoms into each atom's representation. A global attention mechanism, a concept originally developed for natural language processing, accounts for effects of distant atoms as is needed to recognize, for example, a reagent. Pairwise reactivity scores are calculated for each atom pair based on their feature vectors, quantifying the propensity of that (atom, atom) interaction to change. The model is trained to predict which pairs of atoms belong to the reaction center. The top pairs are used to combinatorially enumerate possible bond changes; structural and valence requirements restrict candidates to chemically valid molecules, which are each scored by a Weisfeller–Lehman Difference Network (WLDN) using differences between embedded atom representations of products and reactants as the basis for scoring.

The fully learned approach achieves significantly higher accuracy than the template-based approach and operates orders of magnitude faster, enabling its application to a larger data set of ca. 400 000 reactions. A human benchmarking study showed that the model performs on par with graduate and postdoctoral synthetic chemists.

As described for retrosynthetic prediction, generating molecules can be recast as generating SMILES sequences. Applying translation models to reaction prediction was first demonstrated by Nam and Kim, who trained a sequence-to-sequence model to predict product SMILES strings from reactant SMILES using a combination of real (USPTO) and synthetic (generated from rules) data, tested only on textbook problems.[55] That paradigm was later applied to the larger and more complex USPTO data set, with the important addition of an attention mechanism to account for long-range dependencies.[38] Quantitative performance was nearly identical to Jin et al.,[53] although 15 times as many model parameters were required, demonstrating that one can ignore prior chemical knowledge and apply a purely empirical language model to forward prediction.

Language-based models have unique limitations that warrant mention. Mistakes may be linguistic, rather than chemical, so the second-ranked outcome is not necessarily the next-most chemically plausible. Moreover, model predictions are of the next SMILES character/token to append to the current sequence, not of reactive atoms and bonds, so inferring quantitative reactivity trends requires additional analysis and atom-to-atom mapping; in contrast, template-based and graph-based models can be directly analyzed via scores assigned to different modes of reactivity or (atom, atom) interactions to understand what is perceived as chemically likely.

4. OUTLOOK

4.1. Data Availability

Tabulated reaction data is limited. An important collection is the open-source USPTO data set originally extracted by Daniel Lowe, which now contains reactions through September 2016.[56] Elsevier's commercial Reaxys database comprises a much larger collection of reactions extracted from the chemical literature. However, most entries do not contain atom mapping or information about reaction conditions or yield, nor is any attempt made to include concentrations or equivalence ratios, despite these being specified in the original articles.

The need for open-source data, particularly with detailed specification of reaction conditions, is urgent. Currently, even studies using the same data source may filter and preprocess the data differently. Data standardization and competition-style challenges (akin to the Netflix Prize, ImageNet Competition, or Tox21 Challenge) would accelerate further development of this field.

4.2. Data Sharing

Given the literature bias toward reporting only successful reactions (i.e., with reasonably high yields), access to negative reaction data would present tremendous opportunities for future research. Failed reaction data has proved useful in materials discovery[57] and could offer greater insight into chemical reactivity than can synthetic negative reaction data. If concepts in differential privacy can be extended to the obfuscation of exact molecular structures, perhaps a precompetitive data sharing mechanism among pharmaceutical and chemical companies, whose electronic lab notebooks collectively contain tens of millions of successful and unsuccessful reactions, can be established.

4.3. Data Applicability

We can only ask questions that are answerable by the data. For example, prediction of reaction outcomes and their yields is ill-posed without full specification of reaction conditions, although quantitative yield prediction *is* possible given a suitable problem formulation.[47] For outcome prediction, existing data is insufficient to accurately capture effects of different reaction conditions. Synthetic negative examples necessarily use the same reactants and conditions as positive examples; this can lead to models that mimic patterns in the data instead of learning actual reactivity trends, for example, predicting a Suzuki coupling between an aryl halide and an aryl boronic acid irrespective of the presence of a Pd catalyst. Reactions run under atypical conditions, even those with poor yields, provide complementary information to those run under standard conditions.

4.4. Evaluation

In addition to standardization of data, standardization of evaluation is essential. Forward prediction models have a natural evaluation metric: in the list of predicted outcomes, how highly is the true (recorded) outcome ranked? For retrosynthesis, quantitative evaluation is challenging because experiments are required for definitive proof of success. As a single-step evaluation, an effective program should propose and

DOI: 10.1021/acs.accounts.8b00087
Acc. Chem. Res. 2018, 51, 1281–1289

rank highly the reactants of published reactions when their products are treated as target compounds.

For full pathway planning, evaluating the ability to recover known pathways is misguided because "near-misses" (e.g., using a different halogen) may be forgivable. Programs should also not be evaluated on whether or how quickly pathways are found, as this is orthogonal to the quality of suggestions and neglects consideration of proposals' feasibilities. Segler et al. combine that evaluation with a manual assessment of quality;[40] their evaluation is commendable, but reliance on human chemists is problematic in terms of standardization and scalability. There is no "best" pathway: a discovery-focused medicinal chemist might want a disconnection that enables the synthesis of several diverse analogues, while a process chemist might care more strongly about cost, toxicity of side products, E-factor, etc.

4.5. Interpretability

The increased performance associated with deep neural networks compared to simpler models often comes at the cost of interpretability. Examining model parameters, attention mechanism weights, and effects of input structure modification provides some insight.[58] However, approaches to rationale extraction in other domains (e.g., natural language processing[59]) benefit from fundamental changes to network architectures (e.g., requiring models to select specific sentences to justify their predictions). Similar techniques are applicable to chemistry, although the optimal structure of rationales are likely different.

4.6. Acceptance and Adoption

Although machine learning is firmly established in certain subfields (e.g., property prediction), many practicing chemists remain skeptical of its ability to learn the "art" of organic chemistry. Overcoming this requires not only developing useful tools but also communicating the benefits and drawbacks of machine learning. Basic misconceptions can lead to inaccurate claims about limitations of its utility.[60]

4.7. Conclusion

There are clear opportunities for the application of machine learning and artificial intelligence techniques to organic synthesis and synthesis planning, including countless applications beyond those mentioned here: prediction of regio- and enantioselectivity, rational catalyst design, experimental prioritization through active learning, etc. While few research groups currently work in this area, we expect a rapid increase in the coming years, particularly as practical challenges of data availability and standardization are addressed.

The potential payoff for computer-aided synthesis planning is higher than ever. Combined with advances in automated experimentation,[61,62] sophisticated planning software could one day enable fully autonomous synthesis: a true realization of the "robo-chemist".[63] Further integration of *de novo* molecular design[64,65] and online biological assays would revolutionize small molecule drug discovery.

AUTHOR INFORMATION

Corresponding Authors

*E-mail: whgreen@mit.edu.
*E-mail: kfjensen@mit.edu.

ORCID

Connor W. Coley: 0000-0002-8271-8723
William H. Green: 0000-0003-2603-9694
Klavs F. Jensen: 0000-0001-7192-580X

Notes

The authors declare no competing financial interest.

Biographies

Connor W. Coley received a B.S. in chemical engineering from the California Institute of Technology and an M.S.C.E.P. in chemical engineering from the Massachusetts Institute of Technology (MIT). He is a Ph.D. candidate at MIT working on computer assistance in reaction planning and execution.

William H. Green received a B.A. in chemistry from Swarthmore College and a Ph.D. in physical chemistry from the University of California, Berkeley. He is the Hoyt C. Hottel Professor of Chemical Engineering at MIT where he works primarily on kinetic modeling.

Klavs F. Jensen received a M.Sc. in chemical engineering from the Technical University of Denmark and a Ph.D. in chemical engineering from the University of Wisconsin—Madison. He is the Warren K. Lewis Professor of Chemical Engineering at MIT where he works on flow chemistry and engineering.

ACKNOWLEDGMENTS

This work was supported by the DARPA Make-It program under contract ARO W911NF-16-2-0023. C.W.C. received additional funding from the NSF GRFP under Grant No. 1122374. We thank Luke Rogers, Wengong Jin, Regina Barzilay, and Tommi S. Jaakkola for their roles in the original research supporting this Account.

REFERENCES

(1) Corey, E. J. General Methods for the Construction of Complex Molecules. *Pure Appl. Chem.* **1967**, *14*, 19−38.
(2) Corey, E. J.; Long, A. K.; Rubenstein, S. D. Computer-assisted Analysis in Organic Synthesis. *Science* **1985**, *228*, 408−419.
(3) 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*, 49−53.
(4) Yadav, M. K. On the Synthesis of Machine Learning and Automated Reasoning for an Artificial Synthetic Organic Chemist. *New J. Chem.* **2017**, *41*, 1411−1416.
(5) Baskin, I. I.; Madzhidov, T. I.; Antipin, I. S.; Varnek, A. A. Artificial Intelligence in Synthetic Chemistry: Achievements and Prospects. *Russ. Chem. Rev.* **2017**, *86*, 1127.
(6) Kishimoto, A.; Buesser, B.; Botea, A. AI Meets Chemistry. https://www.research.ibm.com/ai/publications/paper/download/ai_chemistry.pdf, 2018.
(7) Cherkasov, A.; Muratov, E. N.; Fourches, D.; Varnek, A.; Baskin, I. I.; Cronin, M.; Dearden, J.; Gramatica, P.; Martin, Y. C.; Todeschini, R.; Consonni, V.; Kuz'min, V. E.; Cramer, R.; Benigni, R.; Yang, C.; Rathman, J.; Terfloth, L.; Gasteiger, J.; Richard, A.; Tropsha, A. QSAR Modeling: Where Have You Been? Where Are You Going To? *J. Med. Chem.* **2014**, *57*, 4977−5010.
(8) Mitchell, J. B. Machine Learning Methods in Chemoinformatics. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2014**, *4*, 468−481.
(9) Mayr, A.; Klambauer, G.; Unterthiner, T.; Hochreiter, S. DeepTox: Toxicity Prediction Using Deep Learning. *Front. Environ. Sci.* **2016**, *3*, 80.
(10) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. Convolutional Networks on Graphs for Learning Molecular Fingerprints. *NIPS* **2015**, 2224−2232.

DOI: 10.1021/acs.accounts.8b00087
Acc. Chem. Res. 2018, 51, 1281−1289

(11) Kearnes, S.; McCloskey, K.; Berndl, M.; Pande, V.; Riley, P. Molecular graph convolutions: moving beyond fingerprints. *J. Comput.-Aided Mol. Des.* **2016**, *30*, 595−608.
(12) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. arXiv:1704.01212 2017, https://arxiv.org/abs/1704.01212.
(13) Coley, C. W.; Barzilay, R.; Green, W. H.; Jaakkola, T. S.; Jensen, K. F. Convolutional Embedding of Attributed Molecular Graphs for Physical Property Prediction. *J. Chem. Inf. Model.* **2017**, *57*, 1757−1772.
(14) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513−530.
(15) Ihlenfeldt, W.-D.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation of Programs. *Angew. Chem., Int. Ed. Engl.* **1996**, *34*, 2613−2633.
(16) Todd, M. H. Computer-aided organic synthesis. *Chem. Soc. Rev.* **2005**, *34*, 247−266.
(17) 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*, 79−107.
(18) Ravitz, O. Data-driven Computer Aided Synthesis Design. *Drug Discovery Today: Technol.* **2013**, *10*, e443−e449.
(19) Warr, W. A. A Short Review of Chemical Reaction Database Systems, Computer-Aided Synthesis Design, Reaction Prediction and Synthetic Feasibility. *Mol. Inf.* **2014**, *33*, 469−476.
(20) 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**, DOI: 10.1016/j.drudis.2018.02.014.
(21) Corey, E.; Wipke, W. T.; Cramer, R. D., III; Howe, W. J. Computer-Assisted Synthetic Analysis. Facile Man-Machine Communication of Chemical Structure by Interactive Computer Graphics. *J. Am. Chem. Soc.* **1972**, *94*, 421−430.
(22) Szymkuc, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.* **2016**, *55*, 5904−5937.
(23) 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*, 593−602.
(24) Bogevig, 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*, 357−368.
(25) Christ, C. D.; Zentgraf, M.; Kriegl, J. M. Mining Electronic Laboratory Notebooks: Analysis, Retrosynthesis, and Reaction Based Enumeration. *J. Chem. Inf. Model.* **2012**, *52*, 1745−1756.
(26) Coley, C. W. Retrosynthetic template extraction. https://github.com/connorcoley/retrosim/blob/master/retrosim/utils/generate_retro_templates.py#L768, Accessed: 2018-01-31.
(27) Coley, C. W. RDChiral. https://github.com/connorcoley/rdchiral, Accessed: 2018-01−31.
(28) Landrum, G. RDKit: Open-Source Cheminformatics. http://www.rdkit.org, Accessed on 2016-11−20.
(29) Bertz, S. H. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981**, *103*, 3599−3601.
(30) Sheridan, R. P.; Zorn, N.; Sherer, E. C.; Campeau, L.-C.; Chang, C.; 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*, 1604−1616.
(31) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.
(32) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-Like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminf.* **2009**, *1*, 8.
(33) Heifets, A. Automated Synthetic Feasibility Assessment: A Data-driven Derivation of Computational tools for Medicinal Chemistry. Ph.D. thesis, University of Toronto, 2014.
(34) 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*, 252−261.
(35) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.
(36) Flick, A. C.; Ding, H. X.; Leverett, C. A.; Kyne, R. E.; Liu, K. K. C.; Fink, S. J.; O'Donnell, C. J. Synthetic Approaches to the New Drugs Approved During 2015. *J. Med. Chem.* **2017**, *60*, 6480−6515.
(37) 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*, 1103−1113.
(38) Schwaller, P.; Gaudin, T.; Lanyi, D.; Bekas, C.; Laino, T. "Found in Translation": Predicting Outcome of Complex Organic Chemistry Reactions using Neural Sequence-to-Sequence Models. arXiv:1711.04810, 2017, https://arxiv.org/abs/1711.04810.
(39) Segler, M. H. S.; Waller, M. P. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **2017**, *23*, 5966−5971.
(40) Segler, M. H.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604.
(41) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Cent. Sci.* **2017**, *3*, 1237−1245.
(42) Salatin, T. D.; Jorgensen, W. L. Computer-Assisted Mechanistic Evaluation of Organic Reactions. 1. Overview. *J. Org. Chem.* **1980**, *45*, 2043−2051.
(43) Gasteiger, J.; Hutchings, M. G.; Christoph, B.; Gann, L.; Hiller, C.; Löw, P.; Marsili, M.; Saller, H.; Yuki, K. *Organic Synthesis, Reactions and Mechanisms*; Springer, 1987; pp 19−73.
(44) Satoh, H.; Funatsu, K. SOPHIA, a Knowledge Base-guided Reaction Prediction System-utilization of a Knowledge Base Derived from a Reaction Database. *J. Chem. Inf. Model.* **1995**, *35*, 34−44.
(45) Socorro, I. M.; Taylor, K.; Goodman, J. M. ROBIA: a Reaction Prediction Program. *Org. Lett.* **2005**, *7*, 3541−3544.
(46) Elrod, D. W.; Maggiora, G. M.; Trenary, R. G. Applications of Neural Networks in Chemistry. 1. Prediction of Electrophilic Aromatic Substitution Reactions. *J. Chem. Inf. Model.* **1990**, *30*, 477−484.
(47) 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*, 186.
(48) Kayala, M. A.; Azencott, C.-A.; Chen, J. H.; Baldi, P. Learning to Predict Chemical Reactions. *J. Chem. Inf. Model.* **2011**, *51*, 2209−2222.
(49) Kayala, M. A.; Baldi, P. ReactionPredictor: Prediction of Complex Chemical Reactions at the Mechanistic Level using Machine Learning. *J. Chem. Inf. Model.* **2012**, *52*, 2526−2540.
(50) Fooshee, D.; Mood, A.; Gutman, E.; Tavakoli, M.; Urban, G.; Liu, F.; Huynh, N.; van Vranken, D.; Baldi, P. Deep Learning for Chemical Reaction Prediction. *Mol. Syst. Des. Eng.* **2018**, DOI: 10.1039/C7ME00107J.
(51) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725−732.
(52) 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*, 434−443.
(53) Jin, W.; Coley, C.; Barzilay, R.; Jaakkola, T. Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network. *NIPS* **2017**, https://arxiv.org/abs/1709.04555, 2604−2613.
(54) Lei, T.; Jin, W.; Barzilay, R.; Jaakkola, T. Deriving Neural Architectures from Sequence and Graph Kernels. arXiv:1705.09037, 2017, https://arxiv.org/abs/1705.09037.
(55) Nam, J.; Kim, J. Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions. arXiv:1612.09529, 2016, https://arxiv.org/abs/1612.09529.

DOI: 10.1021/acs.accounts.8b00087
Acc. Chem. Res. 2018, 51, 1281−1289

(56) Lowe, D. Chemical reactions from US patents (1976-Sep2016). https://figshare.com/articles/Chemical_reactions_from_US_patents_1976-Sep2016_/5104873, Accessed: 2018-01−31.
(57) Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; Norquist, A. J. Machine-learning-assisted Materials Discovery using Failed Experiments. *Nature* **2016**, *533*, 73−76.
(58) Montavon, G.; Samek, W.; Müller, K.-R. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.* **2018**, *73*, 1−15.
(59) Lei, T.; Barzilay, R.; Jaakkola, T. Rationalizing Neural Predictions. arXiv:1606.04155, 2016, https://arxiv.org/abs/1606.04155.
(60) Skoraczyński, G.; Dittwald, P.; Miasojedow, B.; Szymkuć, S.; Gajewska, E.; Grzybowski, B.; Gambin, A. Predicting the Outcomes of Organic Reactions via Machine Learning: Are Current Descriptors Sufficient? *Sci. Rep.* **2017**, *7*, 3582.
(61) 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*, 795−802.
(62) Hwang, Y.-J.; Coley, C. W.; Abolhasani, M.; Marzinzik, A. L.; Koch, G.; Spanka, C.; Lehmann, H.; Jensen, K. F. A Segmented Flow Platform for On-Demand Medicinal Chemistry and Compound Synthesis in Oscillating Droplets. *Chem. Commun.* **2017**, *53*, 6649−6652.
(63) Peplow, M. The Robo-Chemist. *Nature* **2014**, *512*, 20.
(64) Gómez-Bombarelli, R.; Wei, J. N.; Duvenaud, D.; Hernández-Lobato, J. M.; Sánchez-Lengeling, B.; Sheberla, D.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-driven Continuous Representation of Molecules. *ACS Cent. Sci.* **2018**, *4*, 268−276.
(65) Jin, W.; Barzilay, R.; Jaakkola, T. Junction Tree Variational Autoencoder for Molecular Graph Generation. arXiv:1802.04364, 2018, https://arxiv.org/abs/1802.04364.

DOI: 10.1021/acs.accounts.8b00087
Acc. Chem. Res. 2018, 51, 1281−1289