ACS central science

Research Article

Cite This: ACS Cent. Sci. 2017, 3, 1103-1113

http://pubs.acs.org/journal/acscii

Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models

Bowen Liu,[†] Bharath Ramsundar,[‡] Prasad Kawthekar,[‡] Jade Shi,[†] Joseph Gomes,[†] Quang Luu Nguyen,[†] Stephen Ho,[†] Jack Sloane,[†] Paul Wender,[†,§] and Vijay Pande*[,†,‡,∥]

[†]Department of Chemistry, Stanford University, Stanford, California 94305, United States

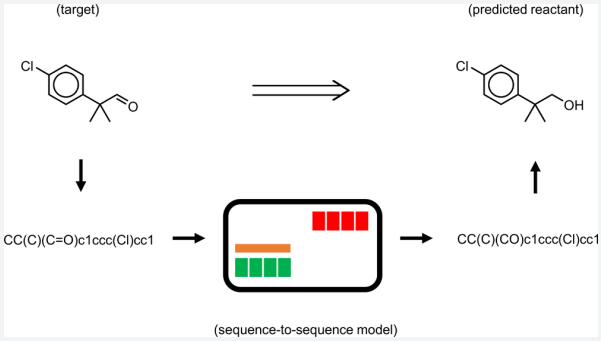
[‡]Department of Computer Science, Stanford University, Stanford, California 94305, United States

[§]Department of Chemical and Systems Biology, Stanford University, Stanford, California 94305, United States

[∥]Department of Structural Biology, Stanford University, Stanford, California 94305, United States

Supporting Information

ABSTRACT: We describe a fully data driven model that learns to perform a retrosynthetic reaction prediction task, which is treated as a sequence-to-sequence mapping problem. The end-to-end trained model has an encoder−decoder architecture that consists of two recurrent neural networks, which has previously shown great success in solving other sequence-to-sequence prediction tasks such as machine translation. The model is trained on 50,000 experimental reaction examples from the United States patent literature, which span 10 broad reaction types that are commonly used by medicinal chemists. We find that our model performs comparably with a rule-based expert system baseline model, and also overcomes certain limitations associated with rule-based expert systems and with any machine learning approach that contains a rule-based expert system component. Our model provides an important first step toward solving the challenging problem of computational retrosynthetic analysis.



INTRODUCTION

Organic synthesis is a critical discipline that directly brings scientific and societal benefits by enabling access to poorly available molecules and, more significantly, to new molecules that have never been studied before. This accessibility fundamentally enables other fields of research such as materials science, environmental science, and drug discovery. Retrosynthetic analysis is a technique widely used by organic chemists to design synthetic routes to "target" molecules, where the target is recursively transformed into simpler precursor molecules until commercially available "starting" molecules are identified.[1−3] It encompasses two related tasks. The first task, reaction prediction, involves predicting how a set of reactants will react to form products. The second task involves planning the optimal series of reaction prediction steps to recursively deconvolute the target molecule into simple or commercially available precursor molecules in a way that minimizes steps, cost, time, and waste.[4,5]

Computational retrosynthetic analysis tools can potentially greatly assist chemists in designing synthetic routes to novel molecules, and would have many applications in drug discovery, medicinal chemistry, materials science, and natural product synthesis. Since the 1960s, chemists have recognized the promise of modern computing in assisting organic synthesis analyses. However, although various algorithms have been developed over the years, their widespread acceptance by mainstream chemists has lagged.[6] In part, this is due to these approaches being applicable to only relatively simple target molecules for which expert chemists could readily deduce synthetic plans without assistance.[7] The first class of algorithms uses reaction rules that are either manually encoded by human experts or automatically derived from a reaction database.[8−23] The key drawback of such rule-based expert systems is that they generally cannot make accurate predictions outside of their knowledge base. As a result, these systems perform poorly when generalizing to new target structures and reaction types. The second class of algorithms uses principles of physical chemistry to predict energy barriers of a reaction based on first principles.[24−30] Although this approach can generalize to novel molecules and reaction types, currently such calculations are often prohibitively expensive to perform for full synthetic planning problems.

The third class of algorithms is based on machine learning techniques, which attempt to address the shortcomings of the rule-based and the physical chemistry approaches by making predictions that generalize better than those of rule-based approaches at a computational cost that is much less than those

Received: July 11, 2017
Published: September 5, 2017

DOI: 10.1021/acscentsci.7b00303
ACS Cent. Sci. 2017, 3, 1103−1113

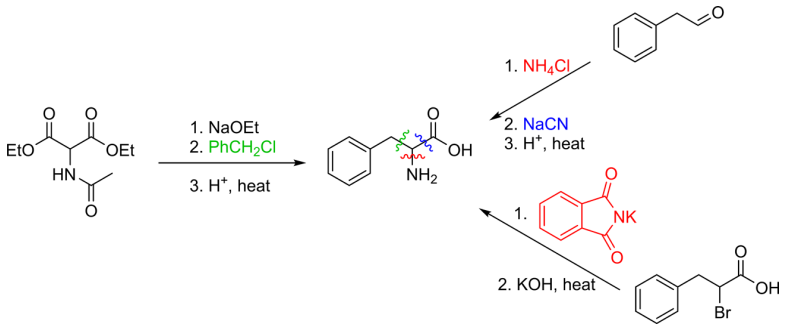


Figure 1. Phenylalanine synthetic scheme.

of physical chemistry approaches.[31−34] More recently, deep learning techniques have been applied to the reaction prediction task. The typical deep learning approach combines a rule-based expert system with a feedforward neural network (NN) component that performs candidate ranking. The NN either ranks the applicability of each rule in the knowledge base to a given example[35,36] or ranks the likelihood of each predicted product obtained by applying all the rules in the knowledge base to a given example.[37]

However, these types of deep learning approaches are fundamentally dependent on the rule-based expert system component and thus inherit some of its major limitations. In particular, these approaches have issues with making accurate predictions outside of the rule-based knowledge base. Additionally, there is a trade-off between defining very general rules that result in a lot of noise and defining very specific rules that are only applicable to a limited set of reactions with very specific reactants and products.[35] The reaction rules are necessarily inadequate representations of the underlying chemistry because they focus on the local molecular environment of the reaction centers only. Furthermore, the rule-based expert system components used by these deep learning approaches do not fully account for stereochemistry; noticeably, none of their reported reaction examples contain molecules with stereocenters.

An alternative deep learning approach that eliminates the rule-based expert system component would overcome these limitations. One way to view the reaction prediction task is to cast it as a sequence-to-sequence prediction problem, where the objective is to map a text sequence that represents the reactants to a text sequence that represents the product, or vice versa. Although molecules are usually represented as 2D or 3D graphs, they can also be equivalently expressed as text sequences in line notation format, such as simplified molecular-input line-entry system (SMILES)[38] or International Chemical Identifier (InChI).[39] The text representation of molecules in various chemoinformatic applications and the relationships between organic chemistry and linguistics have been explored previously.[40−45] Recently, Nam and Kim[46] described a neural sequence-to-sequence (seq2seq) model for the forward reaction prediction task. The model was trained end-to-end on a combination of artificially generated reactions and experimental reactions from an open source patent data set.[47] Given an input SMILES that represents the reactants, the model directly outputs a SMILES that represents the predicted products.

Here, we attempt to solve the retrosynthetic reaction prediction task, which presents additional challenges compared to the forward reaction prediction task, because the input in the retrosynthetic reaction prediction task contains less information and there are many more possible outputs. For forward reaction prediction, the starting materials significantly constrain the reaction types that are possible and limit the number of possible products. On the other hand, for retrosynthetic reaction prediction, there are usually multiple possible ways to disconnect the target molecule, via many different reaction types, to produce a large number of possible starting materials. Indeed, every bond in the target molecule represents a possible retrosynthetic disconnection. Figure 1 shows a synthetic scheme for phenylalanine that illustrates the asymmetry in the input constraints for the forward and retrosynthetic reaction prediction tasks. In the forward direction, each of the three different sets of starting materials and reaction conditions will result in phenylalanine as the single major product. Conversely, in the retrosynthetic direction, the phenylalanine can be disconnected into three different sets of starting materials.

In this work, we describe our initial studies directed at solving the challenging problem of computational retrosynthetic analysis. In particular, we develop a fully data driven seq2seq model that learns to perform the retrosynthetic reaction prediction subtask. For a given target molecule and a specified reaction type, the model predicts the most likely reactants that can react in the specified reaction type to produce the target molecule. The seq2seq model is trained end-to-end on a subset of experimental reactions with labeled reaction types[48] from an open source patent database.[47] We show that the trained seq2seq model performs comparably with a rule-based expert system baseline model on the relatively simple chemistry found in the patent data set.

APPROACH

Problem Definition. Concretely, the retrosynthetic reaction prediction task is shown in Figure 2. Given an input SMILES that represents the target molecule and a specified reaction type, the model predicts the output SMILES which represents the likely reactants that can react in the specified reaction type to form the target molecule.

Data Preparation. We use a filtered patent data set, derived from an open source patent database,[47] which contains 50,000 atom-mapped reactions that have been classified into 10 broad reaction types.[48] This filtered patent data set was originally constructed to represent the typical reaction types found in the medicinal chemist's toolkit. The reaction examples are further preprocessed to eliminate all reagents in order to only contain reactants and products,[48] and then canonicalized. We addition-

DOI: 10.1021/acscentsci.7b00303
ACS Cent. Sci. 2017, 3, 1103−1113

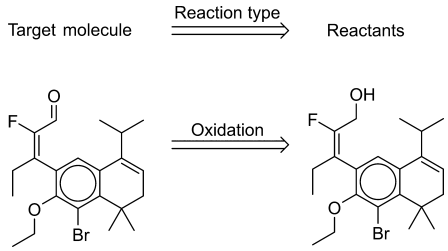


Figure 2. Retrosynthetic reaction prediction task and an example of a possible retrosynthetic disconnection for a target molecule.

ally process this data set so that each reaction example contains a single product by splitting any reactions with multiple products into multiple single product reactions that contain the original reactants. Any resulting reaction examples with trivial products such as inorganic ions and solvent molecules are removed. Table 1 shows the distribution of the 10 reaction classes in the final processed data set. Finally, the data set was split into training, validation, and test data sets (8:1:1).

Table 1. Distribution of Major Reaction Classes within the Processed Reaction Data Set

reaction class	reaction name	no. of examples
1	heteroatom alkylation and arylation	15122
2	acylation and related processes	11913
3	C−C bond formation	5639
4	heterocycle formation	900
5	protections	650
6	deprotections	8353
7	reductions	4585
8	oxidations	814
9	functional group interconversion (FGI)	1834
10	functional group addition (FGA)	227

Model. *Seq2seq Model.* Neural sequence-to-sequence (seq2seq) models map one sequence to another and have recently shown state of the art performance in many tasks such as machine translation.[49,50] It is based on an encoder−decoder architecture that consists of two recurrent neural networks (RNN) and can include an attention mechanism that aligns the target tokens with the source tokens.[49] Figure 3 shows a simple seq2seq encoder−decoder architecture for our retrosynthetic reaction prediction task.

We adapt the open source seq2seq library from Britz et al.[51] for our characterwise seq2seq model. The encoder−decoder architecture consists of long short-term memory (LSTM) cells, which is a variant of RNN cells that more effectively learn long-range dependencies in the sequences.[52] More specifically, the seq2seq model consists of a bidirectional LSTM encoder and a LSTM decoder. Furthermore, an additive attention mechanism is used.[49] The key hyperparameter settings of the seq2seq model are shown in Table S1.

The seq2seq model is trained on the training data set with reaction atom-mapping removed. Each reaction example is split into a source sequence and target sequence. The source sequence consists of a sequence of characters that is derived from splitting the SMILES that correspond to the product into characters, with a reaction type token prepended to the sequence. The source sequence is reversed prior to feeding into the encoder. The target sequence consists of a sequence of characters that is derived from splitting the SMILES that correspond to the reactants into characters. The seq2seq model is evaluated every 4000 training steps on the validation data set, and model training is stopped once the evaluation log perplexity starts to increase.

Finally, the trained seq2seq model is evaluated on the test data set with reaction atom-mapping removed. Each target molecule SMILES example from the test data set is converted into an input sequence of characters, with a reaction type token prepended to the sequence, and reversed prior to feeding into the encoder. A beam search procedure is used for model inference. Figure 4 depicts a partially completed beam search procedure with a beam width of 5 for an example input. For each source sequence input that represents the target molecule, the top N candidate output sequences ranked by overall sequence log probability at each time step during decoding are retained, where N is the width of the beam. The decoding is stopped once the lengths of the candidate sequences reach the maximum decode length of 140 characters. The candidate sequences that contain an end of sequence character are considered to be complete. On average, about 97% of all beam

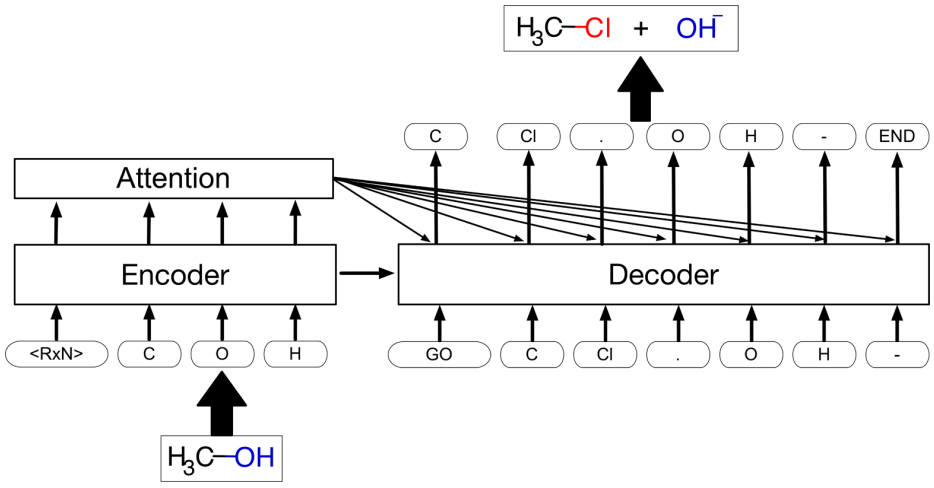


Figure 3. Seq2seq model architecture.

DOI: 10.1021/acscentsci.7b00303
ACS Cent. Sci. 2017, 3, 1103−1113

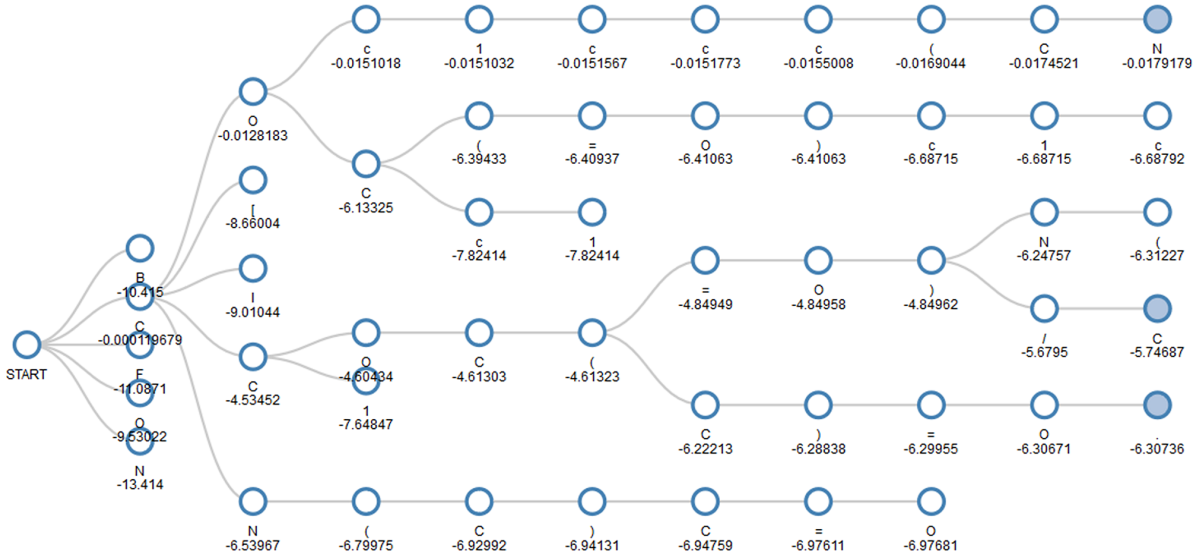


Figure 4. A partially completed beam search procedure with a beam width of 5 for an example input. Note that only the top 5 candidate sequences are retained at each time step. The visualization was produced using the seq2seq model library from Britz et al.[51]

search predicted candidate sequences are complete. These complete candidate sequences represent the reactant sets predicted by the seq2seq model for a particular target molecule, and they are ranked by the overall sequence log probabilities, which consist of the log probabilities of the individual characters in each complete candidate sequence. Additional beam search statistics and performance measures are shown in Table S2.

Baseline Model. The baseline model is a rule-based expert system that applies retrosynthetic reaction rules of a specified reaction type to a target molecule to obtain the reactants. The reaction rules are automatically extracted from the training data set. The rule extraction algorithm is adapted from Coley et al.'s implementation,[37] which was based on the algorithms described by Law et al.[20] and Bogevig et al.[53] For each atom-mapped reaction example in the training data set, the reaction centers are extracted by identifying changes in connectivity between product atoms and the corresponding reactant atoms. The reaction centers are expanded to include immediately neighboring atoms. A SMARTS string that describes the reaction core pattern is generated for the reactants and product, and combined to form a reaction SMARTS string that represents the retrosynthetic reaction rule. Each reaction rule is labeled with the reaction type of the corresponding reaction example that it was extracted from. Overall, 29272 valid rules were extracted from the training data set, which represents a rule coverage of 73.1% in the training data set. A rule is defined to be valid if it is able to regenerate the product from the reactants and the reactants from the product in the reaction example from which the rule is extracted. After filtering out duplicated rules, we end up with 2868 unique retrosynthetic reaction rules, defined by SMARTS strings.

The rule-based expert system is evaluated on the test data set. Each target molecule SMILES example from the test data set is applied by all the rules of the particular reaction type. The resulting top N reactant sets obtained from the successful reaction rules are ranked by the number of occurrences of the corresponding rule of the target reaction class that were observed in the training data set.

All scripts were written in Python (version 3.5), and RDKit (version 2016.09.04)[54] was used for reaction preprocessing and rule extraction. The seq2seq model was built with TensorFlow (version 1.0.1).[55]

RESULTS

Performance on the Test Data Set. Table 2 shows the top-N accuracies of the rule-based expert system baseline and the seq2seq model on the test data set. The top-N accuracy refers to the percentage of examples where the ground truth reactant set, which is the actual patent literature reported reactant set for the corresponding target molecule in the test data set, was found within the top N predictions made by the model. By this metric, we observed that the seq2seq model performs comparably to the baseline model. Although the baseline model does not incorporate any NN component to perform candidate ranking, the performance of the baseline model becomes less sensitive to candidate ranking as N increases. The maximum accuracy of the baseline model, which is the percentage of examples where the ground truth reactant set was found in any of the predictions made by the baseline model, is 69.5%. This represents the maximum possible test accuracy of the baseline model, as well as any deep learning approach that combines a NN component that performs candidate ranking with this rule-based expert system. The reason is that if no reaction rule exists that can produce the ground truth reactant set from the input molecule, then the NN component cannot rank it. The top-50 accuracy of the seq2seq model is higher than this maximum baseline accuracy.

Table 2. Comparison of Top-N Accuracies between the Baseline and Seq2seq Models

	top-N accuracy (%)					
model	top-1	top-3	top-5	top-10	top-20	top-50
baseline	35.4	52.3	59.1	65.1	68.6	69.5
seq2seq	37.4	52.4	57.0	61.7	65.9	70.7

DOI: 10.1021/acscentsci.7b00303
ACS Cent. Sci. 2017, 3, 1103−1113

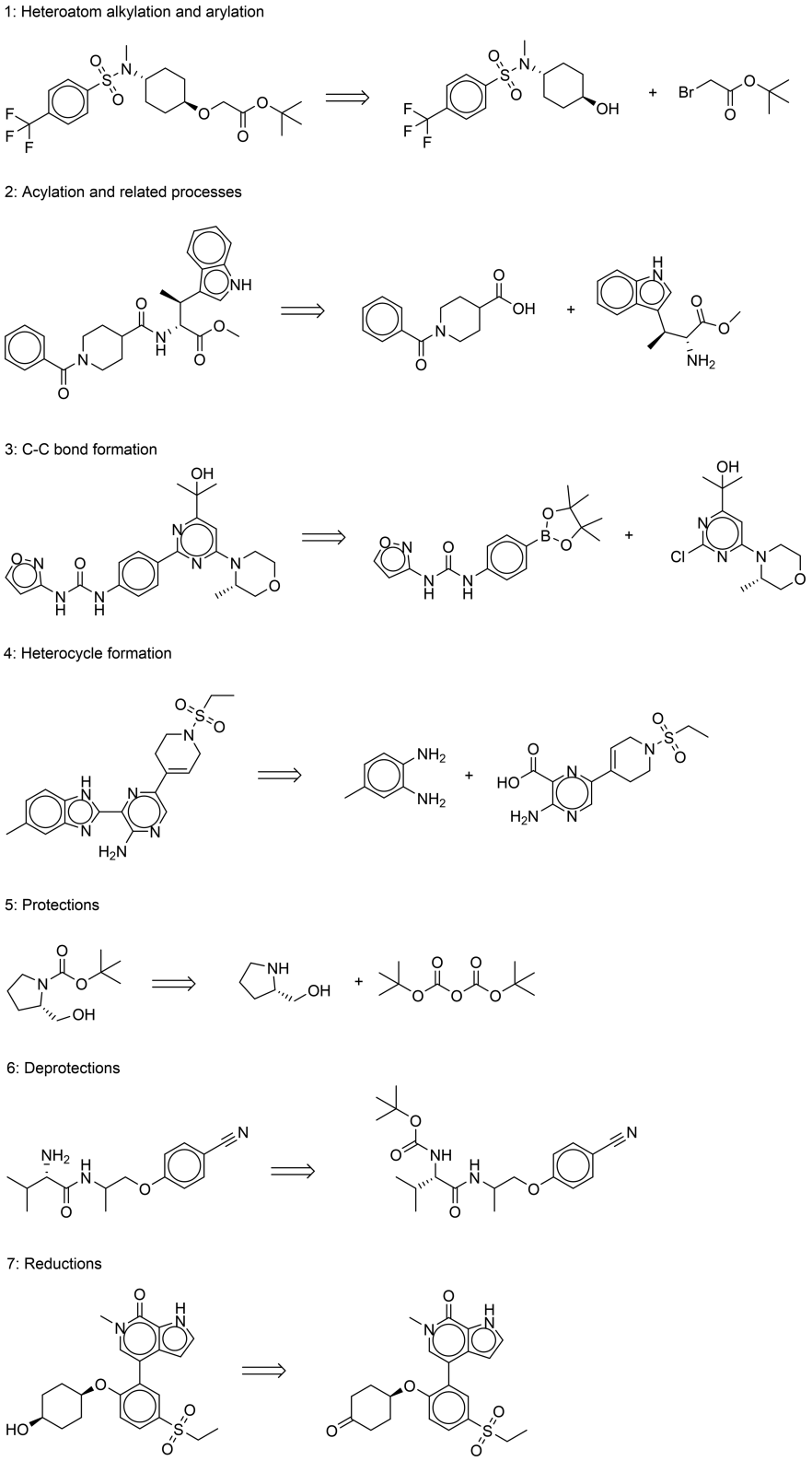


Figure 5. continued

DOI: 10.1021/acscentsci.7b00303
ACS Cent. Sci. 2017, 3, 1103−1113

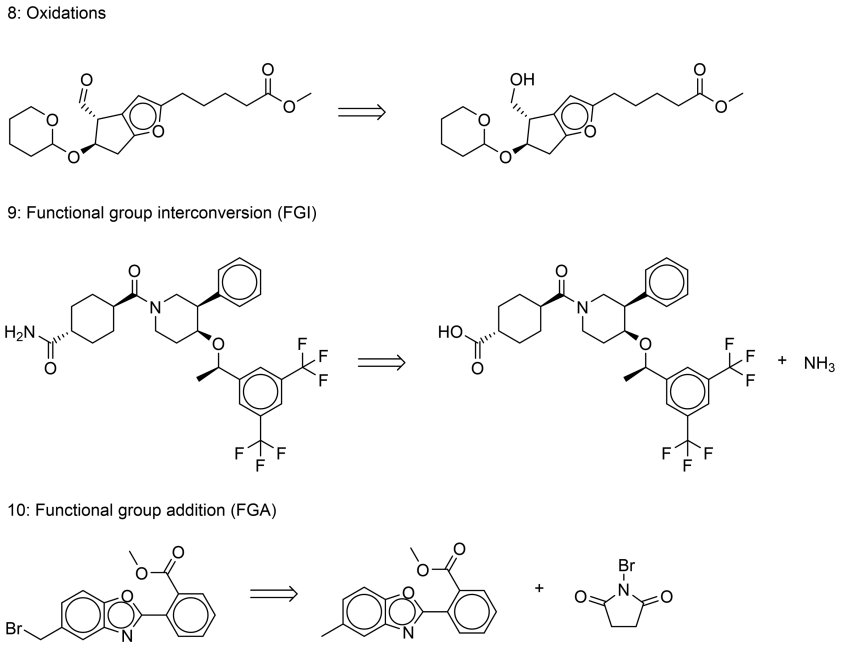


Figure 5. Representative examples of correct seq2seq model predictions for each reaction class.

Table 3. Breakdown of the Top-10 Accuracy of the Baseline and Seq2seq Models by Reaction Class

	reaction class									
	1	2	3	4	5	6	7	8	9	10
top-10 accuracy (%)										
baseline	77.2	84.9	53.4	54.4	6.2	26.9	74.7	68.4	46.7	73.9
seq2seq	57.5	74.6	46.1	27.8	80.0	62.8	67.8	69.1	47.3	56.5
no. of examples	1512	1191	564	90	65	835	459	81	184	23

Some representative examples of correct seq2seq model predictions for each reaction class are shown in Figure 5. The examples are depicted in the retrosynthetic direction.

The detailed top-10 results for the baseline model and the seq2seq model broken down by the reaction classes are shown in Table 3. The name of each reaction class is shown in Table 1.

The baseline model performs significantly better in reaction class 1 (heteroatom alkylation and arylation) and reaction class 2 (acylation and related processes). The common feature of these reaction classes is that the reactions are possible with many different functional groups at the reaction site. For example, in Figure 5, the acylation reaction between a carboxylic acid and an amine would also be possible with another carbonyl compound that has a suitable leaving group, such as an acyl chloride. Additionally, the target molecules in the data set for these reaction classes often have multiple possible reaction sites. The knowledge base in the baseline model contains reaction rules for each of the possible functional groups that were present in the reaction examples from the training data set. Therefore, the baseline model is able to easily enumerate reactant sets that span most of the possible functional group and reaction site combinations. On the other hand, the seq2seq model currently predicts only a few valid reactant sets that contain different possible functional group and reaction site combinations. As a result, the particular ground truth reactant set is more likely to be found in the predicted reactant sets from the baseline model.

The baseline model also performs significantly better in reaction class 4 (heterocycle formation). The key feature of this reaction class is the formation of cyclic and aromatic structures, which results in a large difference between the reactant set SMILES string and the target molecule SMILES string. Also, there is a relatively small number of reaction examples in the training data set for this reaction class. Overall, these two factors cause the seq2seq model to make a lot of grammatical mistakes in the SMILES predictions.

The seq2seq model performs significantly better in reaction class 5 (protections) and reaction class 6 (deprotections). The common feature of reaction classes 5 and 6 is that the reactants have large leaving groups that are not included in the product side. As a result, the very general rules in the baseline model, which only contain the immediate neighborhood of the reaction centers, do not capture the identities of the leaving groups. Conversely, the seq2seq model captures the global molecular environment of all the reaction species and is able to predict the leaving groups correctly.

Error Analysis of the Seq2seq Model. The seq2seq model makes three kinds of prediction errors:

1. The predicted reactant SMILES is grammatically invalid. This is a result of the SMILES text representation of the molecules, which is fragile because single character alterations can completely invalidate the SMILES. Since the seq2seq decoder does not explicitly understand the grammar that underlies the SMILES representation, and

DOI: 10.1021/acscentsci.7b00303
ACS Cent. Sci. 2017, 3, 1103−1113

also formulates predictions one character at a time, it is likely that some of the predicted reactant SMILES are invalid. Table 4 shows a breakdown of this type of error, and Figure 6 shows examples of this type of error.

Table 4. Breakdown of the Grammatically Invalid SMILES Error for Different Beam Sizes

	beam size					
	1	3	5	10	20	50
no. of valid SMILES	4393	12438	19751	37311	71462	167281
no. of invalid SMILES	611	2242	4450	10544	24912	74605
% error	12.2	15.3	18.4	22.0	25.8	30.8

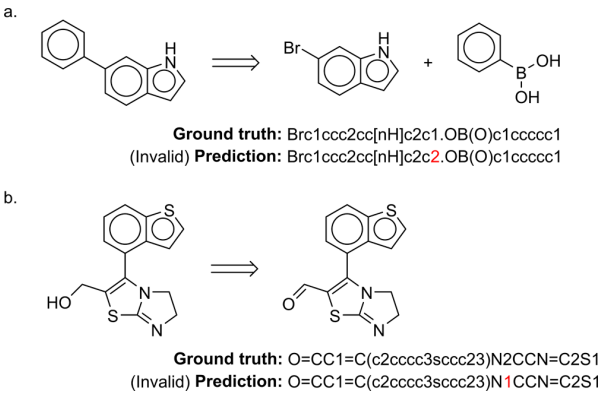


Figure 6. Examples of reactant SMILES that are grammatically invalid: (a) reaction class 3 (C−C bond formation); (b) reaction class 7 (reductions).

2. The predicted reactant SMILES is grammatically valid, but the overall reaction is not chemically plausible. This is a typical error in which the predicted reactant set cannot react in the specified reaction type to produce the target molecule. Many of these errors can also be attributed to the fragile SMILES text representation because small alterations in the SMILES can result in very large differences in the resulting molecule. Figure 7 shows examples of this type of error.
3. The predicted reactant SMILES is grammatically valid and the overall reaction is chemically plausible. Although the predicted reactant set does not match the ground truth reactant set, the predicted reactant set is likely to react in the specified reaction type to produce the target molecule. One reason for this type of error is the possibility of multiple possible functional group combinations or reactants that can react in the same reaction type to form the target molecule. Another reason is the presence of multiple reaction sites in the target molecule that can be disconnected retrosynthetically, so multiple possible reactant sets are chemically plausible. Figure 8 shows examples of this type of error.

Ranking of the Seq2seq Model Predictions. The ranked predictions from the beam search decoding procedure of the seq2seq model correspond well to chemical reactivity. For the top-10 predictions with both the baseline and seq2seq models, Figure 9 depicts a histogram that shows the counts of the highest rank that is assigned to the prediction which matches the ground truth for each example in the test data set. Overall, the higher the rank of the prediction, the more likely that the prediction corresponds to the ground truth. The distribution for the seq2seq model is much more skewed toward the highest ranks compared to the baseline model, which naively ranks by the number of occurrences of the rules that were observed in the training data set.

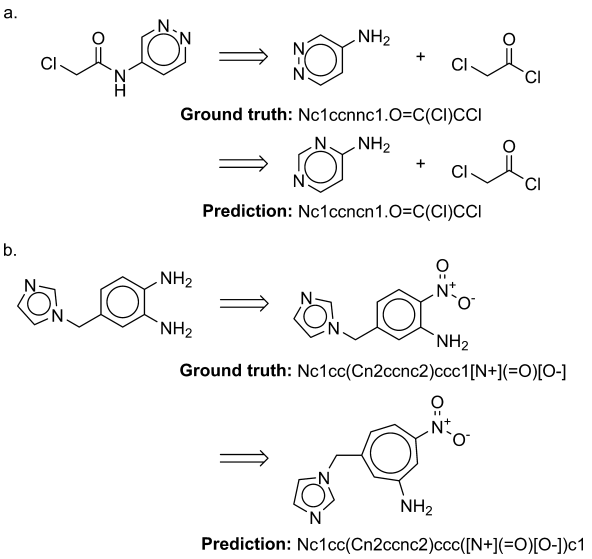


Figure 7. Examples of reactant SMILES that are grammatically valid, but the overall reaction is chemically implausible: (a) reaction class 2 (acylation and related processes); (b) reaction class 7 (reductions).

Analysis of the Seq2seq Model Attention Weights. Figures S1−S10 show the attention weights between characters in the input sequence, which represents the target molecule and specified reaction type, and characters in the output sequence, which represents the predicted reactants, for representative examples of correct seq2seq model predictions in each reaction class. The attention weights provide information on which characters in the input sequence were considered to be more important when a particular character in the output sequence was generated. Generally, we see strong weights that trend diagonally, which monotonically align molecular substructures that are shared between the input and output. Additionally, we see that the input reaction type token generally has high weights with characters in the output sequence that correspond to the neighborhood of the reaction centers.

DISCUSSION

Comparison between the Seq2seq and Rule-Based Baseline Model. The seq2seq model performs comparably to the rule-based baseline model on the processed patent data set, although the models perform differently for certain reaction types. Importantly, the seq2seq model has some significant advantages compared to the baseline model, and by extension to the deep learning based approaches that combine a rule-based expert system with a NN model for candidate ranking.

First, the seq2seq model can be trained in a fully end-to-end manner directly from the training data set. The seq2seq model both implicitly learns the chemical rules and performs candidate ranking via the beam search decoding procedure. Conversely, the typical deep learning approach to reaction prediction combines a rule-based expert system component

DOI: 10.1021/acscentsci.7b00303
ACS Cent. Sci. 2017, 3, 1103−1113

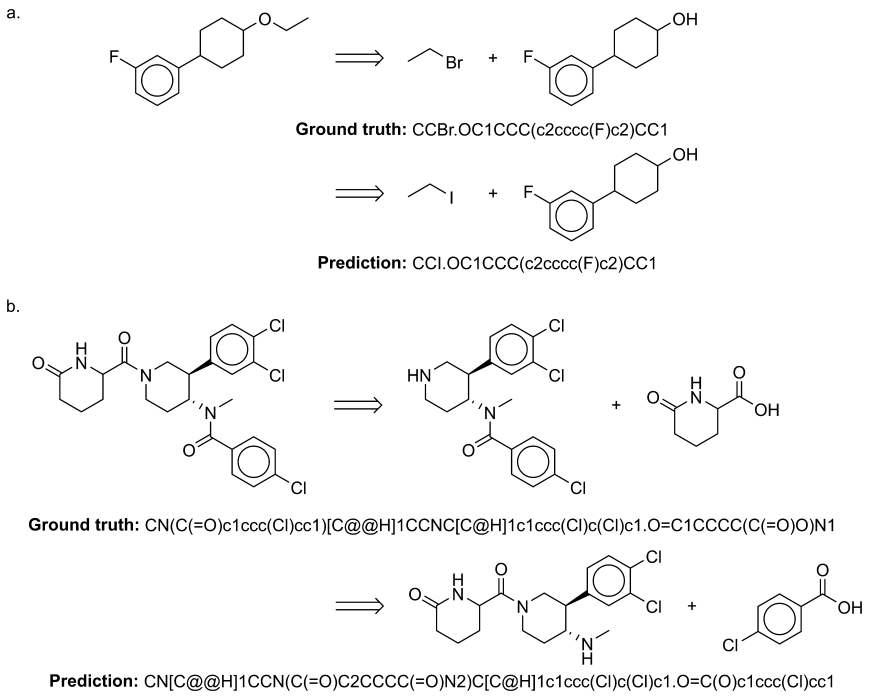


Figure 8. Examples of reactant SMILES that are grammatically valid and the overall reaction is chemically plausible: (a) reaction class 1 (heteroatom alkylation and arylation); (b) reaction class 2 (acylation and related processes).

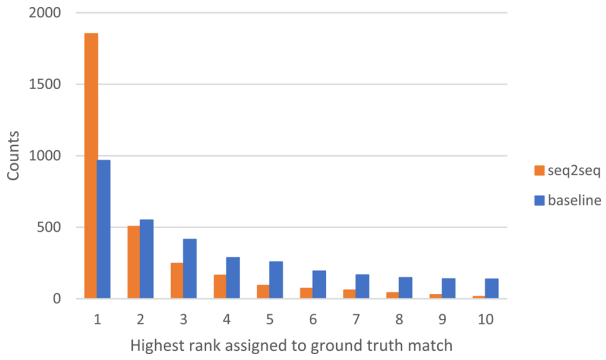


Figure 9. Histogram of the highest rank assigned to the ground truth match in the top-10 predictions of the seq2seq and baseline models for each example. Note that the relative total counts across all the ranks for the seq2seq and baseline models is proportional to their relative top-10 accuracies shown in Table 2.

with a NN model component for candidate ranking, where the individual components need to be independently set up and trained. Also, any rule-based expert system that automatically extracts reaction rules from the reaction data set depend heavily on accurate atom-mapping to describe the correspondence between the reactant and product atoms, which is itself a nontrivial problem.[56] The seq2seq model does not require atom-mapped reaction examples for training.

Second, the seq2seq model scales better to larger training data sets. The efficiency of rule-based expert systems depends on the number of rules in the knowledge base, which is an issue because the size of the knowledge base generally increases as the size of the training data set increases. For the baseline model, as well as any deep learning approaches that use a rule-based expert system and a NN component to rank the likelihood of each predicted molecular species for a given example,[37] the inference cost directly depends on the size of the knowledge base since every rule in the knowledge base must be exhaustively applied. On the other hand, the inference cost of a particular seq2seq model is independent of the size of the training data set and depends primarily on the width of the beam search decoding procedure. For deep learning approaches that use a rule-based expert system and a NN component to rank the applicability of each rule in the knowledge base for a given example,[35,36] a previous study showed that the classification accuracy decreases as the size of the knowledge base increases for a specific training data set size, likely because the number of reaction rules that need to be ranked in the multiclass classification problem also increases.[35] Overall, as training data set size increases, the increasing NN accuracy from more training examples is partially offset by the negative effect on accuracy from the increased number of reaction rules to classify. In order to reduce the number of rules in the knowledge base, previous studies[35,37] removed rare reaction rules that occurred fewer times than a specified threshold. A side effect of this is a reduced rule coverage over the reaction examples, especially for rare reaction types.

Third, the seq2seq model incorporates information about the global molecular environment since the model learns from the complete SMILES for both the reactants and the target molecule in each reaction example. Also, the complete input target molecule SMILES is used to make predictions. However, the baseline model focuses only on the local molecular environment because the automatically extracted reaction rules in the knowledge base only incorporate the immediately neighboring atoms around the reaction centers. This results in an important issue with rule-based expert systems when performing the retrosynthetic reaction prediction task. In particular, the reaction examples in the data set are processed

DOI: 10.1021/acscentsci.7b00303
ACS Cent. Sci. 2017, 3, 1103−1113

into the form depicted in Figure 2, where each reaction consists of a single target molecule and one or more reactants. In all cases, every atom in the target molecule is atom-mapped and can be linked to reactant atoms. Unfortunately, the inverse is not true because, for most reaction classes, not all atoms in the reactants are atom-mapped. These unmapped reactant atoms are the leaving groups which are not incorporated into the target molecule structure in the forward reaction. The issue arises when the leaving groups are large, which occurs commonly in reaction classes 5 (protections) and 6 (deprotections). If very general reaction rules are extracted that only contain the immediate neighborhood of the reaction centers, they will not fully capture the identities of the leaving groups. As a result, for these reaction classes that involve large leaving groups, the rules do not have sufficient information to reproduce the reactants from the target molecule. One solution would be to extract more specific rules that contain a larger neighborhood around the reaction centers in order to capture the identities of the leaving groups. However, the disadvantage with more specific rules is that they are less generalizable to new examples in the test data set. Ultimately, there is a trade-off between defining very general reaction rules and defining very specific reaction rules, which is challenging without manual intervention. Furthermore, a direct benefit of the seq2seq model incorporating the global molecular environment is that the model naturally accounts for stereochemistry. On the other hand, in order for deep learning approaches that combine a rule-based expert system with a NN model component to account for stereochemistry, both the reaction rules in the knowledge base and the descriptors that are used in the NN model must incorporate stereochemistry. Existing algorithms that automatically extract reaction rules do not fully address the issue of stereochemistry.

Molecular Representations. The seq2seq model maps one sequence to another, which restricts us to representing molecules using text sequences. In this study, we predominantly focused on the SMILES text representation because it has characteristics that make it particularly attractive as a text representation for molecules in a product to reactant sequence-to-sequence mapping task. Generally, when a reaction is expressed in a SMILES representation, there are many subsequences that are shared between the reactant and product SMILES strings. The SMILES representation reflects the fact that most reactions modify a particular reaction center, while keeping other parts of the molecule the same. The presence of shared subsequences between the input and output sequences makes the sequence-to-sequence mapping task relatively easier because the input and output sequences are similar. In comparison, our preliminary experiments using the InChI text representation for molecules in the seq2seq model resulted in worse performance when compared to using SMILES. This outcome is likely because InChI has a more complicated hierarchical syntax, which involves arithmetic to obtain the atom connectivity, and also there are fewer shared subsequences between the reactant and product InChI strings in a reaction. The weaker performance of InChI compared with SMILES when applied in a model that decodes text sequences is also in agreement with the observations of Gómez-Bombarelli et al.[40] in their autoencoder architecture for generating molecular structures. Nonetheless, our results have shown that a key weakness with representing molecules using text sequences is the fragility of the representation, since very small changes in the text sequence can result in a very large change in the resulting molecular structure. A more natural way to represent molecules is to directly use graph representations, and our future work will explore model architectures that can take advantage of this feature.

CONCLUSION

Overall, we have created a data driven neural sequence-to-sequence model to solve the retrosynthetic reaction prediction task, which is a critical task in computational retrosynthetic analysis. While the current implementation of the seq2seq model performs comparably to the rule-based expert system baseline, the seq2seq model has fundamental advantages over rule-based expert systems and over any deep learning approach that depends on a rule-based expert system component. The seq2seq model can be trained in an end-to-end manner, scales more efficiently to larger data sets, and naturally incorporates the global molecular environments of the reaction species. We believe that there exists significant room for improvement over the current relatively unoptimized seq2seq model, so further work is likely to lead to greater prediction accuracies.

In following work, we will seek to increase the accuracy of the seq2seq model by exploring architectural variants and by exploring new data sets. Additionally, since the seq2seq model is a very general architecture for mapping input sequences to output sequences, we could include reaction conditions and yields during training in order to predict reaction conditions and yields alongside the reactants during the decoding procedure. Furthermore, an interesting extension to the seq2seq models would be to use one-shot techniques[57−59] to allow our retrosynthetic reaction prediction models to make reasonable predictions for reaction classes where only a few example reactions are available for training. We envision that our seq2seq model and its future variants would act as the single step reaction prediction module in a multistep retrosynthetic analysis tool, whereby a search procedure recursively deconvolutes the target molecule using the chemistry learned by the seq2seq model until simple or commercially available precursor molecules are obtained.

We believe that the approach and model architecture described in this work constitute an important early step toward solving the computational retrosynthetic analysis problem. Although the chemical complexity of the reactions and molecules explored in this work is intentionally simple to facilitate analysis and thus is quite far away from what is typically faced by mainstream synthetic chemists today, we strongly believe that the future evolution of this approach could bridge this gap and result in tools that can become useful for the expert chemist and also more broadly for those less skilled in the science.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acscentsci.7b00303.

Additional tables and figures (PDF)

The neural network model, processed data sets, and evaluation code will be made available at https://github.com/pandegroup/reaction_prediction_seq2seq.git.

AUTHOR INFORMATION

Corresponding Author

*E-mail: pande@stanford.edu.

DOI: 10.1021/acscentsci.7b00303
ACS Cent. Sci. 2017, 3, 1103−1113

ORCID

Bowen Liu: 0000-0003-0609-3087
Bharath Ramsundar: 0000-0001-8450-4262
Paul Wender: 0000-0001-6319-2829

Notes

The authors declare the following competing financial interest(s): V.P. is a consultant and SAB member of Schrodinger, LLC, and Globavir, sits on the Board of Directors of Apeel Inc, Freenome Inc, Omada Health, Patient Ping, and Rigetti Computing, and is a General Partner at Andreessen Horowitz.

■ ACKNOWLEDGMENTS

We thank Franklin Lee for critical reading and feedback on the manuscript. B.L. is supported by the NIH (U19 AI109662). B.R. is supported by the Fannie and John Hertz Foundation. S.H. is supported by a Postdoctoral Fellowship, PF-15-007-01-CDD, from the American Cancer Society. The Pande Group is broadly supported by grants from the NIH (R01 GM062868 and U19 AI109662) as well as gift funds and contributions from Folding@home donors. This research was also supported by the National Science Foundation (P.W.: CHE1265956). We acknowledge the generous support of Dr. Anders G. Frøseth and Mr. Christian Sundt for our work on machine learning.

■ REFERENCES

(1) Corey, E. J.; Cheng, X.-M. *The Logic of Chemical Synthesis*; Wiley: New York, 1989.
(2) Corey, E. J. General Methods for the Construction of Complex Molecules. *Pure Appl. Chem.* **1967**, *14*, 19−38.
(3) Corey, E. J.; Wipke, W. T. Computer-Assisted Design of Complex Organic Syntheses. *Science (Washington, DC, U. S.)* **1969**, *166*, 178−192.
(4) Wender, P. A.; Quiroz, R. V.; Stevens, M. C. Function through Synthesis-Informed Design. *Acc. Chem. Res.* **2015**, *48*, 752−760.
(5) Wender, P. A. Toward the Ideal Synthesis and Molecular Function through Synthesis-Informed Design. *Nat. Prod. Rep.* **2014**, *31*, 433−440.
(6) Ley, S. V.; Fitzpatrick, D. E.; Ingham, R. J.; Myers, R. M. Organic Synthesis: March of the Machines. *Angew. Chem., Int. Ed.* **2015**, *54*, 3449−3464.
(7) Judson, P. *Knowledge-Based Expert Systems in Chemistry: Not Counting on Computers*; RSC Theoretical and Computational Chemistry Series; The Royal Society of Chemistry: 2009.
(8) Szymkuć, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grzybowski, B. A. Computer-Assisted Synthetic Planning: The End of the Beginning. *Angew. Chem., Int. Ed.* **2016**, *55*, 5904−5937.
(9) Pensak, D. A.; Corey, E. J. LHASA—Logic and Heuristics Applied to Synthetic Analysis. In *Computer-Assisted Organic Synthesis*; ACS Symposium Series; American Chemical Society: 1977; Vol. *61*, pp 1−32.
(10) Corey, E. J.; Wipke, W. T.; Cramer, R. D.; Howe, W. J. Techniques for Perception by a Computer of Synthetically Significant Structural Features in Complex Molecules. *J. Am. Chem. Soc.* **1972**, *94*, 431−439.
(11) Corey, E. J.; Howe, W. J.; Orf, H. W.; Pensak, D. A.; Petersson, G. General Methods of Synthetic Analysis. Strategic Bond Disconnections for Bridged Polycyclic Structures. *J. Am. Chem. Soc.* **1975**, *97*, 6116−6124.
(12) Corey, E. J.; Long, A. K.; Greene, T. W.; Miller, J. W. Computer-Assisted Synthetic Analysis. Selection of Protective Groups for Multistep Organic Syntheses. *J. Org. Chem.* **1985**, *50*, 1920−1927.
(13) Salatin, T. D.; Jorgensen, W. L. Computer-Assisted Mechanistic Evaluation of Organic Reactions. 1. Overview. *J. Org. Chem.* **1980**, *45*, 2043−2051.
(14) Gasteiger, J.; Hutchings, M. G.; Christoph, B.; Gann, L.; Hiller, C.; Löw, P.; Marsili, M.; Saller, H.; Yuki, K. A New Treatment of Chemical Reactivity: Development of EROS, an Expert System for Reaction Prediction and Synthesis Design. In *Organic Synthesis, Reactions and Mechanisms*; Springer Berlin Heidelberg: Berlin, Heidelberg, 1987; pp 19−73.
(15) 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.
(16) Satoh, K.; Funatsu, K. A Novel Approach to Retrosynthetic Analysis Using Knowledge Bases Derived from Reaction Databases. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 316−325.
(17) Gelernter, H.; Rose, J. R.; Chen, C. Building and Refining a Knowledge Base for Synthetic Organic Chemistry via the Methodology of Inductive and Deductive Machine Learning. *J. Chem. Inf. Model.* **1990**, *30*, 492−504.
(18) Chen, J. H.; Baldi, P. No Electron Left Behind: A Rule-Based Expert System To Predict Chemical Reactions and Reaction Mechanisms. *J. Chem. Inf. Model.* **2009**, *49*, 2034−2043.
(19) Chen, J. H.; Baldi, P. Synthesis Explorer: A Chemical Reaction Tutorial System for Organic Synthesis Design and Mechanism Prediction. *J. Chem. Educ.* **2008**, *85*, 1699.
(20) 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.
(21) Gothard, C. M.; Soh, S.; Gothard, N. A.; Kowalczyk, B.; Wei, Y.; Baytekin, B.; Grzybowski, B. A. Rewiring Chemistry: Algorithmic Discovery and Experimental Validation of One-Pot Reactions in the Network of Organic Chemistry. *Angew. Chem., Int. Ed.* **2012**, *51*, 7922−7927.
(22) Segler, M. H. S.; Waller, M. P. Modelling Chemical Reasoning to Predict and Invent Reactions. *Chem. - Eur. J.* **2017**, *23*, 6118−6128.
(23) Jorgensen, W. L.; Laird, E. R.; Gushurst, A. J.; Fleischer, J. M.; Gothe, S. A.; Helson, H. E.; Paderes, G. D.; Sinclair, S. CAMEO: A Program for the Logical Prediction of the Products of Organic Reactions. *Pure Appl. Chem.* **1990**, *62*, 1921−1932.
(24) Zimmerman, P. M. Automated Discovery of Chemically Reasonable Elementary Reaction Steps. *J. Comput. Chem.* **2013**, *34*, 1385−1392.
(25) Wang, L.-P.; Titov, A.; McGibbon, R.; Liu, F.; Pande, V. S.; Martínez, T. J. Discovering Chemistry with an Ab Initio Nanoreactor. *Nat. Chem.* **2014**, *6*, 1044−1048.
(26) Wang, L.-P.; McGibbon, R. T.; Pande, V. S.; Martinez, T. J. Automated Discovery and Refinement of Reactive Molecular Dynamics Pathways. *J. Chem. Theory Comput.* **2016**, *12*, 638−649.
(27) Xu, L.; Doubleday, C. E.; Houk, K. N. Dynamics of 1,3-Dipolar Cycloaddition Reactions of Diazonium Betaines to Acetylene and Ethylene: Bending Vibrations Facilitate Reaction. *Angew. Chem.* **2009**, *121*, 2784−2786.
(28) Rappoport, D.; Galvin, C. J.; Zubarev, D. Y.; Aspuru-Guzik, A. Complex Chemical Reaction Networks from Heuristics-Aided Quantum Chemistry. *J. Chem. Theory Comput.* **2014**, *10*, 897−907.
(29) Socorro, I. M.; Taylor, K.; Goodman, J. M. ROBIA: A Reaction Prediction Program. *Org. Lett.* **2005**, *7*, 3541−3544.
(30) Martínez, T. J. Ab Initio Reactive Computer Aided Molecular Design. *Acc. Chem. Res.* **2017**, *50*, 652−656.
(31) Kayala, M. A.; Azencott, C.-A.; Chen, J. H.; Baldi, P. Learning to Predict Chemical Reactions. *J. Chem. Inf. Model.* **2011**, *51*, 2209−2222.
(32) 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.
(33) Carrera, G. V. S. M.; Gupta, S.; Aires-de-Sousa, J. Machine Learning of Chemical Reactivity from Databases of Organic Reactions. *J. Comput.-Aided Mol. Des.* **2009**, *23*, 419−429.
(34) Zhang, Q.-Y.; Aires-de-Sousa, J. Structure-Based Classification of Chemical Reactions without Assignment of Reaction Centers. *J. Chem. Inf. Model.* **2005**, *45*, 1775−1783.

DOI: 10.1021/acscentsci.7b00303
ACS Cent. Sci. 2017, 3, 1103−1113

(35) Segler, M. H. S.; Waller, M. Neural-Symbolic Machine Learning for Retrosynthesis and Reaction Prediction. *Chem. - Eur. J.* **2017**, *23*, 5966−5971.
(36) Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural Networks for the Prediction of Organic Chemistry Reactions. *ACS Cent. Sci.* **2016**, *2*, 725−732.
(37) 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.
(38) Weininger, D. SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Model.* **1988**, *28*, 31−36.
(39) Heller, S.; McNaught, A.; Stein, S.; Tchekhovskoi, D.; Pletnev, I. InChI - the Worldwide Chemical Structure Identifier Standard. *J. Cheminf.* **2013**, *5*, 7.
(40) Gómez-Bombarelli, R.; Duvenaud, D.; Hernández-Lobato, J. M.; Aguilera-Iparraguirre, J.; Hirzel, T. D.; Adams, R. P.; Aspuru-Guzik, A. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *arXiv* **2016**, 1610.02415.
(41) Ikebata, H.; Hongo, K.; Isomura, T.; Maezono, R.; Yoshida, R. Bayesian Molecular Design with a Chemical Language Model. *J. Comput.-Aided Mol. Des.* **2017**, *31*, 379−391.
(42) Jastrzębski, S.; Leśniak, D.; Czarnecki, W. M. Learning to SMILE(S). *arXiv* **2016**, 1602.06289.
(43) Bjerrum, E. J. SMILES Enumeration as Data Augmentation for Neural Network Modeling of Molecules. *arXiv* **2017**, 1703.07076.
(44) Segler, M. H. S.; Kogej, T.; Tyrchan, C.; Waller, M. P. Generating Focussed Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *arXiv* **2017**, 1701.01329.
(45) Cadeddu, A.; Wylie, E. K.; Jurczak, J.; Wampler-Doty, M.; Grzybowski, B. A. Organic Chemistry as a Language and the Implications of Chemical Linguistics for Structural and Retrosynthetic Analyses. *Angew. Chem., Int. Ed.* **2014**, *53*, 8108−8112.
(46) Nam, J.; Kim, J. Linking the Neural Machine Translation and the Prediction of Organic Chemistry Reactions. *arXiv* **2016**, 1612.09529.
(47) Lowe, D. M. *Extraction of Chemical Structures and Reactions from the Literature*; University of Cambridge: 2012.
(48) Schneider, N.; Stiefl, N.; Landrum, G. A. What's What: The (Nearly) Definitive Guide to Reaction Role Assignment. *J. Chem. Inf. Model.* **2016**, *56*, 2336−2346.
(49) Bahdanau, D.; Cho, K.; Bengio, Y. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*; **2014**.
(50) Sutskever, I.; Vinyals, O.; Le, Q. V. Sequence to Sequence Learning with Neural Networks. In *NIPS*; **2014**.
(51) Britz, D.; Goldie, A.; Luong, T.; Le, Q. Massive Exploration of Neural Machine Translation Architectures. *arXiv* **2017**, 1703.03906.
(52) Hochreiter, S.; Schmidhuber, J. LONG SHORT-TERM MEMORY. *Neural Comput.* **1997**, *9*, 1735−1780.
(53) Bøgevig, A.; Federsel, H.-J.; Huerta, F.; Hutchings, M. G.; Kraut, H.; Langer, T.; Löw, P.; Oppawsky, C.; Rein, T.; Saller, H. Route Design in the 21st Century: The IC *SYNTH* Software Tool as an Idea Generator for Synthesis Prediction. *Org. Process Res. Dev.* **2015**, *19*, 357−368.
(54) RDKit: Open-source cheminformatics http://www.rdkit.org.
(55) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. *arXiv* **2015**, 1603.04467.
(56) Chen, W. L.; Chen, D. Z.; Taylor, K. T. Automatic Reaction Mapping and Reaction Center Detection. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **2013**, *3*, 560−593.
(57) Vinyals, O.; Blundell, C.; Lillicrap, T.; Kavukcuoglu, K.; Wierstra, D. Matching Networks for One Shot Learning. In *NIPS*; **2016**.
(58) Santoro, A.; Bartunov, S.; Botvinick, M.; Wierstra, D.; Lillicrap, T. One-Shot Learning with Memory-Augmented Neural Networks. *arXiv* **2016**, 1605.06065.
(59) Altae-Tran, H.; Ramsundar, B.; Pappu, A. S.; Pande, V. Low Data Drug Discovery with One-Shot Learning. *ACS Cent. Sci.* **2017**, *3*, 283−293.

DOI: 10.1021/acscentsci.7b00303
ACS Cent. Sci. 2017, 3, 1103−1113