Evaluating the Performance of a Transformer-based Organic Reaction Prediction Model

William R. Borrelli* and Joshua Schrier*

Department of Chemistry, Fordham University, 441 E. Fordham Rd, The Bronx, NY, 10458, United States

E-mail: wborrelli@fordham.edu; jschrier@fordham.edu

Abstract

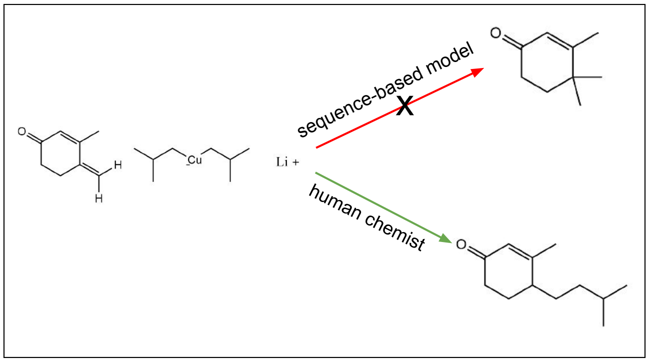
Forward and retrosynthetic organic reaction prediction are challenging applications of artificial intelligence (AI) research in chemistry. IBM's freely available RXN for Chemistry (https://rxn.res.ibm.com) treats reaction prediction as a translation problem, by using transformer-based machine learning models trained on patent data to convert reactant SMILES string sequences into product strings.

Here we characterize the performance of transformer models on 100 undergraduate text-book problems to expose reaction classes where the fundamentals of organic chemistry are violated. The forward prediction model is generally successful in predicting outcomes for substitution reactions and unsuccessful for elimination and organocopper reactions.

For the retrosynthesis model, we found characteristic examples of a lack of atom conservation and nonsensical chemical transformations. We also compared the differences in molecular complexity and synthetic accessibility between predicted and literature reactions to probe how AI plans reactions compared to humans. Forward predictions replicated a similar distribution of differences in molecular complexity as the human reactions from the literature, whereas retrosynthetic predictions resulted in both positive and negative deviations from literature complexity. Finally, we analyzed the atom mapping in test reactions to expose errors in how the model identifies reactive atoms and species.

Keywords: machine-learning, sequence-to-sequence methods, synthesis, Deep Learning, organic chemistry, reaction prediction, chemical reaction

TOC Graphic



Introduction

The rapid and accurate prediction of organic chemistry reactions is an important area of research at the interface of theoretical chemistry and artificial intelligence. High-throughput quantum chemistry calculations can predict reaction outcomes,[1] but the computational cost is prohibitively high. Data-driven approaches offer an alternative method for reaction prediction.[2]

Two prominent types of machine learning models have emerged for the task of organic reaction prediction: template-based methods and template-free methods.[3] Template-based methods rely on extensive databases of rules to guide all possible transformations for a chemical system.[4] A disadvantage of this methodology is the arduous task of hand-curating templates, which becomes increasingly intractable as more reactions are reported in the literature. Methods to automate template generation with machine learning have been successful,[5,6] and offer an alternative to methods that utilize underlying data bases.

Template-free methods, such as graph- and sequence-based methods, offer a means to directly circumvent several of these issues. The former uses a graph representation of the reactants and completes a series of graph edits that lead to the product. Sequence-based models take a text representation of the reactants and translates the string from the language of reactants and reagents to the language of products. These kinds of text-based schemes have recently been applied to prediction of chemical reaction yields,[7] enzymatic reactions,[8] and regio- and stereoselective reactions on carbohydrates.[9] Outside of chemistry, transformer neural networks have become increasingly popular for natural language processing (NLP) tasks, from the transfer learning method ULMFiT[10] to BERT,[11] Google's NLP pre-training neural network technique.

In 2020, Schwaller et al. introduced the Molecular Transformer, a sequence-based model for both forward reaction and retrosynthesis prediction using simplified molecular-input line-entry system (SMILES) strings.[12,13] Their sequence-to-sequence architecture bypasses the need for laborious rule-based methods by treating reaction prediction as a language translation problem, greatly increasing scalability and ease of encoding. When applied to forward reaction prediction, their model achieved remarkable accuracy on several representative data sets. An extension of their Molecular Transformer to retrosynthetic pathway prediction, using a hyper-graph exploration strategy, also achieved impressive performance in a series of test metrics including round-trip accuracy, coverage, class diversity, and Jensen-Shannon divergence. Their predictive model is publicly available on the IBM RXN for Chemistry website[14] and is advertised to chemists and students as a useful tool in pursuits related to organic synthesis. In addition to access through a web-GUI, programmatic access is provided through a RESTful API, and wrappers for this API exist for both Python[15] and Mathematica.[16]

Despite the empirical success of sequence-based models for reaction prediction, without an underlying physical or chemical theory there is nothing to prevent spurious chemical transformations that fail to obey conservation laws or involve implausible mechanisms. Identifying examples of such flaws can help prioritize the addition of new examples to the training data sets that constrain these cases. In this work, we evaluated forward and retrosynthetic prediction performance on a curated data set of undergraduate level organic chemistry reactions. We critically examine the claims in Schwaller et al.'s papers as well as the efficacy of the IBM RXN for Chemistry model for use in teaching or learning organic chemistry.

Methods

Our test sets consist of 100 single-step organic chemistry reactions encompassing ten reaction classes for each the forward prediction and retrosynthesis tests, with a total of 200 test reactions across both models.

Reaction classes examined included substitution, elimination, Grignard, Diels-Alder, Friedel-Crafts, organocopper, acetylide, carbonyl condensation, amine, and aromatic substitution. These reactions were sourced from introductory organic chemistry textbooks[17,18] or other academic resources, motivated by these being representative of the types examples a student would encounter when using IBM RXN as an educational resource.

The fact that these reactions should also be correct and generally “ordinary” enhances the ability to evaluate the performance of the model.

Additionally, to test the validity of the retrosynthetic predictions, all predicted retrosyntheses were run through the forward prediction model. Results were collected in a data set along with metrics returned from IBM RXN for each prediction, such as confidence and optimization scores. Characteristics of the predicted retrosynthetic pathways were assessed using the difference of molecular complexity and synthetic accessibility from reactants to products. Molecular complexity was evaluated using an implementation of Böttcher's 2016 additive definition of molecular complexity (Cm).[19] Our script for computing molecular complexity is available as a GitHub repository[20] as well as through the Wolfram Function Repository[21] as a resource function. Ertl and Schuffenhauer's measure of synthetic accessibility (SA)[22] is readily available in Mathematica and in the Python RDKit. The differences, ΔCm and ΔSA, were computed as the cumulative score of the product(s) minus the cumulative score of the reactants for all literature reactions, as well as all retrosynthetic and forward prediction reactions.

Atom mapping of predicted reactions were computed using RXNMapper[23] in Python 3.6.13 to generate the atom mappings and mapping confidences for all literature and predicted reactions.

The complete data sets and analysis codes (implemented in Mathematica 12.1 and Python 3.6.13) used for this work are available on GitHub.[20]

Results and Discussion

Forward Synthesis Prediction

To test the quality of IBM RXN's forward predictions, we queried 100 total reactions and collected results for up to five top predictions. Top-k accuracies were determined by whether the correct product was returned in the first k ranked prediction results. We adopted a lenient scoring policy with regards to enantioselectivity and racemization, to give the predictive model an advantage. Reactions where racemization is expected were counted as correct if one isomer was provided, or if an achiral representation of the product was predicted. For reactions where some mixture of constitutional isomers is expected (e.g., aromatic substitutions resulting in a mixture of ortho and para products), predictions were considered correct if the second isomer was present in the top-5 predictions collected. The top-k accuracies, shown in Table 1, show considerable variability across the reaction classes. Note that the bolded accuracies are the subsequent highest accuracies recorded as more predictive results are considered.

Substitution reactions were among the most accurate, maintaining an accuracy of 90% regardless of the number of prediction results. Elimination reactions were characteristically poor performing with an accuracy of 30% for top-1 to top-4 predictions, increasing to 40% accuracy given a fifth prediction result. Most reaction classes peaked in accuracy with only one or two prediction outcomes, showing that the increasingly lower confidence predictions provide little in the way of increased accuracy. Table 1 shows the cumulative accuracies for each top-k test across the entirety of the test set. We found that taking only top-1 predictions yielded an accuracy of 61%, which increased to 69% for top-2 predictions. This accuracy remained constant until a fifth prediction was considered, whereby we only saw a slight increase to 70%. These total accuracy scores are lower than those reported by Schwaller et al., who cited a top 1 accuracy between 71% and 91% depending on the specified data set. To understand why our accuracy scores are different than what was originally cited by

Schwaller et al., we must think about the distributions these reaction data sets were sampled from.[24] We constructed our test sets to be representative of undergraduate organic chemistry, similar to what a student might encounter on their final exam. The model itself was trained on patent data originally mined by Lowe,[25] and is therefore different from the types of canonical reactions taught in an undergraduate curriculum. Thakkar et al.[26] showed that patent data is sufficient for the prediction of medicinally important synthetic targets, however it is possible that these reactions do not completely characterize the types of transformations that are taught in the classroom.

Table 1: Top-K Accuracies

Source	Top 1 Accuracy (%)	Top 2 Accuracy (%)	Top 3 Accuracy (%)	Top 4 Accuracy (%)	Top 5 Accuracy (%)
our work (cumulative)	**61**	**69**	69	69	**70**
Schwaller et al.[12]	**76.0-91.0**	**82.0-94.0**	**84.0-95.0**	-	**85.0-96.0**
Substitution	**90**	90	90	90	90
Elimination	**30**	30	30	30	**40**
Grignard	**50**	**70**	70	70	70
Organocopper	**30**	**40**	40	40	40
Friedel-Crafts	**50**	**70**	70	70	70
Diels-Alder	**80**	**90**	90	90	90
Acetylide	**70**	70	70	70	70
Aromatic Substitution	**80**	**90**	90	90	90
Carbonyl Condensation	**70**	**80**	80	80	80
Amine	**60**	60	60	60	60

To investigate where the forward prediction model is the most uncertain, we began by looking at the reactions with the lowest prediction confidence. We defined a low confidence prediction as having a confidence below 0.50, and identified 19 reactions falling below this cutoff. Table 2 shows the mean confidence scores across the reaction classes. Interestingly, the reaction classes with the most low confidence predictions were substitution, organocopper, acetylide, and amine, many of which have the lowest mean confidence scores overall. Six of these low confidence predictions were outright predictive failures, or cases where the model could not successfully provide a predicted product or simply returned the SMILES of the reactants and reagents as output. Eight of these low confidence reactions were actually correct predictions, albeit with low confidence. Other predictions were more troubling, and highlighted a major pitfall of sequence-based models. In reaction 39, shown in Figure 1, the number of carbons predicted to be added to the double bond does not match the number of carbons on the organocopper reagent itself. The model treats the reaction as if the organocopper reagent only adds a methyl group, and moreover adds the new carbon-carbon bond to the incorrect carbon belonging to the ring.

Table 2: Predictive Confidence Scores

Reaction Class	Mean Forward Prediction Confidence	Mean Retrosynthesis Prediction Confidence
Substitution	0.747	0.898
Elimination	0.817	0.558
Grignard	0.886	1.00
Organocopper	0.594	0.808
Friedel-Crafts	0.844	0.930
Diels-Alder	0.741	0.674
Acetylide	0.622	0.794
Aromatic substitution	0.898	0.900
Carbonyl condensation	0.709	0.994
Amine	0.691	0.866

Double elimination reactions are also problematic. In reactions 68 and 69, the model fails to return a prediction. In reaction 20, the model returns a prediction where only a single elimination occurs. To probe this partial success in the case of double eliminations, two alternative predictions were completed, shown in Figure 2. In test reaction 1 (TR1) two equivalents of base were provided explicitly in the SMILES string input to determine if this arose from stoichiometric considerations. Paradoxically, providing this additional information to the model changes the prediction: again, only a single elimination to an alkene is seen, but the product contains no halogens. Test reaction 2 (TR2) was constructed to test whether the model considered the double elimination as a two-step process: the singly eliminated alkene product from reaction 20 was reacted again with base. This results in the correct elimination to an alkyne. This supports our hypothesis that the model treats double elimination as a two-step process proceeding through a single elimination intermediate. However, this still does not resolve the problem of the model failing to generate even single elimination predictions for reactions 68 and 69.

While these errors are interesting cases, prediction failure where confidence is low is generally not problematic—users can be wary of predictions where the model itself is unsure of its answer. Cases where the model has high confidence but is incorrect are more troubling. Twelve reactions had incorrect top-1 predictions despite reaction confidence greater than 0.90

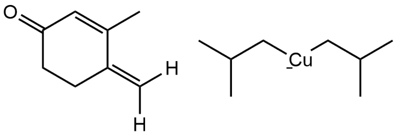
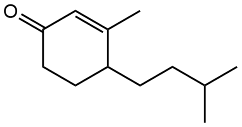
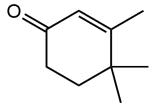
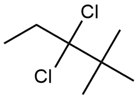
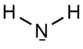
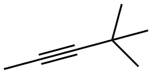
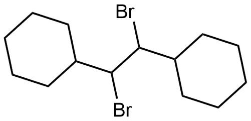
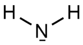
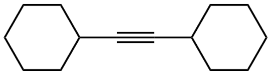
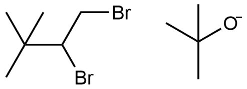
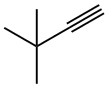
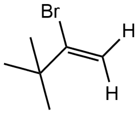
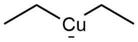
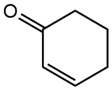
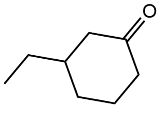
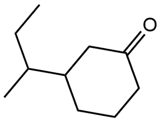
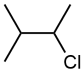
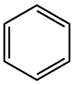
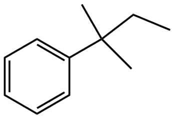
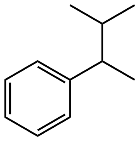
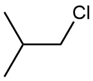
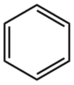
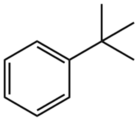
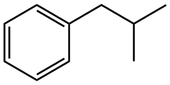
Reaction Number	Specified Reactants/Reagents	True Product	Forward Prediction Product
39	 Li +		
68	 Na+ 		Failure
69	 Na+ 		Failure
20			
40	 Li^+ 		
43	 Cl^- Al^{3+} Cl^- Cl^- 		
50	 Cl^- Al^{3+} Cl^- Cl^- 		

Figure 1: Examples of Problematic Forward Synthesis Prediction Results

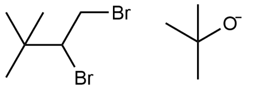
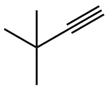
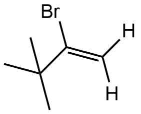
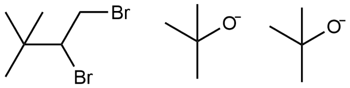
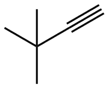
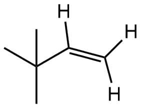
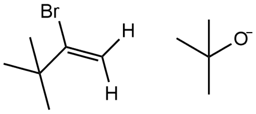
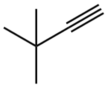
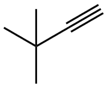
Reaction Number	Specified Reactants/Reagents	True Product	Forward Prediction Product
20			
TR1			
TR2			

Figure 2: Forward Synthesis Reaction 20 and Test Reactions

(see SI for full table).

Two of these (reactions 25 and 27) were cases where the prediction failed, and both were standard Grignard reactions where the model should tend to be successful. Six of these were cases where the model predicted a subsitution over an elimination reaction, showing a distinct favorability for substitution where reaction conditions promote elimination. The model does not heavily consider the differences in alkyl halide-base combinations that are taught as a heuristic for choosing between substitution or elimination mechanisms. Instead, it seems that the model relies on reactions that it knows well from the training data, namely substitutions, and therefore often predicts substitution whenever a base and an alkyl halide are present in the supplied reactants. Reaction 40 is an example where the model confidently predicts the incorrect product from an organocopper reaction, completely misrepresenting the number of carbons that should be added. In both reactions 43 and 50, the carbocation rearrangement for the Friedel-Crafts alkylation is not predicted, leading to incorrect products. These are examples where simple mechanistic considerations suffice to make predictions.

To further investigate why the model makes incorrect predictions, we used the tool RXN-Mapper, introduced by Schwaller et al. in 2021,[23] that provides unsupervised atom mapping from reaction SMILES. The model uses an attention-guided approach, and the authors posit that it should maintain higher mapping accuracy for correctly transcribed or sensible reactions and lower mapping accuracy for incorrect reactions or reactions that were incorrectly transcribed. Figure 3 shows the results of these mappings for synthesis reactions 20, 36, 40, 43, and 50. Reaction 20 is correctly mapped for a single elimination with relatively high confidence, corroborating our hypothesis that the model views double eliminations as a two-step process. Reactions 36 and 40 identify issues with understanding organocopper reagent reactivity, as neither mapping correctly indicates which carbons from the organocopper reagent should take part in the reaction. The mapping of reaction 36 shows only a single carbon of one isobutyl group taking part in the reaction, while the mapping of reaction 40 shows all carbons from both alkyl groups participating. The mappings for reaction 43 and 50 correctly identify that all of the carbons in the alkyl halide reactant should participate in the reaction, however the new carbon-carbon bond is made to the wrong carbon as a rearrangement is not indicated.

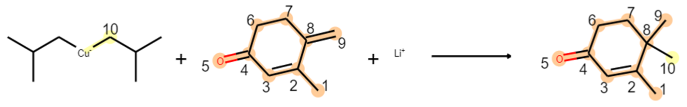
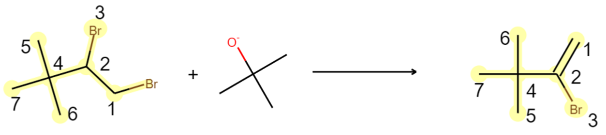
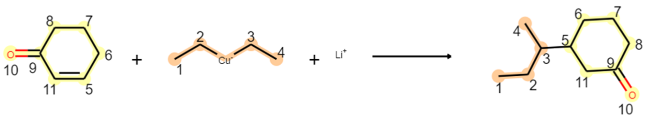
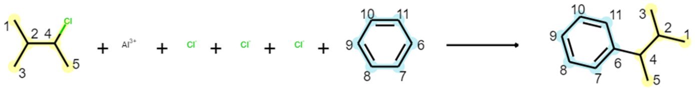
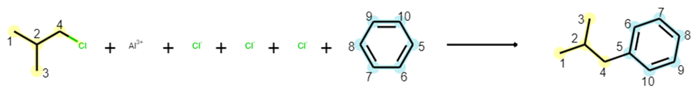
Reaction Number	Reaction Mapping	Mapping Confidence
36		0.39
20		0.60
40		0.71
43		0.29
50		0.36

Figure 3: Synthesis Prediction Atom Mappings

Retrosynthesis Prediction

In testing IBM RXN's retrosynthesis prediction, 100 retrosynthetic targets from the same ten reaction classes were supplied to the model, and up to five prediction results were recorded. To critically analyze the quality of retrosynthesis, we computed changes in molecular complexity (ΔCM) and synthetic accessiblity (ΔSA) for the predicted retrosyntheses and compared them to the results obtained from the literature synthetic routes. We also ran all predicted retrosyntheses through the forward prediction model to assess the model consistency.

All reactions returned a prediction, whose quality we evaluated on several criteria. Table 2 shows the mean retrosynthetic prediction confidence for each reaction class. Similar to the forward predictions, elimination-based retrosyntheses have characteristically low confidence, along with syntheses carried out with a Diels-Alder reaction in the literature.

Filtering on predictions with low (< 0.50) confidence, identified 17 reactions from the top-1 predictions, consisting mostly of elimination and Diels-Alder reaction classes. Reaction information for several low confidence predictions is shown in Figure 4, while the remainder can be found in the supporting information.

Reaction 10, a substitution-based retrosynthesis, predicts a reduction of formaldehyde-d_2 with hydrogen to form the target alcohol. Not only is this a starting material with one carbon when our product has two, curiously, the second deuterium on the original precursor is completely absent in the predicted product. To probe this error, we queried a retrosynthesis of ethanol, which is similar to the target molecule albeit without the deuterium, shown in Figure 5. This resulted in a very similar set of predicted precursors, indicating that the model is simply trying to match these cases to a training example, and giving a chemically unreasonable result in the process. It is disconcerting that the predictive confidence of this ethanol synthesis is 1.0, further emphasizing that this is a reaction the model knows well and wants to apply in the case of reaction 10. Including a single deuterium atom results in the lower quality prediction result, highlighting a significant obstacle to sequence-based methods. Moreover, the predicted synthesis for reaction 10 completely fails when run through the forward prediction model.

Reaction 11 provides potassium tert-butoxide and THF solvent as predicted precursors to make 2-methylpropene. This transformation is not chemically sensical as we would not expect potassium tert-butoxide to undergo elimination on itself. Moreover, the forward prediction for these predicted precursors makes tert-butyl acrylate rather than 2-methylpropene. Cyclohexene is the target for reaction 12, predicted to form from a reduction of benzene with hydrogen. While this is certainly possible, it is problematic that the same retrosynthetic prediction is made when cyclohexa-1,3-diene is the target, as in reaction 20. This indicates a lack of consideration of specific reaction/reagent conditions and consequently the forward prediction confidently fails for this set of precursors. This type of error is similar to the behavior seen in the double elimination cases for forward synthesis. Neither prediction is explicitly wrong, however more consistency within the model's decision-making would result in higher quality and less ambiguous predictions.

In reaction 17, propene is the desired target, however the prediction questionably provides methane as a single precursor. This is a very unlikely transformation, despite the target molecule being one of the most basic alkenes. In the final retrosynthesis, reaction number 92, we see another example of a lack of atom conservation. The target is a simple internal alkyne which could be synthesized using a substitution reaction. Though the set of predicted precursors does attempt a substitution, the set of starting materials has a total of 4 carbons while our target is a five-carbon compound. Seventeen (17) of the predicted retrosynthesis outcomes were not predicted to be successful by the forward prediction model. All of these were cases where the forward prediction failed to return a result, indicating that the predicted synthetic route was unlikely to succeed even by the standards of the model itself. Similarly to the forward prediction results, the performance of the retrosynthetic prediction model was satisfactory. While no failures were explicitly reported, users should be conscious of prediction confidence when considering the validity of each retrosynthesis.

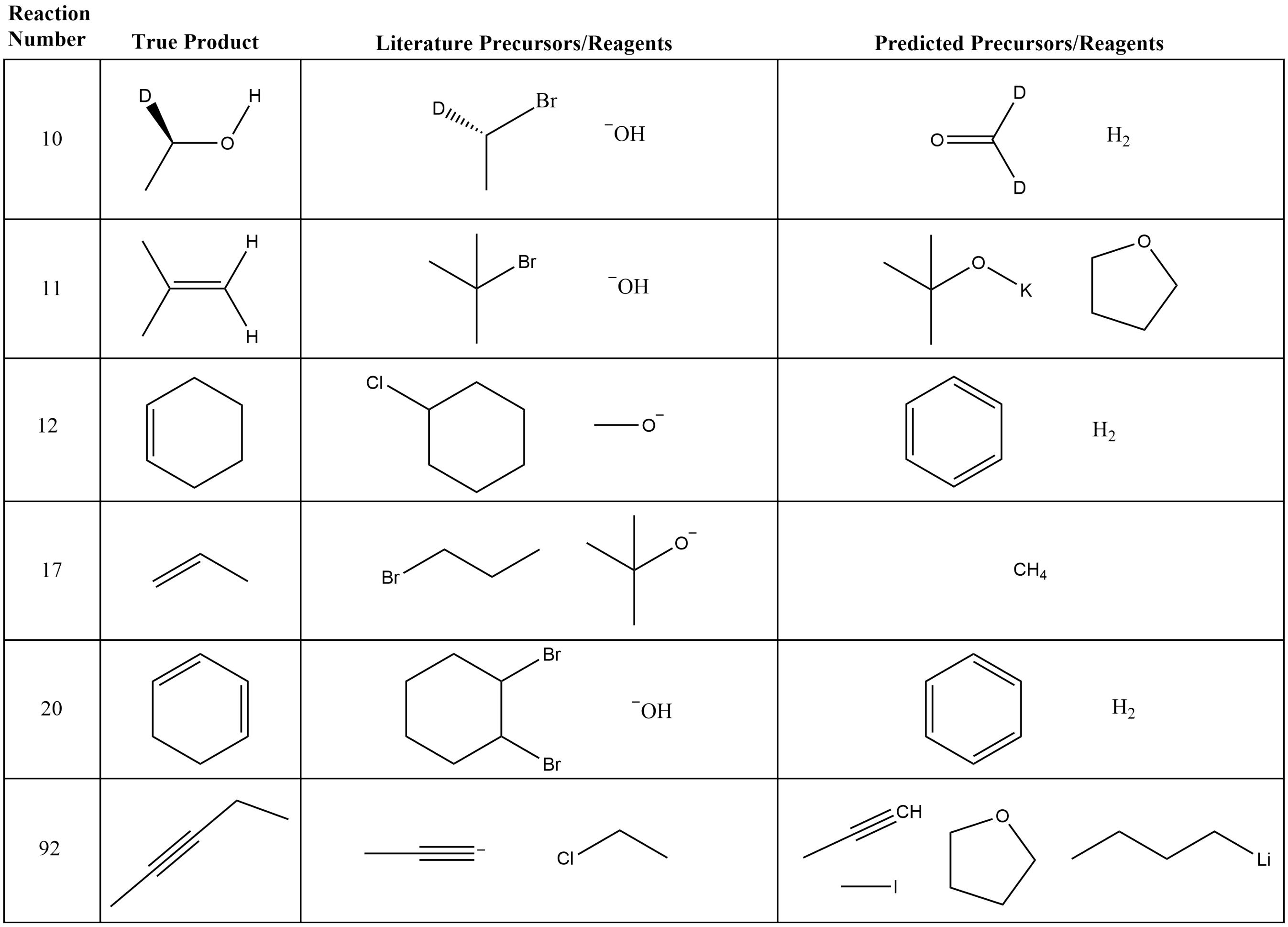


Figure 4: Examples of Problematic Retrosynthesis Prediction Results

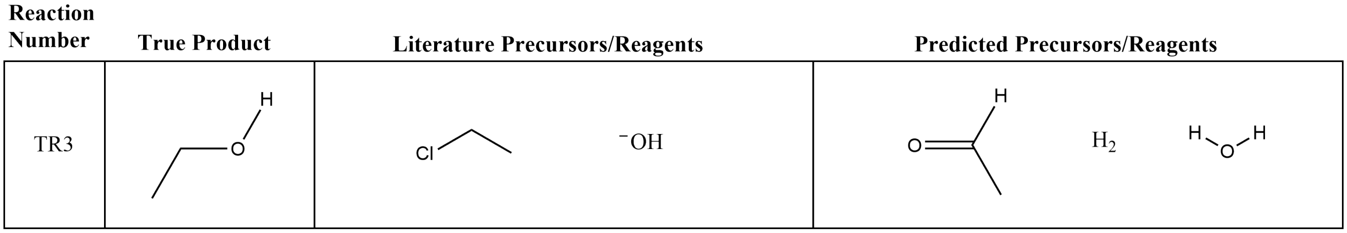


Figure 5: Retrosynthesis Test Reaction

Figure 6 shows the atoms mappings for retrosynthesis reactions 10, 11, 12, 17, 20 and 92, along with their mapping confidences. The mappings for reactions 10 and 17 show the addition of carbon atoms that are unaccounted for in the reactants, and do so with high mapping confidence, despite these transformations being unlikely. Reaction 11 is sensibly mapped and the confidence of 0.45 reflects the poor quality of the predicted transformation. The mappings for reactions 12 and 20 are also reasonable, however they both have low confidence, indicating a poor quality transformation. Finally, the mapping for reaction 92 shows an incorrect characterization of one of the n-butyllithium carbons as a reacting species. The model shows it adds to the original alkyne starting material, along with the carbon from the alkyl halide. The transformation does not make sense and again the low confidence of the mapping supports this.

Human Versus AI Synthesis Planning

Differences in molecular complexity and synthetic accessibility provide another way to compare literature and predicted syntheses and retrosyntheses. In organic synthesis, human experimenters often attempt to build up more complex molecules from simpler precursors, resulting in positive differences in these complexity measures for the reaction. There are cases where chemists will synthesize a more complex molecule and then collapse it into the desired product, or when complexity is temporarily increased due to the presence of protecting groups, however these cases are less applicable for the single-step examples we are considering.[5] Figure 7 shows scatter plots of difference in molecular complexity and synthetic accessiblity for forward synthesis and retrosynthesis prediction versus the literature reported reaction. The dashed line bisecting the plot is the identity line and not a trendline.

For the forward model, the coefficient of determination for difference in synthetic accessibility was a modest ($R^2 = 0.81$), and a weak ($R^2 = 0.24$) for difference in molecular complexity. While the overall distribution of differences in these metrics was similar for predicted and literature reactions, there was a notable peak in predicted reactions with a ΔCm

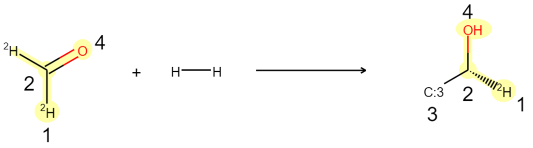
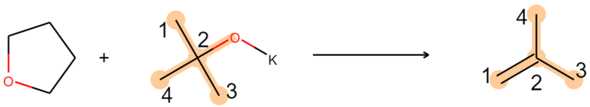
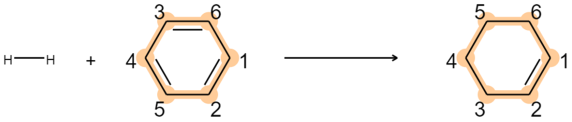
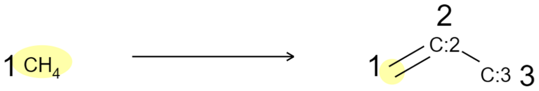
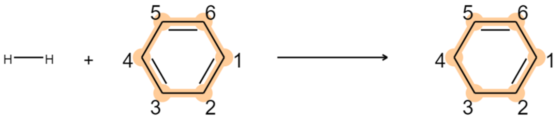
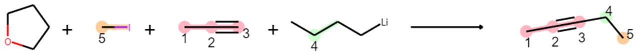
Reaction Number	Reaction Mapping	Mapping Confidence
10		0.84
11		0.45
12		0.25
17		1.00
20		0.22
92		0.20

Figure 6: Retrosynthesis Prediction Atom Mappings

between zero and fifty and a ΔSA beteween zero and ten. The retrosynthetic predictions had a very low coefficient of determination for ΔSA ($R^2 = 0.0021$) and a moderate coefficient of determination for difference in molecular complexity ($R^2 = 0.11$). While Figure 7 shows that predicted retrosyntheses have both positive and negative deviations in both synthetic accessibility and molecular complexity from that of the literature reactions, the distributions of change in both metrics did show a slight shift towards more negative values. If all other factors are equal or comparable, be it reaction yield, reagent cost, stereo- or regioselectivity, these alternative routes could aid in reducing bias in organic synthesis and provide novel variation in areas where human experimenters may be "stuck in a rut".[27]

Conclusion

We evaluated forward and retrosynthetic predictive performance of the IBM RXN for Chemistry model on representative data sets of organic chemistry reactions. Top-k accuracies for forward prediction on our dataset were lower than the figure reported in the original Molecular Transformer paper,[12] likely due to differences in the distributions from which training and test reactions were sampled. We identified representative cases of specific pathologies including failure to generate a prediction and incorrect predictions that fail to obey atom conservation or mechanistic considerations. These failures occur for both low confidence predictions (where they might be expected) and high confidence predictions (where they are particularly deleterious). Similarly for the retrosynthesis model, we identified failures associated with atom nonconservation and nonsensical chemical transformations, as well as a lack of reagent specificity with similar target molecules. We also showed cases where the predicted retrosynthesis completely failed when run through the forward prediction model. We compared differences in synthetic accessibility and molecular complexity for the predicted syntheses and retrosyntheses compared to the literature, and found that while the forward model reproduced similar distributions to the literature, the retrosynthesis model did not.

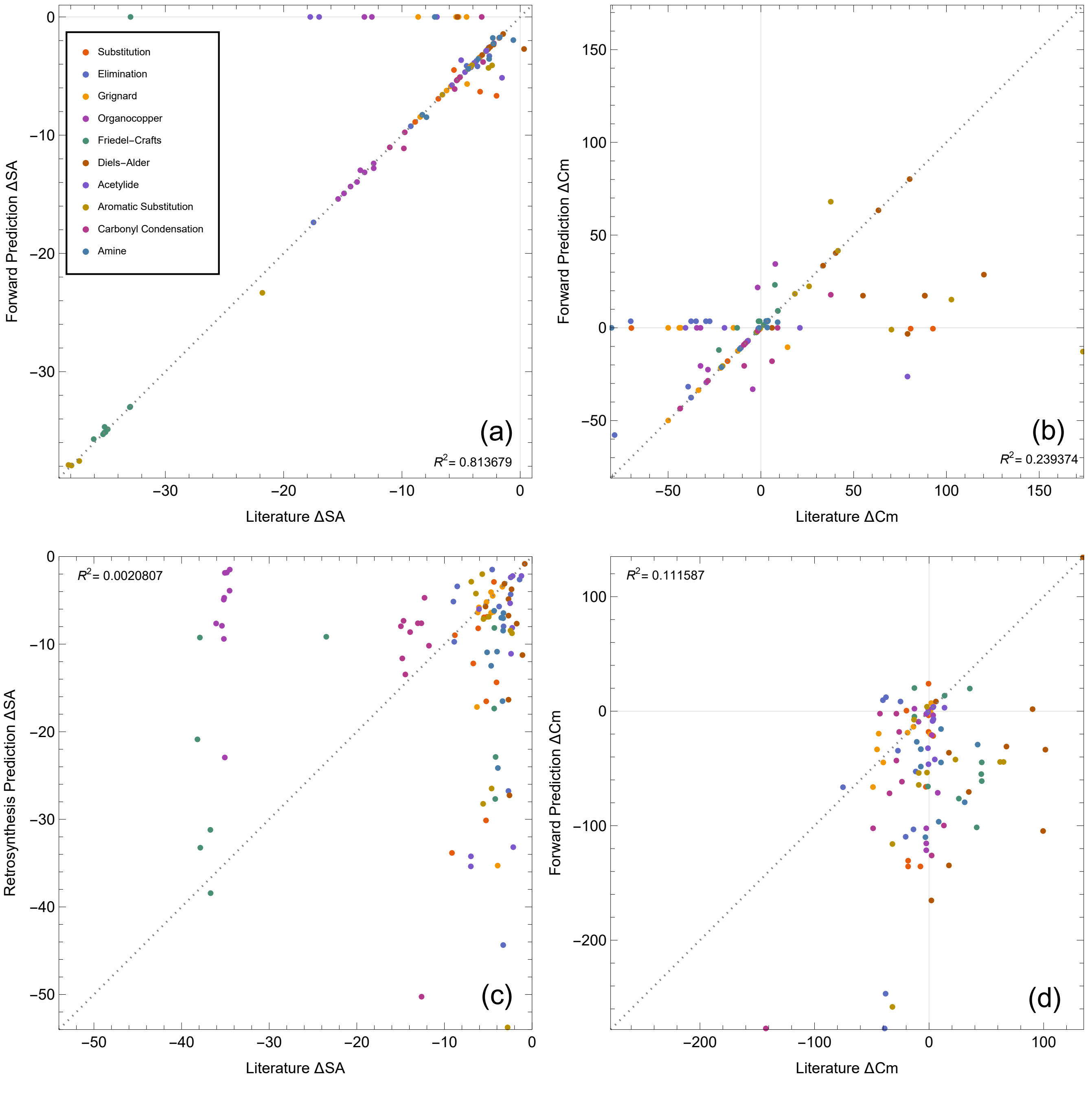


Figure 7: Differences in Synthetic Accessibility and Molecular Complexity

Instead, both positive and negative deviations from the literature complexity were seen. Finally, we used RXNMapper to expose errors in how the model identifies reactive atoms and species, and how mapping confidence can identify poorly predicted chemical transformations. These observations highlight areas for targeted improvement and may provide a useful benchmark for assessing the performance of future organic reaction prediction schemes.

Acknowledgments

We thank Nicholas DiGrandi for stimulating conversations on predictive model performance for classic organic chemistry reactions as well as Dr. Böttcher for his inspiration and assistance in developing our code for computing molecular complexity. The authors acknowledge the support from the National Science Foundation (Grant No. DMR-1928882) and the Henry Dreyfus Teacher-Scholar Award (Grant No. TH-14-010).

Supporting Information

The following code and data sets are openly available from our GitHub repository: https://github.com/wrborrelli/seq_model_SuppInfo

- transformer_eval_SI.pdf: supporting information document
- test_set_WB.xlsx : forward synthesis test set
- test_set_ret_WB.xlsx : retrosynthesis test set
- combDatAllFwds.csv : all forward synthesis model results
- combDatAllRet.csv : all retrosynthesis model results
- combDatFwds.csv : top-1 forward model results
- combDatRet.csv : top-1 retrosynthesis model results

- combFwdDat.csv : top-1 forward results of top-1 retrosynthesis model results

- figs_and_reacs.nb : Mathematica code for generating figures and reactions

- retro_analysis.nb : Mathematica code for retrosynthesis analysis

- synth_analysis.nb : Mathematica code for forward synthesis analysis

- rmap.ipynb : Python code for RXNMapper analysis

- t1CMapConfs.txt : RXNMapper map confidences for top-1 correct predictions

- t1WRmapConfs.txt : RXNMapper map confidences for top-1 incorrect predictions

References

(1) Reactants, products, and transition states of elementary chemical reactions based on quantum chemistry. *Scientific Data* **2020**, *7*, 137.

(2) Schwaller, P.; Laino, T. *Machine Learning in Chemistry: Data-Driven Algorithms, Learning Systems, and Predictions*; ACS Symposium Series; American Chemical Society, 2019; Vol. 1326; pp 4–61.

(3) Valavanidis, A. Synthetic Organic Chemistry and the Emergence of Artificial Intelligence-Driven Technology to Synthesize Target Chemical Compounds. **2020**, *1*, 1–26.

(4) 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. *Angewandte Chemie International Edition* **2016**, *55*, 5904–5937.

(5) Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature* **2018**, *555*, 604–610.

(6) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. Computer-Assisted Retrosynthesis Based on Molecular Similarity. *ACS Central Science* **2017**, *3*, 1237–1245.

(7) Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of chemical reaction yields using deep learning. *Machine Learning: Science and Technology* **2021**, *2*, 015016.

(8) Kreutter, D.; Schwaller, P.; Reymond, J.-L. Predicting enzymatic reactions with a molecular transformer. *Chem. Sci.* **2021**, –.

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

(10) Howard, J.; Ruder, S. *Universal Language Model Fine-tuning for Text Classification*; 2018; pp 328–339.

(11) Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*; 2018.

(12) Schwaller, P.; Laino, T.; Gaudin, T.; Bolgar, P.; Hunter, C. A.; Bekas, C.; Lee, A. A. Molecular Transformer: A Model for Uncertainty-Calibrated Chemical Reaction Prediction. *ACS Central Science* **2019**, *5*, 1572–1583.

(13) Schwaller, P.; Petraglia, R.; Zullo, V.; Nair, V. H.; Haeuselmann, R. A.; Pisoni, R.; Bekas, C.; Iuliano, A.; Laino, T. Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy. *Chemical Science* **2020**, *11*, 3316–3325.

(14) IBM RXN for Chemistry. `https://rxn.res.ibm.com`.

(15) rxn4chemistry, rxn4Chemistry. 2020; `https://github.com/rxn4chemistry/rxn4chemistry`.

(16) Borrelli, W. IBMRxnAPI. 2020; `https://github.com/wrborrelli/IBMRxnAPI`.

(17) Smith, J. G. *Organic Chemistry*; McGraw-Hill Education, 2017.

(18) Stuart Warren, P. W. *Workbook for Organic Synthesis The Disconnection Approach*; John Wiley and Sons, 2009.

(19) Bottcher, T. An Additive Definition of Molecular Complexity. *Journal of Chemical Information and Modeling* **2016**, *56*, 462–470.

(20) Borrelli, W. MolecularComplexityMA. 2020; `https://github.com/wrborrelli/MolecularComplexityMA`.

(21) Borrelli, W. Molecular Complexity. 2021; `https://resources.wolframcloud.com/FunctionRepository/resources/MolecularComplexity`.

(22) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics* **2009**, *1*, 8.

(23) Schwaller, P.; Hoover, B.; Reymond, J.-L.; Strobelt, H.; Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances* **2021**, *7*, eabe4166.

(24) Griffiths, R.-R.; Schwaller, P.; Lee, A. A. Dataset Bias in the Natural Sciences: A Case Study in Chemical Reaction Prediction and Synthesis Design. 2021.

(25) Lowe, D. M. Extraction of Chemical Structures and Reactions from the Literature. Ph.D. Thesis, University of Cambridge, 2012.

(26) Thakkar, A.; Kogej, T.; Reymond, J.-L.; Engkvist, O.; Bjerrum, E. J. Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chemical Science* **2020**, *11*, 154–168.

(27) Jia, X.; Lynch, A.; Huang, Y.; Danielson, M.; Lang'at, I.; Milder, A.; Ruby, A. E.; Wang, H.; Friedler, S. A.; Norquist, A. J.; Schrier, J. Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature* **2019**, *573*, 251–255.