Write-up: IBM RXN for Chemistry Retrosynthesis Prediction Evaluation

William R. Borrelli* and Joshua Schrier*

Department of Chemistry, Fordham University, The Bronx, NY

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

Phone: (626)399-1826; (718) 817-4453

Introduction

Earlier this year, Schwaller et al. introduced a transformer-based model using a hyper-graph exploration strategy for predicting retrosynthetic pathways.[1] Their architecture by-passes the need for laborious rule-based methods by treating reaction prediction as a translation problem, greatly increasing scalability and ease of encoding. In its application to forward reaction prediction, their seq2seq model achieved remarkable accuracy on several representative datasets.[2] To faciliatate the use of their Molecular Transformer in retrosynthetic pathway prediction, a hyper-graph exploration strategy was utilized. To critically assess the performance of their retrosynthetic predictions, several metrics were devised, including round-trip accuracy, coverage, class diversity, and the Jensen-Shannon divergence. Their predictive model is publicly available online on the IBM RXN for Chemistry website.[3] Wrappers for programmatic API access are available in Python[4] as well as through our Mathematica[5] implementation. In this work, we complete a general evaluation of the retrosynthetic prediction performance on a curated dataset of organic chemistry reactions representative of those one would learn in an undergraduate organic chemistry curriculum. Through this analysis we critically examine several of the claims put forth in Schwaller et al.'s paper, including those regarding the round-trip accuracy and coverage of their predictive model, as well as its viability in use for teaching undergraduate organic chemistry.

Methods

Our test set was comprised of 100 single-step retrosynthesis reactions encompassing 16 reaction classes. Many of these reactions were sourced from introductory organic chemistry textbooks,[6,7] with our reasoning being that they should be appropriate examples for a student trying to use IBM Rxn for Chemistry as a resource during his or her academic pursuits. The fact that these reactions should also be correct and generally "ordinary" enhances our ability to evaluate the performance of the model. SMILES for each of the reactions were generated in ChemDraw 19.1 and then canonicalized in Mathematica. Each reaction class was assigned a project ID and a new retrosynthesis was run for each of the reaction products, as well as all of the canonicalized product SMILES. Additionally, all literature reactants were run through the forward prediction. All results were collected in a data set with metrics from each prediction such as confidence and optimization recorded. In order to evaluate the characteristics of the predicted retrosynthetic pathways, two measures relating to molecular complexity were employed. The first of which was based off Böttcher's 2016 work on an additive definition of molecular complexity (Cm).[8] Along with guidance from Dr. Böttcher, Mathematica code for computing complexity in this way was written, and is available as a GitHub repository.[9] Additionally, Ertl and Schuffenhauer's 2009 paper introducing a measure of synthetic accessibility based on molecular complexity[10] is readily available as a metric in Mathematica. The difference in both of these metrics, molecular complexity and synthetic accessiblity, 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. A good retrosynthesis should tend to build up more complex molecules from less complex starting materials, as indicated by a negative ΔSA or ΔCm. Utilizing these measures, as well as the reported prediction measures from the API, we began our analysis by examining how these variables were distributed.

Distribution Analysis

The first of our distribution analyses was focused on the correlations between ΔSA and ΔCm for the literature, retrosynthetic prediction, canonical retrosynthetic prediction, and forward predictions reactions. Figure 1 shows a scatter plot of ΔSA and ΔCm for the retrosynthetic predictions versus literature reactions.

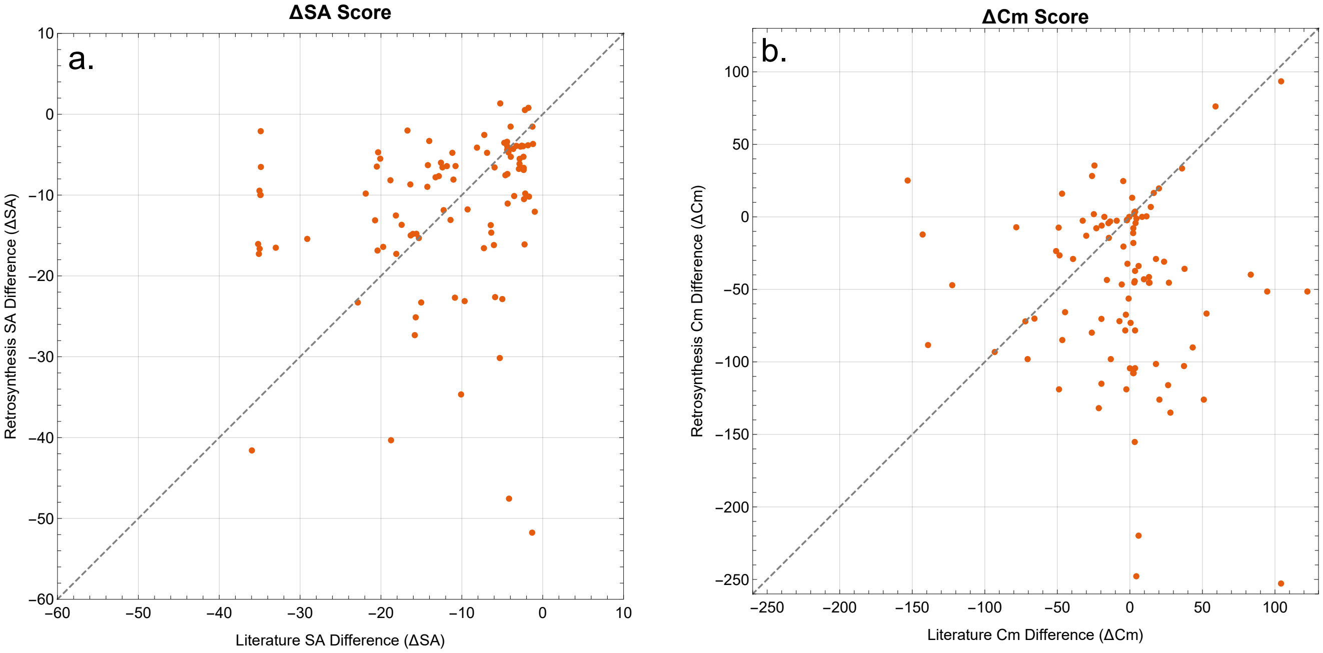


Figure 1: (a) Scatter plot of change in SA for the retrosynthetic prediction vs. literature reactions and (b) Scatter plot of change in Cm for the retrosynthetic prediction vs. literature reactions

The change in synthetic accessibility data is moderately positively correlated with a correlation coefficient of 0.21, whereas the change in molecular complexity data is weakly negatively correlated with a correlation coefficient of -0.073. It should also be noted that the retrosynthetic prediction reactions have several examples with a positive change in synthetic accessibility, whereas the literature reactions do not. Figure 2 shows similar scatter plots comparing the distributions for the forward prediction reactions and the literature reactions.

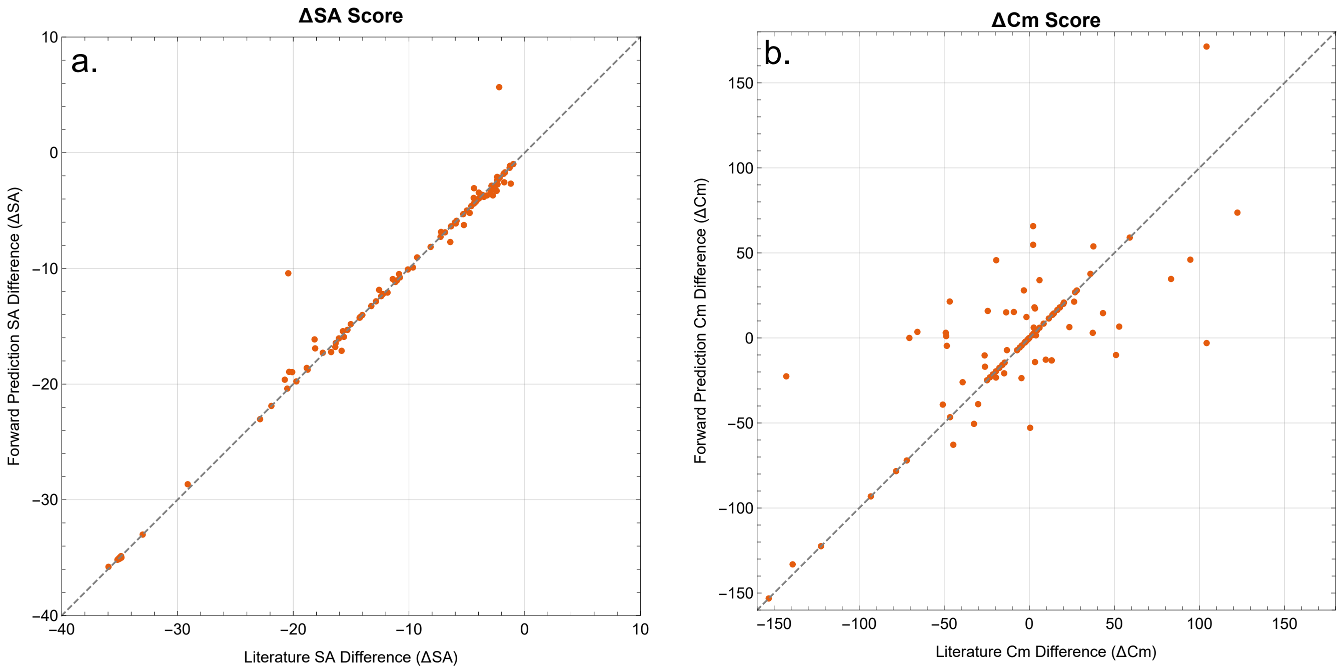


Figure 2: (a) Scatter plot of change in SA for the forward prediction vs. literature reactions and (b) Scatter plot of change in Cm for the forward prediction vs. literature reactions

It can be seen that in the case of the forward prediction reactions versus the literature reactions, there is a much stronger positive correlation in both change in synthetic accessibility and molecular complexity. The correlation coefficients for ΔSA and ΔCm are 0.99 and 0.76 respectively.

Round-Trip and Coverage

In order to adequately evaluate the performance of their retrosynthetic model, Schwaller et al. introduced several metrics, two of which are the round-trip accuracy and coverage. Round-trip accuracy evaluates the percentage of retrosynthetic pathways that successfully produce the desired product when run through the forward prediction model, and the coverage is the percentage of products where at minimum one precursor set is valid. To account for the sensitivity of round-trip scores to the number of alternative retrosynthetic routes tested, we chose to only select alternative sequences with a confidence value within a standard deviation of the mean confidence for all of the given retrosynthetic sequences. In this way, only sequences within a relatively homogeneous confidence interval are taken as to not bring down the round-trip scores by selecting the lowest confidence routes. Coverage percent was also evaluated using only sequences chosen with similar selectivity. These chosen retrosyntheses are hereby referred to as “all reasonable” sequences. Both round-trip and coverage were also computed for solely the highest confidence retrosynthetic sequences for each product. Table 1 shows the round-trip and coverage percents for the top confidence and all reasonable retrosyntheses.

Table 1: Round-Trip and Coverage Percents

Top 1 Confidence RT (%)	All Reasonable RT (%)	All Reasonable Coverage (%)
95.50	70.72	89.16

These values are reasonable compared to what is cited in Schwaller et al.’s original paper, which claims round-trip percents between 71 and 82 percent, along with coverage percents between 92 and 96 percent. We acknowledge the increased round-trip score for the top confidence retrosynthetic pathway as being indicative of the sensitivity of the round-trip to the number of alternative sequences queried.

Reactions with Low Confidence or Positive ΔSA/ΔCm

In this analysis, we sought to examine the reactions that had confidence values lower than 0.1, or a change in synthetic accessibility or molecular complexity greater than zero. Table 2 shows the confidence scores and change in synthetic accessiblity and molecular complexity for 9 reactions with a retrosynthesis confidence below 0.1, several of which have positive changes in synthetic accessibility or molecular complexity.

Table 2: Confidence, ΔSA, and ΔCm

Reaction Number	Retrosynthesis Confidence	Ret. ΔSA	Ret. ΔCm	True ΔSA	True ΔCm
16	0.000349	-12.06	76.16	-0.98	59.03
19	0.0229	0.52	-12.17	-2.20	-142.85
23	0.00992	-6.72	35.40	-2.38	-24.42
68	0.00721	-4.00	93.46	-2.75	104.24
71	0.00283	-3.97	-116.04	-4.38	26.34
72	0.000131	-1.53	33.42	-3.97	35.92
78	0.000198	0.79	0.0	-1.77	8.42
96	0.0000155	1.34	-104.42	-5.25	0.0
98	0.0211	-15.42	-7.42	-29.11	-49.11

Figure 3 shows a 8 reaction subset of retrosyntheses with confidence values less than 0.1, along with their proposed reactants, the true product that was inputted to the model, and the product outputted when the retrosynthesis is run through the forward prediction.

Reaction 16 appears to completely fail to produce the target and instead returns a reactant as the sole product. Reaction number 19 fails to find a retrosynthetic route, yet provides cyclohexanone as a lone reactant for the production of methylene cyclohexane. There are several straightforward routes to this product, such as a dehydration-based synthesis or a Wittig reaction, making this prediction failure troubling. Reaction 23 predicts anti-addition of bromine to the double bond rather than the benzylic bromination desired. A radical bromination, perhaps using NBS, would be preferable as the benzylic position would be favored. Reaction 68 again produces the incorrect cyclic product, choosing not to proceed with a Diels-Alder reaction using a cyclohexyl group functionalized diene. Reaction 71 also forgoes what would be a very simple Diels-Alder synthesis in favor of an attempted reaction of cyclohexene and sulfuric acid. Reaction 96 had the lowest retrosynthetic confidence in our data set, attempting an amide hydrolysis of a very complex starting material. The final example, reaction 98, attempts an alkyne reduction with sodium and liquid ammonia, yet it incorrectly introduces an extra carbon in the chain. The target alkene is octene, yet the retrosynthesis proposes a nonyne reduction. This reaction would produce an alkene with an extra carbon relative to our desired product, which is predicted by the forward model.

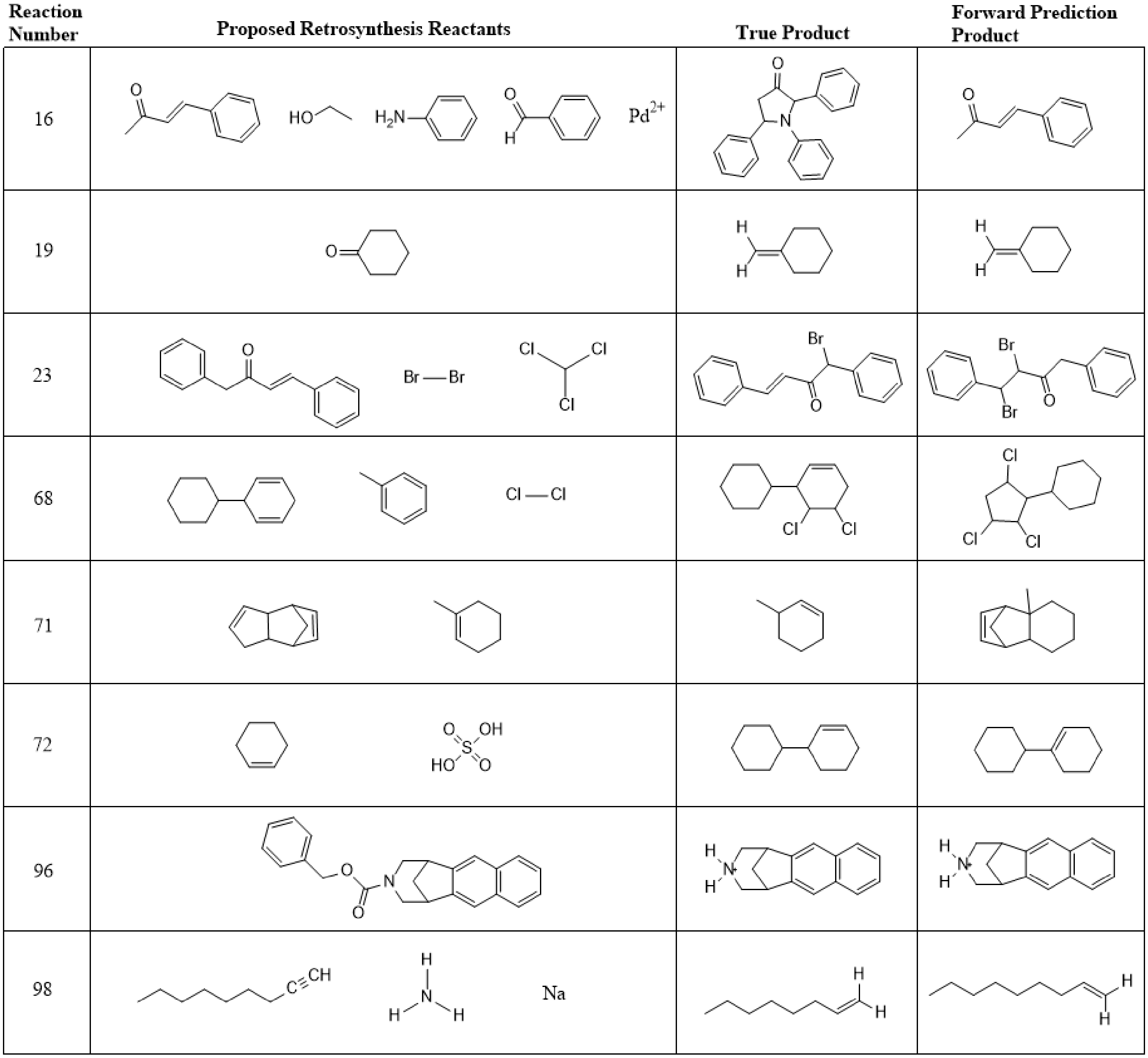


Figure 3: Low Confidence Retrosyntheses Reactants and Products

While the alkyne reduction proposed is a very reasonable pathway, the incorrect number of carbons remains troubling. This is a known disadvantage of transformer-based models as atom conservation is not taken into account in any way. In this case, the literature route utilizes a Hofmann elimination of 2-aminooctane, but a dehydration of 1-octanol or a reduction of octyne are also reasonable and fundamental methods to produce the desired product. It is interesting to note that many of the failed or very low confidence reactions result from products that are easily synthesized with Diels-Alder reactions. This is shown in Figure 4 as Diels-Alder is, on average, among the lowest confidence reaction classes tested.

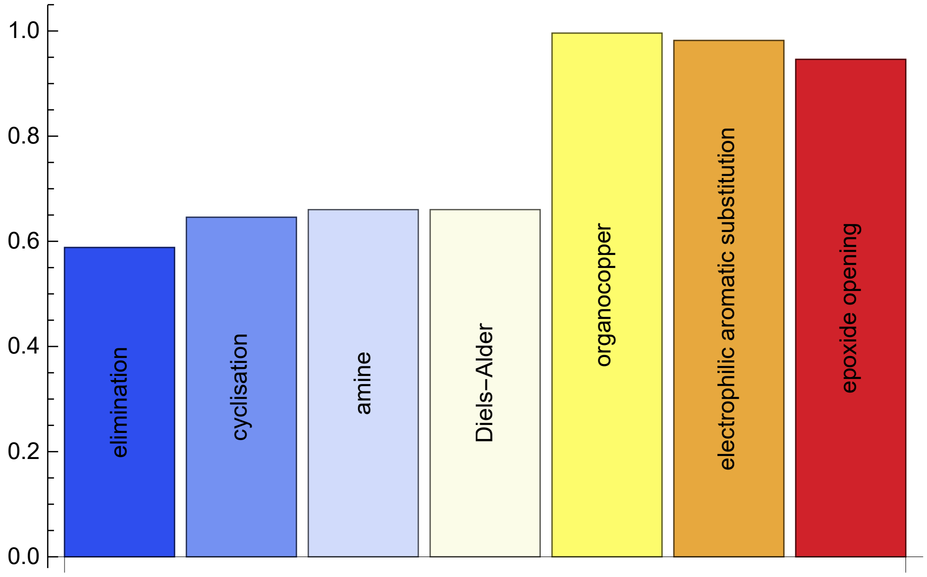


Figure 4: Mean Confidence Values per Reaction Class

Only one reaction, reaction number 42, had a retrosynthesis confidence above 0.1 and a positive change in molecular complexity and failed to produce the desired product when run through the forward prediction. As shown in Figure 5, reaction 42 has a relatively high retrosynthesis confidence of 0.769, and a reasonable forward prediction confidence of 0.507, however it still produces the incorrect product. This example is interesting as all of the other failed reactions have retrosynthesis confidences that are orders of magnitude lower than that of reaction 42. In general, filtering predictions by retrosynthesis confidence is sufficient to separate successes from failures, however this case shows that alternative metrics may be required under certain circumstances. While the best alternative measure may not be a measure of molecular complexity or synthetic accessibility, this work shows that both may have merit.

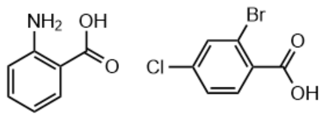
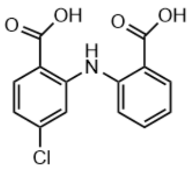
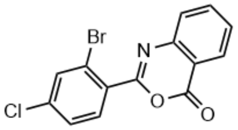
Rxn #	Proposed Retrosynthesis Reactants	True Product	Forward Prediction Product	Ret. Conf.	Fwd. Conf.	Ret. ΔSA	Ret. ΔCm
42				0.769	0.507	-1.54	3.03

Figure 5: Prediction Metrics for Reaction 42

Conclusion

Schwaller et al.'s molecular transformer model as applied to retrosynthetic prediction performs relatively well when utilized for textbook retrosynthetic targets. The cited values for round-trip accuracy and coverage are validated with our curated data set. Despite these successes, examples of predictive failure remain, whether through a disagreement between the true product and the product from the corresponding forward prediction, or in the retrosynthesis itself failing to obey conservation of atoms. We showed that in general one may differentiate between successful and unsuccessful predictions by sorting by retrosynthesis confidence, as well as identified a case where an additional metric may be useful. We presented the use of synthetic accessibility and molecular complexity differences to this end, and show that both aided in identifying a relatively high confidence, incorrectly predicted example.

Supporting Information

The following files are openly available from our GitHub repository: https://github.com/wrborrelli/IBMRxnEvaluation

- WB_IBMRxnRetro_Formatted.nb: Mathematica notebook of the code used for each part of the analysis, abbreviated and annotated for ease of viewing and replication
- ibmRxnRetro.nb: Mathematica notebook containing the raw code used for much of the analysis
- IBMRxn_Eval_Figs.nb: Mathematica notebook containing code used for generating several of the figures used in this paper
- allDatAnaly98.csv: CSV file of the curated dataset containing all reactions and initial retrosynthetic sequences

Acknowledgement

We thank Dr. Thomas Böttcher for his inspiration and assistance in developing our code for computing molecular complexity.

References

(1) 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.

(2) 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.

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

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

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

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

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

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

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

(10) 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.