doi:10.2533/chimia.2019.997 Chimia 73 (2019) 997–1000

Data-driven Chemical Reaction Prediction and Retrosynthesis

Vishnu H Nair, Philippe Schwaller, and Teodoro Laino*

Abstract: The synthesis of organic compounds, which is central to many areas such as drug discovery, material synthesis and biomolecular chemistry, requires chemists to have years of knowledge and experience. The development of technologies with the potential to learn and support experts in the design of synthetic routes is a half-century-old challenge with an interesting revival in the last decade. In fact, the renewed interest in artificial intelligence (AI), driven mainly by data availability, is profoundly changing the landscape of computer-aided chemical reaction prediction and retrosynthetic analysis. In this article, we briefly review different approaches to predict forward reactions and retrosynthesis, with a strong focus on data-driven ones. While data-driven technologies still need to demonstrate their full potential compared to expert rule-based systems in synthetic chemistry, the acceleration experienced in the last decade is a convincing sign that where we use software today, there will be AI tomorrow. This revolution will help and empower bench chemists, driving the transformation of chemistry towards a high-tech business over the next decades.

Keywords: Artificial intelligence · Organic chemistry · Reaction prediction · Retrosynthesis

Vishnu H Nair completed his Bachelor studies at the Indian Institute of Technology Bombay in 2019. He is currently working in the Cognitive Computing and Industry Solutions department at the IBM Research – Zurich Laboratory (ZRL) as an intern on different machine learning methods applied to chemical data.

Philippe Schwaller completed his Bachelor and Master studies in Materials & Engineering at EPFL and obtained an MPhil degree in Physics from the University of Cambridge. Currently, he is a predoctoral researcher at IBM Research – Zurich and pursuing his doctoral studies at the University of Bern in the group of Prof. Reymond. His research focus lies on accelerating organic synthesis using machine learning methods.

Teodoro Laino is a Principal Research Staff Member and the technical leader for Chemistry/Materials at IBM Research – Zurich. He received his degree in Theoretical Chemistry in 2001 (University of Pisa and Scuola Normale Superiore di Pisa) and the doctorate in computational chemistry from the Scuola Normale Superiore di Pisa, Italy supervised by Prof. Michele Parrinello in 2006. From 2006 to 2008, he worked as a post-doctoral researcher in the research group of Prof. Dr. Jürg Hutter at the University of Zurich, where he developed algorithms for *ab initio* and classical molecular dynamics simulations. Since 2008, he has been working in the department of Cognitive Computing and Industry Solutions at the IBM Research – Zurich. The focus of his research is complex molecular modelling for industry-related problems (energy storage, life sciences and nano-electronics) and the application of machine learning/artificial intelligence technologies to chemistry and materials science problems.

*Correspondence: Dr. T. Laino, E-mail: ZRLTEO@ch.ibm.com
IBM Research – Zurich, Säumerstrasse 4, CH-8803 Rüschlikon

1. Introduction

Mastering synthetic organic chemistry is a challenge that involves years of knowledge and practical experience composed of heuristics and hard rules. Over the past 50 years, there have been many attempts to help bench-chemists with the design of synthetic routes[1] but, with few exceptions based on expert systems,[2] most of the contributions did not yield any practical usage. In the last years, AI algorithms applied to organic chemistry problems demonstrated the great value of data-driven technologies, easing the task of constructing highly accurate predictive models. With the current computational power, it takes only a few weeks to build a fully data-driven prediction model compared to the multi-year efforts to build humanly curated rule-based expert systems.[2]

AI algorithms use digital data to encode the chemical rules used to make new predictions in computers. For example, computers can be trained on a set of known chemical reactions to allow them to predict the outcomes of new reactions with high accuracy.[3,4] While many factors have contributed to the recent rise of data-driven methods in the field of organic synthesis, the main pillars are algorithmic developments, the availability of large data sets and broad accessibility to larger computational resources. The emerging methods range from the prediction of a chemical reaction[3] to chemical reactivity prediction,[5] from retrosynthesis[6] to reaction condition optimisation[7] all the way to yield predictions.[8] In this article, we will focus on the developments of reaction prediction and retrosynthesis (see Fig. 1).

2. Reaction Prediction

Predicting the outcome of chemical reactions is a fundamental knowledge upon which modern chemistry is built. It is an essential part of the retrosynthetic analysis used in many fields like drug design and material synthesis. Determining the products of simple reactions may be a straightforward problem for a domain expert with decades of synthetic chemistry experience.

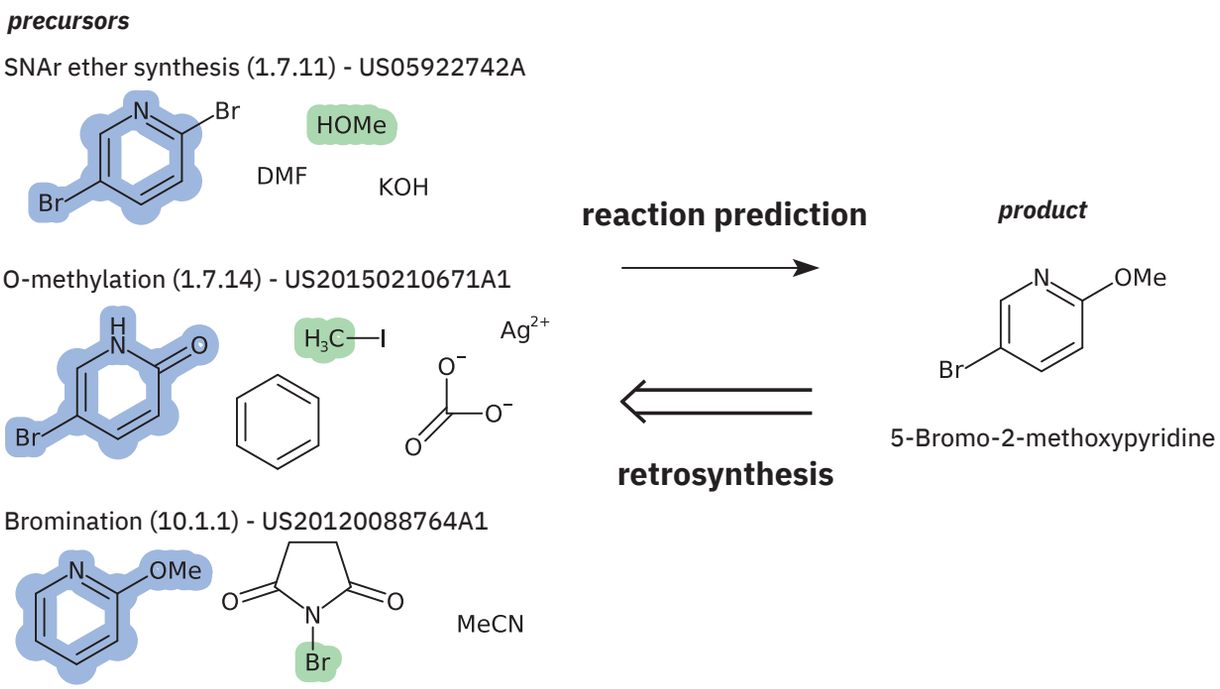

Fig. 1. Reaction prediction and retrosynthesis. Highlighted are the fragments that structurally contribute to the product.

But in the case of more complex reaction schemes, the outcome may be far from trivial as the nature of the reactants, reagents and physical conditions may strongly affect the formation of one product or another. The complexity of these inputs poses a serious problem when human experts try to rationalise the reactivity problem with a set of concise rules. This was the reason why the concept of automating chemical reaction prediction with the aid of computer algorithms became very popular a few decades ago.[1,9,10] Until the recent dawn of AI, the preferred computer-based methods were built around the notion of rule-based expert systems[11] and quantum mechanical simulations.[12] With only one work[7] reporting on the performance of AI algorithms versus humans in the domain of forward chemical reaction prediction and because of its limited statistical significance, it is not clear whether chemical reaction prediction systems today can truly beat humans. Nonetheless, the efforts to reach this goal are progressing at a fast pace.

2.1 *Rule-based Expert Systems*

Initial computer programs, carrying out rule-based systems, were made of hand-coded graph rearrangement patterns fitting a certain template to describe a reaction in its most abstract form (see Fig. 2). The candidate products of a reaction were generated from the rules and ranked according to their likelihood. One of the earlier implementations of such a system is the 'Computer-Assisted Mechanistic Evaluation of Organic Reactions' (CAMEO),[11] which predicts the products when given the reactants and conditions. In CAMEO, each reaction was assessed with the knowledge of reaction mechanisms through identifying the centres of reactive electrophiles and nucleophiles. A modern system, belonging to the same family of rule-based systems, is Chematica, currently known as Synthia.[2] Chematica has around 85'000 reaction templates internally codified, thus achieving a state-of-the-art level of accuracy for forward prediction and retrosynthesis. One of the major limitations of Chematica is the time it takes to compile new reaction templates and maintain the database. To add an entry into the existing database, one must make sure that none of the other rules invalidates itself, which makes the approach less scalable. Also, these systems predict reactions based on the overall chemical transformation and individual steps are not taken into consideration, thereby ignoring relevant chemistry.

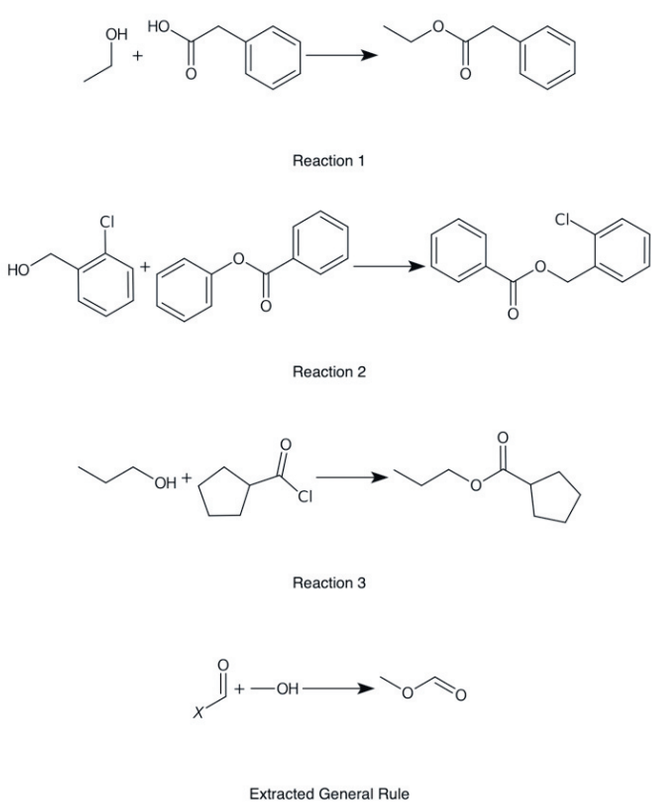

Fig. 2. Procedure of extracting reaction rules for the formation of an ester from reactions of a primary alcohol with an acyl group.

2.2 *Quantum Mechanics-based Systems*

Quantum Mechanical (QM) approaches are theoretically the apt method to tackle the problem of reaction prediction. They rely on modelling the electron distribution over the orbitals of the different atoms in the reaction and simulating the interaction among them to determine the possible orbital transformations that might take place and lead to bond breakage/formation. Even though first principle QM works well for relatively small systems (up to few hundred atoms, depending on the algorithmic complexity), its application to larger systems is still computationally challenging.

Despite the large corpus of approximations that can be exploited, QM calculations are still far from mimicking the results of wet-lab experiments.[12]

2.3 *Machine Learning Systems*

Machine Learning (ML) approaches build a model based on a mathematical representation that learns from the data they are trained on. Handling the large space of chemical compounds requires a data set of considerable size which should ideally be representative of the data on which the model is being tested. Getting hold of publicly available, high-quality reliable data is a challenge in itself. This forces many people to make use of privately curated data sets, *e.g.* Reaxys[13] or Pistachio.[14] One of the challenges related to the data in chemical reactions is that it only showcases reactions that actually take place, have good yield, and are of academic/industrial importance. Similar to the lack of data for unstable compounds as reported by Koichi and Lüthi in this issue,[15] the data sets for chemical reactions do not contain negative examples. For instance, reactions that failed or had a low yield are not commonly reported. This causes important statistical imbalances that need to be included in the overall picture in order to understand the performance of the trained AI models.

Initial works using the ML paradigm revolved around ranking the reaction templates according to the probability of that reaction taking place with the given reactants.[16,17] Reaction templates are predefined schemas composed of general reactant structures and the transformations that can lead to a product. Coley *et al*.[18] deduced that a reaction template might match more than one reactive centre in the reactants and can yield more than one product. Thus, they published a work that generated the products by using all the templates and then ranked the products with a neural network, achieving a top-1 prediction of 71.8%. More recently, the same group presented a model to predict bond changes using the molecular graphs of the reactants as inputs.[5,19] In contrast, Segler and Waller[20] searched for novel reactions by predicting missing links between existing molecules in a chemical space knowledge graph. Inspired by earlier work on a manually generated data set of mechanistic steps,[21] Bradshaw *et al*.[22] predicted electron paths to generate the outcome of a reaction.

Cadeddu *et al*.[23] showed that there are fragments in organic molecules whose distribution is identical to the one of words in natural languages. This paved the way for regarding the reaction prediction task as a translation problem from reactant string to product string, essentially converting it to a problem of machine translation in Natural Language Processing (NLP). The key is that chemical species have to be represented in a text-based form. One example of such a representation is the simplified molecular-input line-entry system (SMILES),[24] which is a line notation of a chemical structure, *e.g.*

`O=C(C)OC1CCCCC1C(=O)O` represents Aspirin

Nam and Kim[25] built upon this approach and applied a sequence-to-sequence model for predicting the reaction outcome. Schwaller *et al*.[3] demonstrated a similar approach in a much wider setup. They showed that an attention-based sequence-to-sequence model, using reactants' and reagents' SMILES (see Fig. 3), performs comparable to other methods that included hand-coded chemical information.

The current state-of-the-art in reaction prediction works is the Molecular Transformer[4] that uses a sequence-to-sequence model where the encoder and decoder contain multi-head attention layers. This helps to extract the local as well as global features of the input string,[26] which is really helpful in the case of SMILES as atoms that are close in the graph need not be close in the SMILES. Compared to other reaction prediction models,[15,19,25,27] which show a comparable performance, Molecular Transformer is chemically agnostic and does not distinguish between the reactants and the reagents/solvents in the precursors. Schwaller *et al*.[4] reported a top-1 accuracy of over 90% and made the corresponding trained model available in the cloud through the IBM RXN platform.[28] Since August 2018, the platform has been used by more than 7000 registered users/chemists, who performed more than 85'000 reaction predictions. The model services can also be accessed through an API. Recently, the team deployed a synthesis route generation framework that uses a transformer-based retrosynthetic model.[29] We describe the current status of synthesis route prediction and data-driven retrosynthetic models in the next section.

```
CC#N.O=C1CCC(=O)N1Br .
COC(=O)c1nccnc1N >>
COC(=O)c1nc(Br)cnc1N
```

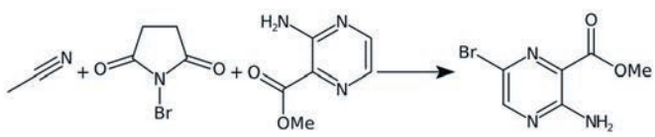

Fig 3. The reaction as depicted by the SMILES above. All the entities to the left of ">>" are the reactants which are the inputs for the models.

3. Retrosynthesis

Retrosynthesis is basically the reverse problem of reaction prediction. The target molecule can be thought of as a recursive combination of reactants by disconnecting (breaking) certain bonds, transforming them into other species (intermediates) and using different chemical strategies until one arrives at commercially available chemicals. The efforts to use computers for this task began in the 1960s[30] with the goal of reducing the workload for chemists and providing them with a deeper situational awareness of the multitude of possible disconnection strategies that may lead to better and more economically advantageous solutions.

3.1 *Template-based Models*

Reaction mechanisms, structural identities and chemical group reactivities were the metrics on which early similarity/template-based models relied. One such tool is LHASA (Logic and Heuristics Applied to Synthetic Analysis),[1] which creates a synthesis tree starting with the desired product and leading to small, much simpler molecules. More recent template-based methods use neural networks[27,31] or molecular similarity algorithms[32] to prioritise disconnections. These methods suggest very effective pathways for types of reactions similar to the data set used for training, but do not entail any kind of learning capable of using the acquired knowledge to solve prediction problems in unfamiliar classes of reactions.

3.2 *Template-free Models*

Since retrosynthesis can be considered to be a sequence of forward reactions in the reverse order, the intuition of using the reaction prediction models with a swapped input-output pair was trivial. Liu *et al*.[6] used a sequence-to-sequence model for the retrosynthesis and attained a top-1 accuracy of 34.1%. Additionally, to the reactant SMILES they included the reaction class information in the input, instead of allowing the model to discern this information based solely on the reaction SMILES information. One critical shortcoming of this sequence-to-sequence model was that the model produced invalid SMILES. This implies that the amount of data was not sufficient for the model to understand the SMILES grammar.

The Transformer model[26] in the NLP domain showed significantly better capabilities compared to a regular sequence-to-

sequence model in understanding the intricacies of word linking and grammar. Encouraged by the success of NLP and the Molecular Transformer[4] for reaction prediction, many groups tried the same architecture on the retrosynthesis problem[32–37] and attained a higher top-1 accuracy than Liu *et al*.[6] Other than the work of Lin *et al*.,[35] the retrosynthesis was confined to a single step, which is not helpful in a practical sense. Recently, Schwaller *et al*.[29] published a multi-step retrosynthesis model which uses the Transformer architecture and a hypergraph exploration algorithm to predict not just the reactant but also the reagents for each retrosynthetic step.

4. Conclusion

In this article, we provide an agile overview of the progress of AI-based data-driven methods for reaction prediction and retrosynthesis, with a particular emphasis on the use of language models. We believe that, even though human chemical experts will continue to be at the forefront of chemical research, advancements in AI along with better computational capabilities, will disrupt the field of synthetic chemistry. Irrespective of the specific AI architectures, chemical reactions data will play a major role in boosting the performance and applicability of AI in the synthetic chemistry domain. Therefore, we encourage bench chemists to record more experiments with negative results and to organize their worldwide community around data sharing platforms and repositories, just like computational chemists use and store data generated through using supercomputers. In the near future, the access to a large quantity of high-quality data will be the key factor for a technological advantage in this specific domain. Regardless of the concerns about the future of artificial intelligence and robotics, these predictive data-driven architectures will fuel the automatic generation of chemical reaction schemes and robot chemists will assist in testing them in wet-lab experiments. AI will not replace chemists, but chemists will learn to use AI as a lab assistant. This will be the core of the next technological revolution, which will be affecting chemistry more than other industrial fields.

Received: November 7, 2019

[1] E. J. Corey, R. D. Cramer, W. J. Howe, *J. Am. Chem. Soc.* **1972**, *94*, 440, DOI: 10.1021/ja00757a022.
[2] T. Klucznik, B. Mikulak-Klucznik, M. P. McCormack, H. Lima, S. Szymkuć, M. Bhowmick, K. Molga, Y. Zhou, L. Rickershauser, E. P. Gajewska, A. Toutchkine, P. Dittwald, M. P. Startek, G. J. Kirkovits, R. Roszak, A. Adamski, B. Sieredzińska, M. Mrksich, S. L. J. Trice, B. A. Grzybowski, *Chem.* **2018**, *4*, 522, DOI: 10.1016/j.chempr.2018.02.002.
[3] P. Schwaller, T. Gaudin, D. Lányi, C. Bekas, T. Laino, *Chem. Sci.* **2018**, *9*, 6091, DOI: 10.1039/C8SC02339E.
[4] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, A. A. Lee, *ACS Cent. Sci.* **2019**, *5*, 1572, DOI: 10.1021/acscentsci.9b00576.
[5] C. W. Coley, W. Jin, L. Rogers, T. F. Jamison, T. S. Jaakkola, W. H. Green, R. Barzilay, K. F. Jensen, *Chem. Sci.* **2019**, *10*, 370, DOI: 10.1039/C8SC04228D.
[6] B. Liu, B. Ramsundar, P. Kawthekar, J. Shi, J. Gomes, Q. Luu Nguyen, S. Ho, J. Sloane, P. Wender, V. Pande, *ACS Cent. Sci.* **2017**, *3*, 1103, DOI: 10.1021/acscentsci.7b00303.
[7] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2018**, *4*, 1465, DOI: 10.1021/acscentsci.8b00357.
[8] D. T. Ahneman, J. G. Estrada, S. Lin, S. D. Dreher, A. G. Doyle, *Science* **2018**, *360*, 186, DOI: 10.1126/science.aar5169.
[9] E. Corey, A. Long, S. Rubenstein, *Science* **1985**, *228*, 408, DOI: 10.1126/science.3838594.
[10] H. Satoh, K. Funatsu, *J. Chem. Inf. Model*. **1995**, *35*, 34, DOI: 10.1021/ci00023a005.
[11] T. D. Salatin, W. L. Jorgensen, *J. Org. Chem.* **1980**, *45*, 2043, DOI: 10.1021/jo01299a001.
[12] D. Mondal, S. Y. Li, L. Bellucci, T. Laino, A. Tafi, S. Guccione, S. D. Lepore, *J. Org. Chem.* **2013**, *78*, 2118, DOI: 10.1021/jo3023439.
[13] Reaxys, *https://www.reaxys.com/#/login*, accessed November 1, **2019**.
[14] NextMove Software | Pistachio, *https://www.nextmovesoftware.com/pistachio.html*, accessed November 1, **2019**.
[15] S. Koichi, H. P. Lüthi, *Chimia.* **2019**, *73*, 990, DOI: 10.2533/chimia.2019.990.
[16] J. N. Wei, D. Duvenaud, A. Aspuru-Guzik, *ACS Cent. Sci.* **2016**, *2*, 725, DOI: 10.1021/acscentsci.6b00219.
[17] M. H. S. Segler, M. P. Waller, *Chem. Eur. J.* **2017**, *23*, 5966, DOI: 10.1002/chem.201605499.
[18] C. W. Coley, R. Barzilay, T. S. Jaakkola, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2017**, *3*, 434, DOI: 10.1021/acscentsci.7b00064.
[19] W. Jin, C. Coley, R. Barzilay, T. Jaakkola, in 'Advances in Neural Information Processing Systems', **2017**, pp. 2607.
[20] M. H. S. Segler, M. P. Waller, *Chem. Eur. J.* **2017**, *23*, 6118, DOI: 10.1002/chem.201604556.
[21] J. H. Chen, P. Baldi, *J. Chem. Inf. Model.* **2009**, *49*, 2034, DOI: 10.1021/ci900157k.
[22] J. Bradshaw, M. J. Kusner, B. Paige, M. H. S. Segler, J. M. Hernández-Lobato, *arXiv:1805.10970* [physics, stat] **2018**.
[23] A. Cadeddu, E. K. Wylie, J. Jurczak, M. Wampler-Doty, B. A. Grzybowski, *Angew. Chem. Int. Ed.* **2014**, *53*, 8108, DOI: 10.1002/anie.201403708.
[24] D. Weininger, *J. Chem. Inf. Model.* **1988**, *28*, 31, DOI: 10.1021/ci00057a005.
[25] J. Nam, J. Kim, *arXiv:1612.09529* [cs] **2016**.
[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, *arXiv:1706.03762* [cs] **2017**.
[27] M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604, DOI: 10.1038/nature25978.
[28] IBM RXN for Chemistry, *https://rxn.res.ibm.com/*, accessed November 1, **2019**.
[29] P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano, T. Laino, *arXiv:1910.08036* [cs, stat] **2019**.
[30] E. J. Corey, *Angew. Chem. Int. Ed. Engl.* **1991**, *30*, 455, DOI: 10.1002/anie.199104553.
[31] A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist, E. J. Bjerrum, 'Datasets and Their Influence on the Development of Computer Assisted Synthesis Planning Tools in the Pharmaceutical Domain', **2019**, DOI: 10.26434/chemrxiv.9897692.v1.
[32] C. W. Coley, L. Rogers, W. H. Green, K. F. Jensen, *ACS Cent. Sci*. **2017**, *3*, 1237, DOI: 10.1021/acscentsci.7b00355.
[33] S. Zheng, J. Rao, Z. Zhang, J. Xu, Y. Yang, *arXiv:1907.01356* [physics] **2019**.
[34] P. Karpov, G. Godin, I. Tetko, 'A Transformer Model for Retrosynthesis', **2019**, DOI: 10.26434/chemrxiv.8058464.v1.
[35] K. Lin, Y. Xu, J. Pei, L. Lai, *arXiv:1906.02308* [q-bio] **2019**.
[36] A. A. Lee, Q. Yang, V. Sresht, P. Bolgar, X. Hou, J. L. Klug-McLeod, C. R. Butler, *Chem. Commun*. **2019**, *55*, 12152, DOI: 10.1039/C9CC05122H.
[37] H. Duan, L. Wang, C. Zhang, J. Li, *arXiv:1908.00727* [physics] **2019**.