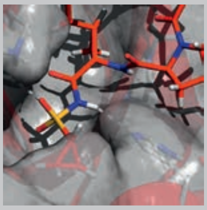
doi:10.2533/chimia.2021.677 Chimia 75 (2021) 677–678 © Swiss Chemical Society



Medicinal Chemistry and Chemical Biology Highlights

Division of Medicinal Chemistry and Chemical Biology

A Division of the Swiss Chemical Society

How AI for Synthesis Can Help Tackle Challenges in Molecular Discovery

Amol Thakkar*[a] and Philippe Schwaller*

*Correspondence: A. Thakkar, E-mail: amol.thakkar@dcb.unibe.ch, Department of Chemistry, Biochemistry and Pharmaceutical Sciences, University of Bern, Dr. P. Schwaller, E-mail: phs@zurich.ibm.com, IBM Research – Zurich, CH-8803 Rüschlikon, Switzerland

Keywords: Artificial intelligence · Machine learning · Reaction prediction · Retrosynthesis

The progress within Chemistry over the past century has led to leaps in technological innovation, and has helped us find solutions to pressing challenges such as how to feed the world's population, treat disease, and find materials for cleaner energy. As the role of the molecular sciences continues to grow, and vast amounts of chemical data are generated, there is an increasing need to explore the accumulated information to extract actionable insights. Initially, data aggregators and providers such as Reaxys and SciFinder enabled search across the breadth of chemical knowledge. However, increasingly there is a demand to go one step beyond and explore chemical space in a more 'intelligent' manner. This entails providing predictions for chemists to facilitate their workflows, and better understand which parts of the chemical space have been extensively explored, have opportunities for growth, contain solutions to pressing challenges of the present day, and how the growing base of information can be exploited. Artificial intelligence (AI) applied to synthesis prediction aims to provide a solution to facilitate the search problem of making a compound and optimizing the process in a more directed manner. To illustrate this, let us consider for instance the discovery and development of a novel bioactive molecule.

The bioactive molecule in question could be derived from a human design, or from one of the many combinatorial, enumerative, or generative methods that have seen a resurgence of interest in recent years.[1] Our primary question is how to make the compound, and once we have made the compound, what is the most optimal method for its synthesis. To answer these questions, we can augment the human chemist's abilities using an AI-based system to efficiently search the space of possibilities. This approach has two fundamental benefits. Firstly, AI models can generate ideas rapidly, and secondly researchers can use their knowledge to build on or find alternatives to ideas that they had not previously considered.

We can view an AI system as a compression of the current state of knowledge aggregated to date, combined with an optimization or search method for navigating through the space to deliver actionable insights. Let us first start with the data necessary for training models for the distinct tasks compromising synthesis prediction. Curated and accessible datasets are the bedrock on which AI-based models are built, and are a key component of enabling reproducibility in the wet lab. Unfortunately, there are few open datasets at present,[2] and efforts are currently underway to improve the accessibility and reporting of chemical reaction data. Recently Toniato and co-workers demonstrated that it is possible to use an unsupervised approach to clean reaction datasets, with the added benefit of improving model performance.[3] Similarly, Thakkar and co-workers highlight that model performance does not necessarily improve with dataset size.[4]

The two main factors influencing a model's performance and application are the diversity of reactions, and the diversity of substrates contained within the dataset. Such an analysis is possible through the use of a reaction fingerprint trained on attention-based neural networks developed by Schwaller and co-workers combined with the visualization power of TMAP by Probst and co-workers.[5,6] The reaction fingerprint has additionally been extended for the prediction of reaction yields, where it was concluded that the model is limited due to the mass scale distribution of reaction yields in the USPTO dataset.[7] Reaction yields are not the only subset of historical/literature/published data following a heavily biased distribution towards high-yielding reactions. The reaction types used favor those for the coupling of sp2 centers, such as the classical cross-coupling reactions. Protections and de-protections for avoiding reactivity conflicts, and amide bond formations also feature as frequently used reaction types. This bias is reflected in the frequency of reaction centers as reported by Schneider and co-workers.[8] In addition, the conditions used and the procedures required to carry out a reaction are usually documented as unstructured text. Vaucher and co-workers tackle this issue by using a custom rule-based natural language processing approach to extract structured synthetic steps and operations.[9] Such an approach could be used for data standardization and is currently used to enable automatization of synthetic steps on a robotic platform.

Rather than viewing the aforementioned data issues as problems that plague the field of AI-driven synthesis, the failure modes of the trained models can be viewed from the perspective of highlighting areas of sparse data, or identifying areas in which digitization efforts may be falling behind. This is derived from the view that the act of training a model is a compression of the current state of knowledge, and is a fast searching method within the current bounds of our digitized knowledge.

Although AI models are biased by the data they have seen during the training/learning phase – a bit similar to humans that will favor reactions that they successfully performed in the past – the amount of data the models can be exposed to is immense. It is impossible for a human chemist to keep up with the ever-growing organic chemistry literature and analyze millions of published chemical reactions to capture the underlying patterns, which is where AI models have the potential to shine. Within days, AI models can be trained on all available historical

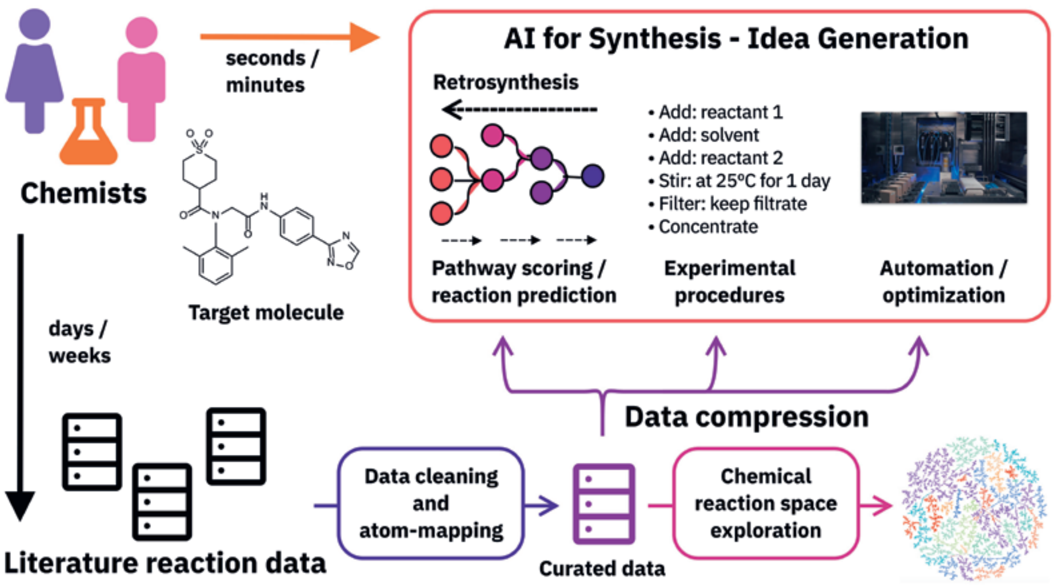


Fig. 1. Representation of the different components composing an AI-based synthesis workflow. This can be used to enable chemists in identifying candidate molecules faster by leveraging insights from large chemical data repositories and translating them into actions to inspire humans, or to enable an automated workflow for the delivery of individual compounds or libraries. The decision space can be explored using tools such as the TMAP.

data to then make informed predictions on new inputs. Trained models can assist chemists with general chemical knowledge, help with idea generation and shrink the gap between experts and less experienced chemists. Even experienced chemists can benefit from potentially unexpected suggestions outside their domain of expertise. Different chemical synthesis-related tasks have been approached using AI models, such as chemical reaction outcome prediction (products/yields),[7] synthesis planning (paths from target to commercially available molecules),[10–13] experimental procedure prediction (actions and conditions to run the reactions),[9] and even atom-mapping labeling (atom reconfiguration in a reaction).[14]

Going back to the bioactive molecule that we would like to make – how can AI models help accelerate its synthesis? AI-driven synthesis planning tools can generate likely routes from the target over several reaction steps to building blocks within minutes.[10–13] Depending on the technique, the reactions can contain reagents and condition information or alternatively, this information can be completed/added in an additional step.[15] Forward reaction prediction models can be used to validate, score and rank the suggestions for the different steps in the route. Other synthetic feasibility scores for molecules based on reaction information have also been developed.[16,17] However, the individual reaction equations in the route are not enough to perform the reactions in an automated manner. Therefore, Vaucher and coworkers developed an AI model that converts arbitrary reaction equations into experimental procedures with standardized steps. Those steps can then be executed by humans or robots alike.[9]

To facilitate the wider adoption of AI models by synthetic chemists, IBM RXN for Chemistry,[12,18] AiZynthFinder by AstraZeneca,[4,13] and ASKCOS by MIT, [11] are leading efforts to make their models accessible through open source code and a graphical user interface where the input molecules can be drawn and the models be used by chemists without coding experience. Besides a fully AI-driven (automatic) synthesis planning mode, RXN for Chemistry has an interactive mode, in which chemists can use their expert knowledge to guide the AI model. In this way, the planning of a new synthesis can be catalyzed by human-AI interaction, and going one step further can be submitted for automation using IBMs RoboRXN.

There remain numerous opportunities to augment the discovery workflow and move towards an era of joint human-computer aided discovery.

Received: June 29, 2021

[1] D. C. Elton, Z. Boukouvalas, M. D. Fuge, P. W. Chung, *Mol. Syst. Des. Eng.* **2019**, *4*, 828, https://doi.org/10.1039/C9ME00039A
[2] D. Lowe, 'Extraction of Chemical Structures and Reactions from the Literature', Doctoral thesis, University of Cambridge, **2012**.
[3] A. Toniato, P. Schwaller, A. Cardinale, J. Geluykens, T. Laino, *Nat. Mach. Intell.* **2021**, *3*, 485, https://doi.org/10.1038/s42256-021-00319-w
[4] A. Thakkar, T. Kogej, J.-L. Reymond, O. Engkvist, E. J. Bjerrum, *Chem. Sci.* **2020**, *11*, 154, https://doi.org/10.1039/C9SC04944D
[5] P. Schwaller, D. Probst, A. C. Vaucher, V. H. Nair, D. Kreutter, T. Laino, J.-L. Reymond, *Nat. Mach. Intell.* **2021**, *3*, 144, https://doi.org/10.1038/s42256-020-00284-w
[6] D. Probst, J.-L. Reymond, *J. Cheminformatics* **2020**, *12*, 12, https://doi.org/10.1186/s13321-020-0416-x
[7] P. Schwaller, A. C. Vaucher, T. Laino, J.-L. Reymond, *Mach. Learn. Sci. Technol.* **2021**, *2*, 015016, https://doi.org/10.1088/2632-2153/abc81d
[8] N. Schneider, D. M. Lowe, R. A. Sayle, M. A. Tarselli, G. A. Landrum, *J. Med. Chem.* **2016**, *59*, 4385, https://doi.org/10.1021/acs.jmedchem.6b00153
[9] A. C. Vaucher, F. Zipoli, J. Geluykens, V. H. Nair, P. Schwaller, T. Laino, *Nat. Commun.* **2020**, *11*, 3601, https://doi.org/10.1038/s41467-020-17266-6
[10] M. H. S. Segler, M. Preuss, M. P. Waller, *Nature* **2018**, *555*, 604, https://doi.org/10.1038/nature25978
[11] C. W. Coley, D. A. Thomas, J. A. M. Lummiss, J. N. Jaworski, C. P. Breen, V. Schultz, T. Hart, J. S. Fishman, L. Rogers, H. Gao, R. W. Hicklin, P. P. Plehiers, J. Byington, J. S. Piotti, W. H. Green, A. J. Hart, T. F. Jamison, K. F. Jensen, *Science* **2019**, *365*, eaax1566, https://doi.org/10.1126/science.aax1566
[12] P. Schwaller, R. Petraglia, V. Zullo, V. H. Nair, R. A. Haeuselmann, R. Pisoni, C. Bekas, A. Iuliano, T. Laino, *Chem. Sci.* **2020**, *11*, 3316, https://doi.org/10.1039/C9SC05704H
[13] S. Genheden, A. Thakkar, V. Chadimová, J.-L. Reymond, O. Engkvist, E. Bjerrum, *J. Cheminformatics* **2020**, *12*, 70, https://doi.org/10.1186/s13321-020-00472-1
[14] P. Schwaller, B. Hoover, J.-L. Reymond, H. Strobelt, T. Laino, *Sci. Adv.* **2021**, *7*, eabe4166, https://doi.org/10.1126/sciadv.abe4166
[15] H. Gao, T. J. Struble, C. W. Coley, Y. Wang, W. H. Green, K. F. Jensen, *ACS Cent. Sci.* **2018**, *4*, 1465, https://doi.org/10.1021/acscentsci.8b00357
[16] C. W. Coley, L. Rogers, W. H. Green, K. F. Jensen, *J. Chem. Inf. Model.* **2018**, *58*, 252, https://doi.org/10.1021/acs.jcim.7b00622
[17] A. Thakkar, V. Chadimová, E. J. Bjerrum, O. Engkvist, J.-L. Reymond, *Chem. Sci.* **2021**, *12*, 3339, https://doi.org/10.1039/D0SC05401A
[18] P. Schwaller, T. Laino, T. Gaudin, P. Bolgar, C. A. Hunter, C. Bekas, A. A. Lee, *ACS Cent. Sci.* **2019**, *5*, 1572, https://doi.org/10.1021/acscentsci.9b00576