Artificial Intelligence: The Future for Organic Chemistry?

Franck Peiretti† and Jean Michel Brunel*‡†

†Faculté de Médecine, Aix Marseille Université, INSERM, INRA, C2VN, 27 Bd Jean Moulin, 13385 Marseille, France
‡Faculté de Pharmacie, U1261, INSERM, UMR-MD1 (Membranes et Cibles Thérapeutiques), IRBA, Aix-Marseille Université, 27 Bd Jean Moulin, 13385 Marseille, France

ABSTRACT: On the basis of a recent article “Predicting reaction performance in C=N cross-coupling using machine learning” that appeared in Science, we had decided to highlight the way forward for artificial intelligence in chemistry. Synthesis of molecules remains one of the most important challenges in organic chemistry, and the standard approach involved by a chemist to solve a problem is based on experience and constitutes a repetitive, time-consuming task, often resulting in nonoptimized solutions. Thus, considering the recent phenomenal progresses that have been made in machine learning, there is little doubt that these systems, once fully operational in organic chemistry, will dramatically speed up development of new drugs and will constitute the future of chemistry.

INTRODUCTION

In 1956, the "Forbidden Planet" movie starred the iconic sci-fi character Robby, the robot whose impressive ability to synthesize almost everything was undoubtedly inspirational for generations of young chemists. Did Robby foreshadow the development of artificial intelligence (AI) in chemistry?

Synthesis of organic molecules remains one of the most important tasks in organic chemistry, and the standard approach involved by a chemist to solve a problem is based on experience, heuristics, and rules of thumb. When chemists manually design a new drug, they need not only to design a target molecule but also need to look at the reaction pathways to synthesize it. Chemists usually work backward, starting with the molecule they want to create and then analyzing by a process known as retrosynthesis in which readily available reagents and sequences of reactions could be used to produce it. This process is time-consuming and often results in nonoptimized solutions or even failure in finding reaction pathways because of human errors.

RESULTS AND DISCUSSION

AI is not a new field of research in organic chemistry, as chemists have been using computers for years in their daily work and are quite willing to accept the aid of a computer (Figure 1).1 Preliminary research studies started for more than five decades ago with the DENDRAL project2 even if for many people it is still just a buzz word associated with no real applications. In this context, Corey and Wipke envisioned that both synthesis and retrosynthesis could be designed by a machine using handcrafted rules known as reaction templates.3 However, a deep chemical expertise was still required, and writing rules remained a time-consuming task.

In the early days of AI approach for chemistry, candidate products were generated from the templates and then were scored according to their plausibility.4,5 However, this kind of approach is fundamentally dependent on the rule-based system component and does not lead to accurate predictions outside of the training domain. Lehn considered: "atoms as letters, molecules as words, supramolecular entities as sentences (...)", and some researchers tried to show that organic chemistry has a structure similar to a natural language and that the concepts of linguistic-based analysis could be used to analyze molecules, their patterns of reactivity, and their organic synthesis.6–8 On the other hand, we can cite the work of Kayala et al. who look at mechanistic steps,9 whereas Liu et al. and Nam et al. use sequence-to-sequence models for retrosynthesis and reaction prediction.10,11

The task of finding pathways in chemistry focuses on the best combination of moves leading to a solution and finds the most promising strategy or pathway. It is a bit like a solitaire game where the game pieces on the board at the beginning of the game would be the precursor molecules, and the only piece that remains at the end of a winning game would be the target molecule. During the game, the movements of the pieces correspond to the reactions applicable to the precursor molecules allowing to progress toward the synthesis of the target molecule; an inadequate combination of moves will prematurely stop the game. In 2016, using fingerprints learned with a neural network algorithm, Wei et al. identified with more than 80% accuracy the reaction type in the scope of alkene and alkyl halide reactions (Figure 2A). This algorithm is able to learn the probabilities of a range of reaction types, and
most importantly, its predictive capabilities can increase with the size of the libraries of training data. This algorithm constituted a step toward the goal of developing a learning machine devoted to the automatic synthesis planning of organic molecules. One year later, a learning machine combining a novel model framework for generating and ranking candidate reaction outcomes and a novel edit-based representation was used by Coley et al. to reproduce in silico the qualitative results of actual experimental reactions. This unique framework combining candidates with more direct relevance to chemical reactivity was improved and expanded to achieve high predictive performance (Figure 2B). In the same year, using the dataset previously used by Jin et al., a team at IBM adopted another innovative approach by creating a tool to solve the forward-reaction prediction problem where the starting materials are known, and the interest is in generating the products. By using such an approach, the team fed chemical components into a neural network trained on a dataset of 395,496 reactions. The neural network then used what it had learned about prior reactions to predict about what would occur under new experimental conditions. The system responded to such requests by offering a list of the most plausible results. Testing showed that the top prediction turned out to be correct 80 percent of the time. Further to these works dealing with forward prediction, Segler et al. focused his attention on retrosynthesis and developed a new deep-learning neural network that considered around 12.4 million of known single-step chemical reactions and allowed the prediction of the reactions that can be used in any single step of a multistep synthesis. This program can deconstruct...
The considered molecule until it ends up with readily available starting reagents. A first neural network determines the search in promising directions by proposing fewer transformations. A second one predicts whether the selected reactions are feasible, whereas a third one samples the transformations during the implementation phase. Simultaneously, by considering the low data regime and calculating theoretical properties of the reactants and reagents, Ahneman et al. predicted reaction performance in palladium Buchwald–Hartwig C–N cross-coupling reactions in the presence of various potentially inhibitory additives with machine learning in multidimensional space using data obtained via high-throughput experimentation. This methodology is based on scripts that extract various molecular descriptors, subsequently used as inputs in the model, whereas reaction yields were considered as outputs. Random forest algorithms operate by randomly sampling the data, and an overall prediction is generated by constructing aggregated decision trees. It was demonstrated that these algorithms led to significantly improved performance over linear regression analysis. Very recently and after more than 10 years of research, Klucznik et al. developed the highly effective Chematica computer program capable of designing novel efficient syntheses of medicinally relevant molecules. This was realized by a subtle analysis of its scoring functions and the comparison with large numbers of logically related criteria, for example, selectivity or reactivity. All these lead Chematica to be able to rapidly propose optimized synthetic pathways for the design of molecules of interest.

All these different intelligent search algorithms have proven their effectiveness in providing chemists with realistic solutions for molecule synthesis. Undoubtedly, the collaboration of all these algorithms would bring an even better predictive relevance to the proposed solutions.

Some chemists may find this provocative, but chemistry considered as an isolated entity without relevant application is not useful and the considerable progresses of AI in this area (summarized above) do not change that.

However, chemistry reaches a whole new dimension when it has applications in other disciplines like in biology or physics of materials. Of course, AI has also infiltrated these disciplines. In biology, for example, the Omics revolution requiring big data mining has been the major driving force behind the development of AI. Beyond that application, in 2015, an AI system developed for biology has successfully inferred the development of AI. Beyond that application, in 2015, an AI system developed for biology has successfully inferred the system-developed for chemistry and biology that would be able to synthesize new molecules effective on new therapeutic targets to cure diseases would be a rather optimistic view of a near future. Already some researchers are turning to AI to design and carry out experiments to prospect for fully automated science. However, in the interest of human intelligence and for the social acceptance of AI (in all areas), we believe that an advisable evolution would be for AI to become the investigator’s collaborator, bringing him new solutions, but also explains to him how it came to these conclusions. The Google’s AI program Alphazero made obsolete the maxim that says that “it takes few minutes to learn to play chess and a human lifetime to become a master”. Indeed, knowing only the rules of the chess game, Alphazero needed just few hours of self-learning to develop an adventurous and unconventional way of playing that allowed it to beat human masters or existing programs. In the same way, and even if the “chemistry game” is more complex than the chess game, the recent phenomenal progresses that have been made in machine learning will undoubtedly dramatically speed up development of new drugs.

**CONCLUSIONS**

Even though today no new drug has been synthesized using AI, experts agree that the commercial issues in this area are enormous. AI will not replace chemists at least in the short term, but AI clearly appears as the future of chemistry.

**AUTHOR INFORMATION**

**Corresponding Author**

E-mail: bruneljm@yahoo.fr.

**ORCID**

Jean Michel Brunel: 0000-0002-9355-8980

**Notes**

The authors declare no competing financial interest.

**REFERENCES**

1. Gray, N. A. B. Artificial intelligence in chemistry. Anal. Chim. Acta 1988, 210, 9–32.
2. Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; Lederberg, J. DENDRAL: A case study of the first expert system for scientific hypothesis formation. Artificial Intelligence 1993, 61, 209–261.
3. Corey, E. J.; Wipke, W. T. Computer-assisted design of complex organic syntheses. Science 1969, 166, 178–192.
4. Baldi, P.; Chauvin, Y. Neural Networks for Fingerprint Recognition. Neural Computation 1993, 5, 402–418.
5. Bromley, J.; Guyon, I.; LeCun, Y.; Säckinger, E.; Shah, R. Signature verification using a “Siamese” time delay neural network. Proceedings of the 6th International Conference on Neural Information Processing Systems; San Francisco, CA, USA, 1993; pp 737–744.
6. Baskin, I. L.; Madzhidov, T. I.; Antipin, I. S.; Varnek, A. A. Artificial intelligence in synthetic chemistry: achievements and prospects. Russ. Chem. Rev. 2017, 86, 1127–1156.
7. Lehn, J. M. Supramolecular Chemistry: Concepts and Perspectives, 1st ed.; Wiley VCH: Weinheim, 1995.
8. Cadeddu, A.; Wylie, E. K.; Jurczak, J.; Wampler-Doty, M.; Graybowski, B. A. Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses. Angew. Chem., Int. Ed. 2014, 53, 8108–8112.
9. Kayala, M. A.; Azencott, C.-A.; Chen, J. H.; Baldi, P. Learning to predict chemical reactions. J. Chem. Inf. Model. 2011, 51, 2209–2222.
10. Liu, B.; Ramsundar, B.; Kawthekar, P.; Shi, J.; Gomes, J.; Luu Nguyen, Q.; Ho, S.; Sloane, J.; Wender, P.; Pande, V. Retrosynthetic Reaction Prediction Using Neural Sequence-to-Sequence Models. ACS Cent. Sci. 2017, 3, 1103–1113.
11. Nam, J.; Kim, J. Linking the neural machine translation and the prediction of organic chemistry reactions. 2016, arXiv:1612.09529.
12. Wei, J. N.; Duvenaud, D.; Aspuru-Guzik, A. Neural networks for the prediction of organic chemistry reactions. ACS Cent. Sci. 2016, 2, 725–732.
13. Coley, C. W.; Barzilay, R.; Jaakkola, T. S.; Green, W. H.; Jensen, K. F. Prediction of organic reaction outcomes using machine learning. ACS Cent. Sci. 2017, 3, 434–443.
14. Jin, W.; Coley, C. W.; Barzilay, R.; Jaakkola, T. Predicting organic reaction outcomes with Weisfeiler-Lehman network. 2017, arXiv:1709.04555v3.
15. Schwaller, P.; Gaudin, T.; Lanyi, D.; Bekas, C.; Laino, T. “Found in translation”: Predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. 2017, arXiv:1711.04810.
16. Segler, M. H. S.; Preuss, M.; Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. Nature 2018, 555, 604–610.
(17) Ahneman, D. T.; Estrada, J. G.; Lin, S.; Dreher, S. D.; Doyle, A. G. Predicting reaction performance in C-N cross-coupling using machine learning. *Science* 2018, 360, 186–190.

(18) Klucznik, T.; Mikulák-Klucznik, B.; McCormack, M. P.; Lima, H.; Szymkuć, S.; Bhowmick, M.; Molga, K.; Zhou, Y.; Rickershauser, L.; Gajewska, E. P.; Touchekine, A.; Dittwald, P.; Startek, M. P.; Kirkovits, G. J.; Roszak, R.; Adamski, A.; Sieredzińska, B.; Mrksich, M.; Trice, S. L. J.; Grybowski, B. A. Efficient syntheses of diverse, medicinally relevant targets planned by computer and executed in the laboratory. *Chem* 2018, 4, 522–532.

(19) Lobo, D.; Levin, M. Inferring regulatory networks from experimental morphological phenotypes: A computational method reverse-engineers planarian regeneration. *PLoS Comput. Biol.* 2015, 11, No. e1004295.

(20) Silver, D.; Schrittwieser, J.; Hubert, T.; Antonoglou, I.; Lai, M.; Guez, A.; Lanctot, M.; Sifre, L.; Kumaran, D.; Graepel, T.; Lillicrap, T.; Simonyan, K.; Hassabis, D. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. 2017, arXiv:1712.01815v1.

(21) Szymkuc, S.; Gajewska, E. P.; Klucznik, T.; Molga, K.; Dittwald, P.; Startek, M.; Bajczyk, M.; Grybowski, B. A. Computer-assisted synthetic planning: The end of the beginning. *Angew. Chem., Int. Ed.* 2016, 55, 5904–5937.

(22) Fleming, N. How artificial intelligence is changing drug discovery. *Nature* 2018, 557, S55–S57.

(23) Schwaller, P.; Gaudin, T.; Lányi, D.; Bekas, C.; Laino, T. “Found in Translation”: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. *Chem. Sci.* 2018, 9, 6091–6098.

(24) Maryasin, B.; Marquetand, P.; Maulide, N. Machine Learning for Organic Synthesis: Are Robots Replacing Chemists? *Angew. Chem., Int. Ed.* 2018, 57, 6978–6980.

(25) Lemonick, S. Is machine learning overhyped? *Chem. Eng. News* 2018, 96 (34).