CHEMBIOMOL-09: Chem. Biol., Org. Chem., Med. Chem., Pharm. Ind., & Mol. Biol. Congress, Paris, France-Galveston, USA, 2023.

Cover for Machine Learning in Organic Chemistry

Shan He
Department of Organic and Inorganic Chemistry, Faculty of Science and Technology, University of the Basque Country UPV/EHU, P.O.Box 644, 48080 Bilbao, Spain.
IKERDATA S.L INVESTIGO Next-generation Researcher.

Abstract. Synthesis of organic molecules is one of the most essential tasks in organic chemistry. The standard methodology started by a chemist solving a problem centered on experience, heuristics, and rules of thumb. Generally, experimentalists often work backward, starting with the molecule desired design and then analyzing the retrosynthesis in which readily available reagents and sequences of reactions could be used to produce it. All this his process is time-consuming and source- consuming, it

can result in non-optimized solutions or even failure in finding reaction pathways because of human errors. In this sense, AI/ML (Artificial Intelligence/Machine Learning) is gaining more and more attention in organic chemistry because it can speed up this process. In this mini-Review provided a guide map to review the digitalization and computerization of organic chemistry principles.

Synthesis of organic molecules is one of the most essential tasks in organic chemistry. The standard methodology started by a chemist solving a problem centered on experience, heuristics, and rules of thumb. If researchers propose a new drug, there are two main steps: target molecule designing and reaction pathways to synthesize it. Generally, experimentalists often work backward, starting with the molecule desired design and then analyzing the retrosynthesis in which readily available reagents and sequences of reactions could be used to produce it. All this his process is time-consuming and source-consuming, it can result in non-optimized solutions or even failure in finding reaction pathways because of human errors.[1] In this sense, AI/ML is gaining more and more attention in organic chemistry because it can speed up this process. In this mini-Review provided a guide map and intrigue chemists to revisit the digitalization and computerization of organic chemistry principles.

Jamison *et al.* [2] proposed a new synthesis routes and automated the synthesis of 15 small molecules. It included warfarin, celecoxib, and quinapril. A typical AI-driven synthesis of the NSAID celecoxib. The author used several robot-controlled common known reaction bays, such as the Claisen condensation of 4-methyl acetophenone with methyl trifluoroacetate which led the first two modules. Then, the final condensation with hydrazine to yield celecoxib in a fully automated synthesis. Similar, they utilized this tool in order to the synthesis planning and execution of a library of ACE inhibitors and analogues of celecoxib with a productivity of 342 to 572 mg/h of the final small molecules. The main disadvantages of this approach is the limitations of the current system. It involved the handling of poorly soluble compounds, batch purifications of final compounds, and subambient temperatures realization, as it was observed in the automated synthesis of bezafibrate. In this scenario, the AI-driven synthesis planning provided an interesting approach via a Bargellini reaction of chloroform, acetone, and phenol, which could not be performed using the flow approach. Nevetheless, the synthetic feasibility of the synthesis route was confirmed by the authors in 76% yield, which highlights the power of AI-driven organic synthesis. [3]

Another example within this topic, Miller *et al.* presented OrbNet Denali, a ML model for an electronic structure that was designed as a direct replacement for density functional theory (DFT) energy calculations. This model is a message-passing graph neural network that utilize symmetry adapted atomic orbital features from quantum calculation to predict the molecule energy. OrbNet Denali was trained on a massive dataset of 2.3×10^6 DFT calculations of molecules and geometries. The model was confirmed in several deep-rooted benchmark datasets, and they found that it provided accuracy that was on par with modern DFT methods while offering a speed of up to three orders of magnitude. For the GMTKN55 benchmark set, OrbNet Denali achieved WTMAD-1 and WTMAD-2 scores of 7.19 and 9.84, on par with modern DFT functionals, respectively. For the Hutchison conformer benchmark set, OrbNet Denali had a median correlation coefficient of $R^2 = 0.90$ compared to the reference DLPNO-CCSD(T) calculation and $R^2 = 0.97$ compared to the ωB97X-D3/def2- TZVP method. Similarly, the model reached chemical accuracy for non-covalent interactions in the S66x10 dataset. For torsional profiles, OrbNet Denali reproduced the torsion profiles of ωB97X-D3/def2-TZVP with an average mean

absolute error of 0.12 kcal/mol for the potential energy surfaces of the diverse fragments in the TorsionNet 500 dataset.[4]

As last example, Cronin and *et al.* developed an ML-guided feedback loop to explore synthetic space with a designed liquid-handling robot.[5] They designated a Suzuki–Miyaura cross-coupling to study the influence of their ML approach in terms of chemical yield prediction. There are 35 different types of organic molecules (reactants, ligands, bases, and solvents) in this space. thus OHE (One-Hot Encoding) defines each specific reaction by a unique 35-digit binary vector. They obtained in training series with 60% of the entire synthetic space (3456 reactions of 5760 possibilities). The neural network model in training series gave a root-mean-square error (RMSE) of 11% for a test set of 1728 transformations. This success of OHE highly relied on the fairly large training set. Furthermore, the trained model can capture the statistical patterns of the training data, specifically the presence of which compound or groups of compounds can increase or decrease the reaction yield, enabling the predicting of the reaction yield.[6]

To conclude, the use of AI/ML is experimenting a revolution in the vast majority of fields. Especially, the advanced in AI/ML prediction inorganic synthesis, the impact of the data-driven research paradigm in synthetic chemistry is apparent. The digitalization, computerization, and especially intellectualization of synthetic transformations will provide a strong momentum to push the frontiers of organic synthesis However, it is worth mentioning that, the knowledge source of organic synthesis is usually giving by human beings. This allows the installation of explicit rules in ML modelling, such as selecting chemically meaningful descriptors or the assignment of transition-state-like geometries. In addition, chemical statistics itself contains rich knowledge, which can be learned by machines to support the performance prediction of a target transformation.[6]

REFERENCE

1. Peiretti, F.; Brunel, J. M. Artificial intelligence: The future for organic chemistry? *ACS omega* **2018,** *3* (10), 13263-13266.
2. Coley, C. W.; Thomas III, D. A.; Lummiss, J. A.; Jaworski, J. N.; Breen, C. P.; Schultz, V.; Hart, T.; Fishman, J. S.; Rogers, L.; Gao, H. A robotic platform for flow synthesis of organic compounds informed by AI planning. *Science* **2019,** *365* (6453), eaax1566.
3. Empel, C.; Koenigs, R. M. Artificial-Intelligence-Driven Organic Synthesis—En Route towards Autonomous Synthesis? *Angewandte Chemie International Edition* **2019,** *58* (48), 17114-17116.
4. Christensen, A. S.; Sirumalla, S. K.; Qiao, Z.; O'Connor, M. B.; Smith, D. G.; Ding, F.; Bygrave, P. J.; Anandkumar, A.; Welborn, M.; Manby, F. R. OrbNet Denali: A machine learning potential for biological and organic chemistry with semi-empirical cost and DFT accuracy. *The Journal of Chemical Physics* **2021,** *155* (20), 204103.
5. Granda, J. M.; Donina, L.; Dragone, V.; Long, D.-L.; Cronin, L. Controlling an organic synthesis robot with machine learning to search for new reactivity. *Nature* **2018,** *559* (7714), 377-381.
6. Zhang, S. Q.; Xu, L. C.; Li, S. W.; Oliveira, J. C.; Li, X.; Ackermann, L.; Hong, X. Bridging Chemical Knowledge and Machine Learning for Performance Prediction of Organic Synthesis. *Chemistry–A European Journal* **2022**, e202202834.