State of the Art and of Outlook of Data Science and Machine Learning in Organic Chemistry

Ricardo Stefani[1*]

[1]Materials Research Laboratory (LEMat). Federal University of Mato Grosso, Campus Araguaia, Barra do Garças - MT, 78600-000, Brazil

**Corresponding author: rstefani@ufmt.br*

Abstract

The use of data science, artificial intelligence, and big data in the field of chemistry has recently grown to speed up the discovery of new materials, drugs, and synthetic substances and the identification of automated compounds. Machine learning and data science are commonly used in organic chemistry to predict biological and physicochemical properties of molecules and are referred to as quantitative structure-–active relationship (QSAR, for biological properties) and quantitative structure-–property relationship (QSPR, for nonbiological properties). In addition, data science and machine learning have advanced the optimization of molecular properties, synthetic pathways, and even the design of novel compounds. These models can learn the underlying patterns of molecular structures to generate new compounds with desirable properties. Hence, machine learning (ML) is extensively used in chemistry, and the field is rapidly adopting state-of-the-art ML algorithms and tools such as deep learning, tensors, and transformers to solve and model chemical problems. The application of data science and ML, particularly deep learning, plays a significant role in advancing research in organic chemistry.

Keywords: *Organic Chemistry, Data Science, Machine Learning, Deep Learning*

1. Introduction

In recent years, data-driven approaches, such as data science (DS), artificial intelligence (AI), and big data, have been changing the way science is done because classical experiments, even the hyphenated or high-throughput approach, can be time-consuming, expensive, and dependent on the availability of equipment and reagents. Therefore, the use of tools and techniques related to DS, ML, and rational designs has attracted the attention of chemists because it can save valuable resources and time. In this sense, DS approaches have recently expanded to accelerate the discovery of new materials, drugs, and synthetic substances and automated compound identification.

The data-driven approach to chemical discovery has been gaining relevance and importance owing to emerging public databases of chemical substances, such as PubChem[1], ZINC[2], ChemBL[3], and PubChemQC[4]. These

open-source databases allowed the development of several machine learning (ML) methodologies and models for the discovery of new advanced materials[5], the study and discovery of bioactive compounds[6,7], and the synthesis of new compounds[8,9]. Therefore, DS can accelerate scientific research, including research in organic chemistry, where it can be used to analyze and make predictions on chemical reactions, structures, physical-chemical and biological properties, and big data analysis.

Although information technology and DS are now widespread, chemists have been using computers and computer science to solve chemistry problems for decades. In the early 1960s, the Chemical Abstract Service developed an algorithm to generate a unique machine descriptor for chemical structures[10]. This algorithm has inspired other algorithms and standards, such as SMILES[11], SYBYL[12], and InChI[13]. Moreover, organic chemists have developed expert systems for chemical structure elucidation[14] and organic synthesis planning[15]. Therefore, chemists have been using computers and algorithms for a long time, and their growing interest in ML and DS is natural. Thus, this review will briefly discuss the state-of-the-art DS and ML applications in chemistry and provide a perspective on how they can change organic chemistry in the following years

2. Historical overview of Data Science and Machine Learning in Organic Chemistry

DS and ML are being increasingly applied in organic chemistry to streamline and automate various research and development tasks such as drug discovery, spectroscopy, synthesis, and chemical data analysis based on the chemical structure and the structures of organic molecules based on their chemical properties.

The prediction of the biological and physicochemical properties of molecules are the most common applications of ML and DS in chemistry because they aid drug and material discoveries. When applied to predicting biological and physical-chemical properties, ML and DS tools are often referred to as quantitative structure-–active relationship (QSAR) when predicting biological properties and quantitative structure-–property relationship (QSPR) when predicting nonbiological properties. QSAR and QSPR are widespread in chemistry, and publications related to these themes often range from simple statistical regression and classification models[16] to sophisticated support vector machines (SVM)[17], artificial neural networks (ANN)[18], and ML ensembles[19]. Although chemometric and statistical methods such as multiple linear regression (MLR), principal component analysis (PCA), and partial least squares (PLS) have been methods of choice for years in QSAR and QSPR studies, ML-based models are now preferred over those classical methods, as the diversity and availability of chemical data increases. For example, pKa, which has been modeled by MLR, PCA, and PLS for years, has been successfully modeled by Mansouri et al.[20] using multiple ML approaches like an ensemble of SVM combined with k-nearest neighbors (kNN), extreme gradient boosting (XGB) and deep neural networks. These models were built using publicly available data; the source code and data are provided on GitHub, and the performance of these models is compared favorably to the commercial products. Other examples of QSAR/QSPR modeling using ML include the discovery of new materials[21,22], the prediction of ionic liquid properties[23], the modeling of drugs[24], the quality

control of fuels[25], and the prediction and modeling of bioconcentration and toxicity of organic compounds[26]. Hence, ML is a valuable tool for QSAR/QSPR to provide confident results and outperform chemometric methods.

Another field in organic chemistry that benefits from DS and ML is computer-assisted structure elucidation (CASE). Earlier, the CASE, such as DENDRAL[14], was solely based on rules, without any AI implementation. Furthermore, systems such as CHEMICS[27] were built on top of sophisticated rules and larger databases without AI aid. In the middle 1990s, systems such as SpecSov[28] and SISTEMAT[29], which combined the database approach and ANN, were released. For the following years, this approach has become the standard for CASE with some developments, such as incorporating genetic algorithms[30] and decision trees[31]. Owing to AI and ML algorithms, there has been considerable development in CASE in the last decade. One of the many challenges in CASE that has benefited from ML is the prediction of the chemical shifts in NMR, which is used to determine the validity of a proposed chemical structure. Message-passing neural networks have been trained to predict and store ^{13}C NMR shifts with high accuracy[32]. Advanced ML algorithms such as SVM and XGBoost have been sed to train ML models to predict the chemical classes of natural products based on the naproc13 database[33]. This is one of the most comprehensive natural products databases publicly available. The ML models predict the probability of a specific ^{13}C NMR spectrum belonging to a particular substance class with an average accuracy of 0.85–0.98, depending on the class. Thus, ML models such as ANN, SVM, and XGBoost tools can be valuable tools to aid structure elucidation. The use of computer and information technology in designing drugs and their synthesis plans has been of interest to chemists for decades and is now improved by ML and DS tools. Early synthesis planning systems such as SYNCHEM2[15] were based on chemical compound databases and decision rules, which limited the scope of the system owing to the existence of hundreds of reaction mechanisms. Therefore, to overcome such limitations, chemists have been applying DS and ML to retrosynthetic analysis, which is a task that requires experience and expertise. The availability of large reaction databases such as Reaxys has allowed the development of powerful ML models to predict and optimize reaction conditions such as temperature, catalysis, and solvents[34]. Because optimizing reaction conditions is not a straightforward task in classical experiments, DS and ML have become valuable tools for organic chemists because they can recommend reaction conditions when other methods, such as quantum chemistry, fail[9,35]. Even computational methods of choice for predicting reaction and reaction conditions, such as quantum chemistry and ab initio methods, are now combined with ML methods to accelerate orbital calculations[9,36]. ML and DS have been applied to design the synthesis of a new metal-organic framework (MOF)[8]. In this study, authors have used automated processes and natural language processing to extract the data available in databases and

in the literature to create a new dataset that was used to train ML models for predicting the final experimental conditions and synthesis protocols of new MOF structures.

Polymer chemistry is another field where ML is accommodating. Bayesian models have been successfully developed to aid the discovery of polymers with high thermal conductivity[37]. However, ML models were trained on a substantially limited number of polymeric properties and predicted the QSPR of thermal conductivity and other polymeric properties. Besides Bayesian models, genetic algorithms have been applied to train ML models to design new polymers[5]. The genetic algorithm was trained to predict and design glass transition temperature and the band gap of polymers. The ML model was able to design 132 polymers with desired properties and good accuracy. Thus, genetic algorithms can be generalized to predict and model other property objectives since corresponding reliable property prediction models can be provided.

3. Outlook and Perspective of Data Science and Machine Learning in Organic Chemistry

DS, ML, and data-driven science have emerged as the 4th paradigm of science[38]. Hence, DS and ML have gained importance in chemistry and correlated disciplines such as materials science[39], biology[40,41], and health sciences[42–44]. Despite the growing interest and research in ML tools for chemistry, data collection and storage are critical steps in ML model designing and validation because the dataset size is considerably important in ML. ML often requires large datasets for training, and insufficient data may cause overfitting or underfitting problems, which leads to poor predictive models. Nevertheless, large chemical datasets are becoming widely available[1–3,45], and access to data for ML design will not be a threat as the importance of open science and FAIR[46] principles increases.

The main strength of ML is the ability to find patterns and relationships that even an experienced researcher may not be able to find. However, the lack of education on the subject and the misinterpretation of the ML nature may threaten the development of advanced and accurate ML models in chemistry. ML algorithms are often referred to as black boxes. However, this is a misinterpretation because a good ML model requires a detailed analysis and adjustment of hyperparameters, which requires the knowledge of how the algorithms and their implementation work. Some chemists also consider ML algorithms as statistical learning or a kind of “advanced chemometrics,” which is not valid, and these misconceptions lead to the inadequate application of ML in chemistry. To overcome such issues, initiatives on chemical education and ML have appeared[47]. The final remarkable weakness of ML is that if the data have many errors, the ML model will not be

reliable[48]. Hence, the accuracy of ML models depends not only on the algorithm but also on the quality of the dataset and the ability of the data scientist to choose and tune the suitable model.

Despite some criticism, ML is increasingly used in chemistry. This field is rapidly adopting state-of-the-art ML algorithms and tools such as deep learning[49], tensors[50,51], and transformers[52,53] to solve and model chemical problems such as drug and polymer design, QSAR and QSPR studies on big data and huge datasets[54,55], and even to boost ab initio calculations[45,56]. As ML applications in chemistry are broad, open-source, and free frameworks and tools to assist chemists in developing ML models, such as OpenChem[57] and ML4Chem[58], have been designed. These frameworks encapsulate other frameworks, such as PyTorch and scikit-learn, making training and testing ML models in chemistry more straightforward.

In conclusion, ML applications in organic chemistry are constantly evolving and greatly accelerating research. Moreover, this interdisciplinary field is playing a central role in changing the way of not only organic chemistry but how chemistry is done. As cutting-edge ML tools and algorithms such as tensors, natural language processing, and transformers become mature and reliable by chemists, ML will be a routine analysis in a chemistry laboratory like any other technique or equipment.

4. References

1. Kim S, Thiessen PA, Bolton EE, *et al.* PubChem substance and compound databases. Nucleic Acids Res 2016; 44(D1): D1202-13.

2. Irwin JJ, Shoichet BK. ZINC - a free database of commercially available compounds for virtual screening. J Chem Inf Model 2005; 45(1): 177-82.

3. Gaulton A, Hersey A, Nowotka ML, *et al.* The ChEMBL database in 2017. Nucleic Acids Res 2017; 45(D1): D945-54.

4. Nakata M, Shimazaki T. PubChemQC Project: a large-scale first-principles electronic structure database for data-driven chemistry. J Chem Inf Model 2017; 57(6): 1300-8.

5. Kim C, Batra R, Chen L, Tran H, Ramprasad R. Polymer design using genetic algorithm and machine learning. Comput Mater Sci 2021; 186.

6. Karydas C, Iatrou M, Kouretas D, *et al.* Prediction of antioxidant activity of cherry fruits from UAS multispectral imagery using machine learning. Antioxidants 2020; 9(2).

7. Idowu SO, Fatokun AA. Artificial intelligence (AI) to the rescue: deploying machine learning to bridge the biorelevance gap in antioxidant assays. Vol. 26, SLAS Technology. SAGE Publications Inc. 2021; pp. 16-25.

8. Luo Y, Zaremba O, Cierpka A, *et al.* MOFs and machine learning mof synthesis prediction enabled by automatic data mining and machine learning. Angwandte Chemie Int Ed 2022; 61: e20220024.

9. Jorner K, Tomberg A, Bauer C, Sköld C, Norrby PO. Organic reactivity from mechanism to machine learning. Vol. 5, Nature Reviews Chemistry. Nature Research 2021; pp. 240-55.

10. Morgan HL. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. J Chem Doc 1965; 4(1): 65.

11. Ritter G, Isenhour TL, Balaban AT, *et al.* Unique description of chemical structures based on hierarchical ordered extended con-nectivities (HOC Procedures). I. Algorithms for finding graph orbits and canonical numbering of atoms. Analiz Metricheskikh svoistv grafov Vychisl Sist 1989; 29(18): 133-86.

12. Ash S, Cline MA, Homer RW, Hurst T, Smith GB. SYBYL line notation (SLN): a versatile language for chemical structure representation †. J Chem Inf Comput Sc 1997; 37: 71-9.

13. Coles SJ, Day NE, Murray-Rust P, Rzepa HS, Zhang Y. Enhancement of the chemical semantic web through the use of InChI identifiers. Org Biomol Chem 2005; 3(10): 1832-4.

14. Lindsay RK, Buchanan BG, Feigenbaum EA, Lederberg J, Lindsay RK. DENDRAL: a case study of the first expert system for scientific hypothesis formation*. Vol. 61, Artificial Intelligence 1993.

15. Benstock JD, Berndt DJ. Graph embedding in SYNCHEM2, an expert system for organic synthesis discovery *. Discrete Appl Math 1988; 19(1-3): 45-63.

16. Kiralj R, Ferreira MMC. Basic validation procedures for regression models in QSAR and QSPR studies: theory and application. J Braz Chem Soc 2009; 20(4): 770-87.

17. Norinder U. Support vector machine models in drug design: applications to drug transport processes and QSAR using simplex optimisations and variable selection. Neurocomputing 2003; 55(1-2): 337-46.

18. Bertinetto C, Duce C, Micheli A, Solaro R, Starita A, Tiné MR. Evaluation of hierarchical structured representations for QSPR studies of small molecules and polymers by recursive neural networks. J Mol Graph Model 2009; 27(7): 797-802.

19. Yao XJ, Panaye A, Doucet JP, *et al.* Comparative study of QSAR/QSPR correlations using support vector machines, radial basis function neural networks, and multiple linear regression. J Chem Inf Comput Sci 2004; 44(4): 1257-66.

20. Mansouri K, Cariello NF, Korotcov A, *et al.* Open-source QSAR models for pKa prediction using multiple machine learning approaches. J Cheminform 2019; 11(1).

21. Yang Y, Lin T, Weng XL, Darr JA, Wang XZ. Data flow modeling, data mining and QSAR in high-throughput discovery of functional nanomaterials. Comput Chem Eng 2011; 35(4): 671-8.

22. Golin AF, Stefani R. Quantitative structure-property relationships of electroluminescent materials: Artificial neural networks and support vector machines to predict electroluminescence of organic molecules. Bull Mater Sci 2013; 36(7).

23. Ding Y, Chen M, Guo C, Zhang P, Wang J. Molecular fingerprint-based machine learning assisted QSAR model development for prediction of ionic liquid properties. J Mol Liq 2021; 326.

24. D'Souza S, Prema KV, Balaji S. Machine learning models for drug–target interactions: current knowledge and future directions. Vol. 25, Drug Discovery Today, Elsevier Ltd. 2020; pp. 748-56.

25. Li R, Herreros JM, Tsolakis A, Yang W. Machine learning-quantitative structure property relationship (ML-QSPR) method for fuel physicochemical properties prediction of multiple fuel types. Fuel 2021; 304.

26. Ai H, Wu X, Zhang L, *et al.* QSAR modelling study of the bioconcentration factor and toxicity of organic compounds to aquatic organisms using machine learning and ensemble methods. Ecotoxicol Environ Saf 2019; 179: 71-8.

27. Funatsu K, Nishizaki M, Sasaki S. Introduction of NOE data to an automated structure elucidation system, CHEMICS. Three-dimensional structure elucidation using the distance geometry method. J Chem Inf Comput Sci 1994; 34: 745-51.

28. Will M, Fachinger W, Richert JR. Fully automated structure elucidation - a spectroscopist's dream comes true. J Chem Inf Comput Sci 1996; 36: 221-7.

29. Emerenciano V, Rodrigues G, Macarp PAT, *et al.* Applications d'intelligence artificielle dans la chimie organique. XVII. Nouveaux programmes du projet SISTEMAT. Spectroscopy 1994; 12: 91-8.

30. Meiler J, Will M. Automated structure elucidation of organic molecules from 13C NMR spectra using genetic algorithms and neural networks. J Chem Inf Comput Sci 2001; 41(6): 1535-46.

31. Masui H, Hong H. Spec2D: A structure elucidation system based on1H NMR and H-H COSY spectra in organic chemistry. Journal of Chemical Information and Modeling. American Chemical Society 2006; pp. 775-87.

32. Han H, Choi S. Transfer learning from simulation to experimental data: NMR chemical shift predictions. J Phys Chem Lett 2021; 12(14): 3662-8.

33. Martínez-Treviño SH, Uc-Cetina V, Fernández-Herrera MA, Merino G. Prediction of natural product classes using machine learning and 13C NMR spectroscopic data. J Chem Inf Model 2020; 60(7): 3376-86.

34. Gao H, Struble TJ, Coley CW, Wang Y, Green WH, Jensen KF. Using machine learning to predict suitable conditions for organic reactions. ACS Cent Sci 2018; 4(11): 1465-76.

35. Zhang SQ, Xu LC, Li SW, *et al.* Bridging chemical knowledge and machine learning for performance prediction of organic synthesis. Chemistry 2022; e202202834.

36. Haghighatlari M, Hachmann J. Advances of machine learning in molecular modeling and simulation. Vol. 23, Current Opinion in Chemical Engineering. Elsevier Ltd. 2019; pp. 51-7.

37. Wu S, Kondo Y, Kakimoto M, *et al.* Machine-learning-assisted discovery of polymers with high thermal conductivity using a molecular design algorithm. NPJ Comput Mater 2019; 5(1).

38. Hey AJG, Tansley DSW, Tolle KM. The fourth paradigm: data-intensive scientific discovery. Proceedings of the IEEE 2011; 99(8): 287.

39. Pollice R, dos Passos Gomes G, Aldeghi M, *et al.* Data-driven strategies for accelerated materials design. Acc Chem Res 2021; 54(4): 849-60.

40. Leonelli S. Process-sensitive naming: trait descriptors and the shifting semantics of plant (data) science. Philos Theory Pract Biol 2022; 14.

41. Abdullah-Zawawi MR, Govender N, Karim MB, Altaf-Ul-Amin M, Kanaya S, Mohamed-Hussein ZA. Chemoinformatics-driven classification of Angiosperms using sulfur-containing compounds and machine learning algorithm. Plant Methods 2022; 18(1).

42. Ehrman TM, Barlow DJ, Hylands PJ. Phytochemical databases of Chinese herbal constituents and bioactive plant compounds with known target specificities. J Chem Inf Model 2007; 47(2): 254-63.

43. Geetha P, Sivaram AJ, Jayakumar R, Gopi Mohan C. Integration of in silico modeling, prediction by binding energy and experimental approach to study the amorphous chitin nanocarriers for cancer drug delivery. Carbohydr Polym 2016; 142: 240-9.

44. Aribisala JO, Aruwa CE, Uthman TO, Nurain IO, Idowu K, Sabiu S. Cheminformatics bioprospection of broad spectrum plant secondary metabolites targeting the spike proteins of omicron variant and wild-type SARS-CoV-2. Metabolites 2022; 12(10).

45. Chen G, Chen P, Hsieh CY, *et al.* Alchemy: a quantum chemistry dataset for benchmarking AI models. arXiv preprint 2019; 1-11.

46. Wilkinson MD, Dumontier M, Aalbersberg IJJ, *et al.* Comment: the FAIR guiding principles for scientific data management and stewardship. Sci Data 2016; 3.

47. Lafuente D, Cohen B, Fiorini G, *et al.* A gentle introduction to machine learning for chemists: an undergraduate workshop using python notebooks for visualization, data processing, analysis, and modeling. J Chem Educ 2021; 98(9): 2892-8.

48. Dobbelaere MR, Plehiers PP, van de Vijver R, Stevens CV, van Geem KM. Machine learning in chemical engineering: strengths, weaknesses, opportunities, and threats. Engineering 2021; 7(9): 1201-11.

49. Mater AC, Coote ML. Deep learning in chemistry. J Chem Inf Model 2019; 59(6): 2545-59.

50. Chen G, Tao L, Li Y. Predicting polymers' glass transition temperature by a chemical language processing model. Polymers (Basel) 2021; 13(11).

51. Gómez-Bombarelli R, Wei JN, Duvenaud D, *et al.* Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 2018; 4(2): 268-76.

52. Kim H, Na J, Lee WB. Generative chemical transformer: neural machine learning of molecular geometric structures from chemical language via attention. Vol. 61, Journal of Chemical Information and Modeling. American Chemical Society 2021; pp. 5804-14.

53. Schwaller P, Hoover B, Reymond JL, Strobelt H, Laino T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. Sci Adv 2021; 7(15).

54. Rodríguez-Martínez X, Pascual-San-José E, Campoy-Quiles M. Accelerating organic solar cell material's discovery: high-throughput screening and big data. Vol. 14, Energy and Environmental Science. Royal Society of Chemistry 2021; pp. 3301-22.

55. Kerner J, Dogan A, von Recum H. Machine learning and big data provide crucial insight for future biomaterials discovery and research. Vol. 130, Acta Biomaterialia. Acta Materialia Inc. 2021; pp. 54-65.

56. Botu V, Ramprasad R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. Int J Quantum Chem 2015; 115(16): 1074-83.

57. Korshunova M, Ginsburg B, Tropsha A, Isayev O. OpenChem: a deep learning toolkit for computational chemistry and drug design. J Chem Inf Model 2021; 61(1): 7-13.

58. Kuntz D, Wilson AK. Machine learning, artificial intelligence, and chemistry: how smart algorithms are reshaping simulation and the laboratory. Pure Appl Chem. 2022; 94(8): 1019-54.