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 grown in recent years to speed up the discovery of new materials, drugs, synthetic substances and automate compound identification. Machine learning and data science are commonly used in organic chemistry to predict biological and physical-chemical properties of molecules and are referred to as QSAR (for biological properties) and QSPR (for non-biological properties). In addition, data science and machine learning have advanced the optimization of molecular properties, synthetic pathways, and even design of novel compounds. These models can learn the underlying patterns of molecular structures and generate new compounds with desirable properties. Machine Learning use is increasing 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 machine learning, particularly deep learning, is playing a significant role in advancing research in organic chemistry

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

1. Introduction

In recent years, the so-called data-driven approach, which involves data science, artificial intelligence, and big data has been changing the way science is done, since classic experiments, even in the hyphenated or high-throughput approach, can be time consuming, costly and dependant on the availability of equipment and reagents. Therefore, the use of tools and techniques related to Data Science, machine learning and rational designed have attracted attention of chemists, since it can save valuable resources and time. In this sense, approaches using data science have been grown in recent years with the aim of accelerating the discovery of new materials, drugs, synthetic substances and automated compound identification, among other tasks.

The data-driven approach to chemical discovery has been gaining relevance and importance as public databases of chemical substances emerge, such as PubChem(1), ZINC(2), ChemBL(3) and PubChemQC(4). The availability of such databases allowed the development of several Machine Learning 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, data science can accelerate scientific research, including research in organic chemistry, where it can be used to analyse and make predictions on chemical reactions, structures, physical-chemical, biological properties and big data analysis.

Although information technology and data science are now widespread, chemists have been using computers and computer science to solve chemistry problems for decades. In the early 1960s, Chemical Abstract Service developed an algorithm to generate a unique machine descriptor for chemical structure(10). This algorithm has inspired other algorithms and standards such as SMILES(11), SYBYL(12) and InChI(13). Moreover, organic chemists have also 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 it is natural their growing interest in machine learning and data science. Thus, this paper will briefly review the state of the art of data science and machine learning applications in chemistry and discuss a perspective on how it can change organic chemistry during the next years.

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

Data Science (DS) and Machine Learning (ML) are being increasingly applied in the field of Organic Chemistry to streamline and automate various research and development tasks such as drug discovery, spectroscopy, synthesis and chemical data analysis based on chemical structure, as well as the structures of organic molecules based on their chemical properties.

The prediction of biological and physical-chemical properties of molecules are the most common applications of machines learning and data science to chemistry, since they are useful to aid drug and material discoveries. When applied to the prediction of biological and physical-chemical properties, the use of ML and DS tools are often referred as QSAR when predicting biological properties and QSPR when predicting non-biological properties. QSAR and QSPR are so widespread in chemistry, that publications related to this theme often range from simple statistical regression and classification models(16) to sophisticated SVM(17), ANN(18) and ML ensembles(19). Although chemometric and statistical methods like Multiple Linear Regression (MLR), Principal Component Analysis (PCA) and Partial Least Squares (PLS), have been methods of choice for years for QSAR and QSPR studies, ML-based models are now preferred over these classical methods, as the diversity and availability of chemical data increases. pKa, for example, which has been modelled by MLR, PCA and PLS for years, now it has been successfully modelled by Mansouri et al.(20) using multiple ML approaches like an ensemble of SVM combined with k-nearest neighbours (kNN), extreme gradient boosting (XGB) and deep neural networks (DNN). The models were built using publicly available data, The source code and data are provided on GitHub and the performance of models compared favourably to the commercial products. Other examples of QSAR/QSPR modelling using ML include discovery of new materials(21,22), the prediction of ionic liquid properties(23), the modelling of drugs(24), quality control of fuels(25), the prediction and modelling of bioconcentration and toxicity of organic compounds(26). Hence, ML is a valuable tool for QSAR/QSPR, provides confident results and clearly outperforms chemometric methods.

Another field in organic chemistry that benefits of DS and ML is the computer-assisted structure elucidation (CASE). Early CASE such as DENDRAL(14) were solely based on rules, without any AI implementation. Further systems such as CHEMICS(27) were built on top of sophisticated rules and larger databases, but still without any AI aid. On the middle 1990's systems such as SpecSov(28) and SISTEMAT(29), which combined database approach and ANN were released. For the next years, this approach has become the standard for CASE with some developments, such as the incorporation of genetic algorithms(30) and decision trees(31). Due to AI and ML algorithms, there was a great development of CASE in the last decade. One of the many challenges in CASE that has been benefited from ML is the prediction of chemical shifts in NMR, which is used to check the validity of a structural propose. Message Passing Neural Networks (MPNN) 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 used to train ML models to predict the chemical classes of natural products(33). Models were trained based on the naproc13 database. This is one of most comprehensive natural products databases publicly available. The ML models predict the probability of a certain ^{13}C NMR spectrum belong to a substance certain class with an average accuracy ranging from 0.85 to 0.98, depending on the class. Thus, ML models such as ANN, SVM and XGBoost tools can be valuable tools to aid the task of structure elucidation.

The use of computer and information technology to drug and synthesis planning have also been of chemists' interest 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 system, since hundreds of reaction mechanisms do exist. 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 reaction conditions such as temperature, catalysis, solvents and other conditions(34). As finding reaction conditions is not a straightforward task, DS and ML have become a valuable tool for organic chemists, since it can recommend reaction conditions when other methods such as quantum chemistry fails(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 machine learning methods to accelerate calculations.(9,36). Machine Learning and Data Science have been applied to design the synthesis of new Metallo-Organic Framework (MOF)(8), in this work, authors have used automated processes and natural language processing to extract the data available in databases and in the literature to create a dataset that was used to train ML models to predict the final conditions and synthesis protocols to new MOF structures. Another field where machine learning is helpful is the polymer chemistry. Bayesian models have successfully developed to aid the discovery of polymers with high thermal conductivity(37). The ML models were trained on a substantially limited amount of polymeric properties data and have predicted QSPR with respect to 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 as well as band gap of polymers. The ML model was able to design 132 polymers with the desired properties and good accuracy. Thus, genetic algorithms can be generalised to predict and model other properties objectives, since corresponding reliable property prediction models can be provided.

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

Data Science, machine learning and data-driven science have emerged as the 4th paradigm of science(38), Hence, DS and ML has gaining importance not only in chemistry, but also in 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 is a critical step in ML model design and validation, since the size of dataset is of utmost importance 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 as Open-Science and FAIR(46) principles importance increases, the access to data for ML design will not be a threat.

The main strength of ML is the ability to find patterns and relationships where even an experienced researcher may not find. However, lack of education on the subject and misunderstanding of the nature of ML may be a threat for the development of advanced and accurate ML models in chemistry. ML algorithms are often referred as black boxes. This is not true, since a good ML model requires detailed analysis and adjustment of hyperparameters, which requires knowledge of how the algorithms and their implementation work. Some chemists also consider ML algorithms as statistical learning or a kind of “advanced chemometrics”, ML is neither and this kind of misunderstanding leads to the inadequate application of ML in chemistry. To overcome such issues, initiatives on chemical education and ML have appeared(47). A final remarkable weakness of ML is that if the data have many errors, the ML model will not be trustable(48). Hence, the accuracy of ML models depends not only on the algorithm, but also on the quality of the dataset and on the ability of the data scientist to choose and tune the right model.

Despite some criticism, ML use is increasing in chemistry and the 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 like 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 application in

chemistry are broad, opensource and free frameworks and tools to assist chemists in the development of ML models, such as OpenChem(57) and ML4Chem(58) have been designed. These frameworks encapsulate other frameworks such as pyTorch and sciKit-learn, making the task of training and testing ML models in Chemistry much easier.

In conclusion, ML applications in organic chemistry is in constant evolution and is greatly acceleration research. Moreover, this interdisciplinary field is playing a central role in changing the way 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 trusted by chemists, ML will be a routine in chemistry laboratory like any other equipment.

4. References

1. Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, 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, Patricia Bento A, Chambers J, Mendez D, 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 Jan 1;186.

6. Karydas C, Iatrou M, Kouretas D, Patouna A, Iatrou G, Lazos N, et al. Prediction of antioxidant activity of cherry fruits from UAS multispectral imagery using machine learning. Antioxidants. 2020 Feb 1;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. p. 16–25.

8. Luo Y, Zaremba O, Cierpka A, Andreo J, Wuttke S, Friederich P, et al. MOFs and Machine Learning MOF Synthesis Prediction Enabled by Automatic Data Mining and Machine Learning.

Angwandte Chemie Int Ed [Internet]. 2022;61:e20220024. Available from: https://doi.org/10.33774/chemrxiv-2021-kgd0h

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. p. 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 [Internet]. 1965;4(1):65. Available from: https://pubs.acs.org/sharingguidelines

11. Ritter G, Isenhour TL, Balaban AT, Moloc I, Read RC, Cornell DG, 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 [Internet]. 1989;29(18):133–86. Available from: https://pubs.acs.org/sharingguidelines

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 [Internet]. 1997;37:71–9. Available from: https://pubs.acs.org/sharingguidelines

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 May 21;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 (1979). 1988;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 [Internet]. 2009;20(4):770–87. Available from: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0103-50532009000400021&lng=en&nrm=iso&tlng=en

17. Norinder U. Support vector machine models in drug design: applications to drug transport processes and QSAR using simplex optimisations and variable selection. Neurocomputing [Internet]. 2003 Sep [cited 2012 Aug 7];55(1–2):337–46. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0925231203003746

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 [Internet]. 2009 Apr [cited 2012 Jul 28];27(7):797–802. Available from: http://www.ncbi.nlm.nih.gov/pubmed/19150251

19. Yao XJ, Panaye a, Doucet JP, Zhang RS, Chen HF, Liu MC, 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 [Internet]. 2004;44(4):1257–66. Available from: http://www.ncbi.nlm.nih.gov/pubmed/15272833

20. Mansouri K, Cariello NF, Korotcov A, Tkachenko V, Grulke CM, Sprankle CS, et al. Open-source QSAR models for pKa prediction using multiple machine learning approaches. J Cheminform. 2019 Sep 18;11(1).

21. Yang Y, Lin T, Weng XL, Darr J a., Wang XZ. Data flow modeling, data mining and QSAR in high-throughput discovery of functional nanomaterials. Comput Chem Eng [Internet]. 2011 Apr [cited 2013 Mar 1];35(4):671–8. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0098135410001602

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. Bulletin of Materials Science. 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 Mar 15;326.

24. D'Souza S, Prema K v., Balaji S. Machine learning models for drug–target interactions: current knowledge and future directions. Vol. 25, Drug Discovery Today. Elsevier Ltd; 2020. p. 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 Nov 15;304.

26. Ai H, Wu X, Zhang L, Qi M, Zhao Y, Zhao Q, 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 Sep 15;179:71–8.

27. Kimito Funatsu, Minoru Nishizaki, Shin-ichi Sasaki. 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 [Internet]. 1996;36:221–7. Available from: https://pubs.acs.org/sharingguidelines

29. Emerenciano V, Rodrigues G, Macarp PAT, Vesti SA, Borges JHG, Gastmans JP, 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. In: Journal of Chemical Information and Modeling. American Chemical Society; 2006. p. 775–87.

32. Han H, Choi S. Transfer Learning from Simulation to Experimental Data: NMR Chemical Shift Predictions. Journal of Physical Chemistry Letters. 2021 Apr 15;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 Jul 27;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 Nov 28;4(11):1465–76.

35. Zhang SQ, Xu LC, Li SW, Oliveira JCA, Li X, Ackermann L, et al. Bridging Chemical Knowledge and Machine Learning for Performance Prediction of Organic Synthesis. Chemistry – A European Journal. 2022 Nov 27;

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

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

38. HEY AJG, TANSLEY DSW, TOLLE KM. The Fourth Paradigm: Data-Intensive Scientific Discovery. Proceedings of the IEEE [Internet]. 2011;99(8):287. Available from: http://research.microsoft.com/en-us/collaboration/fourthparadigm/

39. Pollice R, dos Passos Gomes G, Aldeghi M, Hickman RJ, Krenn M, Lavigne C, et al. Data-Driven Strategies for Accelerated Materials Design. Acc Chem Res. 2021 Feb 16;54(4):849–60.

40. Leonelli S. Process-Sensitive Naming: Trait Descriptors and the Shifting Semantics of Plant (Data) Science. Philosophy, Theory, and Practice in Biology. 2022 Oct 21;14(0).

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 Dec 1;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 [Internet]. 2007;47(2):254–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/17381164

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 [Internet]. 2016;142:240–9. Available from: http://dx.doi.org/10.1016/j.carbpol.2016.01.059

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 Oct 1;12(10).

45. Chen G, Chen P, Hsieh CY, Lee CK, Liao B, Liao R, et al. Alchemy: A Quantum Chemistry Dataset for Benchmarking AI Models. arXiv preprint [Internet]. 2019 Jun 22;1–11. Available from: http://arxiv.org/abs/1906.09427

46. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. Comment: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3.

47. Lafuente D, Cohen B, Fiorini G, García AA, Bringas M, Morzan E, 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 Sep 14;98(9):2892–8.

48. Dobbelaere MR, Plehiers PP, van de Vijver R, Stevens C v., van Geem KM. Machine Learning in Chemical Engineering: Strengths, Weaknesses, Opportunities, and Threats. Engineering. 2021 Sep 1;7(9):1201–11.

49. Mater AC, Coote ML. Deep Learning in Chemistry. J Chem Inf Model. 2019;

50. Chen G, Tao L, Li Y. Predicting polymers' glass transition temperature by a chemical language processing model. Polymers (Basel). 2021 Jun 1;13(11).

51. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, et al. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. ACS Cent Sci. 2018 Feb 28;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. p. 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 andbig data. Vol. 14, Energy and Environmental Science. Royal Society of Chemistry; 2021. p. 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. p. 54–65.

56. Botu V, Ramprasad R. Adaptive machine learning framework to accelerate ab initio molecular dynamics. Int J Quantum Chem. 2015 Aug 1;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 Jan 25;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 and Applied Chemistry. 2022 Aug 1;94(8):1019–54.