Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Open data and algorithms for open science in AI-driven molecular informatics

Henning Otto Brinkhaus[1], Kohulan Rajan[1], Jonas Schaub[1], Achim Zielesny[2] and Christoph Steinbeck[1]

Abstract

Recent years have seen a sharp increase in the development of deep learning and artificial intelligence-based molecular informatics. There has been a growing interest in applying deep learning to several subfields, including the digital transformation of synthetic chemistry, extraction of chemical information from the scientific literature, and AI in natural product-based drug discovery. The application of AI to molecular informatics is still constrained by the fact that most of the data used for training and testing deep learning models are not available as FAIR and open data. As open science practices continue to grow in popularity, initiatives which support FAIR and open data as well as open-source software have emerged. It is becoming increasingly important for researchers in the field of molecular informatics to embrace open science and to submit data and software in open repositories. With the advent of open-source deep learning frameworks and cloud computing platforms, academic researchers are now able to deploy and test their own deep learning models with ease. With the development of new and faster hardware for deep learning and the increasing number of initiatives towards digital research data management infrastructures, as well as a culture promoting open data, open source, and open science, AI-driven molecular informatics will continue to grow. This review examines the current state of open data and open algorithms in molecular informatics, as well as ways in which they could be improved in future.

Addresses

[1] Institute for Inorganic and Analytical Chemistry, Friedrich Schiller University Jena, Lessingstr. 8, 07743 Jena, Germany

[2] Institute for Bioinformatics and Chemoinformatics, Westphalian University of Applied Sciences, August-Schmidt-Ring 10, 45665 Recklinghausen, Germany

Corresponding author: Steinbeck, Christoph (christoph.steinbeck@uni-jena.de)

Current Opinion in Structural Biology 2023, **79**:102542

This review comes from a themed issue on **Artificial Intelligence (AI) Methodology in Structural Biology (2023)**

Edited by **Andreas Bender**, **Chris de Graaf** and **Noel O'Boyle**

For complete overview of the section, please refer the article collection - Artificial Intelligence (AI) Methodology in Structural Biology (2023)

Available online 17 February 2023

https://doi.org/10.1016/j.sbi.2023.102542

Introduction

Considerable improvements in artificial intelligence (AI) research through the introduction of deep neural networks promise to transform society [1–4] and the way research is conducted [5,6]. However, in most areas of molecular informatics, the amount of training data available is insufficient for the use of today's most powerful deep neural network architectures, which demonstrate superior performance only by training with large amounts of data [7]. In addition, a thorough assessment of a model's true predictive performance in practice is a rare exception (e.g. the Critical Assessment of Protein Structure Prediction (CASP) [8]).

Because of this lack of accessible experimental data [9,10], machine learning predictions in chemistry are generally too error-prone to realize the potential of the new methods at this time. This necessitates a change in the way chemists publish their data and the type of data published [11,12]. The call for open data, open source, and open science (ODOSOS) in chemistry is not new [13,14], but with the advent of more powerful data-driven algorithms, it has never been more important.

Journals and funders demanding the deposition of research data and the necessary establishment of suitable research data infrastructures will inevitably alleviate the data shortage problem in the future [15,16]. The German government, for example, has recently decided to implement and long-term-fund a national research data infrastructure (Nationale Forschungsdateninfrastruktur, NFDI) [17] with 30 consortia in all areas of science, collaboratively developing open research data management (RDM) e-infrastructures, coordinated by an umbrella process and a joint directorate. One of those consortia is NFDI4Chem which is building an RDM e-infrastructure for chemistry that follows FAIR data principles [18] to make chemical data findable, accessible, interoperable, and reusable [19,20]. One flagship project of NFDI4Chem is nmrXiv, an open

and FAIR repository and analysis platform for NMR spectroscopy data [21].

In recent years, advances in artificial intelligence and data-driven applications in molecular informatics have provided a glimpse into the magnitude of future accomplishments, which have made open data a necessity for machine learning algorithms. Here, we attempt to present some of the major milestones of the past years and discuss obstacles that are yet to be overcome to enable similar AI-driven progress in (nearly) every area of chemistry.

The importance of openly available resources and data

One cause of the dissatisfying data shortage situation has been the lack of a culture of data deposition and sharing in chemistry in the past, where at least from the early 1990s onwards, with the advent of the internet, widespread data deposition and sharing would have been possible. There have been notable exceptions, such as the crystallography community, that have developed data deposition cultures even earlier. Both small molecules and biomacromolecule structures have been and are being deposited in the Protein Data Bank (PDB) [22,23] and the Cambridge Crystallographic Database (CCD) [24]. Of particular note, the open PDB in combination with the openly available protein sequence information (for multiple sequence alignments) formed the basis for the outstanding success of the AlphaFold protein 3D structure prediction system [5]. Similarly, open databases such as PubChem [25], ChEMBL [26], ChEBI [27], Drugbank [28], the Human Metabolome Database (HMDB) [29], the Collection of Open Natural Products (COCONUT) [30], the Natural Products Atlas [31], the Natural Products Magnetic Resonance Database [32], and ZINC [33] fundamentally broaden the research opportunities [34]. The PubChem database is used by millions of users every month [35]. An example for the usage of the referenced databases is the creation of a classifier that determines whether a Natural Product (NP) originates from funghi, plants, or bacteria based on its chemical structure with data obtained from the COCONUT database [36]. The ZINC database has recently been used for the in silico determination of drug candidates that inhibit the main protease of SARS-CoV-2 [37].

Another crucial aspect is the availability of open software libraries to handle and process chemical information, like the Chemistry Development Kit (CDK) [38], Indigo [39], RDKit [40], or OpenBabel [41], as well as the recently published Python-based Informatics Kit for Analysing Chemical Units (PIKAChU) [42]. Without these open-source projects, the research community would lack basic tools for programmatically reading, modifying, and processing chemical information. Accordingly, they are fundamental for every researcher in the field of molecular informatics.

Molecular string representations, such as DeepSMILES [43] and SELFIES [44], enable processing chemical structures using models like transformers that are designed to process sequential data. Recently, a study investigated the performance of transformers on different tasks using SMILES, DeepSMILES, and SELFIES. The amount of returned invalid chemical structures could be decreased when using DeepSMILES and especially SELFIES compared to SMILES, although the overall best performance was achieved using SMILES [45].

Without open libraries such as Tensorflow [46] and Pytorch [47] for the implementation and training of neural networks as well as the ubiquitous availability of Graphical Processing Units (GPU) and Tensor Processing Units (TPU) in cloud environments [48], the big leaps in molecular AI research would not have been possible.

An approach to the protein folding problem - AlphaFold

The problem of protein folding is considered one of the fundamental challenges of molecular biology because a large number of degrees of freedom of bonds and atoms in a protein leads to a combinatorial explosion in the number of possible low-energy arrangements [49]. In 2020, the *DeepMind* team announced a widely recognised breakthrough in the prediction of spatial protein 3D structures from their amino acid sequence with their deep learning-based system *AlphaFold* [5]. The system participated in the 13th and 14th Critical Assessment of Protein Structure Prediction (CASP) competition [8], outperforming all competitors. Since then, it has been made openly available and used to fill the open *AlphaFold Protein Structure Database* [50] which contains more than 200 million predicted protein 3D structures, covering nearly every known protein on earth [51]. Within a short period of time, the structures of 98.5% of the human proteome have been predicted using *AlphaFold*, while the previous decades of experimental research yielded 17% [52]. The system was trained on structural data openly deposited in the Protein Data Bank [22,23], which was founded and announced in 1971 [53]. The success story of *AlphaFold* illustrates what is possible today when researchers are able to access the data that scientists have produced over the course of 50 years.

It is important to mention that challenges like the prediction of the relative positions of protein domains and their changes when an external stimulus is applied remain partially unsolved. Additionally, the transition

from disordered to ordered domain states cannot be elucidated using *AlphaFold*, and it is limited to structures with less than 2700 amino acids [54]. Nevertheless, the high impact of its accurate protein structure predictions is indisputable [55]. For example, the predicted structural information about nucleoporins has been combined with cryo-electron tomography (cryo-ET) to generate a model that precisely explains 90% of the scaffold of the human nuclear pore complex (NPC) [56]. Another example is the identification of tens of thousands of unknown potential binding sites for iron-sulfur clusters and zinc ions in more than 360,000 proteins [57].

Digital transformation of synthetic chemistry

Similar to other fields, the foundation for successful machine learning applications in synthetic organic chemistry is the availability of extensive experimental data [58]. Recently, Strieth-Kalthoff et al. demonstrated the benefit that emerges from the usage of real experimental data for machine learning-based chemical yield predictions [12] while the prediction of reaction outcomes and yields remains a challenge in general [59]. Nonetheless, there have been impressive developments using attention-based deep learning methods to explore the chemical reaction space [60]. Schwaller et al. trained a transformer to predict chemical reaction outcomes with state-of-the-art results [61]. The resulting model which is referred to as *molecular transformer* was then used in combination with hypergraph exploration to automatically plan retrosynthesis routes [62]. Since then, the *molecular transformer* has been extended to predict the products of enzymatic reactions [63]. Based on the aforementioned retrosynthesis planning system, Probst et al. have published a biocatalysed synthesis planning system [64].

Schwaller et al. have also shown that the attention matrix weights of transformers that have been trained on unlabelled chemical reaction data can be used to determine accurate atom mappings [65]. Additionally, they demonstrated that attention-based models are highly suitable for the classification of chemical reactions [66]. Similar model architectures were successfully used to generate specific synthesis instructions [67] and to determine the yield of a given chemical reaction formula [68]. Andronov et al. successfully demonstrated the prediction of reagents based on given reaction SMILES strings using transformers. They were then able to use the reagent prediction model to fill in missing reagents in incomplete reaction data from US patents leading to an improved state-of-the-art model [61] for the prediction of reaction products [69]. Recently, Rohrbach et al. demonstrated the translation of synthesis protocols in the literature into a standardized chemical language, which could then be executed by their automated synthesis system [70].

Again, the described advances are exemplary cases of the synergy of deep learning-based models and the availability of training data. There are datasets extracted from US patents [66,71–74], the scientific literature [75], and high-throughput experiments (HTE) [76] available [60]. Recently, the Open Reaction Database (ORD) has been launched as a platform to replace unstructured reaction data in the supporting information of publications [77]. If it is accepted by the research community, the ORD may become a part of the solution to problems caused by the aforementioned lack of data and report bias [11,12]. Providing structured data in standardized formats may become a key step towards the digital transformation of synthetic chemistry.

Extraction of chemical information from the scientific literature

Besides enforcing FAIR data publication standards today and in the near future, it is important to tackle the damage that has already been done by publishing chemical data almost exclusively in a human-readable form with unstructured text and images in the past decades. The advances in the fields of natural language processing (NLP) [78–80] and computer vision (CV) [81–83] have made a new generation of chemical literature mining tools possible. These can be considered AI-driven solutions that enable further AI-driven advances by making concealed data accessible in structured, machine-readable formats.

The field of optical chemical structure recognition (OCSR) deals with the translation of images of chemical structures as they are published in the scientific literature into machine-readable representations of the underlying molecular graph [84,85]. In the past two years, a variety of deep learning-based OCSR methods [86–89] has been published, where *DECIMER Image-Transformer* [90], *Img2Mol* [91] and *SwinOCSR* [92] provide openly available source code and trained models. For the segmentation of chemical structure images from whole pages, the open-source tool *DECIMER Segmentation* can be used [93]. With the publication of the open-source depiction generation tool RanDepict, efforts have been made to standardize and diversify the training data for deep learning-based OCSR tools [94]. The newest version of DECIMER was trained on more than 400 Million images using the latest Tensor Processing Units [95] available on the Google cloud platform. Currently, DECIMER performs with an accuracy rate of above 90% and is regarded as an important point of reference for future work [85]. Without open databases like PubChem, where one can download over 100 million chemical structures for free, this would not have been possible.

Since its original release in 2016, the chemical literature mining toolkit *ChemDataExtractor* [96] has been

continuously developed [97,98]. The highly adaptable toolkit uses user-defined models of the information to be extracted in a pipeline with readers for different publisher formats and a system for interdependency resolution with a set of parsers and a sophisticated chemical named entity recognition system [99] to extract chemical information in a structured data format [97]. In the past years, ChemDataExtractor has been extensively used to automatically generate databases about refraction indices and dielectric constants [100], battery material properties [101], properties of semiconductors for building solar cells [102], magnetic properties [103], as well as UV/Vis spectra [104].

In addition to the technical obstacles, scientific publishers hinder literature mining essentially by hiding publications behind paywalls and limiting the number of publications that can be downloaded and used even if a subscription is available. Some publishers like Elsevier offer markup versions of their publications for text mining purposes to academic researchers [105], but there is a long way to go to truly make all published chemical information available. In 2018, an international group of research funders announced the initiative *Plan S* which requires scientists who benefit from their funding to publish in open-access journals [106]. Recently, the US government announced that they will require all publicly funded research to be openly accessible from 2026 on [107]. With RDM e-infrastructures being established as the mandatory scientific data publication standard, the kind of literature mining methods described herein will become obsolete in the future. For now, they are indispensable for artificially intelligent data-driven applications.

AI in natural product-based drug discovery

The field of drug discovery has shifted towards implementing approaches based on the analysis of large amounts of data and deep learning [108]. As a result of the growing demand for efficient new drugs, the field has experienced rapid growth in the last few years. NP are attractive to drug developers due to their availability and their potential affinity to protein drug targets [109,110].

There have been significant advances in various areas of the field, such as the prediction of biochemical effects of NP based on their molecular structure [111], in the field of genome mining for the discovery of bioactive compounds [112], the mining of mass spectrometry-based metabolomics data [113], and integrative approaches that combine metabolomics and genomics data [114].

The initial hope that large-scale data analysis in the different omics-related research fields would boost the drug discovery rate has not yet materialised [115], but the methods are progressing continuously. The open access to databases and repositories such as MetaboLights [116], the HMDB [29], the Metabolomics Workbench [117], and METASPACE [118] is crucial for the identification of metabolites and NP [112]. In 2021, the Paired Omics Data Platform (PODP) was launched as a community-driven platform that provides linked metabolome and genome data according to the FAIR principles [119].

NP-based drug discovery has greatly benefited from models developed for NLP [120]. For example, in 2021, Huang et al. published *MolTrans*, a state-of-the-art deep learning-based framework for the in silico prediction of Drug–Protein Interactions (DPI) [121]. In the following year, Wang et al. presented their structure-aware multimodal deep DPI prediction model *STAMP-DPI*, which outperforms *MolTrans*. The tool has been published along a large high-quality training and benchmarking dataset [122]. The adaptation of sequence models like the transformer [78] for AI-based drug design requires large amounts of well-curated, high-quality data.

The recent development in the field of deep generative models helps researchers generate molecules with desired properties [123], but a model that can generalise well and can generate molecules with desirable properties requires a large amount of training data. When dealing with artificially generated structures, it is also necessary to consider their synthetic accessibility. To successfully use deep learning on published NP structures, well-curated data is essential. Published data resources are often incomplete, inaccessible, or no longer available [124] which makes available resources like the Natural Products Atlas [31], LOTUS [125], and COCONUT [30] even more important.

The development of deep learning-based models has assisted the advancement of drug discovery overall, with more advancements being made in the development of models and increasing access to open data and open databases helping this field grow. We hope that the research community will continue to actively contribute to openly available data sources to enable further progress in the field.

Conclusions

The developments of the past years demonstrate the potential of data-driven machine learning applications in the field of molecular informatics in an impressive manner [5,65,70]. An obvious requirement to benefit from this development is the availability of open structured experimental data [11,12]. The integration of open data infrastructures will enable AI to be used in nearly every field of chemistry. The application of deep learning methodologies and the sharing of code and data in the field of chemistry are still in their early stages and

require more community standards to be developed. Many of the models are still being trained from scratch using in-house servers and GPUs, which is a time-consuming and restrictive process. The rapid growth of the field will be enabled by the sharing of already-trained models and curated data with the public. When sharing code or data, high quality and data standards must be maintained [126]. Using the public cloud infrastructures will readily allow researchers to take advantage of the latest developments in hardware and software, which will lead to faster growth and a reduction in energy consumption. There are several initiatives working continuously to implement open data, open-source, and open science in their individual research area [13,14,17,18,20,21,77,106,107,127,128]. Fueled by the availability of more and more open research data, AI-powered molecular informatics will be a key driver of the digital transformation of chemistry in the coming years.

Author contributions

HOB, KR and JS conducted the literature review and wrote the article. CS and AZ conceived the study and supervised the work. All authors read and approved the final manuscript.

Declaration of competing interest

AZ is co-founder of GNWI—Gesellschaft für naturwissenschaftliche Informatik mbH, Dortmund, Germany.

Data availability

No data was used for the research described in the article.

Acknowledgements

HOB, JS and CS acknowledge funding by the Carl-Zeiss-Foundation. KR acknowledges the funding by the German Research Foundation within the framework CRC1127 ChemBioSys.

List of abbreviations

AI Artificial Intelligence
CASP Critical Assessment of protein Structure Prediction
CCD Cambridge Crystallographic Database
CDK Chemistry Development Kit
cryo-ET cryo-Electron Tomography
COCONUT COlleCtion of Open Natural Products
CV Computer Vision
DECIMER Deep lEarning for Chemical ImagE Recognition
DPI Drug–Protein Interaction
FAIR Findable, Accessible, Interoperable, and Reusable
GPU Graphics Processing Unit
HTE High-Throughput Experiments
HMDB Human Metabolome DataBase
NFDI Nationale ForschungsDatenInfrastruktur (National Research Data Infrastructure)
NFDI4Chem National Research Data Infrastructure for Chemistry
NLP Natural Language Processing
NP Natural Products
NP-MRD Natural Products Magnetic Resonance Database
NPC Nuclear Pore Complex
OCSR Optical Chemical Structure Recognition
ODOSOS Open Data, Open Source, and Open Science
ORD Open Reaction Database
PDB Protein DataBank
PODP Paired Omics Data Platform
PIKAChU Python-based Informatics Kit for Analysing CHemical Units
RDM Research Data Management
TPU Tensor Processing Unit

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest

1. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, *et al.*: **Mastering the game of Go without human knowledge**. *Nature* 2017, **550**: 354–359.

2. Gupta A, Anpalagan A, Guan L, Khwaja AS: **Deep learning for object detection and scene perception in self-driving cars: survey, challenges, and open issues**. *Array* 2021, **10**:100057.

3. Ramesh A, Dhariwal P, Nichol A, Chu C, Chen M: **Hierarchical text-conditional image generation with CLIP latents**. *arXiv [csCV]* 2022.

4. Rombach Robin, Blattmann Andreas, Lorenz Dominik, Esser Patrick, Ommer B: **High-resolution image synthesis with latent diffusion models**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2022: 10684–10695.

5. * Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, *et al.*: **Highly accurate protein structure prediction with AlphaFold**. *Nature* 2021, **596**:583–589.

The first ever state-of-the-art neural network model capable of predicting the structure of a protein with great accuracy. The authors of this paper demonstrate that the model is capable of predicting the structure of a protein even when there are no similar structures available.

6. Kirkpatrick J, McMorrow B, Turban DHP, Gaunt AL, Spencer JS, Matthews AGDG, Obika A, Thiry L, Fortunato M, Pfau D, *et al.*: **Pushing the frontiers of density functionals by solving the fractional electron problem**. *Science* 2021, **374**:1385–1389.

7. Chuang KV, Gunsalus LM, Keiser MJ: **Learning molecular representations for medicinal chemistry**. *J Med Chem* 2020, **63**:8705–8722.

8. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J: **Critical assessment of methods of protein structure prediction (CASP)-Round XIV**. *Proteins* 2021, **89**:1607–1617.

In this paper, the authors mention that the deep-learning-based method has a higher accuracy for predicting protein structures compared to all previous approaches and it is on par with experimental results in terms of accuracy.

9. Bajorath J: **State-of-the-art of artificial intelligence in medicinal chemistry**. *Future Sci OA* 2021, **7**. FSO702.

10. Tripathi MK, Nath A, Singh TP, Ethayathulla AS, Kaur P: **Evolving scenario of big data and Artificial Intelligence (AI) in drug discovery**. *Mol Divers* 2021, **25**:1439–1460.

11. Cole JM: **The chemistry of errors**. *Nat Chem* 2022, **14**:973–975.
The author of this paper states that in order to improve predictions with machine-learning methods, it is necessary to have access to a substantial amount of experimental data.

12. Strieth-Kalthoff F, Sandfort F, Kühnemund M, Schäfer FR, Kuchen H, Glorius F: **Machine learning for chemical reactivity: the importance of failed experiments**. *Angew Chem Int Ed Engl* 2022, **61**:e202204647.

13. Guha R, Howard MT, Hutchison GR, Murray-Rust P, Rzepa H, Steinbeck C, Wegner J, Willighagen EL: **The Blue Obelisk-interoperability in chemical informatics**. *J Chem Inf Model* 2006, **46**:991–998.

14. O'Boyle NM, Guha R, Willighagen EL, Adams SE, Alvarsson J, Bradley J-C, Filippov IV, Hanson RM, Hanwell MD, Hutchison GR, *et al.*: **Open data, open source and open standards in chemistry: the blue obelisk five years on**. *J Cheminf* 2011, **3**:37.

15. * Schymanski EL: **Bolton EE: FAIR chemical structures in the journal of cheminformatics**. *J Cheminf* 2021, **13**:50.
The purpose of this paper is to emphasize the importance of openly sharing chemical information and structural data in the field of cheminformatics. Providing data according to FAIR principles enhances the journal's commitment to open science, and both readers and authors will benefit from such an initiative.

16. Zdrazil B, Guha R: **Diversifying cheminformatics**. *J Cheminf* 2022, **14**:25.

17. Hartl N, Wössner E, Sure-Vetter Y: **Nationale Forschungsdateninfrastruktur (NFDI)**. *Informatik-Spektrum* 2021, **44**:370–373.

18. Steinbeck C, Koepler O, Bach F, Herres-Pawlis S, Jung N, Liermann J, Neumann S, Razum M, Baldauf C, Biedermann F, *et al.*: **NFDI4Chem-Towards a national research data infrastructure for chemistry in Germany**. *Research Ideas and Outcomes* 2020, **6**:e55852.

19. Rzepa HS: **The long and winding road towards FAIR data as an integral component of the computational modelling and dissemination of chemistry**. *Isr J Chem* 2022, **62**:e202100034.

20. Herres-Pawlis S, Liermann JC, Koepler O: **Research data in chemistry – results of the first NFDI4Chem community survey**. *Z Anorg Allg Chem* 2020, **646**:1748–1757.
With the growing demand for open and FAIR data, research data management is becoming increasingly important in chemistry. NFDI4Chem is a first-of-its-kind national research data infrastructure initiative aimed at providing a public research data management platform for researchers to store and share the data they produce in the field of chemistry.

21. NFDI4Chem: **nmrXiv - open, FAIR and Consensus-Driven NMR spectroscopy data repository and analysis platform**. In *nmrXiv - Open, FAIR and Consensus-Driven NMR spectroscopy data repository and analysis platform*; 2022.
An ongoing important project of NFDI4Chem is nmrXiv, an open and FAIR repository and analysis platform for NMR spectroscopy data. This is the first open platform of its kind made available to the public.

22. wwPDB consortium: **Protein Data Bank: the single global archive for 3D macromolecular structure data**. *Nucleic Acids Res* 2019, **47**:D520–D528.

23. Burley SK, Bhikadiya C, Bi C, Bittrich S: **RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental**. *Nucleic acids* 2021.

24. Groom CR, Bruno IJ, Lightfoot MP, Ward SC: **The Cambridge structural database**. *Acta Crystallogr B Struct Sci Cryst Eng Mater* 2016, **72**:171–179.

25. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, *et al.*: **PubChem in 2021: new data content and improved web interfaces**. *Nucleic Acids Res* 2021, **49**:D1388–D1395.

26. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, *et al.*: **The ChEMBL database in 2017**. *Nucleic Acids Res* 2017, **45**:D945–D954.

27. Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C: **ChEBI in 2016: improved services and an expanding collection of metabolites**. *Nucleic Acids Res* 2016, **44**:D1214–D1219.

28. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, *et al.*: **DrugBank 5.0: a major update to the DrugBank database for 2018**. *Nucleic Acids Res* 2018, **46**:D1074–D1082.

29. Wishart DS, Guo A, Oler E, Wang F, Anjum A, Peters H, Dizon R, Sayeeda Z, Tian S, Lee BL, *et al.*: **HMDB 5.0: the human metabolome database for 2022**. *Nucleic Acids Res* 2022, **50**:D622–D631.

30. * Sorokina M, Merseburger P, Rajan K, Yirik MA: **Steinbeck C: COCONUT online: collection of open natural products database**. *J Cheminf* 2021, **13**:2.
The COlleCtion of Open Natural prodUcTs (COCONUT) is currently the largest open database available for natural products. The database currently contains data from 53 natural products databases, and it is one of the most active databases in the field.

31. van Santen JA, Poynton EF, Iskakova D, McMann E, Alsup TA, Clark TN, Fergusson CH, Fewer DP, Hughes AH, McCadden CA, *et al.*: **The Natural Products Atlas 2.0: a database of microbially-derived natural products**. *Nucleic Acids Res* 2022, **50**:D1317–D1323.

32. Wishart DS, Sayeeda Z, Budinski Z, Guo A, Lee BL, Berjanskii M, Rout M, Peters H, Dizon R, Mah R, *et al.*: **NP-MRD: the natural products magnetic resonance database**. *Nucleic Acids Res* 2022, **50**:D665–D677.

33. Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, Moroz YS, Mayfield J, Sayle RA: **ZINC20-A free ultralarge-scale chemical database for ligand discovery**. *J Chem Inf Model* 2020, **60**:6065–6073.

34. Wegner JK, Sterling A, Guha R, Bender A, Faulon J-L, Hastings J, O'Boyle N, Overington J, Van Vlijmen H, Willighagen E: *Cheminformatics. Commun ACM* 2012, **55**:65–75.

35. Kim S, Cheng T, He S, Thiessen PA, Li Q, Gindulyte A, Bolton EE: **PubChem protein, gene, pathway, and taxonomy data collections: bridging biology and chemistry through target-centric views of PubChem data**. *J Mol Biol* 2022, **434**:167514.

36. Capecchi A, Reymond J-L: **Classifying natural products from plants, fungi or bacteria using the COCONUT database and machine learning**. *J Cheminf* 2021, **13**:82.

37. Mathpal S, Joshi T, Sharma P, Joshi T, Pundir H, Pande V, Chandra S: **A dynamic simulation study of FDA drug from zinc database against COVID-19 main protease receptor**. *J Biomol Struct Dyn* 2022, **40**:1084–1100.

38. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, Kuhn S, Pluskal T, Rojas-Chertó M, Spjuth O, *et al.*: **The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching**. *J Cheminf* 2017:9.

39. Pavlov D, Rybalkin M, Karulin B, Kozhevnikov M, Savelyev A, Churinov A: **Indigo: universal cheminformatics API**. *J Cheminf* 2011, **3**. P4.

40. Landrum G, Tosco P, Kelleyet B, *et al.*: ***rdkit: 2022_03_3 (Q1 2022) Release.*** 2022, https://doi.org/10.5281/zenodo.7541264.

41. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR: **Open Babel: an open chemical toolbox**. *J Cheminf* 2011, **3**:33.

42. Terlouw BR, Vromans SPJM, Medema MH: **PIKAChU: a Python-based informatics kit for analysing chemical units**. *J Cheminf* 2022, **14**:34.

The PIKAChU cheminformatics library is a pure python library that can be used for many cheminformatics analyses using python. It is the first-ever cheminformatics library to be written entirely in Python.

43. O'Boyle N, Dalke A: *DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures*. 2018, https://doi.org/10.26434/chemrxiv.7097960.v1.

44. * Krenn M, Ai Q, Barthel S, Carson N, Frei A, Frey NC, Friederich P, Gaudin T, Gayle AA, Jablonka KM, *et al.*: **SELFIES and the future of molecular string representations**. *Patterns Prejudice* 2022, **3**:100588.

String representations are widely used in deep learning models in chemistry in order to learn and interpret chemical structures. The most widely used string representation, SMILES, is not designed for deep learning tasks. To address this issue, the authors of this paper have developed a new machine-readable string representation, SELFreferencIng Embedded Strings (SELFIES). Moreover, the paper examines the different string representations that are currently available, their shortcomings, and the potential for future development.

45. Rajan K, Steinbeck C, Zielesny A: **Performance of chemical structure string representations for chemical image recognition using transformers**. *Digital Discovery* 2022, **1**:84–90.

46. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, Corrado GS, Davis A, Dean J, Devin M, *et al.*: **TensorFlow: large-scale machine learning on heterogeneous distributed systems**. *arXiv [csDC]* 2016.

47. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, *et al.*: **PyTorch: an imperative style, high-performance deep learning library**. *arXiv [csLG]* 2019.

48. You Y, Zhang Z, Hsieh C, Demmel J, Keutzer K: **Fast deep neural network training on distributed systems and cloud TPUs**. *IEEE Trans Parallel Distr Syst* 2019, **30**:2449–2462.

49. Levinthal C: *How to fold graciously. Mossbauer spectroscopy in biological systems proceedings*. University of Illinois Bulletin; 1969.

50. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, *et al.*: **AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models**. *Nucleic Acids Res* 2022, **50**:D439–D444.

51. * Callaway E: **"The entire protein universe": AI predicts shape of nearly every known protein**. *Nature* 2022, **608**:15–16.

Thus far, AlphaFold, a deep-learning algorithm, has predicted nearly 200 million protein structures for over one million species, meaning that nearly every protein on the planet has a structure that has been predicted by AlphaFold. This article highlights the significance of this contribution to the natural sciences.

52. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, Bridgland A, Cowie A, Meyer C, Laydon A, *et al.*: **Highly accurate protein structure prediction for the human proteome**. *Nature* 2021, **596**:590–596.

53. **Data, Crystallography: protein data bank**. *Nat New Biol* 1971.

54. David A, Islam S, Tankhilevich E, Sternberg MJE: **The AlphaFold database of protein structures: a biologist's guide**. *J Mol Biol* 2022, **434**:167336.

55. Varadi M, Velankar S: **The impact of AlphaFold Protein Structure Database on the fields of life sciences**. *Proteomics* 2022.

56. Mosalaganti S, Obarska-Kosinska A, Siggel M, Taniguchi R, Turoňová B, Zimmerli CE, Buczak K, Schmidt FH, Margiotta E, Mackmull M-T, *et al.*: **AI-based structure prediction empowers integrative structural analysis of human nuclear pores**. *Science* 2022, **376**. eabm9506.

57. Wehrspan ZJ, McDonnell RT, Elcock AH: **Identification of iron-sulfur (Fe-S) cluster and zinc (Zn) binding sites within proteomes predicted by DeepMind's AlphaFold2 program dramatically expands the metalloproteome**. *J Mol Biol* 2022, **434**:167377.

58. Segler MHS, Preuss M, Waller MP: **Planning chemical syntheses with deep neural networks and symbolic AI**. *Nature* 2018, **555**:604–610.

59. Davies IW: **The digitization of organic synthesis**. *Nature* 2019, **570**:175–181.

60. Schwaller P, Vaucher AC, Laplaza R, Bunne C, Krause A, Corminboeuf C, Laino T: **Machine intelligence for chemical reaction space**. *Wiley Interdiscip Rev Comput Mol Sci* 2022, https://doi.org/10.1002/wcms.1604.

61. Schwaller P, Laino T, Gaudin T, Bolgar P, Hunter CA, Bekas C, Lee AA: **Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction**. *ACS Cent Sci* 2019, **5**:1572–1583.

This is a first-of-its-kind data-driven deep-learning model for the prediction of chemical reactions. Using the model presented in this publication, the authors were able to achieve an accuracy of over 98% in classification. In addition, they explain that the learned representations could be used as fingerprints of reactions.

62. Schwaller P, Petraglia R, Zullo V, Nair VH, Haeuselmann RA, Pisoni R, Bekas C, Iuliano A, Laino T: **Predicting retrosynthetic pathways using transformer-based models and a hyper-graph exploration strategy**. *Chem Sci* 2020, **11**:3316–3325.

63. Kreutter D, Schwaller P, Reymond J-L: **Predicting enzymatic reactions with a molecular transformer**. *Chem Sci* 2021, **12**:8648–8659.

64. Probst D, Manica M, Nana Teukam YG, Castrogiovanni A, Paratore F, Laino T: **Biocatalysed synthesis planning using data-driven learning**. *Nat Commun* 2022, **13**:964.

65. Schwaller P, Hoover B, Reymond J-L, Strobelt H, Laino T: **Extraction of organic chemistry grammar from unsupervised learning of chemical reactions**. *Sci Adv* 2021:7.

66. Schwaller P, Probst D, Vaucher AC, Nair VH, Kreutter D, Laino T, Reymond J-L: **Mapping the space of chemical reactions using attention-based neural networks**. *Nat Mach Intell* 2021, **3**:144–152.

67. Vaucher AC, Zipoli F, Geluykens J, Nair VH, Schwaller P, Laino T: **Automated extraction of chemical synthesis actions from experimental procedures**. *Nat Commun* 2020, **11**:3601.

68. Schwaller P, Vaucher AC, Laino T, Reymond J-L: **Prediction of chemical reaction yields using deep learning**. *Mach Learn: Sci Technol* 2021, **2**, 015016.

69. Andronov M, Voinarovska V, Andronova N, Wand M, Clevert D-A, Schmidhuber J: **Reagent prediction with a molecular transformer improves reaction data quality**. *ChemRxiv* 2022, https://doi.org/10.26434/chemrxiv-2022-sn2kr.

70. Rohrbach S, Šiaučiulis M, Chisholm G, Pirvan P-A, Saleeb M, Mehr SHM, Trushina E, Leonov AI, Keenan G, Khan A, *et al.*: **Digitization and validation of a chemical synthesis literature database in the ChemPU**. *Science* 2022, **377**:172–180.

71. Coley CW, Barzilay R, Jaakkola TS, Green WH, Jensen KF: **Prediction of organic reaction outcomes using machine learning**. *ACS Cent Sci* 2017, **3**:434–443.

72. Jin W, Coley C, Barzilay R, Jaakkola T: **Predicting organic reaction outcomes with Weisfeiler-Lehman network**. *Adv Neural Inf Process Syst* 2017, **30**.

73. 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.

74. Dai H, Li C, Coley C: **Dai: retrosynthesis prediction with conditional graph logic network**. *Adv Neural Inf Process Syst* 2019, **32,** b.

75. Jiang S, Zhang Z, Zhao H, Li J, Yang Y, Lu B-L, Xia N: **When SMILES smiles, practicality judgment and yield prediction of chemical reaction via deep chemical Language Processing**. *IEEE Access* 2021, **9**:85071–85083.

76. Nielsen MK, Ahneman DT, Riera O, Doyle AG: **Deoxyfluorination with sulfonyl fluorides: navigating reaction space with machine learning**. *J Am Chem Soc* 2018, **140**:5004–5008.

77. Kearnes SM, Maser MR, Wleklinski M, Kast A, Doyle AG, Dreher SD, Hawkins JM, Jensen KF, Coley CW: **The open reaction database**. *J Am Chem Soc* 2021, **143**:18820–18826.

78. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I: **Attention is all you need**. *Adv Neural Inf Process Syst* 2017, **30**.

79. Devlin J, Chang M-W, Lee K, Toutanova K: **BERT: pre-training of deep bidirectional transformers for language understanding**. *arXiv [csCL]* 2018.

80. Brown T, Mann B, Ryder N, *et al.*: **Language models are few-shot learners**. *Adv Neural Inf Process Syst* 2020, **33**: 1877–1901.

81. Tan M, Le Q: **EfficientNetV2: smaller models and faster training**. In *Proceedings of the 38th international conference on machine learning*. Edited by Meila M, Zhang T; 18–24 Jul 2021: 10096–10106. PMLR.

82. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S, Guo B: **Swin Transformer: Hierarchical Vision Transformer using Shifted Windows**. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 2021:10012–10022.

83. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, *et al.*: **An image is worth 16x16 words: transformers for image recognition at scale**. *arXiv [csCV]* 2020.

84. * Rajan K, Brinkhaus HO, Zielesny A: **Steinbeck C: a review of optical chemical structure recognition tools**. *J Cheminf* 2020, **12**:60.

This paper summarizes the currently available rule-based and deep learning-based tools for optical chemical structure recognition and explains the need for more open-source and deep learning-based tools in this field. This is the first review published in the field of OCSR.

85. Musazade F, Jamalova N, Hasanov J: **Review of techniques and models used in optical chemical structure recognition in images and scanned documents**. *J Cheminf* 2022, **14**:61.

86. Oldenhof M, Arany A, Moreau Y, Simm J: **ChemGrapher: optical graph recognition of chemical compounds by deep learning**. *J Chem Inf Model* 2020, **60**:4506–4517.

87. Weir H, Thompson K, Woodward A, Choi B, Braun A, Martínez TJ: **ChemPix: automated recognition of hand-drawn hydrocarbon structures using deep learning**. *Chem Sci* 2021, **12**:10622–10633.

88. Yoo S, Kwon O, Lee H: **Image-to-Graph transformers for chemical structure recognition**. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*; 2022:3393–3397.

89. Zhang X-C, Yi J-C, Yang G-P, Wu C-K, Hou T-J, Cao D-S: **ABC-Net: a divide-and-conquer based deep learning architecture for SMILES recognition from molecular images**. *Briefings Bioinf* 2022:23.

90. Rajan K, Zielesny A, Steinbeck C: **DECIMER 1.0: deep learning for chemical image recognition using transformers**. *J Cheminf* 2021, **13**:61.

91. Clevert D-A, Le T, Winter R, Montanari F: **Img2Mol - accurate SMILES recognition from molecular graphical depictions**. *Chem Sci* 2021, https://doi.org/10.1039/D1SC01839F.

92. Xu Z, Li J, Yang Z, Li S, Li H: **SwinOCSR: end-to-end optical chemical structure recognition using a Swin Transformer**. *J Cheminf* 2022, **14**:41.

93. * Rajan K, Brinkhaus HO, Sorokina M, Zielesny A: **Steinbeck C: DECIMER-Segmentation: automated extraction of chemical structure depictions from scientific literature**. *J Cheminf* 2021, **13**:20.

DECIMER -Segmentation is the first and only open deep learning-based algorithm available for the segmentation of chemical structures in the literature. This tool is capable of detecting and segmenting chemical structures from literature with an accuracy of more than 90%. In this field, this is the first paper to explain such an algorithm.

94. Brinkhaus HO, Rajan K, Zielesny A, Steinbeck C: **RanDepict: random chemical structure depiction generator**. *J Cheminf* 2022, **14**:31.

95. Norrie T, Patil N, Yoon DH, Kurian G, Li S, Laudon J, Young C, Jouppi N, Patterson D: **The design process for google's training chips: TPUv2 and TPUv3**. *IEEE Micro* 2021, **41**:56–63.

96. Swain MC, Cole JM: **ChemDataExtractor: a toolkit for automated extraction of chemical information from the scientific literature**. *J Chem Inf Model* 2016, **56**:1894–1904.

97. Mavračić J, Court CJ, Isazawa T, Elliott SR, Cole JM: **ChemDataExtractor 2.0: autopopulated ontologies for materials science**. *J Chem Inf Model* 2021, **61**:4280–4289.

98. Zhu M, Cole JM: **PDFDataExtractor: a tool for reading scientific text and interpreting metadata from the typeset literature in the portable document format**. *J Chem Inf Model* 2022, **62**: 1633–1643.

99. Isazawa T, Cole JM: **Single model for organic and inorganic chemical named entity recognition in ChemDataExtractor**. *J Chem Inf Model* 2022, **62**:1207–1213.

100. Zhao J, Cole JM: **A database of refractive indices and dielectric constants auto-generated using ChemDataExtractor**. *Sci Data* 2022, **9**:192.

101. Huang S, Cole JM: **A database of battery materials auto-generated using ChemDataExtractor**. *Sci Data* 2020, **7**:260.

102. Beard EJ, Cole JM: **Perovskite- and dye-sensitized solar-cell device databases auto-generated using ChemDataExtractor**. *Sci Data* 2022, **9**:329.

103. Court CJ, Cole JM: **Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction**. *Sci Data* 2018, **5**:180111.

104. Beard EJ, Sivaraman G: **Vázquez-Mayagoitia Á, Vishwanath V, Cole JM: comparative dataset of experimental and computational attributes of UV/vis absorption spectra**. *Sci Data* 2019, **6**:307.

105. Van Noorden R: **Elsevier opens its papers to text-mining**. *Nature* 2014, **506**:17.

106. Else H: **A guide to Plan S: the open-access initiative shaking up science publishing**. *Nature* 2021, https://doi.org/10.1038/d41586-021-00883-6.

This article describes a first-of-its-kind global initiative of scientific funders to require all research conducted under their funding to be published openly. An initiative like this of key stakeholders in the scientific system holds great promise to improve the open availability of research data and findings in the near future.

107. Tollefson J, Van Noorden R: **US government reveals big changes to open-access policy**. *Nature* 2022, **609**:234–235.

108. Jiménez-Luna J, Grisoni F, Weskamp N, Schneider G: **Artificial intelligence in drug discovery: recent advances and future perspectives**. *Expet Opin Drug Discov* 2021, **16**:949–959.

109. Atanasov AG, Zotchev SB, Dirsch VM: **International natural product sciences taskforce, supuran CT: natural products in drug discovery: advances and opportunities**. *Nat Rev Drug Discov* 2021, **20**:200–216.

110. Saldívar-González FI, Aldas-Bulos VD, Medina-Franco JL, Plisson F: **Natural product drug discovery in the artificial intelligence era**. *Chem Sci* 2022, **13**:1526–1546.

111. Jeon J, Kang S, Kim HU: **Predicting biochemical and physiological effects of natural products from molecular structures using machine learning**. *Nat Prod Rep* 2021, **38**:1954–1966.

112. Bauman KD, Butler KS, Moore BS, Chekan JR: **Genome mining methods to discover bioactive natural products**. *Nat Prod Rep* 2021, **38**:2100–2129.

113. Jarmusch SA, van der Hooft JJJ, Dorrestein PC, Jarmusch AK: **Advancements in capturing and mining mass spectrometry data are transforming natural products research**. *Nat Prod Rep* 2021, **38**:2066–2082.

114. Caesar LK, Montaser R, Keller NP, Kelleher NL: **Metabolomics and genomics in natural products research: complementary tools for targeting new chemical entities**. *Nat Prod Rep* 2021, **38**:2041–2065.

115. Cech NB, Medema MH, Clardy J: **Benefiting from big data in natural products: importance of preserving foundational skills and prioritizing data quality**. *Nat Prod Rep* 2021, **38**: 1947–1953.

116. Haug K, Cochrane K, Nainala VC, Williams M, Chang J, Jayaseelan KV, O'Donovan C: **MetaboLights: a resource evolving in response to the needs of its scientific community**. *Nucleic Acids Res* 2020, **48**:D440–D444.

117. Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Burant C, Edison A, Fiehn O, Higashi R, Nair KS, *et al.*: **Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools**. *Nucleic Acids Res* 2016, **44**: D463–D470.

118. Alexandrov T, Ovchinnikova K, Palmer A, Kovalev V, Tarasov A, Stuart L, Nigmetzianov R, Fay D, Gaudin M, Lopez CG, *et al.*: **METASPACE: a community-populated knowledge base of spatial metabolomes in health and disease**. *bioRxiv* 2019, https://doi.org/10.1101/539478.

119. Schorn MA, Verhoeven S, Ridder L, Huber F, Acharya DD, Aksenov AA, Aleti G, Moghaddam JA, Aron AT, Aziz S, *et al.*: **A community resource for paired genomic and metabolomic data mining**. *Nat Chem Biol* 2021, **17**:363–368.

120. Walters WP, Barzilay R: **Critical assessment of AI in drug discovery**. *Expet Opin Drug Discov* 2021, **16**:937–947.

121. Huang K, Xiao C, Glass LM, Sun J: **MolTrans: molecular interaction transformer for drug-target interaction prediction**. *Bioinformatics* 2021, **37**:830–836.

122. Wang P, Zheng S, Jiang Y, Li C, Liu J, Wen C, Patronov A, Qian D, Chen H, Yang Y: **Structure-aware multimodal deep learning for drug-protein interaction prediction**. *J Chem Inf Model* 2022, **62**:1308–1317.

123. Nigam A, Pollice R, Krenn M, Gomes GDP, Aspuru-Guzik A: **Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (STONED) algorithm for molecules using SELFIES**. *Chem Sci* 2021, **12**:7079–7090.

124. Sorokina M, Steinbeck C: **Review on natural products databases: where to find data in 2020**. *J Cheminf* 2020, **12**:20.

125. Rutz A, Sorokina M, Galgonek J, Mietchen D, Willighagen E, Gaudry A, Graham JG, Stephan R, Page R, Vondrášek J, *et al.*: **The LOTUS initiative for open knowledge management in natural products research**. *Elife* 2022:11.

126. Artrith N, Butler KT, Coudert F-X, Han S, Isayev O, Jain A, Walsh A: **Best practices in machine learning for chemistry**. *Nat Chem* 2021, **13**:505–508.

127. UniProt Consortium: **UniProt: the universal protein knowledgebase in 2021**. *Nucleic Acids Res* 2021, **49**:D480–D489.

128. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chen L, Crichlow GV, Christie CH, Dalenberg K, Di Costanzo L, Duarte JM, *et al.*: **RCSB Protein Data Bank: powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences**. *Nucleic Acids Res* 2021, **49**:D437–D451.