Steinbeck *Journal of Cheminformatics* (2025) 17:44
https://doi.org/10.1186/s13321-025-00990-w

Journal of Cheminformatics

COMMENT Open Access

The evolution of open science in cheminformatics: a journey from closed systems to collaborative innovation

Christoph Steinbeck[1*]

Abstract

Cheminformatics has significantly transformed over the past four decades, evolving from a field dominated by proprietary systems to one increasingly embracing open science principles. In its early years, cheminformatics was characterised by commercial software and restricted data access, limiting collaboration and reproducibility. The advent of open-source software in the late 1990s and early 2000s, including tools such as the Chemistry Development Kit (CDK) and RDKit, played a crucial role in democratising computational chemistry. Open data initiatives, such as PubChem and NMRShiftDB, further enhanced accessibility by providing freely available chemical information, fostering transparency and interoperability and introducing key standards, such as the International Chemical Identifier (InChI), revolutionised data integration and retrieval across diverse platforms. Community-driven efforts, including the Blue Obelisk movement and Open Notebook Science, have promoted open methodologies and collaborative research. More recently, national data infrastructure projects like NFDI4Chem have aimed to standardise research data management in cheminformatics, ensuring the long-term sustainability of open science practices. The increasing adoption of the FAIR (Findable, Accessible, Interoperable, Reusable) principles has further reinforced data sharing and reuse in computational chemistry. Challenges remain, particularly in overcoming resistance to data sharing and ensuring sustainable funding for open projects. However, the trajectory of cheminformatics demonstrates that embracing openness enhances scientific integrity and accelerates discovery and innovation.

Introduction

Cheminformatics has experienced a remarkable transformation over the last four decades. This evolution mirrors broader changes in the scientific landscape, where open science has emerged as a cornerstone of innovation and collaboration. From its relatively insular beginnings in the 1980s to its current role as an early adopter of open data, open source software, and open collaboration, cheminformatics has undergone a journey shaped by technological advancements, shifting paradigms, and the collective push for transparency in science.

The closed beginnings: 1980s and early 1990s

Cheminformatics in the 1980s was primarily a domain of proprietary systems. Computational tools for tasks such as molecular modelling, structure search, and quantitative structure–activity relationship (QSAR) studies were developed by commercial entities. Companies like Tripos (with the SYBYL platform) and MDL Information Systems (creators of the widely used SD file format) dominated the landscape. These revolutionary tools came with significant limitations: high costs, restricted accessibility, little interoperability, and hard to extend and learn from. While companies were willing to pay these costs with the prospect of generating higher revenue, academic

*Correspondence:
Christoph Steinbeck
christoph.steinbeck@uni-jena.de
[1] Institute for Inorganic and Analytical Chemistry, Friedrich-Schiller-University, Jena, Germany

groups would rarely be able to afford to use such commercial software. Even if, for example, through collaborative agreements, academics had access to commercial software at affordable or no cost, they were restricted in the ways they shared their results computed with that software.

The culture of this era was heavily influenced by intellectual property concerns, competitive secrecy, and ideas that the source code could not be maintained in an academic setting. Research outputs were often siloed within organisations, with limited exchange of data or methodologies. The notion of openly sharing software or datasets was largely absent, as the prevailing business model incentivised exclusivity in academia.

The emergence of open source: late 1990s and early 2000s

The late 1990s and early 2000s marked the beginning of a paradigm shift spurred by the broader open-source movement in software development. Inspired by the success of projects like Linux, developers in cheminformatics began to recognise the potential of collaborative software development. This period saw the birth of foundational open-source projects such as JChemPaint (1997) [1], Jmol (1997) [2], the Chemical Markup Language project (1998) [3] and the Jumbo software [4], the Chemistry Development Kit (CDK; 2000) [5], RDKit (2003) and Open Babel (2001) [6], which provided freely accessible tools for molecular manipulation and data conversion.

These tools democratised access to computational chemistry, enabling researchers without significant funding to perform high-quality cheminformatics work. Adopting permissive licenses facilitated widespread use and modification while fostering a culture of collaboration and mutual improvement. At the same time, the scientific community began to appreciate the benefits of transparency, reproducibility, and the ability to audit code—a significant departure from the black-box nature of proprietary software.

Open data and the rise of collaboration: 2000s to early 2010s

Alongside open-source software, the push for open data began to gain traction. The realisation that data—structures, reactions, properties, and spectra—are as valuable as the algorithms processing them drove efforts to make chemical information widely accessible—initiatives such as PubChem [7]. PubChem was launched in 2004 by the National Center for Biotechnology Information (NCBI) at the U.S. National Institutes of Health (NIH) as part of the NIH Molecular Libraries Program. Its initial goal was to provide an open repository of small molecule data, including chemical structures, biological activities, and associated metadata, to stimulate drug discovery research. Over time, PubChem grew into one of the largest open-access databases in the world, integrating information from multiple contributors ranging from government agencies to academic laboratories and commercial suppliers. By centralising and freely sharing this breadth of chemical data, PubChem played a significant role in democratising access to research-grade chemical information, thereby driving forward open science initiatives in cheminformatics and enhancing collaborative research across disciplines.

While open-access databases are the rule in the biosciences, they remain rare in chemistry until today. An open-access database that arose during the early 2000s is NMRShiftDB [8, 9], an open access of NMR chemical shifts assigned to organic chemical structures. It was published even earlier than PubChem and originated in the lab of one of the authors. Thanks to the great effort of its leading developer, it is still alive today and has just celebrated its 20th birthday [10].

A seminal development which fostered linking and identity matching of chemical structures across data sources was the International Chemical Identifier (InChI) [11], which was conceived in the early 2000s through a partnership between the International Union of Pure and Applied Chemistry (IUPAC) and the National Institute of Standards and Technology (NIST). It aimed to provide a uniform, non-proprietary textual identifier that could encode a chemical structure and enable unambiguous linking and retrieval of information across databases and software tools. As a standardised, machine-readable representation, the InChI has become a cornerstone of cheminformatics, playing a crucial role in database interoperability, facilitating the sharing and reuse of chemical data, and supporting sophisticated search strategies that accelerate scientific discovery.

This period also saw the birth of Open Notebook Science (ONS) [12], an ideal that advocated for complete transparency in research. The term was popularized by chemist Jean-Claude Bradley, whose efforts showcased the power of openly sharing experimental protocols and data in real-time. Although not universally adopted, ONS highlighted the possibilities of crowdsourcing problem-solving and maximising the impact of scientific discoveries.

A group of open-source, open data and open-standards enthusiasts in chemistry formed the Blue Obelisk movement to promote these ideas [13].

This was complemented by expressions from other members of the scientific community and, during these days, certainly in many other fields of science. An example is the Panton Principles, first publicised in 2010 by members of the Open Knowledge Foundation (including Jonathan Gray, Jordan Hatcher, Rufus Pollock, and

Peter Murray-Rust), which articulate guidelines for making scientific data fully open to promote transparency, reproducibility, and innovation. Within cheminformatics, these principles underscore the importance of placing chemical data in the public domain or under open licenses, allowing unrestricted reuse, redistribution, and modification. By advocating explicit legal frameworks and urging the removal of access barriers, the Panton Principles have encouraged broader participation in data curation, contributed to a culture of data sharing, and strengthened the foundations of open science in chemistry.

This period also featured a strong push for integration of different cheminformatics platforms. Platforms like Taverna [14], KNIME [15], and Bioclipse [16] integrated multiple solutions, making it more straightforward to collaborate on more complex cheminformatics workflows.

Despite promising developments in the late twentieth century, the lack of big open data in chemistry remains one of the biggest roadblocks for cheminformatics.

Here, national data management initiatives like the German National Research Data Initiative (NFDI) can play a pivotal role in advancing the ideals of open science and pave the ground for more scientific opportunities. Within the broader framework of the NFDI, domain-specific consortia, such as NFDI4Chem [17, 18], enhance the discoverability, integration, and sharing of research data in chemistry. These coordinated efforts unite researchers, data stewards, and infrastructure providers to establish common standards, best practices, and cutting-edge digital tools for managing chemical data. By harmonising data formats and creating robust platforms for data exchange, NFDI4Chem streamlines research workflows and accelerates innovation in key areas like materials science, drug discovery, and environmental chemistry. Ultimately, these national initiatives reinforce transparency, reproducibility, and collaboration by making high-quality datasets accessible to the scientific community and ensuring that research outputs remain interoperable over the long term.

The maturation of open science: early 2010s to today

The last decade has solidified open science as a core principle in cheminformatics. The interplay between open-source tools, open data, and cloud computing has revolutionised how researchers collaborate. Projects like RDKit have become indispensable in academia and industry, demonstrating the scalability and robustness of community-driven development.

Simultaneously, the FAIR data principles (Findable, Accessible, Interoperable, and Reusable) have gained widespread acceptance. These guidelines, championed by the European Open Science Cloud and other initiatives, have provided a framework for structuring and sharing chemical data. Databases such as ChEMBL [19], which integrates bioactivity data from medicinal chemistry literature, exemplify the FAIR ethos and its role in accelerating drug discovery. Adopting the FAIR principles was only one of the building blocks towards greater reproducibility in cheminformatics. Version-controlled computational Workflows, such as those introduced through workflow tools like KNIME [20] and Jupyter Notebooks [21], greatly enhanced reproducibility. Community-driven validation Efforts like the Open Chemistry Challenge allowed for cross-validation of datasets and methods. Pre-registration and Open Notebook Science fostered more transparency in cheminformatics research.

The adoption of Semantic Web technologies and Linked Open Data (LOD) principles in cheminformatics began gaining momentum in the early 2000s, inspired by Tim Berners-Lee's vision of a "web of data" [22]. Researchers could make chemical information more interoperable, machine-readable, and reusable across disparate databases and services by employing standardised ontologies (e.g., Web Ontology Language, OWL) and structured data formats (e.g., Resource Description Framework, RDF). Initiatives like Linked Open Drug Data (LODD) [23] and the representation of resources such as ChEMBL in Linked Data formats [24] showcased how open standards could unify chemical structures, properties, and bioactivity data while also connecting them to related domains in life sciences. This interconnected ecosystem facilitates discovery and collaboration, as chemical entities are seamlessly linked to other biological and biomedical resources, fueling open science through improved data transparency and cross-domain integration.

Moreover, cheminformatics has embraced machine learning and artificial intelligence, fields that thrive on large, open datasets. The success of predictive modelling efforts hinges on access to diverse, high-quality data, underscoring the importance of openness in the modern era.

Wikipedia [25], launched in 2001, quickly became a cornerstone of open knowledge by harnessing collaborative editing to create and maintain articles on various topics, including chemical information. Its volunteer-driven model, exemplified by WikiProject Chemistry (established in 2002), has contributed to curating chemical compound data, providing a freely accessible and continually updated resource that supports education, research, and public engagement. Building on Wikipedia's success, Wikidata [26] was introduced in 2012 as a centralised, structured knowledge base designed to store machine-readable data across numerous domains—chemistry included. By unifying chemical identifiers, properties, and related metadata, Wikidata reinforces open science principles in cheminformatics, enabling researchers, educators, and developers to automate data analysis and cross-reference chemical information more efficiently. Together, Wikipedia and Wikidata have profoundly influenced open science by democratising access to chemical knowledge and fostering collaborative data curation on a global scale.

Educational resources in open cheminformatics

As open science principles have gained traction, the accessibility of cheminformatics education has expanded significantly. Open educational initiatives have been crucial in democratising knowledge making cheminformatics more accessible to students, researchers, and professionals worldwide.

Massive Open Online Courses (MOOCs) have provided structured learning opportunities for cheminformatics enthusiasts. Platforms like Coursera and edX offer courses with cheminformatics content developed by qualified institutions and individuals. While not in all cases free and open, these courses allow learners worldwide to gain hands-on experience in cheminformatics methods, tools, and applications at affordable costs, often including interactive coding exercises and real-world case studies.

Beyond formal courses, community-driven tutorials and documentation have emerged as invaluable educational resources. The RDKit project, for instance, maintains an extensive collection of tutorials and example workflows that help beginners and advanced users apply cheminformatics techniques in their research. Similarly, the Chemistry Development Kit (CDK) and Open Babel communities actively provide documentation, workshops, and forums for knowledge sharing and troubleshooting.

Another crucial component of open cheminformatics education is the role of collaborative knowledge platforms like Wikipedia and Wikidata. These platforms provide freely accessible, continuously updated information on chemical data, cheminformatics algorithms, and software tools. Volunteers from the cheminformatics community contribute regularly, ensuring these resources remain relevant and reliable. Wikidata, in particular, has become an essential resource for structured chemical data, aiding researchers and educators alike.

Together, these educational initiatives exemplify the impact of open science in cheminformatics, lowering barriers to entry and fostering a global, interconnected learning community. As cheminformatics continues to evolve, these open educational resources' ongoing development and support will be essential in training the next generation of researchers and professionals.

Challenges and the path ahead

Despite these advances, challenges remain. Resistance to data sharing persists in academia, particularly and partly understandably in industry, where competitive and regulatory pressures are significant. Ensuring the long-term sustainability of open projects is another critical issue, as many rely on the voluntary efforts of contributors or precarious funding. Various aspects could help to ensure the long-term viability of Open Science:

Institutional support from universities, non-university research institutions and government agencies increasingly fund open projects. Examples include Germany's National Research Data Infrastructure for Chemistry, NFDI4Chem [18], and NIH's data-sharing policies. Community-driven projects like RDKit thrive through industry-academia partnerships and crowdfunding.

Public–private collaborations through pharmaceutical consortia, such as the Innovative Medicines Initiative (IMI), provide funding for open data initiatives.

Integrating AI and machine learning research requires open datasets and open-source toolkits, fueling AI-driven drug discovery and attracting further funding and interest.

The trajectory is clear. The community has demonstrated that openness aligns with scientific integrity and catalyses innovation. Integrating cheminformatics into broader initiatives such as the Human Cell Atlas or global pandemic response efforts further illustrates its indispensable role in collaborative science (Table 1).

Conclusion

The development of open science in cheminformatics represents a microcosm of the broader scientific revolution. From closed systems in the 1980s to today's interconnected, collaborative ecosystem, the journey reflects a collective effort to democratise knowledge and accelerate discovery. The milestones along the way—open-source tools, open data repositories, and the adoption of open science principles—serve as a testament to what is possible when the barriers of exclusivity are dismantled.

The shift toward open science has significantly impacted academia and the pharmaceutical industry. Increased access to open datasets has accelerated discovery and improved reproducibility in academia. For instance, ChEMBL's open bioactivity data has facilitated thousands of drug discovery studies.

In the pharmaceutical industry, open science has fostered pre-competitive collaboration. One example is the Open Targets initiative [27], a partnership between EMBL-EBI, GSK, and others, which leverages open data

Table 1 Key Milestones in Open Science for Cheminformatics

Year	Milestone	Description
1980s	Proprietary Software Dominance	Commercial software like SYBYL and MDL tools dominate the field with closed access
Late 1990s	Early Open-Source Projects	Introduction of JChemPaint, Jmol, Chemical Markup Language, and the Chemistry Development Kit (CDK)
The early 2000s	Open Data Emergence	PubChem, NMRShiftDB, and InChI revolutionise chemical data sharing
Mid-2000s	Blue Obelisk Movement	Advocacy for open data, open-source software, and open standards
2010s	FAIR Principles & National Data Infrastructures	NFDI4Chem and other initiatives focus on long-term data accessibility and reproducibility
2020s	AI & Machine Learning with Open Data	Increased integration of cheminformatics with AI, fueled by open databases and collaborative research

for target validation in drug discovery. This initiative has reduced redundant efforts, accelerated hypothesis generation, and improved international collaboration.

As cheminformatics continues to evolve, its commitment to openness still has a significant potential to evolve, not only for the benefit of the discipline itself but for the scientific community.

Disclosure

Large language models were used to assist the author in researching the timeline, generating ideas, and generating text suggestions on individual historical events, toolkits, and databases. The complete text of this article reflects the author's in-depth knowledge of the subject matter and methods of the work described here.

Acknowledgements
The author wishes to thank Dr. Egon Willighagen for his comments on the manuscript.

Author contribution
C.S. performed the research and wrote and reviewed the manuscript.

Funding
Not applicable.

Availability of data and materials
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The author declares no competing interests.

Received: 18 February 2025 Accepted: 19 March 2025
Published online: 03 April 2025

References

1. Krause S, Willighagen E, Steinbeck C (2000) JChemPaint—using the collaborative forces of the internet to develop a free editor for 2D chemical structures. Molecules 5:93–98
2. Hanson RM (2010) Jmol– a paradigm shift in crystallographic visualization. J Appl Crystallogr 43:1250–1260
3. Murray-Rust P, Rzepa HS (1999) Chemical markup, XML, and the Worldwide Web. 1. Basic principles. J Chem Inf Comput Sci 39:928–942
4. Murray-Rust P (1997) JUMBO: an object-based XML browser. World Wide Web j 2:197–206
5. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, Kuhn S, Pluskal T, Rojas-Chertó M, Spjuth O et al (2017) The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J Cheminform 9:33
6. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open babel: an open chemical toolbox. J Cheminform 3:33
7. Bolton EE, Wang Y, Thiessen PA, Bryant SH: PubChem: Integrated platform of small molecules and biological activities. In: Annual Reports in Computational Chemistry. Elsevier; 2008:217–241.
8. Steinbeck C, Krause S, Kuhn S (2003) NMRShiftDB - Constructing a free chemical information system with open-source components. J Chem Inf Comput Sci 43:1733–1739
9. Steinbeck C, Kuhn S (2004) NMRShiftDB—compound identification and structure elucidation support through a free community-built web database. Phytochemistry 65:2711–2717
10. Kuhn S, Kolshorn H, Steinbeck C, Schlörer N (2024) Twenty years of nmrshiftdb2: a case study of an open database for analytical chemistry. Magn Reson Chem 62:74–83
11. Heller S, McNaught A, Stein S, Tchekhovskoi D, Pletnev I (2013) InChI - the worldwide chemical structure identifier standard. J Cheminform 5:7
12. Bradley J-C (2007) Open notebook science using blogs and wikis. Nat Prec. https://doi.org/10.1038/npre.2007.39
13. O'Boyle NM, Guha R, Willighagen EL, Adams SE, Alvarsson J, Bradley J-C, Filippov IV, Hanson RM, Hanwell MD, Hutchison GR et al (2011) Open data, open source and open standards in chemistry: the blue obelisk five years on. J Cheminform 3:37
14. Oinn T, Greenwood M, Addis M, Alpdemir MN, Ferris J, Glover K, Goble C, Goderis A, Hull D, Marvin D et al (2006) Taverna: lessons in creating a workflow environment for the life sciences. Concurr Comput 18:1067–1100
15. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Sieb C, Thiel K, Wiswedel B (2008) KNIME: the konstanz information miner. In: Analysis D (ed) Machine learning and applications. Berlin Heidelberg, Springer, pp 319–326
16. Spjuth O, Helmus T, Willighagen EL, Kuhn S, Eklund M, Wagener J, Murray-Rust P, Steinbeck C, Wikberg JES (2007) Bioclipse: an open source workbench for chemo- and bioinformatics. BMC Bioinformatics 8:59
17. Steinbeck C, Koepler O, Bach F, Herres-Pawlis S, Jung N, Liermann J, Neumann S, Razum M, Baldauf C, Biedermann F et al (2020) NFDI4Chem-towards a national research data infrastructure for chemistry in Germany. Res Ideas Outcomes 6:e55852
18. Herres-Pawlis S, Koepler O, Steinbeck C (2019) NFDI4Chem: shaping a digital and cultural change in Chemistry. Angew Chem Int Ed Engl 58:10766–10768
19. Zdrazil B, Felix E, Hunter F, Manners EJ, Blackshaw J, Corbett S, de Veij M, Ioannidis H, Lopez DM, Mosquera JF et al (2024) The ChEMBL Database in

2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res 52:D1180–D1192
20. Fillbrunn A, Dietz C, Pfeuffer J, Rahn R, Landrum GA, Berthold MR (2017) KNIME for reproducible cross-domain analysis of life science data. J Biotechnol 261:149–156
21. Kluyver T, Ragan-Kelley B, Pérez F, Granger BE, Bussonnier M, Frederic J, Kelley K, Hamrick J, Grout J, Corlay S: Jupyter Notebooks-A publishing format for reproducible computational workflows. In: Positioning and Power in Academic Publishing: Players, Agents and Agendas. Edited by Loizides F, Schmidt B. IOS Press; 2016:87–90.
22. Berners-Lee T, Hendler J, Lassila O (2001) Web semantic. Sci Am 284:34–43
23. Samwald M, Jentzsch A, Bouton C, Kallesøe CS, Willighagen E, Hajagos J, Marshall MS, Prud'hommeaux E, Hassenzadeh O, Pichler E et al (2011) Linked open drug data for pharmaceutical research and development. J Cheminform 3:19
24. Willighagen EL, Waagmeester A, Spjuth O, Ansell P, Williams AJ, Tkachenko V, Hastings J, Chen B, Wild DJ (2013) The ChEMBL database as linked open data. J Cheminform 5:23
25. Reagle JM Jr: Good faith collaboration: The culture of Wikipedia. MIT Press; 2010. https://doi.org/10.7551/mitpress/8051.001.0001
26. Vrandečić D, Krötzsch M (2014) Wikidata. Commun ACM 57:78–85
27. Hulcoop DG, Trynka G, McDonagh EM (2025) Open targets: 10 years of partnership in target discovery. Nat Rev Drug Discov 24:153–154

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.