Journal of Cheminformatics

COMMENT

Open Access

Milestones in chemoinformatics: global view of the field

Jürgen Bajorath[1,2*]

Abstract

Over the past ~25 years, chemoinformatics has evolved as a scientific discipline, with a strong foundation in pharmaceutical research and scientific roots that can be traced back to the late 1950s. It covers a wide methodological spectrum and is perhaps best positioned in the greater context of chemical information science. Herein, the chemoinformatics discipline is delineated, characteristic (and partly problematic) features are discussed, and a global view of the field is provided, emphasizing key developments.

Keywords Chemoinformatics, Drug discovery, Information science, Chemical data, Molecular design

Introduction

Discussing chemoinformatics (in the US mostly referred to as cheminformatics) as a scientific discipline immediately raises a question: what is it? Or, in other words, how should one best define it? The term itself first appeared in the literature in 1998, when the late Frank K. Brown introduced chemoinformatics as follows: "*The use of information technology and management has become a critical part of the drug discovery process. Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and organization*" [1]. This original definition reflects a strong foundation of chemoinformatics in drug discovery and also emphasizes a close link between chemoinformatics and information science. Notably, the related term chemical informatics (which might be a little dubious semantically: can informatics be "chemical"?) was also used at that time and more generally understood as the application of information technology to chemistry. In a similar vein, chemoinformatics has been designated as the "*application of informatics methods to solve chemical problems*" [2] or the "*manipulation of information about chemical structures*" [3]. Hence, going beyond drug discovery, chemoinformatics might be positioned in the broader context of chemical information science, covering all computational methods for the representation and analysis of chemical structures and data and retrieval of chemical information from any source [4, 5]. Importantly, in chemoinformatics, information associated with chemical structures mostly relates to molecular properties, in particular, biological activities [3, 6]. From the early days of chemoinformatics on, molecular similarity analysis has been a central theme in the field [7, 8], and similarity of compounds was for the most part quantified as an indicator of similar properties [7, 9]. In addition, molecular diversity analysis, based on the assessment of distance relationships in chemical reference spaces, substantially impacted combinatorial chemistry and compound library design [10]. Taking these aspects into consideration, the chemoinformatics spectrum should also cover methods for deriving and navigating chemical space, the prediction of biological activity (and other molecular properties) from chemical structure, and compound design [6,

*Correspondence:
Jürgen Bajorath
bajorath@bit.uni-bonn.de
[1] Department of Life Science Informatics and Data Science, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, University of Bonn, Friedrich-Hirzebruch-Allee 5/6, 53115 Bonn, Germany
[2] Lamarr Institute for Machine Learning and Artificial Intelligence, University of Bonn, Friedrich-Hirzebruch-Allee 5/6, 53115 Bonn, Germany

11], rendering the boundaries between chemoinformatics, molecular modeling, and drug design rather fluid.

In drug discovery, in addition to data analysis and warehousing, chemoinformatics covers all computational approaches contributing to the efficiency and quality of hit identification and early-phase compound development such as virtual screening, data mining via machine learning, molecular design, and quantitative structure-activity relationship (QSAR) methods. Furthermore, more specialized approaches such as computational polypharmacology or the prediction of adverse drug side effects also fall into the chemoinformatics spectrum.

Furthermore, computational synthesis planning was from the beginning a part of chemoinformatics [12]. Accordingly, scientific origins of chemoinformatics can be traced back to the late 1950s and early 1960, long before the discipline was formally established, when substructure-based similarity, QSAR, and synthesis design methods were beginning to be introduced [3, 9, 12].

Discussion

Two characteristic features of the chemoinformatics field are its conceptual link to chemical information science and wide methodological sprectrum, as described above. A third one is its data-oriented nature, essentially paralleling the development and growth of bioinformatics. Unprecedented volumes of compound structures and data became first available in the heydays of combinatorial chemistry during the 1990s. The ensuing need to manage large amounts of chemical data, quickly exceeding the capacity of traditional data infrastructures in drug discovery environments, was one of the driving forces behind the inception and development of chemoinformatics as a discipline [1, 11]. The early data deluge in the 1990s was followed by exponential growth in publicly available compounds and activity data beginning in the mid 2000s [13] and, more recently, the wealth of data originating from artificial intelligence (AI)-driven approaches such as generative modeling [14], advanced synthesis design [15], or screening of ultra-large virtual libraries [16]. The generation of computational workflows and infrastructures for data processing and design of databases have always been an integral part of chemoinformatics, as much as the analysis of chemical space and the search for new active compounds using a variety of methods including machine learning. Notably, machine learning has been a key component of this evolving field since the early 1990s.

The foundation of chemoinformatics in drug discovery, as reflected by Frank Brown's original definition, largely resulted from the need of the pharmaceutical industry to conduct drug discovery research in the presence of strong data growth, build more extensive database structures, and increase the efficiency of compound and data processing, especially with the advent of high-throughput technologies [11]. Therefore, much of the groundbreaking work in chemoinformatics including method development was carried out in drug discovery settings, often with limited publication opportunities, due to the dominant proprietary nature of pharmaceutical research. This had consequences for the development of chemoinformatics as a scientific discipline. Academic institutions and major funding agencies perceived chemoinformatics mostly as an industrial affair, leading to reluctance of (traditionally conservative) chemistry departments to establish new faculty positions and integrate chemoinformatics into their curricula. Equally important, the early pharma-centric view of chemoinformations also led to the absence of stable public funding sources for this evolving field, expecting pharma companies to pay the price for what they were for the most part interested in. While collaborations between industry and academia are a part of the chemoinformatics culture, reluctance of accepting chemoinformatics as an evolving discipline in academic settings and the absence of stable funding sources have made it difficult to this date for young investigators to pursue an academic career in this field and have also limited opportunities for chemoinformatics education. While the academic chemoinformatics community expanded over time, partly through investigators moving from drug discovery to academia, compared to bioinformatics, it remained small, with limited impact on the development of new academic initiatives and structures, especially in chemistry. Over time, chemoinformatics has entered new territories such as material and nano science or industrial process control, but this has not substantially changed its image as a niche discipline (albeit a scientifically exciting one!), which must be taken into consideration when judging the development of this field.

Any of the scientific areas comprising the chemoinformatics spectrum has its milestone events, as discussed in separate contributions composing this special issue. Moreover, viewing the chemoinformatics field globally, there are a number of developments and initiatives that have been -and continue to be- critically important for its development. While pharma companies have their own and mostly proprietary data, academic research in chemoinformatics and published methodological advances largely depend on the availability of data in the public domain. Accordingly, major public repositories of compounds and activity data such as ChEMBL [17], BindingDB [18], PubChem [19], or ZINC [20], complemented by large protein information resources such as UniProt [21], have been critically important for advancing chemoinformatics. Furthermore, open science

initiatives involving collaborative efforts between industry and academia are crucial for further increasing the impact of chemoinformatics. Noteworthy examples include the pharma-driven generation of joint public compound data sets for protein kinase research [22, 23], the European Research EU-OPENSCREEN platform for collaborative biological screening and hit identification [24], or Boehringer Ingelheim's open innovation portal *opnMe.com* that makes extensively characterized tool compounds available for academic research [25]. Equally important to databases and open science initiatives are open source collections of software tools and libraries such as RDKit [26], Scikit-learn [27], PyTorch [28], TensorFlow [29], or Keras [30], including major contributions from companies such as Google [29, 30].

For the evolving chemoinformatics field, establishing a publication culture was another essential requirement. The *Journal of Chemical Information and Computer Sciences* (JCICS), which succeeded the *Journal of Chemical Documentation* in 1975, became for long the core publication venue, with a strong focus on chemical information. In 2005, it was transformed into the *Journal of Chemical Theory and Computation* and the *Journal of Chemical Information and Modeling* that substantially widened its scope compared to JCICS. Other core journals for chemoinformatics include *Molecular Informatics* (succeeding *QSAR & Combinatorial Science* in 2010) and the *Journal of Cheminformatics* that was launched in 2009, became an open access journal in 2015, and arguably has the strongest impact on the field at present.

Education in chemoinformatics represents a critical issue, as mentioned above. Trained chemoinformaticians are in high demand by the pharmaceutical and chemical industries, but educational opportunities are limited. Beginning in the early 2000s, master programs in chemoinformatics were initiated [31, 32], the first one at the University of Sheffield (Prof. Peter Willett), another one at the University of Strasbourg, and one in the US at Indiana University, accompanied by programs at smaller schools, also for undergraduate education [31, 32]. However, most of the bachelor and master programs in chemoinformatics have not been sustainable, with the exception of the program at the University of Strasbourg (Prof. Alexandre Varnek), which has further expanded over the years. While chemoinformatics courses are offered in a few chemistry or chemical engineering programs, or as a specialization in bioinformatics curricula, most of the scientific training is carried out during PhD studies in a limited number of academic centers worldwide. Furthermore, interfacing education and research, schools in chemoinformatics established in Obernai/Strasbourg (Prof. Alexandre Varnek) [33] and Tokyo/Nara (Prof. Kimito Funatsu) [34], bringing together students and investigators from academia and industry, have been cornerstones for the further development of the chemoinformatics field and recently been complemented by an online school covering Latin America (Prof. José L. Medina-Franco) [35].

Conclusion and outlook

Chemoinformatics has evolved as a scientific discipline at interfaces between drug discovery, chemistry, computer science, and information technology. A hallmark of this field is its wide methodological spectrum. A global view of chemoinformatics reveals a number of developments that have been of critical importance for shaping this field and sustaining its development. Although the methodological foundations of and challenges for chemo- and bioinformatics are very similar, despite the diversity of applications, chemoinformatics has essentially remained to be a niche discipline, much smaller than bioinformatics on a global scale. This is largely a consequence of its roots in pharma environments and the ensuing reluctance of chemistry departments to integrate chemoinformatics into their traditional curricula, in contrast to theoretical chemistry. However, the times are changing. In particular, in the AI era, it is no longer conceivable how next generations of chemists (or other scientists) might be able to function without at least basic skills in informatics and data science. Hence, while research in chemoinformatics continues to be as exciting as it has been over the years, if not more so, there are good reasons to anticipate that the field and its academic presence will further expand in the near future and increase its impact on chemistry and beyond.

Abbreviations

AI	Artificial Intelligence
JCICS	Journal of Chemical Information and Computer Sciences
QSAR	Quantitative Structure-Activity Relationship

Author contributions

J.B. wrote the manuscript.

Funding

There are no external funders of this work.

Availability of data and materials

No datasets were generated or analysed during the current study.

Declarations

Competing interests

The author declare no competing interests.

Received: 11 October 2024 Accepted: 28 October 2024
Published online: 05 November 2024

References

1. Brown FK (1998) Chemoinformatics: what is it and how does it impact drug discovery? Ann Rep Med Chem 33:375–384

2. Gasteiger J (2006) The central role of chemoinformatics. Chemometr Intell Lab Syst 82:200–209. https://doi.org/10.1016/j.chemolab.2005.06.022
3. Willett P (2011) Chemoinformatics: a history. Wiley Interdiscip Rev Comput Mol Sci 1:46–56. https://doi.org/10.1002/wcms.1
4. Willett P (2008) From chemical documentation to chemoinformatics: 50 years of chemical information science. J Inf Sci 34:477–499. https://doi.org/10.1177/0165551507084631
5. Bajorath J (2015) Entering new publication territory in chemoinformatics and chemical information science. F1000Res. https://doi.org/10.12688/f1000research.6101.1
6. Bajorath J (2004) Understanding chemoinformatics: a unifying approach. Drug Discov Today 9:13–14. https://doi.org/10.1016/s1359-6446(04)02916-2
7. Johnson M, Maggiora GM (eds) (1990) Concepts and applications of molecular similarity. Wiley, New York
8. Willett P (2009) Similarity methods in chemoinformatics. Ann Rev Inf Sci Technol 43:3–71
9. Eckert H, Bajorath J (2007) Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. Drug Discov Today 12:225–233. https://doi.org/10.1016/j.drudis.2007.01.011
10. Van Drie JH, Lajiness MS (1998) Approaches to virtual library design. Drug Discov Today 3:274–283. https://doi.org/10.1016/S1359-6446(98)01186-6
11. Hann M, Green R (1999) Chemoinformatics—a new name for an old problem? Curr Opin Chem Biol 3:379–383. https://doi.org/10.1016/S1367-5931(99)80057-X
12. Ihlenfeldt WD, Gasteiger J (1996) Computer-assisted planning of organic syntheses: the second generation of programs. Angew Chem Int Ed 34:2613–2633. https://doi.org/10.1002/anie.199526131
13. Hu Y, Bajorath J (2017) Entering the 'big data'era in medicinal chemistry: molecular promiscuity analysis revisited. Future Sci OA 3:FSO179. https://doi.org/10.4155/fsoa-2017-0001
14. Zeng X, Wang F, Luo Y et al (2022) Deep generative molecular design reshapes drug discovery. Cell Rep Med 3:100794. https://doi.org/10.1016/j.xcrm.2022.100794
15. Johansson S, Thakkar A, Kogej T et al (2019) AI-assisted synthesis prediction. Drug Discov Today Technol 32:65–72. https://doi.org/10.1016/j.ddtec.2020.06.002
16. Crunkhorn S (2022) Screening ultra-large virtual libraries. Nat Rev Drug Discov 21:10–1038. https://doi.org/10.1038/d41573-022-00002-8
17. Gaulton A, Bellis LJ, Bento AP et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107. https://doi.org/10.1093/nar/gkr777
18. Liu T, Lin Y, Wen X, Jorissen RN, Gilson MK (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic Acids Res 35:D198–D201. https://doi.org/10.1093/nar/gkl999
19. Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Bryant SH (2009) PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 37:W623–W633. https://doi.org/10.1093/nar/gkp456
20. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. J Chem Inf Model 45:177–182. https://doi.org/10.1021/ci049714+
21. UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47:D506–D515. https://doi.org/10.1093/nar/gky1049
22. Drewry DH, Wells CI, Andrews DM et al (2017) Progress towards a public chemogenomic set for protein kinases and a call for contributions. PLoS ONE 12:e0181585. https://doi.org/10.1371/journal.pone.0181585
23. Wells CI, Al-Ali H, Andrews DM et al (2021) The kinase chemogenomic set (KCGS): an open science resource for kinase vulnerability identification. Int J Mol Sci 22:566. https://doi.org/10.3390/ijms22020566
24. Brennecke P, Rasina D, Aubi O et al (2019) EU-OPENSCREEN: a novel collaborative approach to facilitate chemical biology. SLAS Discov 24:398–413. https://doi.org/10.1177/2472555218816
25. Gollner A, Köster M, Nicklin P et al (2022) opnMe.com: a digital initiative for sharing tools with the biomedical research community. Nat Rev Drug Discov 21:475–476. https://doi.org/10.1038/d41573-022-00071-9
26. RDKit: cheminformatics and machine learning software; 2021. http://www.rdkit.org/
27. Pedregosa F, Varoquaux G, Gramfort A et al (2011) Scikit-learn: machine learning in Python. J Mach Learn Res 12:2825–2830
28. Paszke A, Gross S, Massa F et al (2019) PyTorch: an imperative style, high-performance deep learning library. Adv Neural Inf Proc Syst 32:8026–8037
29. Abadi M, Agarwal A, Barham P et al. TensorFlow: large-scale machine learning on heterogeneous systems; 2015. http://tensorflow.org
30. Keras developerguide; 2020. https://keras.io/guides/
31. Warr W Cheminformatics education (online). https://www.warr.com/CheminformaticsEducationfinal.pdf
32. Wild DJ, Wiggins GD (2006) Challenges for chemoinformatics education in drug discovery. Drug Discov Today 11:436–439. https://doi.org/10.1016/j.drudis.2006.03.010
33. 9th Strasbourg Summer School in Chemoinformatics; 2024. https://infochim.chimie.unistra.fr/-Strasbourg-Summer-School-in-Chemoinformatics-2024-.html
34. 8th Autumn School of Chemoinformatics in Nara; 2023. http://www-dsc.naist.jp/dsc_naist/en/autumn_school2023/
35. Gonzalez-Ponce K, Horta Andrade C, Hunter F, Kirchmair J, Martinez-Mayorga K, Medina-Franco JL, Rarey M, Tropsha A, Varnek A, Zdrazil B (2023) School of cheminformatics in Latin America. J Cheminform 15:82. https://doi.org/10.1186/s13321-023-00758-0

Publisher's note