Indo. J. Chem. Sci.13 (3) 2024

Indonesian Journal of Chemical Science

http://journal.unnes.ac.id/sju/index.php/ijcs

Bibliometric Analysis of Chemistry Publication Trends on Scopus in The Era of Big Data

Andi Budi Bakti ✉

Chemistry Department, Faculty of Mathematics and Natural Sciences, Bandung Institute of Technology, Ganesha street No. 10, Lb. Siliwangi, Coblong, Bandung, West Java, 40132, Indonesia

Article Info

Accepted : 24-09-2024
Approved : 24-10-2024
Published : 25-11-2024

Keywords:
Bibliometric
Big Data
Chemistry
Scopus
VOSviewer

Abstract

In the 21st century, chemistry has seen a major shift in data management practices. Analyzing research trends is crucial for maintaining relevance and innovation. This study aims to: (1) identify the contributions of researchers and country, and assess their impact; (2) map the network of interactions among publications, references, topics, researchers, and institutions; and (3) predict future directions and recommend research opportunities in chemistry. This study employs bibliometric analysis and science mapping techniques with VOSviewer. The study finds that: (1) Jacqueline. M. Cole had made important contribution, pioneering the application of data-mining to discover material and autometic text-mining to build databases. Her research has opened new perspective and inspired other researcher. The United States is a leader for research in big data, so many researchers emerged in this field. Because having the most data centers and the best university for studying data science and AI; (2) big data research in chemistry is not yet deeply and continuously explored, with researchers often working in small groups and focusing on topics such as "machine learning," "chemometrics," "cheminformatics," and "deep learning"; and (3) "data mining" is underexplored, presenting new research opportunities and directions for chemistry in the era of big data.

✉ Correspondence address:
Ganesha street No. 10, Lb. Siliwangi, Coblong, Bandung, 40132
E-mail: andibudi055@gmail.com

p-ISSN 2252-6951
e-ISSN 2502-6844

Introduction

Nowadays, advancements in information technology (IT) have led to the rapid generation of vast amounts of data. The evolution of mobile devices, digital sensors, virtual communication, as well as computing and electronic storage, has been remarkably swift. This large volume of data, which exhibits the common characteristics of the 4Vs—Volume, Variety, Velocity, and Veracity—is referred to as "Big Data." The term Volume refers to the size of the data, Velocity denotes the speed at which data is created and processed, Variety relates to the sources and types of data (Lancy, 2001), and Veracity addresses the reliability of the data (Gantz & Reinsel, 2011). As of 2020, the total amount of data created, captured, copied, and consumed globally reached 64.2 zettabytes, and it is projected to reach 180 zettabytes by 2025 (Taylor, 2023). A significant portion of this data is generated primarily from the Internet of Things (IoT), multimedia, and social media (Yaqoob et al., 2016). With the global data volume increasing by 40% annually, it presents a considerable challenge for researchers and practitioners to address this evolving issue.

In chemistry, the rise of big data is closely linked to advancements in sophisticated instruments or sensors, and simultaneous experimental methods. In biochemistry, among the various analytical instruments used for metabolomics—the study of metabolite collections within biological systems—mass spectrometers (MS) are the most frequently employed. Due to their high sensitivity and capability to analyze numerous samples simultaneously, MS generates substantial amounts of data. For instance, in metabolomics, LC-MS can produce over 10 GB of data for each analysis of 30 samples (Guo et al., 2022). Furthermore, a significant amount of real-time data is also generated by IoT sensors, where machines or objects connected to the internet continuously stream data to local servers or cloud storage (Purnama & Sejati, 2023). In addition to sophisticated instrumentation, the miniaturization of testing and laboratory automation facilitates chemical testing within biological systems through High-Throughput Screening (HTS). HTS is characterized by the ability to test a range of 104 to 105 data points per day (Mayr & Bojanic, 2009).

In addition to experimental sources, the volume of big data in chemistry is also increasing due to advancements in computational simulations and databases containing a wide range of chemical literature and information. Widely recognized computational simulations such as molecular dynamics (MD) provide information on the positions and velocities of all particles at each time step, which is used to explore both microscopic and macroscopic properties of model systems. The data generated from these simulations are vast and complex, as the data is continuously collected at each time step (Yeguas & Casado, 2014). Chemical information, derived from both experimental and computational results, is aggregated into databases. Publicly accessible databases such as PubChem and ChEMBL, as well as commercial databases like SciFinder, contain extensive amounts of data extracted from tens of thousands of articles. Similarly, in industry, repositories collect substantial quantities of data. For instance, AstraZeneca International's bioscience information system contains over 150 million data points derived from experiments conducted before 2008 (Tetko et al., 2016).

In the 21st century, chemistry, like other sciences, has undergone a significant shift in the ways data are collected, stored, and utilized. The collection of vast amounts of data has become commonplace in supporting discoveries in chemistry through computer algorithms. Indeed, many experiments now generate so much raw data that a scientist could not feasibly review it all manually within a single lifetime. In this era of data, there has been a transformation in chemical discoveries from a trial-and-error approach to a data-driven approach supported by machines through artificial intelligence (AI) (Duke et al., 2024). In the 19th century, chemists recognized the importance of collecting chemical data, leading to the creation of catalogs such as the Beilstein Handbook of Organic Chemistry (Luckenbach, 1981), the Gmelin Handbook of Inorganic Chemistry (Mague, 1984), and the standardized IUPAC Color Book (Hartshorn, 2017). With the advent of computers in the 20th century, chemists began to compile chemical information in electronic formats and employed search techniques through the Chemical Abstracts Service (CAS) (Morgan, 1965).

Data literacy is now considered a fundamental skill due to the increasing prevalence of data interactions in daily life. Individuals frequently make decisions based on data and manage their personal information. Consequently, there is a renewed push to introduce data literacy in schools, aiming to enhance data literacy across society through education (Wolff et al., 2016). In the field of chemistry, the term "Data-driven Chemistry" has emerged, though it lacks a clear definition in various books and publications. However, data-driven approaches have been applied in organic chemistry through methods such as Linear Free Energy Relationships (LFER), chemometrics, cheminformatics, and machine learning (Williams et al., 2021). At the University of Edinburgh, for instance, data-driven chemistry is offered as a course in the undergraduate chemistry program, which includes the introduction to programming languages such as Python and their applications in chemistry, included topics such as data classification, statistical analysis, 3D visualization, and curve fitting.

The shift in various aspects of chemistry due to big data requires a better understanding of current trends in chemical research. By gaining a deep insight into these trends, researchers can adapt to rapid changes and ensure that their work remains relevant and innovative in the era of big data. Bibliometric analysis is a systematic study of scientific literature designed to identify patterns, trends, and impacts within a particular field (Passas, 2024). The advent of scientific databases such as Scopus and Web of Science has facilitated the acquisition of large-scale bibliometric data, supported by bibliometric software tools like Gephi, Leximancer, and VOSviewer. Bibliometric analysis techniques are divided into two categories: (1) performance analysis and (2) science mapping. Essentially, performance analysis records the contributions of research constituents, while science mapping focuses on the relationships among these constituents. Results from science mapping techniques can be enhanced in the form of network metrics, clustering, or visualization (Donthu et al., 2021).

Based on the above discussion, bibliometric analysis plays a crucial role. Through this analysis, chemists can map the development of chemical research overtime, identifying patterns of changes in research focus and methodologies during spesific periods. From the identified patterns, predicted trends in chemical research for the future. Additionally, by reviewing previous findings, identified research gaps such as underexplored areas or unsolved problems, as well as evaluated the impact of research on both theoritical and practical aspects. If bibliometric analysis is not conducted, research may lose relevance and innovation, resulting in outcomes unreflected current needs. This can lead to repeat the same studies, overlook important gaps and reduce the impact of research. Thus, this study performs a bibliometric analysis of trends in chemical research in the era of big data using science mapping techniques and the VOSviewer visualization software. The objectives of this study are: (1) to identify the contributions of researchers and affiliations and assess the impact of their research based on publications and citations, (2) to map the network of interactions among publications, references, topics, researchers, and affiliations through clustering and visualization, and (3) to predict future research directions in chemistry and provide recommendations for research opportunities based on identified patterns.

Method

This study employs bibliometric analysis methods. Bibliometric analysis involves four stages: (1) defining the objectives and scope of the bibliometric study, (2) selecting techniques for bibliometric analysis, (3) collecting data for bibliometric analysis, and (4) performing the bibliometric analysis and reporting the findings (Donthu et al., 2021). The scope of this research encompasses all aspects of big data in chemistry, including the collection, management, storage, analysis, and interpretation of chemical big data from both experimental and computational results. The bibliometric analysis technique employed is science mapping, which includes citation analysis (relationships between publications), co-citation analysis (relationships between references), bibliographic coupling (relationships between citing publications), co-word analysis (relationships between topics), and co-author analysis (relationships between authors).

In this study, the researchers utilized bibliometric data sourced from the Scopus database (www.scopus.com). Scopus is one of the largest databases providing a collection of reputable publications. Data collection was conducted on August 24, 2024, through a search for documents containing the title, abstract, and keywords "Big Data" AND Chemistry. The search criteria were restricted based on subject area: Chemistry; document type: Article; publication stage: Final; source type: Journal; and language: English. The search results can be processed directly using Microsoft Excel software to analyze statistics such as the number of publications per year, publications by each author, publications by each institution, publications by each country, publications by each sponsor, citations per author, citations per publication, and citations per year. Additionally, the search results can be exported as a .csv file, which can be analyzed using VOSviewer. The information required for each mapping analysis is detailed in Table 1, so it is important to ensure these requirements are met during the export process.

Table 1. Data Requirements and Units of Analysis for Science Mapping Techniques (Donthu et al., 2021)

Analysis Technique	Units of Analysis	Data Requirements
Citation Analysis	Document	Author Name, Citation, Title, Journal, DOI, Reference
Co-citation Analysis	Document	Reference
Bibliographic Coupling	Document	Author Name, Title, Journal, DOI, Reference
Co-word Analysis	Word	Title, Abstract, Author's Keyword, Indexed Keyword, Full Text
Co-author Analysis	Author, Affiliation, Country	Author, Affiliation (Institution and Country)

Results and Discussion

The search results reveal that only 145 articles meet the criteria. These articles, published between 2003 and 2024, involve 703 authors and have a total of 5,945 citations. According to number of citations, Table 2 shows list ten articles with the most citations. In 2015 stands out as the most frequently cited year, with 1,344 citations from 10 articles. This indicates that articles published in 2015, particularly about the management of computational big data through AI to understand theoretical chemistry, provided a strong foundation and attracted wide attention for many subsequent studies. In 2021, articles focusing on the management of big data through AI for drug discovery were cited by many articles in a relatively short time. This shows that this area of research is highly relevant and in demand, as the emergence of new diseases or unresolved old diseases necessitates effective drug discovery. One method for evaluating the quality of publications is by counting how often they are cited by other researchers. A highly-cited work indicates that it is frequently referenced in discussions among researchers. In other words, publications with numerous citations suggest that they have significant relevance and contribution to the advancement of knowledge in their field.

Table 2. List 10 articles with the most citations

No	Title	Year	Citation	Referensi
1	Big data meets quantum chemistry approximations: The Δ-machine learning approach	2015	603	(Ramakrishnan et al., 2015)
2	Artificial intelligence to deep learning: machine intelligence approach for drug discovery	2021	466	(Gupta et al., 2021)
3	Managing the computational chemistry big data problem: The ioChem-BD platform	2015	419	(Álvarez-Moreno et al., 2015)
4	Machine learning molecular dynamics for the simulation of infrared spectra	2017	369	(Gastegger et al., 2017)
5	Flexible and durable wood-based triboelectric nanogenerators for self-powered sensing in athletic big data analytics	2019	366	(Luo et al., 2019)
6	ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature	2016	322	(Swain & Cole, 2016)
7	The Evolution of Chemical High-Throughput Experimentation to Address Challenging Problems in Pharmaceutical Synthesis	2017	211	(Krska et al., 2017)
8	Visualization of very large high-dimensional data sets as minimum spanning trees	2020	166	(Probst & Reymond, 2020)
9	Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery	2015	163	(Pyzer-Knapp et al., 2015)
10	Machine learning based toxicity prediction: From chemical structural description to transcriptome analysis	2018	137	(Y. Wu & Wang, 2018)

The most-cited author is Anatole von Lilienfeld from the Vector Institute and at University of Toronto, Canada, with 661 citations across two publications published in 2015 and 2021. But the researcher with the most publications related to big data in chemistry is Jacqueline M. Cole (2016, 2020, 2024) from the University of Cambridge. Another prominent researcher is Hao Zhu (2020, 2021, 2023) from Rowan University. In other word, work's Anatole provide a strong foundation to be cited, as focused on the theoretical field like a physics-based understanding of chemical compound space using machine learning, quantum and statistical mechanics, and high-performance computing. Meanwhile research's Cole and Zhu are concerned the specific applied context. Cole is a pioneer in the application of data-mining to discover material and automatic text-mining to build databases. Zhu is an expert in the use of public big data and molecular structure information to predict the chemical efficacy and toxicity. Interestingly, neighboring countries such as Singapore also have researchers engaged in this topic. Markus Kraft from Nanyang Technological University, Singapore, is among the researchers contributing to this field. Although Kraft only a co-author, his collaboration with UK researchers on the validation of thermodynamic big data could open up opportunities for similar researcher to develop in Singapore.

Table 3. List 10 authors with the most citations

No	Author	Year	Citation	Ref.
1	O. Anatole Von Lilienfeld	2015, 2021	661	(Lemm et al., 2021; Ramakrishnan et al., 2015)
2	Pavlo. O Dral	2015, 2019	612	(Ramakrishnan et al., 2015; X. Wu et al., 2019)
3	Raghunathan Ramakrishnan	2015, 2021	609	(Ramakrishnan et al., 2015; Senthil et al., 2021)
4	Jacqueline M. Cole	2016, 2020, 2024	391	(Cole, 2020; Jung et al., 2024; Swain & Cole, 2016)
5	Jean-Louis Reymond	2016, 2020	255	(Probst & Reymond, 2020; Tetko et al., 2016)
6	Paola Gramatica	2010, 2011	156	(Bhhatarai & Gramatica, 2011; Li & Gramatica, 2010)
7	Hao Zhu	2020, 2021, 2023	115	(Chung et al., 2023; Jia et al., 2021; Yan et al., 2020)
8	Xian Liu	2014, 2021	74	(Liu et al., 2014; Wang et al., 2021)
9	Sonia Arrasate	2018, 2020	47	(Bediaga et al., 2018; Santana et al., 2020)
10	Humbert Gonzalez-diaz	2018, 2020	47	(Bediaga et al., 2018; Santana et al., 2020)

Research on big data in the field of chemistry began in 2003, with only one article published according to Scopus. There were no publications from 2004 to 2009. However, the field experienced a peak in 2022, with 20 publications emerging that year. Figure 1 displays the number of publications from 2003 to 2024. From 2018 to 2022, many studies focused on the development of data analysis methods and tools, involving the integration of data and AI. This indicates a strong push for innovation and extensive exploration to adopt new technologies, resulting in numerous experimental and discovery-oriented researchs. Subsequently, from 2023 to 2024, there was a decline due to a shift in research focus and approaches toward the concrete application of previously theoritical findings, such as SARS-CoV detection. The complexity of applied research, requiring adequate facilities, technological and funding readiness, plays a significant role in trend applied studies.

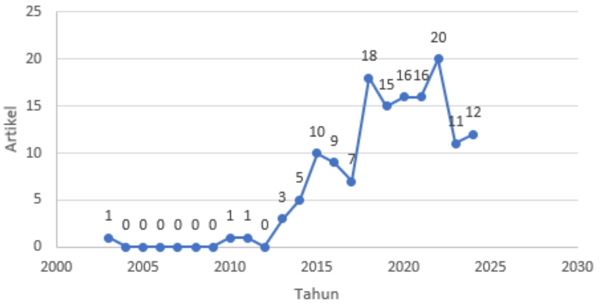

Figure 1. Number of Articles Per Year from 2003 to 2024

The country with the highest number of big data research publications in chemistry is the United States, with 38 articles, followed by China with 32 articles. Figure 2 illustrates the top 10 countries that have facilitated the most research on big data in chemistry. Both countries are developed nations that allocate substantial budgets to Research and Development (R&D). With significant funding, they possess advanced computational infrastructure capable of handling large data volumes. This infrastructure supports researchers in performing complex analyses and processing big data efficiently. The United States have the most data centers (5.381) and the best university (MIT) for studying data science and AI. Thus, the direction of research will refer to the US, where many researchers and their innovation emerge due to good facilities. As a result, researchers in other countries will conduct studies based on findings and approaches that have already proven effective. In Asia, China stands out as the most serious Asian country in studying big data in the field of chemistry. Although China does not have facilities as good as the US, China provides substantial funding for its researchers to conduct studies and researches in abroad. The National Natural Science Foundation of China (NSFC) is the most frequent sponsor, funding a total of 20 articles. Meanwhile, the Chinese Academy of Sciences is the institution that has facilitated the most research, with eight articles published under its support.

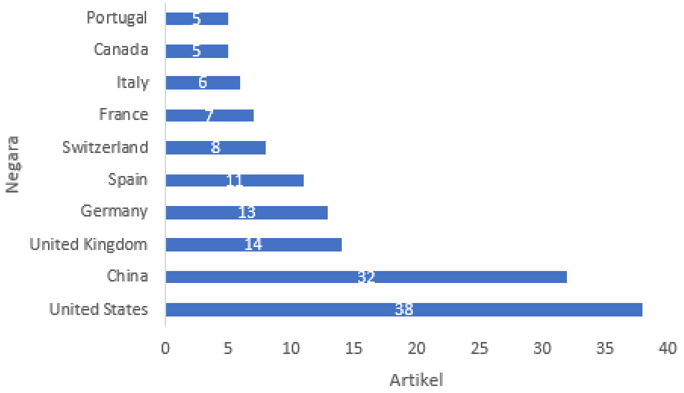

Figure 2. Number of Articles by Country from 2003 to 2024

Citation analysis aims to explore the relationships between publications by identifying the most influential publications in a research field (Donthu et al., 2021). Based on the citation analysis results, no publication has shown significant influence on the advancement of big data research in chemistry. The publication by Ramakrishnan (Ramakrishnan et al., 2015), which has the highest number of connections with other publications, is cited by studies conducted by Choi (Choi et al., 2018) and Senthil (Senthil et al., 2021). Despite Ramakrishnan's publication being cited 603 times, it is not referenced by other research specifically examining big data in chemistry. Additionally, Zabolotna (Zabolotna et al., 2022) is linked to two publications by Lin (Lin et al., 2019) and Humbeck (Humbeck et al., 2018), but Zabolotna is cited by these works rather than citing them. These findings suggest that researchers in this field are still somewhat fragmented and not yet fully interconnected, indicating that a continuous body of knowledge has not yet been established. In other words, the research in this area has not yet been thoroughly examined and integrated across different researchers.

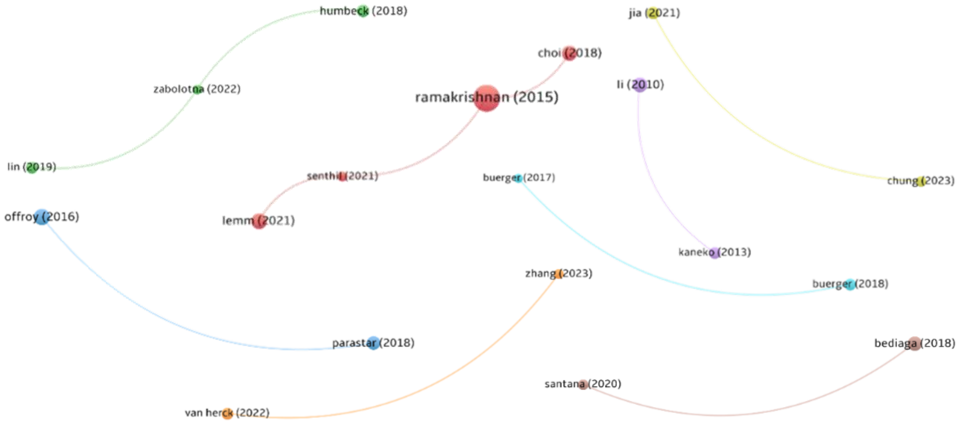

Figure 3. Visualization of Citation Analysis Results

Co-citation analysis aims to examine the relationships between references to understand the development of basic themes in a research field (Donthu et al., 2021). References are interconnected because they are cited by the same documents. The more a reference is linked to other references, the more relevant it is within the research field. The publication by Rogers & Hahn (Rogers & Hahn, 2010) on "molecular fingerprints" is identified as a central node or the most relevant reference, appearing in various research contexts. This is due to the capability of molecular fingerprints to numerically represent molecules, making them amenable to various analytical methods including machine learning, and their applications in areas such as drug discovery and computational simulations. Based on the co-citation results, three clusters reflecting core themes within the context of big data in chemistry have been identified cluster 1 (red) reflecting themes such as "enhancement of DFT methods," "application of machine learning," and "molecular enumeration", cluster 2 (green) describing themes related to "development of molecular fingerprints" and "drug discovery", cluster 3 (blue) focusing on "application of deep learning." Overall, the fundamental themes in big data research in chemistry revolve around the integration of computational methods with artificial intelligence.

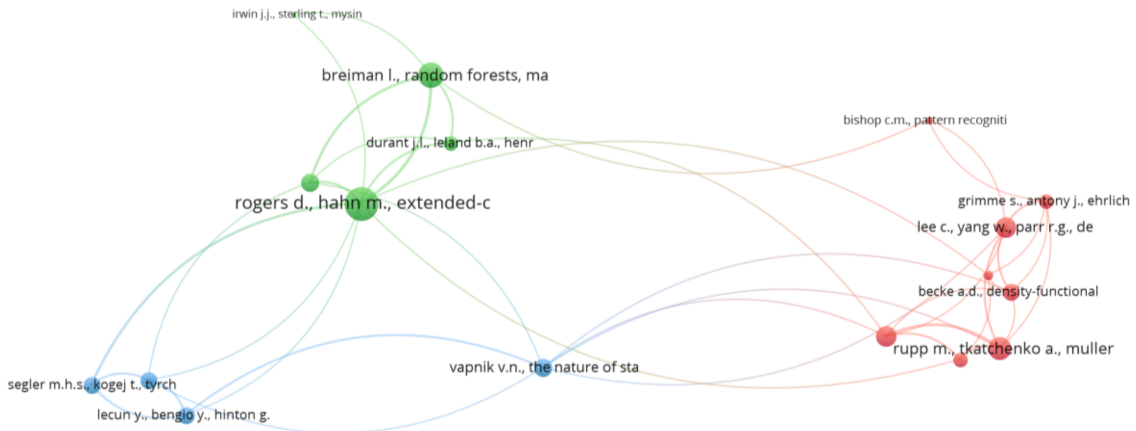
irwin j.j., sterling t., mysin
breiman l., random forests, ma
durant j.l., leland b.a., henr
rogers d., hahn m., extended-c
bishop c.m., pattern recogniti
grimme s., antony j., ehrlich
lee c., yang w., parr r.g., de
becke a.d., density-functional
rupp m., tkatchenko a., muller
vapnik v.n., the nature of sta
segler m.h.s., kogej t., tyrch
lecun y., bengio y., hinton g.

Figure 4. Visualization of Co-citation Analysis Results

Bibliographic coupling aims to analyze the relationships between publications to understand the periodic or current development of themes within a research field (Donthu et al., 2021). Publications are connected through shared references, indicating that they focus on similar themes. Figure 5 presents the visualization of the bibliographic coupling results. Publication by Harari (Harari et al., 2024) is minimally connected with other publications due to its focus on a novel topic. It explores the identification and analysis of SARS-CoV-2 virus mutations using millions of genomes and language models. The novelty of the topic results in limited references to similar works. Despite being published some time ago, publication by Wei (Wei et al., 2018) is also minimally connected with other publications. It discusses the use of nano-scale X-ray spectroscopy to study the morphological evolution and composition of cathode particles in lithium-ion batteries, employing big data to identify significant minor phases in operating batteries. The low level of connection suggests that this research theme has not been widely explored. The results indicate that both new and underexplored research areas, such as those highlighted, present opportunities for the emergence of unique and innovative follow-up studies.

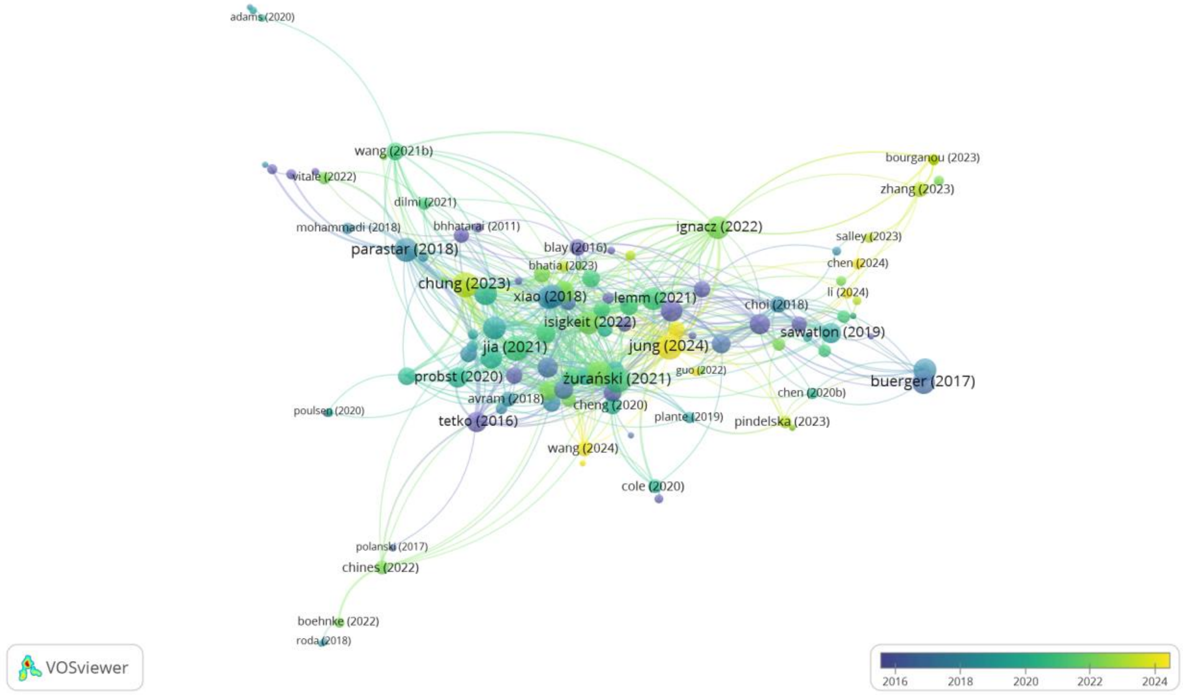

Figure 5. Visualization of Bibliographic Coupling Results

Co-word analysis aims to explore the relationships between current or emerging topics in a research field by focusing on the content within the publications themselves (Donthu et al., 2021). The results of the co-word analysis identified eight clusters: Cluster 1 (cheminformatics, data curation, drug discovery, quantitative structure-activity relationship, upper-division undergraduate, virtual screening, workflow), Cluster 2 (artificial intelligence, computational chemistry, deep learning, density functional theory, dimensionality reduction, emerging technologies, new directions in chemistry research), Cluster 3 (artificial neural networks, ChemBL, ligand-based virtual screening, machine learning, multitarget models, perturbation theory), Cluster 4 (analytical chemistry, chemometrics, mass spectrometry, omics, Raman spectroscopy), Cluster 5 (big data, bioinformatics, cloud computing), Cluster 6 (enthalpy formation, error-cancelling balance, validation), Cluster 7 (inkjet printing, silver nanoparticles), and Cluster 8 (data mining). Figure 6 shows the visualization of the co-word analysis results.

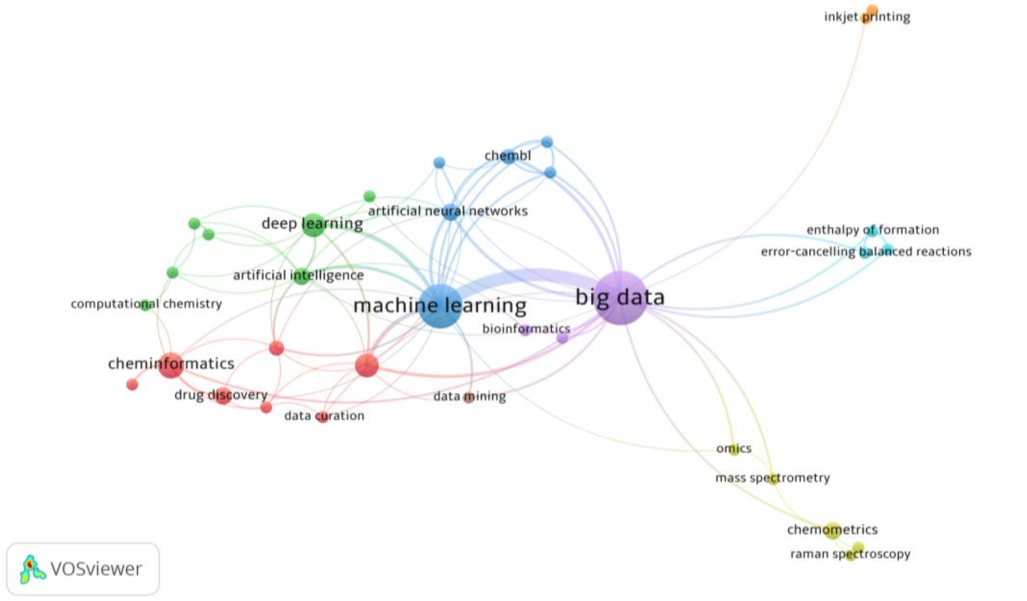

Figure 6. Visualization of Co-word Analysis Results

Based on the co-word analysis, it is observed that the topic of "machine learning" is closely related to "big data," indicating that big data analysis methods frequently employ artificial intelligence, particularly machine learning techniques. Additionally, statistical methods such as "chemometrics" and computer science approaches like "cheminformatics" are also connected to big data in chemistry. Interestingly, the topic of "data mining" forms its own distinct cluster, separate from other topics. Data mining (DM) is a subdomain of artificial intelligence (AI) defined as a process aimed at extracting knowledge from data and presenting findings comprehensively to users (Schuh et al., 2019).This suggests that the topic of data mining has not been widely utilized or explored in the context of chemical research. Consequently, this represents a new research opportunity and opens up new directions for chemistry research in the era of big data, aligning with its more recent emergence compared to other topics.

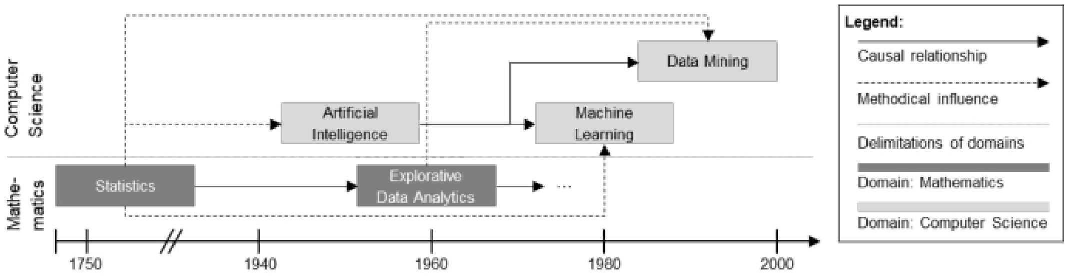

Figure 7. Historical Development of Data Analysis (Schuh et al., 2019)

In cluster 1, it is evident that drug discovery involving big data can be more efficient and effective when employing the appropriate and systematic steps of "cheminformatics." Cheminformatics is a broad term encompassing the design, creation, organization, management, retrieval, analysis, dissemination, visualization, and use of chemical information (Arulmozhi & Rajesh, 2011). "Data curation" serves as the initial step to ensure that the data to be analyzed is accurate and well-structured. Following this, "Quantitative Structure-Activity Relationship" (QSAR) is utilized—one of the cheminformatics methods that is useful for predicting biological activity by analyzing its relationship with chemical structure. Finally, "virtual screening" is conducted to evaluate thousands of designs computationally based on the developed models. In cluster 2, “deep learning” (DL) plays a crucial role in advancing computational chemistry due to its ability to analyze large amounts of data generated from computational methods such as Density Functional Theory (DFT). A branch of artificial intelligence (AI) known as machine learning uses algorithms to enable robots to learn from data and improve over time (Vrontis et al., 2022). Artificial neural networks with many hidden layers are employed in DL, a type of machine learning (ML), to help computers learn from vast volumes of data (Howard, 2019). The integration of AI and computation allows for studies that were previously difficult to access, thus accelerating new discoveries in chemistry.

In cluster 3, “artificial neural networks” are an integral part of “machine learning”. One example of leveraging machine learning algorithms is in “ligand-based virtual screening”, which is useful for accurately evaluating the potential interactions between compounds and biological targets such as proteins. The data stored in “ChemBL” can be utilized for this evaluation. The effectiveness of the evaluation can be enhanced with “multitarget models”, which allow for the simultaneous assessment of a compound's potential against multiple targets. To achieve optimal results, “perturbation theory” is used to predict how small changes in variables can affect the overall outcomes. In cluster 4, Mass Spectrometry (MS) instruments significantly contribute to generating large volumes of data in measurements. Generally, -omics studies such as genomics, proteomics, metabolomics, lipidomics, and glycomics produce big data from MS’s measurement. This data can be analyzed using chemometrics, which involves applying statistical and mathematical procedures to evaluate experimental data beyond univariate approaches (Inobeme et al., 2022). Chemometrics is frequently used in analytical chemistry, for instance, in measurements using Raman spectrometers to test food authenticity (Xu et al., 2020).

In cluster 5, a large amount of biomedical data, such as images and signals including omics data, has been accumulating over time. Bioinformatics can be used to analyze this big data. Bioinformatics involves using computer science to collect, store, analyze, and disseminate biological data and information. In addition to methods like bioinformatics, handling big data can be facilitated by infrastructure through internet-based services known as cloud computing. By using these services, large volumes of data can be stored, processed, and analyzed with high cost flexibility and scalability. Cloud computing also enables data access from various locations and supports team collaboration in data analysis. In cluster 6 and 7, The topics are specifically interrelated, which prevents the emergence of patterns that could describe something.

In cluster 8, the focus is solely on data mining. This indicates that in chemistry, data mining has not been explored in depth, and many data mining techniques are still in the exploration phase, with their applications in the chemical context not yet fully optimized. A better understanding of how data mining can be used to identify patterns and hidden relationships within large chemical datasets can provide deeper insights and accelerate discoveries that were previously difficult to achieve or not previously considered. By studying deeper into data mining, the expansion of cluster 8 can occur, offering researchers opportunities to drive innovation and discover new applications that integrate data mining with cheminformatics and bioinformatics, thereby accelerating progress in the field of chemistry.

Co-author analysis aims to examine social interactions or relationships among authors and their affiliations, and their impact on the development of research fields (Donthu et al., 2021). Based on the co-author analysis, it is observed that researchers do not collaborate with other researchers outside their specific study areas. This is evident from the lack of connections between clusters. Similarly, there is a lack of collaboration among institutions. This indicates that authors and institutions tend to work independently or collaborate only in small groups with minimal cross-group interaction. The negative impact of such a lack of collaboration is that big data research in chemistry may face limited perspectives, resulting in reduced innovation and slower progress in discoveries or problem-solving. However, international collaborations are emerging, with the United States being the most active in collaborating with other countries. China follows, although it more frequently collaborates with countries in the same region, such as Japan, Hong Kong, Australia, and Singapore. Figure 8 displays the visualization of co-author analysis by country.

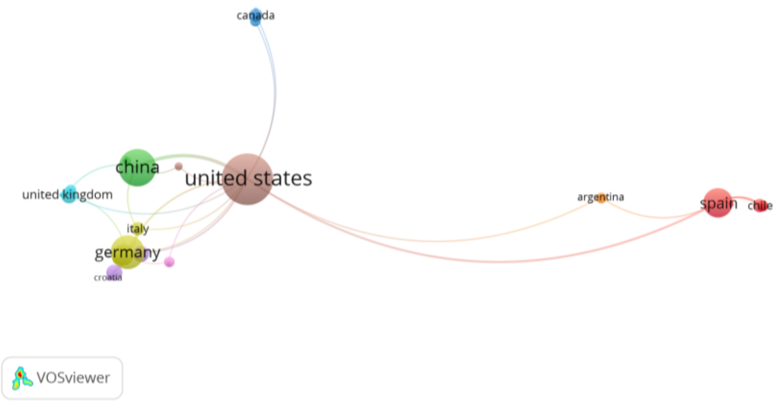

Figure 8. Visualization of Co-author Analysis by Country

Conclusion

This study was conducted to examine trends in big data research in chemistry as recorded in the Scopus database using bibliometric analysis. The bibliometric analysis technique employed is science mapping using the VOSviewer visualization software. The study's findings indicate that (1) the most impactful researcher is Pavlo O. Dral, while the most contributing researchers are Jacqueline M. Cole and Hao Zhu. The United States is the leading country in facilitating big data research in chemistry, followed by China. (2) Citation analysis reveals that researchers in this field are still fragmented, leading to research that is not fully explored in a continuous and in-depth manner. Co-citation analysis identifies that the core theme of big data research in chemistry is the integration of computation with artificial intelligence. Co-author analysis shows that researchers do not collaborate extensively with others, working independently or only in small groups. This lack of collaboration results in limited perspectives, hindering innovation and slowing down research progress. Co-word analysis identifies frequently discussed topics such as "machine learning," "chemometrics," "cheminformatics," and "deep learning." (3) The topic of "data mining" has not been extensively explored within the context of chemical research, presenting a new research opportunity and the potential to open new directions in chemistry research in the big data era.

References

Álvarez-Moreno, M., de Graaf, C., López, N., Maseras, F., Poblet, J. M., & Bo, C. (2015). Managing the Computational Chemistry Big Data Problem: The ioChem-BD Platform. *Journal of Chemical Information and Modeling*, *55*(1), 95–103. https://doi.org/10.1021/ci500593j

Arulmozhi, V., & Rajesh, R. (2011). Chemoinformatics &#x2014; A quick review. *2011 3rd International Conference on Electronics Computer Technology*, 416–419. https://doi.org/10.1109/ICECTECH.2011.5942128

Bediaga, H., Arrasate, S., & González-Díaz, H. (2018). PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS Combinatorial Science*, *20*(11), 621–632. https://doi.org/10.1021/acscombsci.8b00090

Bhhatarai, B., & Gramatica, P. (2011). Prediction of Aqueous Solubility, Vapor Pressure and Critical Micelle Concentration for Aquatic Partitioning of Perfluorinated Chemicals. *Environmental Science & Technology*, *45*(19), 8120–8128. https://doi.org/10.1021/es101181g

Choi, S., Kim, Y., Kim, J. W., Kim, Z., & Kim, W. Y. (2018). Feasibility of Activation Energy Prediction of Gas-Phase Reactions by Machine Learning. *Chemistry – A European Journal*, *24*(47), 12354–12358. https://doi.org/10.1002/chem.201800345

Chung, E., Russo, D. P., Ciallella, H. L., Wang, Y.-T., Wu, M., Aleksunes, L. M., & Zhu, H. (2023). Data-Driven Quantitative Structure–Activity Relationship Modeling for Human Carcinogenicity by Chronic Oral Exposure. *Environmental Science & Technology*, *57*(16), 6573–6588. https://doi.org/10.1021/acs.est.3c00648

Cole, J. M. (2020). A Design-to-Device Pipeline for Data-Driven Materials Discovery. *Accounts of Chemical Research*, *53*(3), 599–610. https://doi.org/10.1021/acs.accounts.9b00470

Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296. https://doi.org/10.1016/j.jbusres.2021.04.070

Duke, R., McCoy, R., Risko, C., & Bursten, J. R. S. (2024). Promises and Perils of Big Data: Philosophical Constraints on Chemical Ontologies. *Journal of the American Chemical Society*, *146*(17), 11579–11591. https://doi.org/10.1021/jacs.3c11399

Gantz, J., & Reinsel, D. (2011). *Extracting value from chaos*.

Gastegger, M., Behler, J., & Marquetand, P. (2017). Machine learning molecular dynamics for the simulation of infrared spectra. *Chemical Science*, *8*(10), 6924–6935. https://doi.org/10.1039/C7SC02267K

Guo, J., Yu, H., Xing, S., & Huan, T. (2022). Addressing big data challenges in mass spectrometry-based metabolomics. *Chemical Communications*, *58*(72), 9979–9990. https://doi.org/10.1039/D2CC03598G

Gupta, R., Srivastava, D., Sahu, M., Tiwari, S., Ambasta, R. K., & Kumar, P. (2021). Artificial intelligence to deep learning: machine intelligence approach for drug discovery. *Molecular Diversity*, *25*(3), 1315–1360. https://doi.org/10.1007/s11030-021-10217-3

Harari, S., Miller, D., Fleishon, S., Burstein, D., & Stern, A. (2024). Using big sequencing data to identify chronic SARS-Coronavirus-2 infections. *Nature Communications*, *15*(1), 648. https://doi.org/10.1038/s41467-024-44803-4

Hartshorn, R. (2017). Research Data, Big Data, and Chemistry. *Chemistry International*, *39*(3), 2–4. https://doi.org/10.1515/ci-2017-0301

Howard, J. (2019). Artificial intelligence: Implications for the future of work. *American Journal of Industrial Medicine*, *62*(11), 917–926. https://doi.org/10.1002/ajim.23037

Humbeck, L., Weigang, S., Schäfer, T., Mutzel, P., & Koch, O. (2018). CH *I* PMUNK: A Virtual Synthesizable Small-Molecule Library for Medicinal Chemistry, Exploitable for Protein–Protein Interaction Modulators. *ChemMedChem*, *13*(6), 532–539. https://doi.org/10.1002/cmdc.201700689

Inobeme, A., Nayak, V., Mathew, T. J., Okonkwo, S., Ekwoba, L., Ajai, A. I., Bernard, E., Inobeme, J., Mariam Agbugui, M., & Singh, K. R. (2022). Chemometric approach in environmental pollution analysis: A critical review. *Journal of Environmental Management*, *309*, 114653. https://doi.org/10.1016/j.jenvman.2022.114653

Jia, X., Ciallella, H. L., Russo, D. P., Zhao, L., James, M. H., & Zhu, H. (2021). Construction of a Virtual Opioid Bioprofile: A Data-Driven QSAR Modeling Study to Identify New Analgesic Opioids. *ACS Sustainable Chemistry & Engineering*, *9*(10), 3909–3919. https://doi.org/10.1021/acssuschemeng.0c09139

Jung, S. G., Jung, G., & Cole, J. M. (2024). Automatic Prediction of Peak Optical Absorption Wavelengths in Molecules Using Convolutional Neural Networks. *Journal of Chemical Information and Modeling*, *64*(5), 1486–1501. https://doi.org/10.1021/acs.jcim.3c01792

Krska, S. W., DiRocco, D. A., Dreher, S. D., & Shevlin, M. (2017). The Evolution of Chemical High-Throughput Experimentation To Address Challenging Problems in Pharmaceutical Synthesis. *Accounts of Chemical Research*, *50*(12), 2976–2985. https://doi.org/10.1021/acs.accounts.7b00428

Lancy, D. (2001). *3D Data Management: Controlling Data Volume, Velocity and Variety*.

Lemm, D., von Rudorff, G. F., & von Lilienfeld, O. A. (2021). Machine learning based energy-free structure predictions of molecules, transition states, and solids. *Nature Communications*, *12*(1), 4468. https://doi.org/10.1038/s41467-021-24525-7

Li, J., & Gramatica, P. (2010). Classification and Virtual Screening of Androgen Receptor Antagonists. *Journal of Chemical Information and Modeling*, *50*(5), 861–874. https://doi.org/10.1021/ci100078u

Lin, A., Horvath, D., Marcou, G., Beck, B., & Varnek, A. (2019). Multi-task generative topographic mapping in virtual screening. *Journal of Computer-Aided Molecular Design*, *33*(3), 331–343. https://doi.org/10.1007/s10822-019-00188-x

Liu, X., Xu, Y., Li, S., Wang, Y., Peng, J., Luo, C., Luo, X., Zheng, M., Chen, K., & Jiang, H. (2014). In Silicotarget fishing: addressing a "Big Data" problem by ligand-based similarity rankings with data fusion. *Journal of Cheminformatics*, *6*(1), 33. https://doi.org/10.1186/1758-2946-6-33

Luckenbach, R. (1981). The Beilstein Handbook of Organic Chemistry: the first hundred years. *Journal of Chemical Information and Computer Sciences*, *21*(2), 82–83. https://doi.org/10.1021/ci00030a006

Luo, J., Wang, Z., Xu, L., Wang, A. C., Han, K., Jiang, T., Lai, Q., Bai, Y., Tang, W., Fan, F. R., & Wang, Z. L. (2019). Flexible and durable wood-based triboelectric nanogenerators for self-powered sensing in athletic big data analytics. *Nature Communications*, *10*(1), 5147. https://doi.org/10.1038/s41467-019-13166-6

Mague, J. (1984). Gmelin Handbook of Inorganic Chemistry. 8th Edition Rh. *Organometallics*, *3*(6), 948–948. https://doi.org/10.1021/om00084a900

Mayr, L. M., & Bojanic, D. (2009). Novel trends in high-throughput screening. *Current Opinion in Pharmacology*, *9*(5), 580–588. https://doi.org/10.1016/j.coph.2009.08.004

Morgan, H. L. (1965). The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, *5*(2), 107–113. https://doi.org/10.1021/c160017a018

Passas, I. (2024). Bibliometric Analysis: The Main Steps. *Encyclopedia*, *4*(2), 1014–1025. https://doi.org/10.3390/encyclopedia4020065

Probst, D., & Reymond, J.-L. (2020). Visualization of very large high-dimensional data sets as minimum spanning trees. *Journal of Cheminformatics*, *12*(1), 12. https://doi.org/10.1186/s13321-020-0416-x

Purnama, S., & Sejati, W. (2023). Internet of Things, Big Data, and Artificial Intelligence in The Food and Agriculture Sector. *International Transactions on Artificial Intelligence (ITALIC)*, *1*(2), 156–174. https://doi.org/10.33050/italic.v1i2.274

Pyzer-Knapp, E. O., Li, K., & Aspuru-Guzik, A. (2015). Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery. *Advanced Functional Materials*, *25*(41), 6495–6502. https://doi.org/10.1002/adfm.201501919

Ramakrishnan, R., Dral, P. O., Rupp, M., & von Lilienfeld, O. A. (2015). Big Data Meets Quantum Chemistry Approximations: The Δ-Machine Learning Approach. *Journal of Chemical Theory and Computation*, *11*(5), 2087–2096. https://doi.org/10.1021/acs.jctc.5b00099

Rogers, D., & Hahn, M. (2010). Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, *50*(5), 742–754. https://doi.org/10.1021/ci100050t

Santana, R., Zuluaga, R., Gañán, P., Arrasate, S., Onieva Caracuel, E., & González-Díaz, H. (2020). PTML Model of ChEMBL Compounds Assays for Vitamin Derivatives. *ACS Combinatorial Science*, *22*(3), 129–141. https://doi.org/10.1021/acscombsci.9b00166

Schuh, G., Reinhart, G., Prote, J.-P., Sauermann, F., Horsthofer, J., Oppolzer, F., & Knoll, D. (2019). Data Mining Definitions and Applications for the Management of Production Complexity. *Procedia CIRP*, *81*, 874–879. https://doi.org/10.1016/j.procir.2019.03.217

Senthil, S., Chakraborty, S., & Ramakrishnan, R. (2021). Troubleshooting unstable molecules in chemical space. *Chemical Science*, *12*(15), 5566–5573. https://doi.org/10.1039/D0SC05591C

Swain, M. C., & Cole, J. M. (2016). ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature. *Journal of Chemical Information and Modeling*, *56*(10), 1894–1904. https://doi.org/10.1021/acs.jcim.6b00207

Taylor, P. (2023, November 16). *Amount of data created, consumed, and stored 2010-2020, with forecasts to 2025.* Statista.

Tetko, I. V., Engkvist, O., Koch, U., Reymond, J., & Chen, H. (2016). BIGCHEM: Challenges and Opportunities for Big Data Analysis in Chemistry. *Molecular Informatics*, *35*(11–12), 615–621. https://doi.org/10.1002/minf.201600073

Vrontis, D., Christofi, M., Pereira, V., Tarba, S., Makrides, A., & Trichina, E. (2022). Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review. *The International Journal of Human Resource Management*, *33*(6), 1237–1266. https://doi.org/10.1080/09585192.2020.1871398

Wang, L., Zhao, L., Liu, X., Fu, J., & Zhang, A. (2021). SepPCNET: Deeping Learning on a 3D Surface Electrostatic Potential Point Cloud for Enhanced Toxicity Classification and Its Application to

Suspected Environmental Estrogens. *Environmental Science & Technology*, *55*(14), 9958–9967. https://doi.org/10.1021/acs.est.1c01228

Wei, C., Xia, S., Huang, H., Mao, Y., Pianetta, P., & Liu, Y. (2018). Mesoscale Battery Science: The Behavior of Electrode Particles Caught on a Multispectral X-ray Camera. *Accounts of Chemical Research*, *51*(10), 2484–2492. https://doi.org/10.1021/acs.accounts.8b00123

Williams, W. L., Zeng, L., Gensch, T., Sigman, M. S., Doyle, A. G., & Anslyn, E. V. (2021). The Evolution of Data-Driven Modeling in Organic Chemistry. *ACS Central Science*, *7*(10), 1622–1637. https://doi.org/10.1021/acscentsci.1c00535

Wolff, A., Gooch, D., Cavero Montaner, J. J., Rashid, U., & Kortuem, G. (2016). Creating an Understanding of Data Literacy for a Data-driven Society. *The Journal of Community Informatics*, *12*(3). https://doi.org/10.15353/joci.v12i3.3275

Wu, X., Dral, P. O., Koslowski, A., & Thiel, W. (2019). Big data analysis of *ab Initio* molecular integrals in the neglect of diatomic differential overlap approximation. *Journal of Computational Chemistry*, *40*(4), 638–649. https://doi.org/10.1002/jcc.25748

Wu, Y., & Wang, G. (2018). Machine Learning Based Toxicity Prediction: From Chemical Structural Description to Transcriptome Analysis. *International Journal of Molecular Sciences*, *19*(8), 2358. https://doi.org/10.3390/ijms19082358

Xu, Y., Zhong, P., Jiang, A., Shen, X., Li, X., Xu, Z., Shen, Y., Sun, Y., & Lei, H. (2020). Raman spectroscopy coupled with chemometrics for food authentication: A review. *TrAC Trends in Analytical Chemistry*, *131*, 116017. https://doi.org/10.1016/j.trac.2020.116017

Yan, X., Sedykh, A., Wang, W., Yan, B., & Zhu, H. (2020). Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nature Communications*, *11*(1), 2519. https://doi.org/10.1038/s41467-020-16413-3

Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management*, *36*(6), 1231–1247. https://doi.org/10.1016/j.ijinfomgt.2016.07.009

Yeguas, V., & Casado, R. (2014). Big Data issues in Computational Chemistry. *2014 International Conference on Future Internet of Things and Cloud*, 389–392. https://doi.org/10.1109/FiCloud.2014.69

Zabolotna, Y., Bonachera, F., Horvath, D., Lin, A., Marcou, G., Klimchuk, O., & Varnek, A. (2022). Chemspace Atlas: Multiscale Chemography of Ultralarge Libraries for Drug Discovery. *Journal of Chemical Information and Modeling*, *62*(18), 4537–4548. https://doi.org/10.1021/acs.jcim.2c00509