Estd. 1989
Asian Journal of Chemistry

Asian Journal of Chemistry; Vol. 36, No. 12 (2024), 2744-2754

ASIAN JOURNAL OF CHEMISTRY

https://doi.org/10.14233/ajchem.2024.32653

REVIEW

Unveiling the Molecular World: A Narrative Review on Data Science and Visualization in Chemical Sciences

Eashwar Sai Komarla Rajasekhar[1], Abdur Rahman Junaid Nayeem[2], Vishal Satish Patil[3], Kamati Mounika[4], Satish Lahu Patil[5], Shruti Srivastava[6] and Gaurav Tiwari[7,*]

[1]Department of Data Science and Artificial Intelligence, Indian Institute of Technology-Bhilai, Kutela Bhata, Bhilai-491001, India
[2]Department of Medicinal Chemistry, NIPER-Kolkata, Chunilal Bhawan, Maniktala Main Road, Kolkata-700054, India
[3]Department of Pharmacognosy, Navsahyadri Institute of Pharmacy, Naigaon, Bhor-412213, India
[4]Department of Pharmacy Practice, MB School of Pharmaceutical Sciences, MB University (Erstwhile: Sree Vidyanikethan College of Pharmacy), Sri Sainath Nagar, Rangampeta, Tirupati-517102, India
[5]Pharmacist, Datta Medical Stores, Bhusawal-425201, India
[6]Department of Pharmaceutical Chemistry, Amity Institute of Pharmacy, Amity University, Sector-125, Noida-201301, India
[7]Department of Pharmaceutics, PSIT-Pranveer Singh Institute of Technology (Pharmacy), Bhauti, Kanpur-209305, India

*Corresponding author: E-mail: gt.karspbhs1@gmail.com

Received: 4 September 2024; *Accepted*: 10 November 2024; *Published online*: 30 November 2024; AJC-21813

This narrative review explores the transformative role of data science and visualization in modern chemistry. It begins by contextualizing the importance of chemistry across various domains, highlighting the emergence of data-driven approaches as catalysts for innovation and discovery. Harnessing big data in chemistry is discussed, emphasizing the diverse sources of chemical data and the need for robust analysis strategies. The review then delves into the power of machine learning (ML) algorithms in chemical discovery, highlighting their ability to accurately forecast molecular characteristics and significantly expedite pharmaceutical development. Visualizing chemical structures and dynamics is explored, with an emphasis on the role of visualization techniques in elucidating complex molecular phenomena. Integrative approaches in cheminformatics are examined, illustrating how interdisciplinary collaboration enables comprehensive analysis of chemical data. Challenges and opportunities in data-driven chemistry are addressed, alongside future perspectives on intelligent chemical systems. The review concludes by underscoring the transformative impact of data science and visualization on chemistry, advocating for continued investment and interdisciplinary collaboration to drive scientific innovation.

Keywords: Data science, Data visualization, Machine learning, Cheminformatics, Big data, Chemical discovery, Intelligent systems.

INTRODUCTION

Chemistry, often referred to as the "central science," is fundamental to understanding and addressing many of the world's most pressing challenges. Its applications span across numerous fields, each benefiting significantly from chemical research and innovations. It serves as the foundation of contemporary society, influencing every aspect of our existence, from the pharmaceuticals that cure us to the materials that construct the natural world. Its profound impact is evident across diverse fields, including medicine, drug discovery and development. Indeed, chemistry serves as the fundamental language of the natural world, enabling us to decipher the intricate mechanisms governing molecular interactions and transformations [1].

The advent of data science and visualization technologies has profoundly influenced scientific research, particularly in chemistry. These tools have empowered chemists to analyze extensive datasets and visualize molecular structures in novel ways, enabling deeper insights into chemical phenomena. The advancements in these fields have empowered scientists to peer into the molecular world with unprecedented clarity, unveiling hidden patterns, elucidating complex structures and predicting novel properties with remarkable accuracy [2].

The significance of this intersection cannot be overstated. Data science, with its arsenal of statistical methods [3], machine learning (ML) algorithms [4] and computational tools [5], offers a powerful lens through which to analyze and interpret vast troves of chemical data. From high-throughput screening assays to molecular simulations, data science enables researchers to extract valuable insights from the deluge of information generated by modern experimental and computational techniques [6-8].

Similarly, visualization techniques serve as indispensable tools for translating abstract data into intuitive representations, allowing researchers to explore and interact with complex chemical systems in meaningful ways [9]. Whether through molecular modeling software, interactive dashboards or immersive virtual reality environments, visualization techniques empower scientists to uncover hidden relationships, identify novel patterns and communicate their findings with clarity and precision [10]. Together, data science and visualization serve as the cornerstone of modern chemical research, driving innovation and breakthroughs across diverse applications. Leveraging data-driven methodologies, scientists are able to expedite advancements, refine chemical processes and develop new materials with customized properties to tackle pressing societal challenges [11,12].

In this narrative review, we embark on a journey into the heart of the chemical landscape, exploring the transformative role of data science and visualization in unlocking the secrets of the molecular world. Through a comprehensive examination of recent advancements, case studies and future perspectives, we aim to illuminate the profound impact of these interdisciplinary approaches on the field of chemistry and inspire new avenues of exploration and discovery.

Harnessing big data in chemistry: In the realm of chemistry, the exponential growth of data has emerged as a defining characteristic of the modern scientific landscape. This surge in data availability is propelled by a confluence of factors, including advancements in experimental techniques, computational methods and data-sharing initiatives. Experimental techniques such as high-throughput screening, mass spectrometry and X-ray crystallography yield vast quantities of data, providing detailed insights into molecular structures, properties and interactions [13,14]. Similarly, the computational methods, ranging from quantum chemistry simulations to molecular dynamics simulations, generate immense volumes of data, enabling researchers to explore chemical phenomena at unprecedented levels of detail [15]. Furthermore, the proliferation of data-sharing initiatives, such as open-access databases and collaborative research platforms, has democratized access to chemical data, fueling innovation and collaboration within the scientific community [16].

The sources of chemical data are as diverse as the field itself, encompassing a myriad of repositories, databases and experimental measurements. Databases such as PubChem [17], ChemSpider [18] and the Cambridge Structural Database house [19] vast collections of chemical compounds, properties and structures, providing invaluable resources for chemical research and discovery. Literature repositories, such as PubMed [20] and the Chemical Abstracts Service [21,22], contain a wealth of information gleaned from scientific publications, including experimental results, computational models and theoretical insights. Additionally, experimental measurements, spanning spectroscopic data, chromatographic profiles and chemical synthesis pathways, contribute valuable data points to the ever-expanding landscape of chemical information [23,24].

Table-1 presents a detailed overview of the various applications of data science in chemistry, illustrating the interplay between data science methods and visualization techniques. It highlights the impact of these applications on research, showcasing how advancements in data science are accelerating discoveries and optimizing processes across different chemical disciplines. The table categorizes each application by the data

TABLE-1
APPLICATIONS OF DATA SCIENCE IN CHEMISTRY. THE TABLE LISTS KEY APPLICATIONS WHERE DATA SCIENCE IS EMPLOYED IN THE FIELD OF CHEMISTRY. FOR EACH APPLICATION, THE DATA SCIENCE METHOD USED, THE VISUALIZATION TECHNIQUE APPLIED AND THE IMPACT ON RESEARCH ARE DESCRIBED

Application	Data science method	Visualization technique	Impact on research
Drug discovery	ML	Heatmaps, compound activity maps	Accelerates identification of potential drug candidates and optimizes lead compounds.
Materials science	High-throughput screening & ML	Ternary diagrams, principal component analysis plots	Speeds up the discovery of new materials with tailored properties.
Chemical synthesis prediction	Neural networks, Bayesian inference	Reaction pathways, decision trees	Improves accuracy in predicting chemical reaction outcomes and designing new synthesis routes.
Quantum chemistry	Quantum ML, kernel methods	Energy surface plots, contour maps	Enhances computational efficiency in predicting molecular properties and behaviours.
Spectroscopy data analysis	Pattern recognition, clustering algorithms	Spectral overlays, dendrograms	Automates spectral interpretation, increasing precision in structural elucidation.
Environmental chemistry	Geospatial analysis, time-series analysis	Geographic information system (GIS) maps, time-series plots	Aids in modeling pollutant dispersion and assessing environmental impacts.
Catalyst design	Genetic algorithms, reinforcement learning	Optimization landscapes, convergence graphs	Identifies and optimizes catalysts, leading to more efficient industrial processes.
Chemical informatics	Data mining, natural language processing (NLP)	Chemical space visualizations, network diagrams	Enhances chemical data integration and retrieval, improving research efficiency.

science method employed, the visualization technique used to interpret data and the resulting impact on research outcomes.

With the abundance of chemical data comes significant responsibility, as its sheer volume and complexity pose challenges for data management and analysis. Robust strategies are crucial to maintain the integrity, accessibility and interoperability of chemical data across various platforms. Standardized formats, metadata schemas and ontologies must be developed to facilitate seamless data integration and exchange. Additionally, effective analysis is vital for extracting meaningful insights from this vast data. This requires employing diverse statistical methods, ML algorithms and visualization techniques to uncover hidden patterns, correlations and trends [25,26].

Given these challenges and opportunities, interdisciplinary collaboration and innovation are essential to fully unlock the potential of big data in chemistry. By fostering partnerships between chemists, data scientists and computational researchers, we can develop advanced tools, techniques and methodologies to address complex chemical problems and drive scientific discovery forward [27]. For instance, Ferrero *et al.* [28] provided strategic recommendations for pharmaceutical chemists, drug designers and researchers to promote a digital culture shift and data science transformation within their organizations. Keith *et al*. [29] conducted a thorough review demonstrating the integration of computational chemistry and ML with data science to facilitate insightful predictions in molecular and material modeling, retrosynthesis, catalysis and drug design.

Moreover, by embracing open science principles and promoting data sharing and transparency, we can accelerate the pace of innovation and facilitate reproducibility and collaboration within the scientific community. By taking this approach, we can leverage the power of big data to unravel the mysteries of the molecular world and open new frontiers in chemistry [30-32]. Fig. 1 illustrates the various applications of data science in chemistry, highlighting key areas where computational tools and data-driven techniques are transforming the field. The visual representation is designed to provide a clear and concise overview of the integration of data science into chemical research and development, focusing on predictive modeling, molecular visualization, big data analytics, chemical informatics and ML. A corresponding icon, displaying the versatility and importance of data science in modern chemistry, represents each area. More efficient research, better decision-making, and ground-breaking discoveries in the discipline of chemistry are being made possible through the integration of data science tools and methodology, as seen in this figure.

Predictive modeling represents the use of data science to predict chemical properties, reactions and behaviours based on the existing data. Predictive modeling helps chemists anticipate outcomes and optimize the experimental conditions. The chemical informatics involves the application of information technology to manage and analyze chemical data. It also includes the database management, cheminformatics and the use of algorithms to process chemical information. Big data analytics emphasizes the analysis of large datasets in chemistry. Big Data Analytics allows researchers to identify patterns, trends and correlations in vast amounts of chemical data, leading to new insights and discoveries.

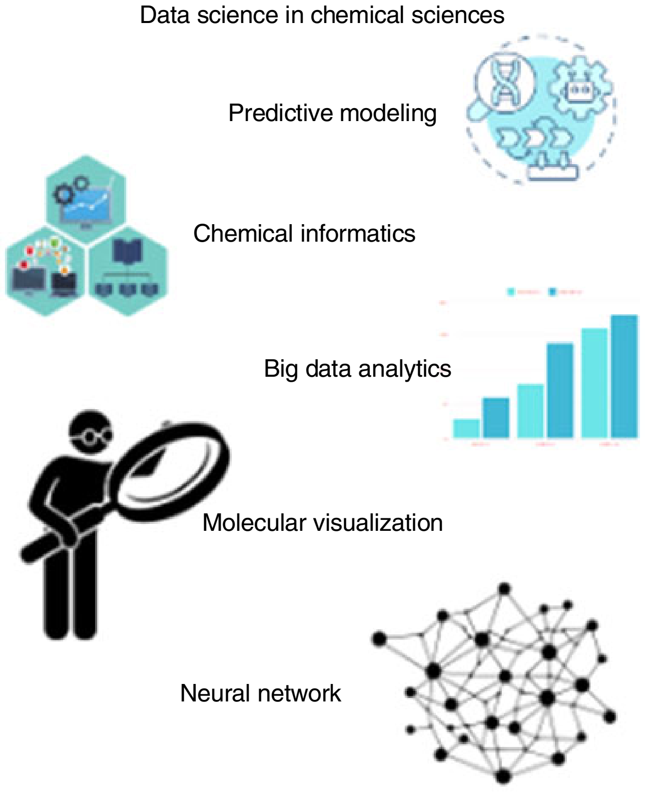


Fig. 1. Applications of data science in chemical sciences

Molecular visualization depicted by a magnifying glass, focuses on the use of visualization techniques to explore and analyze molecular structures, dynamics and interactions. These tools are essential for understanding complex chemical systems at the atomic level. The neural network icon represents ML applications in chemistry. The ML algorithms are used to analyze chemical data, identify patterns and make predictions, playing a crucial role in accelerating research and innovation.

Power of machine learning (ML) in chemical discovery: The disruptive power of ML algorithms in chemistry ushers in a new phase of discovery and innovation, offering exceptional opportunities to untangle the intricacies of chemical systems and accelerate scientific advancement. At the heart of this revolution lies the capacity of ML algorithms to uncover hidden patterns [33], predict molecular properties [34] and streamline the drug discovery [35] process with remarkable accuracy and efficiency. Through a diverse array of applications, ranging from quantitative structure-activity relationship (QSAR) modelling to virtual screening, prediction of physico-chemical, biological and toxicity properties, ML has emerged as a powerful tool for deciphering the molecular basis of biological phenomena, designing novel therapeutics and optimizing chemical processes [36,37].

A particularly significant application of ML in chemistry is QSAR modeling, where the goal is to quantify the relationship between chemical structure and biological activity. Nearly two decades ago, Svetnik *et al.* [38] demonstrated the effectiveness of the Random Forest algorithm in QSAR modeling by developing predictive models for six cheminformatics datasets, displaying its high accuracy. Recent developments in QSAR modeling have been explored by Tropsha *et al.* [39] who highlighted advances such as deep generative and reinforcement learning in molecular design, deep learning models for organic synthetic and retrosynthetic planning and the use of deep QSAR models in structure-based virtual screening.

In a notable application of ML, Button *et al.* [40] introduced DINGOS (Design of Innovative New Chemical Entities Generated by Optimization Strategies), a system that combines rule based methods with ML. This approach, trained on successful synthetic routes from chemical patents, enabled the generation of feasible syntheses for new chemical entities. DINGOS proposed synthetic routes for four approved drugs, with over 50% of the predicted compounds showing biological activity and four computer-generated compounds were successfully synthesized according to its proposed routes. Furthermore, Lind & Anderson [41] employed Random Forest models to predict drug activity against cancer cells by integrating recent screening data with models trained on the mutational status of 145 oncogenes and compound structural descriptors. This advancement has transformed drug discovery by allowing the rapid screening of large virtual compound libraries, prioritizing compounds with the highest potential biological activity and significantly improving the efficiency of the drug development pipeline.

Another key area where ML has made significant strides is virtual screening. By employing predictive models trained on diverse datasets of molecular structures and their associated activities, ML algorithms can efficiently sift through extensive chemical databases to identify promising drug candidates with desirable pharmacological profiles. This capability was demonstrated by Carpenter & Huang [42] who reviewed ML-based virtual screening methods for anti-Alzheimer's drug discovery and proposed a workflow for conducting such screenings. This high-throughput method allows researchers to efficiently prioritize compounds for further validation, streamlining the drug development process and conserving both time and resources. Moreover, ML algorithms are invaluable for predicting a range of molecular properties critical for drug designs, including solubility, stability and toxicity. Schapin *et al.* [43] examined ML models for predicting small molecule properties in drug discovery, highlighting key research directions such as collaborative partnerships, data sharing, data expansion, multi-task learning and decision support. By analyzing extensive datasets of chemical structures and their physico-chemical properties, ML models can accurately forecast the behaviour of new compounds, aiding in the design of safer and more effective drugs. This predictive capability accelerates the development of new therapeutics and reduces risks associated with drug toxicity and side effects.

Overall, the power of ML in chemical discovery lies in its ability to utilize vast amounts of chemical data to uncover insights, predict the molecular properties and accelerate the drug discovery process. Advanced algorithms and computational techniques enable researchers to explore new opportunities, design novel therapeutics and address pressing societal challenges with unprecedented speed and precision. As ML technology continues to evolve, its impact on chemistry is set to grow, ushering in a new era of discovery and innovation in the molecular sciences [44].

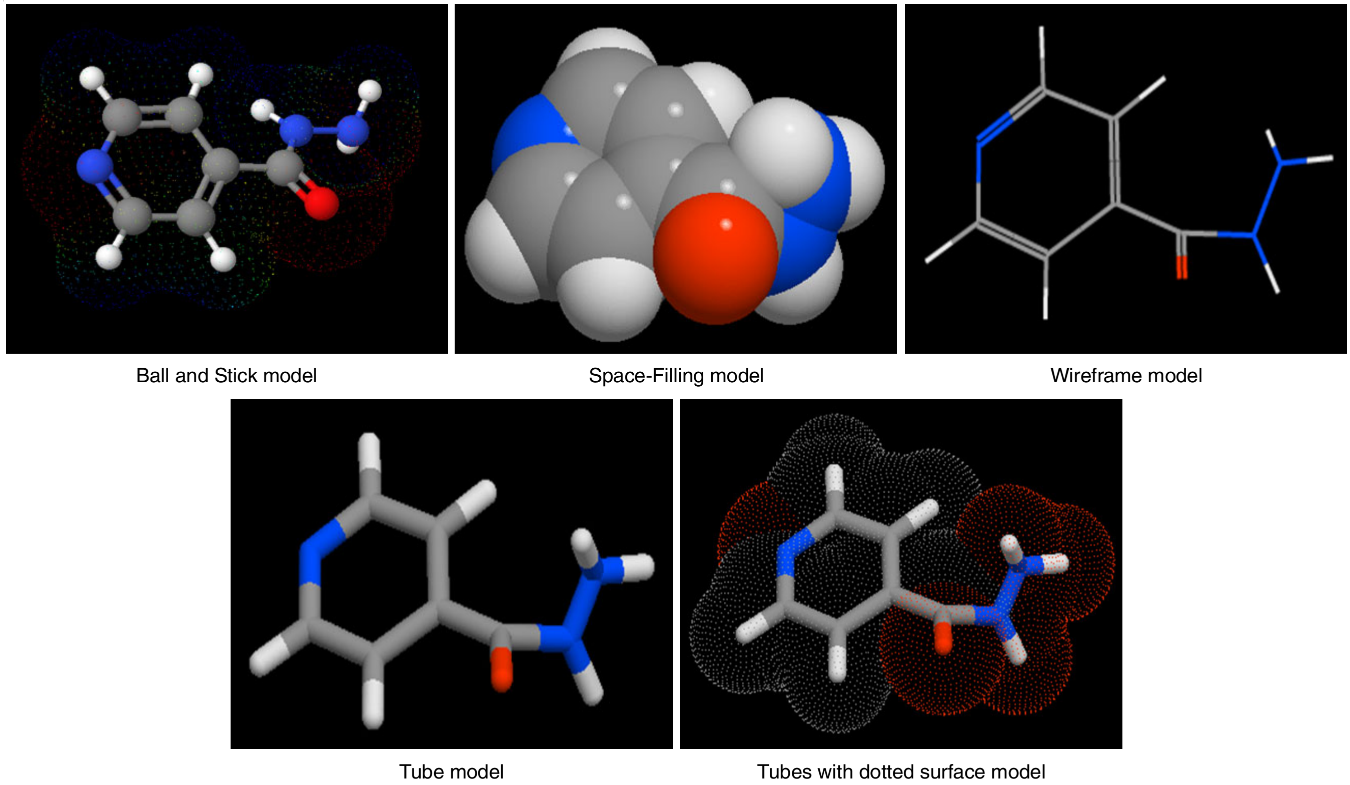
Visualizing chemical structures and dynamics: Visualizing chemical structures and dynamics is crucial for elucidating the complex molecular world, providing researchers with critical insights into the structure-function interactions that govern chemical behavior. At the heart of molecular visualization are the principles of representation, interaction and interpretation, which form the foundation for conveying complex molecular information in a clear and intuitive manner.

A diverse array of visualization techniques allows researchers to elucidate the spatial arrangements of atoms and molecules, explore dynamic processes such as molecular motion and reaction pathways and uncover underlying patterns and correlations within the data. These techniques not only deepen our understanding of molecular systems but also facilitate the development of new hypotheses and the design of innovative experiments [45]. In this context, Belghit *et al.* [46] explored various approaches and technologies used in visualizing molecular dynamics simulations, highlighting strategies tailored for this purpose. Their review discussed the advantages, limitations and future challenges in the field, emphasizing the importance of scientific and technological advances in visualizing complex molecular dynamics. They also explored the potential of multiscale molecular representations, visual abstraction and aggregation, all of which are vital for gaining deeper insights into molecular behaviour and interactions. Furthermore, they underscored the pressing need for ongoing exploration and development in these areas. The representative principle is the fundamental to molecular visualization since it allows for the easy interpretation and analysis of otherwise unknown chemical structures through the use of illustrations. This may include rendering molecular structures as ball-and-stick models, space filling representations or schematic diagrams, each providing unique perspectives on the spatial organization and connectivity of atoms within a molecule. By selecting appropriate representations and rendering styles, researchers can highlight key features such as functional groups, binding sites or structural motifs, thereby facilitating the interpretation of complex molecular structures and interactions [47,48].

Fig. 2 highlights the visualization of molecular structures, emphasizing spatial organization and atomic connectivity. Various models are used: the Ball-and-Stick model illustrates the 3D arrangement and bond angles; the Space-Filling model shows the overall shape and volume; the Wireframe model emphasizes atomic connectivity; and the Ribbon diagram or Tubes focus on the folding patterns and backbone structures in complex molecules like proteins.

Ball-and-stick model: This representation uses spheres to represent atoms and sticks to represent the bonds between them. The model makes it easier to visualize the 3D structure of the molecule by highlighting both the spatial arrangement of atoms and the bond angles.

Space-filling model (Corey-Pauling-Koltun model): Atoms are depicted as spheres that are scaled to their van der Waals radii, creating a dense representation where spheres overlap slightly. The spheres are typically colour-coded according to the element they represent (*e.g.* white for hydrogen, black for carbon, blue for nitrogen, red for oxygen). The general



Ball and Stick model

Space-Filling model

Wireframe model

Tube model

Tubes with dotted surface model

Fig. 2. Visualization representation of molecular structures in spatial organization and connectivity of atoms within a molecule

volume and form of the molecule, along with its potential interactions with other molecules, can be better understood with this model.

Surface rendering: This method involves creating a continuous surface over the molecule, often representing areas of similar electrostatic potential or solvent-accessible regions. This offers a distinct perspective on the external configuration of the molecule, helping in the comprehension of its interactions with other molecules, including ligand binding or protein-protein interactions.

Wireframe model: The molecule is represented as a network of lines (wires) connecting atoms. This model emphasizes the connectivity and topology of the molecule, allowing for a clear view of how atoms are linked. It is particularly advantageous for visualizing extensive or intricate molecules when detailed surface information may mask connection.

Tubes: Tubes, which are simplified versions of ribbon diagrams, trace the backbone of a molecule. This model focuses on the connectivity of polymer chain, emphasizing the path without showing side chains or detailed atomic positions. These molecular visualization representations are essential for understanding the spatial organization and connectivity of atoms within a molecule. They offer insights into molecular structure, functionality, and potential interactions with other molecules, which is essential in disciplines such as chemistry, biochemistry and molecular biology.

Interaction is equally crucial in molecular visualization, allowing researchers to manipulate and explore chemical structures in real-time, thereby enhancing their understanding of molecular behaviour and dynamics. Interactive visualization tools enable users to rotate, zoom and manipulate molecular models, offering a hands-on approach to exploring complex chemical systems and phenomena. Additionally, interactive features such as molecular docking, energy minimization and molecular dynamics simulations allow researchers to simulate and visualize dynamic processes like protein-ligand binding, enzyme catalysis and molecular conformational changes, providing critical insights into the mechanisms underlying biological function and disease [49]. Building on this framework, Fernandes *et al.* [50] illustrated the power of interactive visualization through the development of BioSIM Augmented Reality (BioSIMAR), a free online platform that transforms the way users interact with 3D molecular models in a virtual environment. BioSIMAR allows for interactive exploration of molecular structures, offering a hands-on approach to understand the chemical concepts. The platform operates seamlessly on any device with a camera and internet access, without the need for additional software. This innovative tool significantly enhances the ability to grasp the characteristics and behaviours of atoms and molecules.

A wide range of visualization tools and software packages are available to researchers, each offering distinct functionalities for visualizing chemical structures and dynamics. For example, molecular modeling software such as PyMOL [51], VMD [52] and ChimeraX [53] provide powerful tools for rendering and analyzing molecular structures. These tools enable users to visualize protein structures, analyze protein-ligand interactions and generate high-quality images and animations for publication and presentation. Similarly, the molecular dynamics simulation packages like GROMACS [54], AMBER [55] and NAMD [56] allow researchers to simulate and visualize the dynamic behaviour of biomolecular systems, offering detailed insights into atomic-scale motions and interactions [57].

Table-2 provides a comprehensive overview of various molecular visualization techniques used in chemistry. Each technique is briefly described, followed by its primary application in chemical research and the impact it has on advancing the field. The techniques include both traditional models, such as the Ball-and-Stick model and Space-Filling model, as well as more advanced visualizations like surface rendering and electron density isosurfaces.

In conclusion, data visualization techniques are critical for advancing our understanding of chemical structures and dynamics. By employing principles of representation, interaction and interpretation, researchers can decode the complexities of chemical systems, leading to discoveries that propel scientific progress and innovation. These advanced visualization tools and software packages enable academicians and researchers to visualize, analyze and communicate complex molecular information with unprecedented clarity and precision, thereby unlocking new insights and discoveries in chemistry and beyond [58,59].

From data to insights: Cheminformatics operates at the intersection of chemistry, computer science and statistics, leveraging data analysis and algorithmic methods to extract meaningful insights from chemical data. By integrating principles from these diverse fields, cheminformatics empowers scientists to analyze chemical information with exceptional accuracy, unlocking new possibilities for breakthroughs in drug development, materials science and chemical informatics. This field employs a variety of data analysis techniques including clustering, classification and network analysis, each offering distinct capabilities for extracting knowledge from chemical datasets and supporting informed decision-making in research and development [60].

TABLE-2
SUMMARY OF MOLECULAR VISUALIZATION TECHNIQUES. THIS TABLE SUMMARIZES KEY MOLECULAR VISUALIZATION TECHNIQUES USED IN CHEMISTRY, DETAILING THEIR DESCRIPTIONS, APPLICATIONS AND THE IMPACT THEY HAVE ON RESEARCH. THE TECHNIQUES RANGE FROM BASIC STRUCTURAL MODELS TO COMPLEX SURFACE RENDERINGS AND DYNAMIC SIMULATIONS, ILLUSTRATING THEIR BROAD UTILITY IN CHEMICAL AND BIOCHEMICAL STUDIES

Visualization technique	Description	Application	Impact on research	Example
Ball-and-Stick model	3D representation where atoms are depicted as spheres and bonds as sticks, showing molecular geometry.	Structural chemistry	Helps in understanding molecular geometry and bond angles, essential in structural analysis.	Ball-and-Stick model of water (H_2O), showing O as a red sphere and H as white spheres with sticks representing the O-H bonds.
Space-Filling model	3D model where atoms are represented by spheres proportional to their van der Waals radii, filling the space.	Molecular visualization	Provides a realistic view of molecular volume and space occupation, aiding in steric analysis.	Space-Filling model of methane (CH_4), showing carbon surrounded by hydrogen atoms as spheres sized by van der Waals radii.
Ribbon diagram	3D depiction of protein structures emphasizing the folding patterns of secondary structures like α-helices and β-sheets.	Structural biology	Enhances understanding of protein folding, stability and function, critical in bioinformatics and drug design.	Ribbon diagram of hemoglobin, highlighting α-helices and β-sheets, with heme groups shown in red.
Surface rendering	Visualization that displays the molecular surface, showing the accessible surface area and potential interaction sites.	Drug design, protein-ligand interactions	Aids in visualizing the topography of biomolecules, critical for understanding binding interactions.	Surface rendering of HIV protease, showing pockets and grooves where inhibitors might bind.
Wireframe model	Simplified 3D model using lines to represent bonds, often used for large macromolecules.	Structural analysis	Allows quick visualization of large biomolecules, useful in initial structure assessment.	Wireframe model of DNA, depicting the double helix with lines for the sugar-phosphate backbone and nitrogenous bases.
Electron density isosurfaces	3D surfaces representing regions of constant electron density, used in crystallography and quantum chemistry.	X-ray crystallography, quantum chemistry	Facilitates interpretation of experimental data and validation of molecular structures.	Electron density isosurfaces of benzene from X-ray crystallography, showing electron density around the carbon atoms.
Electrostatic potential maps	Visualization of the electrostatic potential on the molecular surface, indicating charge distribution.	Drug Design, Enzyme Catalysis	Assists in identifying reactive sites and understanding molecular interactions.	Electrostatic potential map of acetylcholinesterase, highlighting regions of positive and negative potential for interaction with substrates.
Molecular dynamics trajectories	Animated visualization showing the movement of atoms in a molecule over time during simulations.	Computational Chemistry	Provides insights into molecular flexibility, stability and interactions, crucial for dynamic studies.	Molecular dynamics simulation of a drug binding to a kinase, showing ligand interactions and conformational changes over time.

In line with this approach, Humer *et al.* [61] developed the ChemInformatics model explorer (CIME), a web-based platform designed to facilitate the inspection of chemical datasets, visualization of model explanations, comparison of interpretability methods and exploration of compound subgroups. This model-agnostic tool, operable on both servers and workstations, allows users to interactively navigate chemical spaces using both broad and detailed visualization techniques. CIME enhances the collaboration between chemistry and data science professionals, optimizing cheminformatics workflows [61]. Similarly, Saldivar-González *et al.* [62] introduced an electronic handbook on GitBook that guides users through Python programming, focusing on the analysis, representation and visualization of chemical data. The manual covers molecular representations of low molecular weight organic compounds, methods for acquiring data from public databases like ChEMBL and PubChem and techniques for analyzing and visualizing chemical information using concepts such as chemical space. This freely accessible GitBook aims to promote open science and support learning for students and professionals interested in chemical data analysis and visualization.

Clustering analysis, a fundamental technique in cheminformatics, allows researchers to identify patterns and groupings within large datasets based on structural or property similarities. By applying algorithms such as hierarchical clustering or k-means, researchers can partition chemical compounds into distinct clusters, revealing hidden relationships and associations. This approach is particularly valuable in drug discovery, where clustering can identify structurally similar compounds with similar pharmacological profiles, facilitating targeted drug design and the optimization of chemical libraries [63].

Sharma [64] classified antihypertensive medicines into six significant categories based on their ring structures, revealing significant developments within each group. Similarly, Voicu *et al*. [65] employed clustering techniques to identify analogous structures among 23 anticancer compounds based on their molecular fingerprints, illustrating the utility of statistical and cheminformatics tools in drug candidate selection.

Classification analysis is another essential tool, enabling researchers to categorize chemical compounds based on predefined criteria or properties. By training models on labeled datasets, researchers can develop predictive models capable of classifying new compounds accurately. This technique is particularly useful in toxicity prediction, helping to identify potentially hazardous compounds and prioritize them for further evaluation, thereby streamlining the drug development process and reducing toxicological risks [66]. Leveraging this strategy, Djoumbou *et al.* [67] developed ClassyFire, a versatile chemical ontology tool that classifies compounds into a taxonomy of over 4,800 categories based on their structures and features. This taxonomy, which spans up to 11 levels, uses clear, computable rules and consensus-based nomenclature to define each category based on the distinct structural properties of compound. Furthermore, Dong *et al.* [68] developed advanced ChemSAR, an online platform for developing structure-activity relationship (SAR) classification models for small molecules. Accessible across various operating systems and devices, ChemSAR provides features such as chemical structure validation, computation of molecular descriptors and predictive model generation. Users can interpret models through feature importance, tree visualization and report generation, making ChemSAR a comprehensive tool for SAR classification that benefits both cheminformatics and biomedical research.

Network analysis offers a powerful approach for representing and interpreting chemical data by constructing networks of chemical entities based on the structural or functional similarities. This method helps uncover underlying patterns and connections within the data, offering insights into the modular organization of chemical space and aiding in the identification of novel chemical scaffolds and lead compounds. Applications include elucidating relationships between compounds, biological targets and pharmacological activities, which guides the rational design of new therapeutics and molecular probes [69]. For example, Ruf & Danger [70] analyzed a complex network of astrochemical data related to interstellar ice analogs, identifying key transformations and refining existing knowledge with their structural annotations compared to the PubChem database.

In their review, Amara *et al.* [71] established clear nomenclature and formalism to clarify terminology related to various networks in metabolomics. They provided an overview of current network-based methods for mass spectrometry data analysis and discussed future advancements. Their review included network applications to biochemical reactions, mass spectrometry features, chemical structural similarities and metabolite correlations as well as the use of knowledge networks and the integration of multiple networks for simultaneous analysis and interpretation [71]. In conclusion, cheminformatics is a dynamic and interdisciplinary field that merges chemistry, computer science and statistics to analyze and interpret chemical data with exceptional accuracy. By employing techniques such as clustering, classification and network analysis, researchers can gain valuable insights that advance drug design, materials discovery and chemical informatics. The integration and interoperability of data are crucial, as they facilitate the comprehensive analysis of chemical information across various scales and domains, enhancing our understanding of chemical systems and phenomena [72].

Challenges and opportunities in data driven chemistry: Data driven approaches in chemistry hold immense potential for accelerating scientific discovery and innovation. However, they also present several challenges that must be addressed to fully realize their benefits. Key among these challenges are issues related to data quality, reproducibility and privacy. Ensuring the reliability and integrity of chemical data is crucial, as errors or inconsistencies can compromise research outcomes and lead to misleading conclusions. Additionally, reproducibility the ability to independently verify and replicate research findings is essential for building trust and confidence in scientific results. Furthermore, safeguarding privacy in the collection and dissemination of confidential chemical data is vital for protecting proprietary information and personal data [73,74].

For instance, Tetko *et al.* [75] examined the complexities of visualizing millions of compounds by integrating chemical and biological data. They discussed the potential of advanced

ML methods for mining "Big Data" in areas such as polypharmacology prediction, target identification and target resolution in phenotypic assays. They also addressed the challenge of securely sharing information without disclosing chemical structures, which is critical for enabling collaborative data exchange among multiple parties. Additionally, they highlighted the importance of education in advancing "Big Data" for progress in data-driven chemistry [75]. Himanen *et al.* [76] explored the evolving landscape of data-driven materials science and chemistry, focusing on the role of materials data infrastructures within the open science framework.

To mitigate these challenges, researchers must implement robust strategies for data standardization, validation and ethical guidelines. Data standardization involves establishing common formats, metadata schemas and ontologies to ensure consistency and interoperability across different datasets and platforms. Adhering to standardized data formats and metadata conventions enhances data quality, facilitates integration and promotes interoperability between datasets and research tools [77]. Validation methods, including cross-validation and external validation, are essential for assessing the precision and reliability of predictive models, ensuring that research outcomes are robust and reproducible [78]. Moreover, following ethical guidelines and data privacy regulations is crucial for protecting the integrity and confidentiality of proprietary chemical information, thereby fostering trust and accountability within the scientific community [79,80].

Despite these challenges, data-driven approaches offer significant opportunities for addressing long-standing issues in chemistry and driving scientific progress. Leveraging the vast amounts of chemical data available allows researchers to design more effective drugs, optimize chemical processes and discover novel materials with tailored properties [81,82]. For example, ML algorithms can process extensive chemical datasets and predict pharmacological characteristics, enabling the prioritization of drug candidates for further experimental testing and accelerating therapeutic discovery. Similarly, data driven methods can optimize chemical processes by identifying key parameters influencing reaction outcomes, leading to more efficient and sustainable processes. By analyzing chemical data at scale, researchers can uncover new patterns, correlations and relationships, facilitating the discovery of innovative materials with unique characteristics and functionalities [83].

Table-3 summarizes key challenges and future directions in data driven chemistry. It highlights issues such as data integration, quality and handling of big data, as well as interpreting complex models, ethical concerns and the integration of AI and ML. Each challenge is described along with its impact on research and examples of current efforts to address these issues. In conclusion, while the data driven approaches in chemistry present challenges related to data quality, reproducibility and privacy, they also offer exceptional opportunities for overcoming long standing issues and advancing scientific progress. By adopting rigorous strategies for data standardization, validation and ethical practices, researchers can navigate these challenges and fully harness the potential of data driven methods. With thoughtful consideration of both the obstacles and opportunities, data driven chemistry has the potential to transform our understanding, analysis and manipulation of chemical systems, leading to groundbreaking discoveries and innovations in the molecular sciences [84,85].

TABLE-3
CHALLENGES AND FUTURE DIRECTIONS IN DATA-DRIVEN CHEMISTRY. THIS TABLE OUTLINES SIGNIFICANT CHALLENGES AND FUTURE DIRECTIONS IN THE FIELD OF DATA-DRIVEN CHEMISTRY. IT DETAILS VARIOUS ISSUES INCLUDING DATA INTEGRATION, QUALITY ASSURANCE, BIG DATA MANAGEMENT, MODEL INTERPRETATION, ETHICAL CONSIDERATIONS, AI INTEGRATION AND TRAINING NEEDS

Challenge/future direction	Description	Impact on research	Example
Data integration and standardization	Integrating diverse datasets from various sources and ensuring consistent data formats and standards.	Facilitates comprehensive analyses and comparisons across different studies.	Efforts to standardize chemical databases such as PubChem and ChEMBL for better data integration.
Data quality and reproducibility	Ensuring the accuracy and reproducibility of data across different experiments and datasets.	Enhances reliability of research findings and supports robust conclusions.	Initiatives like the FAIR (Findable, Accessible, Interoperable, Reusable) data principles for improving data quality.
Handling big data	Managing and analyzing large-scale chemical datasets efficiently using advanced computational tools.	Improves the ability to derive insights from complex and voluminous data.	Use of high-performance computing (HPC) and cloud-based platforms for large-scale chemical simulations and data processing.
Interpreting complex models	Understanding and interpreting results from complex data-driven models and algorithms.	Enhances the ability to derive actionable insights from sophisticated models.	Application of explainable AI (XAI) techniques to make ML models more interpretable in chemical research.
Ethical and privacy concerns	Addressing privacy and ethical issues related to the use and sharing of chemical data.	Ensures responsible use of data and compliance with regulations.	Development of frameworks for secure data sharing and privacy-preserving techniques in chemical research.
Integration of AI and ML	Leveraging artificial intelligence and ML techniques to enhance data analysis and predictions.	Advances predictive capabilities and automates complex data analysis tasks.	Implementation of deep learning models for predicting chemical properties and reactions.
Training and education	Enhancing training and education for researchers in data-driven methodologies and tools.	Promotes effective use of data-driven techniques and tools in research.	Development of educational programs and online courses focusing on data science and computational methods in chemistry.

Future perspectives towards intelligent chemical systems: The future of chemistry is set for a transformative shift towards intelligent chemical systems, driven by advancements in data science, visualization and emerging technologies. This new era will see a paradigm shift towards autonomous decision making and adaptive behaviour in chemical systems, made possible through the integration of AI, quantum computing and ML techniques [86,87]. These intelligent chemical systems will autonomously analyze, predict and optimize chemical processes with unprecedented efficiency and accuracy, revolutionizing our approach to studying and manipulating chemical systems [88,89]. AI's ability to process vast amounts of data, identify patterns and make informed decisions presents significant potential for advancing chemical research and applications. AI algorithms can efficiently handle extensive datasets of chemical compounds, predict their properties and behaviours and recommend optimal experimental conditions or synthesis routes. This not only accelerates the drug discovery process but also aids in designing novel materials with tailored properties. Additionally, AI-powered robotic systems can automate laboratory tasks such as compound synthesis, screening and analysis, enabling researchers to focus on higher-level tasks and creative problem-solving [90,91].

Alongside AI, quantum computing represents another frontier in chemistry research, offering the potential to address complex chemical challenges that exceed the capabilities of classical computing [92]. Quantum algorithms can simulate molecular and material behaviours at atomic resolution, facilitating the exploration of new quantum chemistry frontiers, the design of more efficient catalysts and the optimization of molecular structures with exceptional precision [93]. Furthermore, quantum ML techniques can leverage the unique properties of quantum systems to enhance AI algorithms, leading to faster and more accurate predictions of molecular properties and behaviours [94-96].

Machine learning (ML), as a cornerstone of intelligent chemical systems, will continue to play a pivotal role in shaping the future of chemistry research and applications. By training ML models on large chemical datasets, researchers can develop predictive models that uncover hidden patterns, optimize chemical processes and guide rational design efforts. Techniques such as deep learning and reinforcement learning promise to advance our understanding of complex chemical phenomena, helping researchers decode molecular interactions, reaction mechanisms and material properties [97-99]. In conclusion, the future of chemistry offers boundless opportunities for innovation and discovery, driven by the convergence of data science, visualization and emerging technologies. By embracing interdisciplinary collaboration and fostering a culture of innovation, we can harness the transformative potential of intelligent chemical systems to address pressing societal challenges, drive scientific progress and shape the future of chemistry research and applications. As we move towards a future powered by intelligent chemical systems, let us remain committed to expanding the boundaries of knowledge and exploring new frontiers in the molecular sciences.

Conclusion

In conclusion, this review has highlighted the transformative impact of data science and visualization on the field of chemistry, emphasizing their crucial role in advancing scientific understanding, accelerating discovery and driving innovation. By integrating principles from chemistry, computer science and statistics, cheminformatics has enabled researchers to gain new insights into chemical systems and phenomena. This integration has guided the design of more effective drugs, optimized chemical processes and facilitated the discovery of novel materials with tailored properties. The use of big data, ML algorithms and advanced visualization techniques has revolutionized our approach to studying and manipulating chemical systems, empowering chemists to explore new frontiers and tackle complex challenges in the molecular sciences. Looking ahead, the opportunities presented by data science and visualization in chemistry are both vast and multifaceted. Realizing this potential will require sustained investment in research, education and infrastructure, as well as a commitment to fostering interdisciplinary collaboration and innovation. By embracing data driven research principles and cultivating a culture of curiosity and exploration, we can unlock new insights, drive scientific progress and shape the future of chemistry research and applications. As we move forward, let us seize the opportunities offered by data science and visualization to advance chemical research and fuel innovation, striving to unravel the mysteries of the molecular world and explore new frontiers in the pursuit of scientific discovery.

ACKNOWLEDGEMENTS

The authors are grateful to Department of Data Science and Artificial Intelligence, IIT-Bhilai, Department of Medicinal Chemistry, NIPER-Kolkota, Inaia and their respective institutes.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interests regarding the publication of this article.

REFERENCES

1. M. Yadav, R. Srivastava, F. Naaz, R. Verma and R.K. Singh, *Curr. Pharm. Des.*, **28**, 232 (2022); https://doi.org/10.2174/1381612827666211102101617
2. W.L. Williams, L. Zeng, T. Gensch, M.S. Sigman, A.G. Doyle and E.V. Anslyn, *ACS Cent. Sci.*, **7**, 1622 (2021); https://doi.org/10.1021/acscentsci.1c00535
3. H. Hwang and L. Ryan, *Biom. J.*, **62**, 270 (2020); https://doi.org/10.1002/bimj.201900034
4. I.H. Sarker, *SN Comput. Sci.*, **2**, 160 (2021); https://doi.org/10.1007/s42979-021-00592-x
5. C. Wang, M.H. Chen, E. Schifano, J. Wu and J. Yan, *Stat. Interface*, **9**, 399 (2016); https://doi.org/10.4310/SII.2016.v9.n4.a1
6. M. Breinig, F.A. Klein, W. Huber and M. Boutros, *Mol. Syst. Biol.*, **11**, 846 (2015); https://doi.org/10.15252/msb.20156400
7. J. Bajorath, *Nat. Rev. Drug Discov.*, **1**, 882 (2002); https://doi.org/10.1038/nrd941
8. A. Glielmo, B.E. Husic, A. Rodriguez, C. Clementi, F. Noé and A. Laio, *Chem. Rev.*, **121**, 9722 (2021); https://doi.org/10.1021/acs.chemrev.0c01195

9. D. Kuták, P. Vázquez, T. Isenberg, M. Krone, M. Baaden, J. Byška, B. Kozlíková and H. Miao, *Comput. Graph. Forum*, **42**, e14738 (2023); https://doi.org/10.1111/cgf.14738
10. J. Crossley-Lewis, J. Dunn, C. Buda, G.J. Sunley, A.M. Elena, I.T. Todorov, C.W. Yong, D.R. Glowacki, A.J. Mulholland and N.L. Allan, *J. Mol. Graph Model*., **125**, 108606 (2023); https://doi.org/10.1016/j.jmgm.2023.108606
11. E.M. Williamson and R.L. Brutchey, *Inorg. Chem.*, **62**, 16251 (2023); https://doi.org/10.1021/acs.inorgchem.3c02697
12. S.K. Niazi and Z. Mariam, *Pharmaceuticals*, **17**, 22 (2023); https://doi.org/10.3390/ph17010022
13. X. Dai and L. Shen, *Front. Med.*, **9**, 911861 (2022); https://doi.org/10.3389/fmed.2022.911861
14. X. Liu, L. Abad, L. Chatterjee, I.M. Cristea and M. Varjosalo, *Mass Spectrom. Rev.*, (2024); https://doi.org/10.1002/mas.21887
15. L. Böselt, M. Thürlemann and S. Riniker, *J. Chem. Theory Comput.*, **17**, 2641 (2021); https://doi.org/10.1021/acs.jctc.0c01112
16. R. Cárdenas, J. Martínez-Seoane and C. Amero, *Molecules*, **25**, 4783 (2020); https://doi.org/10.3390/molecules25204783
17. S. Kim, P.A. Thiessen, E.E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B.A. Shoemaker, J. Wang, B. Yu, J. Zhang and S.H. Bryant, *Nucleic Acids Res.*, **44(D1)**, D1202 (2016); https://doi.org/10.1093/nar/gkv951
18. H.E. Pence and A. Williams, *J. Chem. Educ.*, **87**, 1123 (2010); https://doi.org/10.1021/ed100697w
19. C.R. Groom, I.J. Bruno, M.P. Lightfoot and S.C. Ward, *Acta Crystallogr. B Struct. Sci. Cryst. Eng. Mater.*, **72**, 171 (2016); https://doi.org/10.1107/S2052520616003954
20. P.O. Williamson and C.I.J. Minter, *J. Med. Libr. Assoc.*, **107**, 16 (2019); https://doi.org/10.5195/jmla.2019.433
21. American Chemical Society National Historic Chemical Landmarks, Chemical Abstracts Service (CAS); http://www.acs.org/content/acs/en/education/whatischemistry/landmarks/cas.html (accessed on August 2, 2024).
22. G. Asche, *World Pat. Inf.*, **48**, 16 (2017); https://doi.org/10.1016/j.wpi.2016.11.004
23. R. Guha, D.T. Nguyen, N. Southall and A. Jadhav, *Curr. Protoc. Chem. Biol.*, **4**, 193 (2012); https://doi.org/10.1002/9780470559277.ch110262
24. L. Liu, B.F. Jones, B. Uzzi and D. Wang, *Nat. Hum. Behav.*, **7**, 1046 (2023); https://doi.org/10.1038/s41562-023-01562-4
25. A.T. Rosário and J.C. Dias, *Int. J. Inform. Manage. Data Insights*, **3**, 100203 (2023); https://doi.org/10.1016/j.jjimei.2023.100203
26. M.L. Heacock, A.R. Lopez, S.M. Amolegbe, D.J. Carlin, H.F. Henry, B.A. Trottier, M.L. Velasco and W.A. Suk, *Environ. Sci. Technol.*, **56**, 7544 (2022); https://doi.org/10.1021/acs.est.1c08383
27. A.H. Cheng, C.T. Ser, M. Skreta, A. Guzmán-Cordero, L. Thiede, A. Burger, A. Aldossary, S.X. Leong, S. Pablo-García, F. Strieth-Kalthoff and A. Aspuru-Guzik, *Faraday Discuss*., (2024); https://doi.org/10.1039/D4FD00153B
28. E. Ferrero, S. Brachat, J.L. Jenkins, P. Marc, P. Skewes-Cox, R.C. Altshuler, C. Gubser-Keller, A. Kauffmann, E.K. Sassaman, J.M. Laramie, B. Schoeberl, M.L. Borowsky and N. Stiefl, *PLOS Comput. Biol.*, **16**, e1008126 (2020); https://doi.org/10.1371/journal.pcbi.1008126
29. J.A. Keith, V. Vassilev-Galindo, B. Cheng, S. Chmiela, M. Gastegger, K.-R. Müller and A. Tkatchenko, *Chem. Rev.*, **121**, 9816 (2021); https://doi.org/10.1021/acs.chemrev.1c00107
30. R.T. Thibault, O.B. Amaral, F. Argolo, A.E. Bandrowski, A.R. Davidson and N.I. Drude, *PLoS Biol.*, **21**, e3002362 (2023); https://doi.org/10.1371/journal.pbio.3002362
31. National Academies of Sciences, Engineering, and Medicine; Policy and Global Affairs; Board on Research Data and Information; Committee on Toward an Open Science Enterprise; Open Science by Design: Realizing a Vision for 21st Century Research. Washington (DC): National Academies Press (USA) (2018).
32. A. Salazar, B. Wentzel, S. Schimmler, R. Gläser, S. Hanf and S.A. Schunk, *Chem. Eur. J.*, **29**, e202202720 (2023); https://doi.org/10.1002/chem.202202720
33. R.P. França, A.C.B. Monteiro, R. Arthur and Y. Iano, eds.: V. Piuri, S. Raj, A. Genovese and R. Srivastava, Hybrid Computational Intelligence for Pattern Analysis, In: Trends in Deep Learning Methodologies, Academic Press, pp. 63-87 (2021); https://doi.org/10.1016/B978-0-12-822226-3.00003-9
34. K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O.A. von Lilienfeld, K.-R. Müller and A. Tkatchenko, *J. Phys. Chem. Lett.*, **6**, 2326 (2015); https://doi.org/10.1021/acs.jpclett.5b00831
35. D. Paul, G. Sanap, S. Shenoy, D. Kalyane, K. Kalia and R.K. Tekade, *Drug Discov. Today*, **26**, 80 (2021); https://doi.org/10.1016/j.drudis.2020.10.010
36. C. Selvaraj, I. Chandra and S.K. Singh, *Mol. Divers.*, **26**, 1893 (2022); https://doi.org/10.1007/s11030-021-10326-z
37. B.J. Neves, R.C. Braga, C.C. Melo-Filho, J.T. Moreira-Filho, E.N. Muratov and C.H. Andrade, *Front. Pharmacol.*, **9**, 1275 (2018); https://doi.org/10.3389/fphar.2018.01275
38. V. Svetnik, A. Liaw, C. Tong, J.C. Culberson, R.P. Sheridan and B.P. Feuston, *J. Chem. Inf. Comput. Sci.*, **43**, 1947 (2003); https://doi.org/10.1021/ci034160g
39. A. Tropsha, O. Isayev, A. Varnek, G. Schneider and A. Cherkasov, *Nat. Rev. Drug Discov.*, **23**, 141 (2023); https://doi.org/10.1038/s41573-023-00832-0
40. A. Button, D. Merk, J.A. Hiss and G. Schneider, *Nat. Mach. Intell.*, **1**, 307 (2019); https://doi.org/10.1038/s42256-019-0067-7
41. A.P. Lind and P.C. Anderson, *PLoS One*, **14**, e0219774 (2019); https://doi.org/10.1371/journal.pone.0219774
42. K.A. Carpenter and X. Huang, *Curr. Pharm. Des.*, **24**, 3347 (2018); https://doi.org/10.2174/1381612824666180607124038
43. N. Schapin, M. Majewski, A. Varela-Rial, C. Arroniz and G.D. Fabritiis, *Artif. Intellig. Chem.*, **1**, 100020 (2023); https://doi.org/10.1016/j.aichem.2023.100020
44. R. Han, H. Yoon, G. Kim, H. Lee and Y. Lee, *Pharmaceuticals*, **16**, 1259 (2023); https://doi.org/10.3390/ph16091259
45. O.M.H. Salo-Ahen, I. Alanko, R. Bhadane, A.M.J.J. Bonvin, R.V. Honorato, S. Hossain, A.H. Juffer, A. Kabedev, M. Lahtela-Kakkonen, A.S. Larsen, E. Lescrinier, P. Marimuthu, M.U. Mirza, G. Mustafa, A. Nunes-Alves, T. Pantsar, A. Saadabadi, K. Singaravelu and M. Vanmeert, *Processes*, **9**, 71 (2020); https://doi.org/10.3390/pr9010071
46. H. Belghit, M. Spivak, M. Dauchez, M. Baaden and J. Jonquet-Prevoteau, *Front. Bioinform.*, **4**, 1356659 (2024); https://doi.org/10.3389/fbinf.2024.1356659
47. B. Kozlíková, M. Krone, M. Falk, N. Lindow, M. Baaden, D. Baum, I. Viola, J. Parulek and H.C. Hege, *Comput. Graph. Forum*, **36**, 178 (2017); https://doi.org/10.1111/cgf.13072
48. Z. Liu, B. Kerr, M. Dontcheva, J. Grover, M. Hoffman and A. Wilson, *Comput. Graph. Forum*, **36**, 527 (2017); https://doi.org/10.1111/cgf.13208
49. H.M. Deeks, R.K. Walters, S.R. Hare, M.B. O'Connor, A.J. Mulholland and D.R. Glowacki, *PLoS One*, **15**, e0228461 (2020); https://doi.org/10.1371/journal.pone.0228461
50. H.S. Fernandes, N.M.F.S.A. Cerqueira and S.F. Sousa, *J. Chem. Educ.*, **98**, 1789 (2021); https://doi.org/10.1021/acs.jchemed.0c01317
51. S. Rosignoli and A. Paiardini, *Biomolecules*, **12**, 1764 (2022); https://doi.org/10.3390/biom12121764
52. J. Hsin, A. Arkhipov, Y. Yin, J.E. Stone and K. Schulten, Curr Protoc Bioinformatics. Chapter 5: Unit 5.7(2008); https://doi.org/10.1002/0471250953
53. E.C. Meng, T.D. Goddard, E.F. Pettersen, G.S. Couch, Z.J. Pearson, J.H. Morris and T.E. Ferrin, *Protein Sci.*, **32**, e4792 (2023); https://doi.org/10.1002/pro.4792
54. M.J. Abraham, T. Murtola, R. Schulz, S. Páll, J.C. Smith, B. Hess and E. Lindahl, *SoftwareX*, **1–2**, 19 (2015); https://doi.org/10.1016/j.softx.2015.06.001

55. D.A. Case, T.E. Cheatham III, T. Darden, H. Gohlke, R. Luo, K.M. Merz Jr., A. Onufriev, C. Simmerling, B. Wang and R.J. Woods, *J. Comput. Chem.*, **26**, 1668 (2005); https://doi.org/10.1002/jcc.20290
56. J.C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R.D. Skeel, L. Kalé and K. Schulten, *J. Comput. Chem.*, **26**, 1781 (2005); https://doi.org/10.1002/jcc.20289
57. J.E. Stone, W.R. Sherman and K. Schulten, *IEEE Int. Symp. Parallel Distrib. Process. Workshops PhD Forum*, **2016**, 1048 (2016); https://doi.org/10.1109/IPDPSW.2016.121
58. L.L. Jones, *J. Chem. Educ.*, **90**, 1571 (2013); https://doi.org/10.1021/ed4001206
59. R.J. Petillion and W.S. McNeil, *J. Chem. Educ.*, **97**, 1536 (2020); https://doi.org/10.1021/acs.jchemed.9b01105
60. D.S. Wishart, Curr Protoc Bioinformatics. Chapter 14: Unit 14.1(2007); https://doi.org/10.1002/0471250953.bi1401s18
61. C. Humer, H. Heberle, F. Montanari, T. Wolf, F. Huber, R. Henderson, J. Heinrich and M. Streit, *J. Cheminform.*, **14**, 21 (2022); https://doi.org/10.1186/s13321-022-00600-z
62. F.I. Saldivar-González, D.L. Prado-Romero, R. Cedillo-González, A.L. Chávez-Hernández, J.F. Avellaneda-Tamayo, A. Gómez-García, L. Juárez-Rivera and J.L. Medina-Franco, *J. Chem. Educ.*, **101**, 2549 (2024); https://doi.org/10.1021/acs.jchemed.4c00041
63. W. Li, *J. Chem. Inf. Model.*, **46**, 1919 (2006); https://doi.org/10.1021/ci0600859
64. R. Sharma, *Inform. Med. Unlocked*, **19**, 100303 (2020); https://doi.org/10.1016/j.imu.2020.100303
65. A. Voicu, N. Duteanu, M. Voicu, D. Vlad and V. Dumitrascu, *J. Cheminform.*, **12**, 3 (2020); https://doi.org/10.1186/s13321-019-0405-0
66. M. Mathea, W. Klingspohn and K. Baumann, *Mol. Inform.*, **35**, 160 (2016); https://doi.org/10.1002/minf.201501019
67. Y. Djoumbou Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner and D.S. Wishart, *J. Cheminform.*, **8**, 61 (2016); https://doi.org/10.1186/s13321-016-0174-y
68. J. Dong, Z.J. Yao, M.F. Zhu, N.N. Wang, B. Lu, A.F. Chen, A.P. Lu, H. Miao, W.B. Zeng and D.S. Cao, *J. Cheminform.*, **9**, 27 (2017); https://doi.org/10.1186/s13321-017-0215-1
69. D. Hevey, *Health Psychol. Behav. Med.*, **6**, 301 (2018); https://doi.org/10.1080/21642850.2018.1521283
70. A. Ruf and G. Danger, *Anal. Chem.*, **94**, 14135 (2022); https://doi.org/10.1021/acs.analchem.2c01271
71. A. Amara, C. Frainay, F. Jourdan, T. Naake, S. Neumann, E.M. Novoa-del-Toro, R.M. Salek, L. Salzer, S. Scharfenberg and M. Witting, *Front. Mol. Biosci.*, **9**, 841373 (2022); https://doi.org/10.3389/fmolb.2022.841373
72. J. Bajorath, A.L. Chávez-Hernández, M. Duran-Frigola, E. Fernández-de Gortari, J. Gasteiger, E. López-López, G.M. Maggiora, J.L. Medina-Franco, O. Méndez-Lucio, J. Mestres, R.A. Miranda-Quintana, T.I. Oprea, F. Plisson, F.D. Prieto-Martínez, R. Rodríguez-Pérez, P. Rondón-Villarreal, F.I. Saldívar-Gonzalez, N. Sánchez-Cruz and M. Valli, *J. Cheminform.*, **14**, 82 (2022); https://doi.org/10.1186/s13321-022-00661-0
73. J.I. Martinez Alvarado, J.M. Meinhardt and S. Lin, *Tetrahedron Chem*, **1**, 100012 (2022); https://doi.org/10.1016/j.tchem.2022.100012
74. National Academies of Sciences, Engineering and Medicine. Data Matters: Ethics, Data, and International Research Collaboration in a Changing World: Proceedings of a Workshop. Washington, DC: The National Academies Press (2018); https://doi.org/10.17226/25214
75. I.V. Tetko, O. Engkvist, U. Koch, J.L. Reymond and H. Chen, *Mol. Inform.*, **35**, 615 (2016); https://doi.org/10.1002/minf.201600073
76. L. Himanen, A. Geurts, A.S. Foster and P. Rinke, *Adv. Sci. (Weinh.)*, **6**, 1900808 (2019); https://doi.org/10.1002/advs.201900808
77. D. Mittal, R. Mease, T. Kuner, H. Flor, R. Kuner and J. Andoh, *Gigascience*, **12**, giad049 (2022); https://doi.org/10.1093/gigascience/giad049
78. E. Lopez, J. Etxebarria-Elezgarai, J.M. Amigo and A. Seifert, *Anal. Chim. Acta*, **1275**, 341532 (2023); https://doi.org/10.1016/j.aca.2023.341532
79. H.H.H. Aldboush and M. Ferdous, *Int. J. Financial Studies*, **11**, 90 (2023); https://doi.org/10.3390/ijfs11030090
80. J. Medina, A.W. Ziaullah, H. Park, I.E. Castelli, A. Shaon, H. Bensmail and F. El-Mellouhi, *Matter*, **5**, 3614 (2022); https://doi.org/10.1016/j.matt.2022.10.007
81. K. Gao, D.D. Nguyen, M. Tu and G.W. Wei, *J. Chem. Inf. Model.*, **60**, 5682 (2020); https://doi.org/10.1021/acs.jcim.0c00599
82. T.K. Patra, *ACS Polym. Au*, **2**, 8 (2022); https://doi.org/10.1021/acspolymersau.1c00035
83. A. Blanco-González, A. Cabezón, A. Seco-González, D. Conde-Torres, P. Antelo-Riveiro, Á. Piñeiro and R. Garcia-Fandino, *Pharmaceuticals*, **16**, 891 (2023); https://doi.org/10.3390/ph16060891
84. R. Mercado, S.M. Kearnes and C.W. Coley, *J. Chem. Inf. Model.*, **63**, 4253 (2023); https://doi.org/10.1021/acs.jcim.3c00607
85. J. Bajorath, *Future Sci. OA*, **4**, FSO320 (2018); https://doi.org/10.4155/fsoa-2018-0057
86. A.J.S. Hammer, A.I. Leonov, N.L. Bell and L. Cronin, *JACS Au*, **1**, 1572 (2021); https://doi.org/10.1021/jacsau.1c00303
87. C. Liu, Y. Chen and F. Mo, *Natl. Sci. Open*, **3**, 20230037 (2023); https://doi.org/10.1360/nso/20230037
88. V.P. Ananikov, *Artif. Intellig. Chem.*, **2**, 100075 (2024); https://doi.org/10.1016/j.aichem.2024.100075
89. X.Y. Tai, H. Zhang, Z. Niu, S.D.R. Christie and J. Xuan, *Energy and AI*, **2**, 100036 (2020); https://doi.org/10.1016/j.egyai.2020.100036
90. K.S. Vidhya, A. Sultana, M. Naveen Kumar and H. Rangareddy, *Cureus*, **15**, e47486 (2023); https://doi.org/10.7759/cureus.47486
91. Y. Xu, X. Liu, X. Cao, C. Huang, E. Liu, S. Qian, X. Liu, Y. Wu, F. Dong, C.-W. Qiu, J. Qiu, K. Hua, W. Su, J. Wu, H. Xu, Y. Han, C. Fu, Z. Yin, M. Liu, R. Roepman, S. Dietmann, M. Virta, F. Kengara, Z. Zhang, L. Zhang, T. Zhao, J. Dai, J. Yang, L. Lan, M. Luo, Z. Liu, T. An, B. Zhang, X. He, S. Cong, X. Liu, W. Zhang, J.P. Lewis, J.M. Tiedje, Q. Wang, Z. An, F. Wang, L. Zhang, T. Huang, C. Lu, Z. Cai, F. Wang and J. Zhang, *Innovation*, **2**, 100179 (2021); https://doi.org/10.1016/j.xinn.2021.100179
92. A. Pyrkov, A. Aliper, D. Bezrukov, Y.C. Lin, D. Polykovskiy, P. Kamya, F. Ren and A. Zhavoronkov, *Drug Discov. Today*, **28**, 103675 (2023); https://doi.org/10.1016/j.drudis.2023.103675
93. P.K. Barkoutsos, F. Gkritsis, P.J. Ollitrault, I.O. Sokolov, S. Woerner and I. Tavernelli, *Chem. Sci.*, **12**, 4345 (2021); https://doi.org/10.1039/D0SC05718E
94. B. Huang, N.O. Symonds and O.A. von Lilienfeld, eds.: W. Andreoni and S. Yip, Quantum Machine Learning in Chemistry and Materials, In: Handbook of Materials Modeling, Springer, Cham, pp 1883–1909 (2020); https://doi.org/10.1007/978-3-319-44677-6_67
95. M. Sajjan, J. Li, R. Selvarajan, S.H. Sureshbabu, S.S. Kale, R. Gupta, V. Singh and S. Kais, *Chem. Soc. Rev.*, **51**, 6475 (2022); https://doi.org/10.1039/D2CS00203E
96. A. Ajagekar and F. You, *Korean J. Chem. Eng.*, **39**, 811 (2022); https://doi.org/10.1007/s11814-021-1027-6
97. G. Huang, Y. Guo, Y. Chen and Z. Nie, *Materials*, **16**, 5977 (2023); https://doi.org/10.3390/ma16175977
98. A. Akinpelu, M. Bhullar and Y. Yao, *J. Phys. Condens. Matter*, **36**, 453001 (2024); https://doi.org/10.1088/1361-648X/ad6bdb
99. X. Jiang, S. Luo, K. Liao, S. Jiang, J. Ma, J. Jiang and Z. Shuai, *Cell Rep. Phys. Sci.*, **5**, 102049 (2024); https://doi.org/10.1016/j.xcrp.2024.102049