Chemistry Africa
https://doi.org/10.1007/s42250-025-01343-8

REVIEW

Artificial Intelligence in Computational and Materials Chemistry: Prospects and Limitations

David B. Olawade[1,2,3] · Oluwaseun Fapohunda[4] · Sunday Oluwadamilola Usman[5] · Abiola Akintayo[6] · Ayokunle O. Ige[7] · Yemi A. Adekunle[8,9] · Adedapo O. Adeola[10]

Received: 8 February 2024 / Accepted: 26 May 2025
© Crown 2025

Abstract
Computational chemistry, at the intersection of theoretical chemistry and computer science, employs various models to analyze molecular structures and properties, enabling the understanding and prediction of intricate chemical processes. The integration of artificial intelligence (AI) has revolutionized several fields, particularly in materials chemistry, with applications spanning drug discovery, materials design, and quantum mechanics. However, challenges related to quantum system complexity, model interpretability, and data quality remain a few of the Achilles' heel of AI applications. This paper provides an overview of AI's evolution in computational and materials chemistry, focusing on several applications. AI's transformative potential in materials chemistry is emphasized, facilitating precise material property predictions, crucial for industries reliant on materials innovation. In materials chemistry, AI has led to substantial advancements, enabling the rapid discovery of materials with tailored properties. Yet, the challenges of modeling complex quantum systems, achieving model interpretability, and accessing high-quality data remain. The integration of AI into computational and materials chemistry promises to reshape the field, revolutionizing chemical research, materials design, and technological innovation. In order to harness AI's full potential, transparent AI models, advanced quantum simulations, optimized data utilization, scalable computing, interdisciplinary collaboration, and ethical AI practices are essential.

Keywords Artificial intelligence · Computational chemistry · Modelling tools · Machine learning · Quantum chemistry · Materials chemistry

✉ David B. Olawade
d.olawade@uel.ac.uk

✉ Adedapo O. Adeola
adedapo.adeola@concordia.ca

[1] Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London, UK

[2] Department of Research and Innovation, Medway NHS Foundation Trust, Gillingham ME7 5NY, UK

[3] Department of Public Health, York St John University, London, UK

[4] Department of Chemistry and Biochemistry, University of Arizona, Tucson, USA

[5] Department of Systems and Industrial Engineering, University of Arizona, Tucson, USA

[6] Department of Chemistry and Biochemistry, The University of Texas at Dallas, Richardson, USA

[7] School of Computer Sciences, Universiti Sains Malaysia, Gelugor, Pulau Pinang 11800, Malaysia

[8] Department of Pharmaceutical and Medicinal Chemistry, College of Pharmacy, Afe Babalola University, Ado-Ekiti, Nigeria

[9] Centre for Natural Products Discovery, School of Pharmacy and Biomolecular Sciences, Faculty of Science, Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK

[10] Department of Chemistry and Biochemistry and the Centre for NanoScience Research, Concordia University, Montreal, QC H4B 1R6, Canada

Published online: 11 June 2025

1 Introduction

The development of artificial intelligence (AI) in computational and materials chemistry may be traced back to the early days of computer science when researchers first started using computers to model and predict the features of chemical systems [1]. Researchers started simulating and predicting the characteristics of chemical systems in the 1960s and 1970s using straightforward rule-based systems, such as expert systems. These early AI-based techniques had limited functionality and were mostly employed for straightforward tasks like predicting the boiling points of compounds. Researchers started utilizing more sophisticated AI approaches, like neural networks and genetic algorithms in the 1980s and 1990s, which led to the development of AI in computational chemistry [2]. Compared to the prior rule-based systems, these techniques were more potent and enabled more intricate simulations and predictions.

With the introduction of deep learning in the early 2000s, it became substantially easier to analyze and predict chemical properties [3]. Deep learning algorithms, including deep neural networks, may learn from vast datasets and produce predictions that are more accurate than those made by traditional AI techniques. Researchers are currently employing AI to carry out a wide range of activities, including drug discovery, materials design, and quantum chemistry. In recent years, the application of AI in computational and materials chemistry has continued to develop [4]. The use of active learning and transferred learning as well as the integration of physical models with AI are also becoming more common in the field, allowing for more effective data use and improved generalization of AI models to new systems.

This review discusses the challenges of integrating artificial intelligence (AI) into computational chemistry. It highlights the complexities that arise from the quantum nature of chemical systems, the interpretability issues in AI models, and high-quality data as a crucial requirement. By highlighting the historical progression from basic rule-based systems to sophisticated deep learning algorithms, the paper justifies the relevance of exploring AI's potential in the field. This paper also emphasizes the significant benefits that AI can bring to drug discovery, material design, and quantum chemistry, underlining the urgency of addressing the associated challenges.

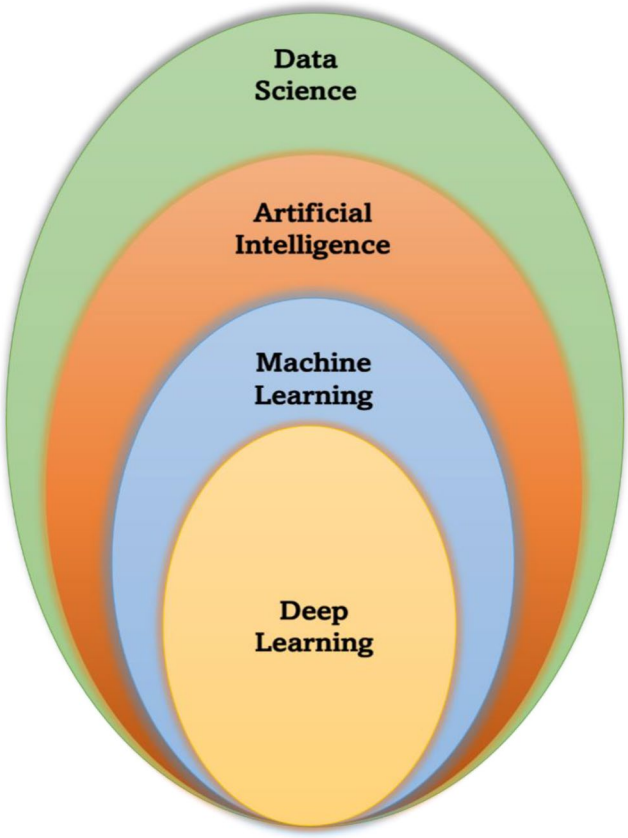

Fig. 1 Subset of data science

2 Description of AI Models

AI models have the potential to emulate human intelligence. These models are highly adept at detecting the pattern of datasets without human intervention. An AI model that accurately represents a dataset can make highly efficient and reliable predictions about future events. For instance, linear models such as logistic and linear regression (LR) are two famous ML computationally efficient and highly interpretable algorithms, which are categorized under the supervised learning models or techniques. They are considered to be the most straightforward ML predictive models. LR predictive models have been most recently employed in the prediction of clinical outcomes of COVID-19 mortality. Similarly, they are presently employed in the prediction of chronic diseases such as hypertension (HTN) and diabetes mellitus (DM), stroke risk, and in the prediction of acute myeloid leukemia outcomes based on the gene signature of a patient [5]. In many cases, AI models have proven to be highly effective in resolving complex problems, resulting in superior results, and saving time and cost. A typical example is the use of traditional methods for material screening, which is time-consuming. In clear contrast to the forgoing, AI models result in a shorter screening time, and enable predictive modeling, data-driven analysis and accelerated material screening. Research shows that the use of AI for material screening resulted in the synthesis of a large amount (four-fold) of halide perovskite single crystals through the inverse temperature crystallization process (ITC) [6]. Figure 1 depicts the various subsets of AI, which include machine learning (ML) which has gained prominence in the last two decades.

Table 1 Description of commonly used machine learning models

S/*N*	ML Model type	Description
1	Supervised learning	This learning method involves knowing both input and output. Models created using this type of learning utilize a combination of input and output data, also known as ground truth, to subsequently forecast future events. The Model's objective is to establish a precise correlation between the input and output data [7].
2	Unsupervised learning	This approach is based on unlabeled datasets, where inputs are provided without corresponding outputs. The primary objective of the model is to deduce patterns and structures from the input data, to comprehend the underlying distribution of the datasets [8].
3	Semi-supervised learning	In this learning technique, both labeled and unlabeled datasets are utilized, with the ratio of unlabeled to labeled data being significant. Typically, there is a smaller portion of labeled data, with the majority of the datasets being unlabeled [9].

Table 1 provides a categorization of the various kinds of ML models, which are facets of the AI model by default.

The selection of the appropriate ML model is determined by the nature of the datasets under consideration. Generally, most AI models deployed in computational chemistry use supervised learning, where the data label is readily available [10–12], among others. However, in cases where the data label is insufficient or sometimes unavailable, semi-supervised and unsupervised learning has been leveraged. For instance, a self-organizing map (SOM) is a clustering method used in unsupervised learning to map molecular representations [13]. Also, a transformer neural network for atomicmapping of products and reactants without data labelling has been used [14]. In the realm of computational chemistry, it is often the case that only input datasets are available, with predictions required to be made regarding future events.

One of the limitations of ML models is the issue of manual feature extraction, which is based on heuristics and often necessitates domain knowledge [15]. For this reason, recent works have adopted deep learning models such as Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) in computational chemistry. CNN can automatically extract spatial features, while RNN models can extract temporal features. A convolutional neural network is a deep feed-forward artificial neural network, often composed of numerous layers of distinct neural networks containing multiple neurons [16]. CNN works by extracting features over a specified number of convolutional filters. Another limitation is its inability to capture non-linear relationships/trends, which characterize image processing, video, speech, document reading, and handwriting recognition. Many real-world phenomena exhibit non-linear relationships/trends, meaning the expected output has no direct relationship to the input data. For instance, in performing such complex predictions involving non-linear datasets, several CNN-based optical characters and handwriting systems have been developed and deployed by Microsoft [17]. Similarly, CNN models have been employed in the image-processing-based intelligent defect diagnosis of rolling element bearings [18]. Generally, CNN relies on sparse interactions, equivariant representations, and parameter sharing. The architecture of the CNN layers is shown in Fig. 2.

CNN is indeed a powerful tool and one of the most commonly used deep learning methods in computational chemistry since it can automatically extract features. For instance, CNN has been used to decode the structure-odor relationship of chemical compounds [20]. Also, a graph CNN has been utilized to predict chemical reactivity [21, 22]. The authors

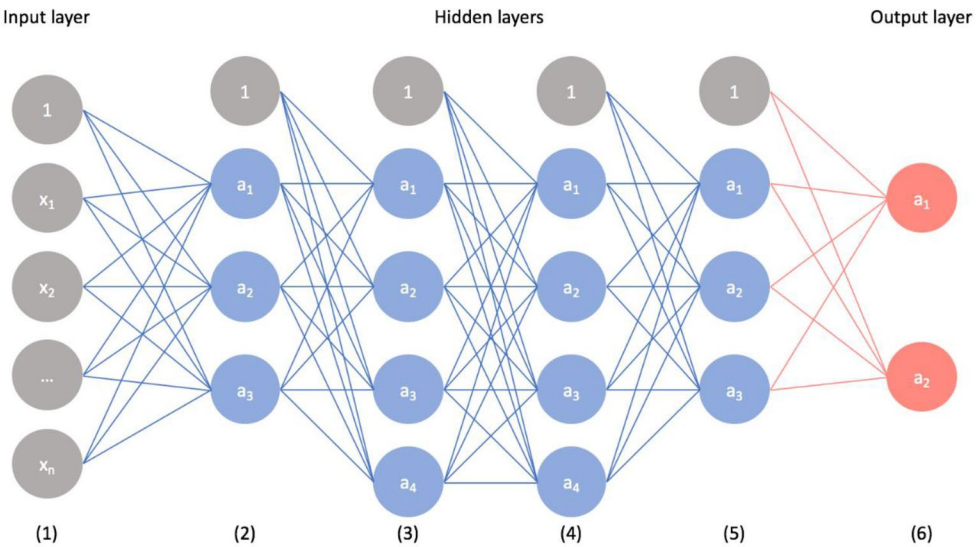

Fig. 2 Virtual representation of convolutional neural network. Adapted with permission from [19]

trained their model on thousands of reaction precedents, and the model was able to achieve 85% reaction prediction accuracy. However, in some cases whereby temporal features are important to improve the quality of features learnt by AI models, RNN models can also be leveraged. RNN models can capture temporal information from sequential data and retain temporal memory. RNNs consist of the input layer, hidden layers with multiple nodes, and the output layer.

Generally, RNN experiences explosive and disappearing gradient issues. For this reason, variants of RNNs such as Long-Short Term Memory (LSTM) and Gated Recurrent Unit (GRU), among others have been leveraged in computational chemistry. For instance, LSTM with Computational Fluid Dynamics has been employed for source localization of chemicals [23]. Some other works have combined machine learning models with deep learning models [24], where the authors proposed an ensemble model of SVM, CNN and RNN to extract chemical-protein relationships.

3 Types of Datasets Used To Train AI Models in Computational Chemistry

The success of AI models in computational chemistry is deeply tied to the availability of high-quality, diverse datasets. These datasets allow AI algorithms, particularly machine learning and deep learning models, to identify patterns within chemical systems, enabling them to make predictions, simulate reactions, and discover new materials. The types of datasets used in computational chemistry are varied and come from multiple sources, each serving distinct purposes.

3.1 Quantum Chemistry Datasets

Quantum chemistry datasets are crucial for modeling molecular properties based on quantum mechanics. These datasets contain atomic configurations, molecular geometries, and electronic structures, which AI models use to predict energy states, electronic properties, and molecular reactivity. Examples such as QM7 and QM9, provide quantum mechanical properties for small organic molecules, this also includes atomization energies and polarizabilities, which help AI models in simulating molecular properties without resorting to computationally intensive quantum mechanical calculations [25]. These datasets are used in drug discovery and materials science, enabling the development of new molecules and reactions through predictive modeling [26]. Similarly, the ANI-1 dataset, which offers data from high-quality density functional theory (DFT) calculations for organic molecules, provides detailed potential energy surfaces for a wide range of organic molecules. These datasets enable AI models to comprehend quantum interactions at the atomic scale, enhancing the precision of predictions regarding molecule interactions, reaction processes, and electrical characteristics in chemical systems [26]. By analyzing these quantum chemistry datasets, AI models facilitate the advancement of research in areas such as catalysis and molecular design, hence enhancing predictions in chemical systems and fostering innovation across diverse industries, including pharmaceuticals and renewable energy [21, 26].

3.2 Chemical Reaction Datasets

Chemical reaction datasets include both experimental and simulated reaction data, such as reaction pathways, transition states, and kinetics. AI models use this data to predict reaction outcomes or optimize reaction conditions. The USPTO Reaction Dataset contains millions of chemical reactions from U.S. patent filings, which are widely used for reaction prediction tasks. Other datasets like Reaxys and SciFinder provide experimental reaction data from scientific literature, offering a vast array of real-world chemical reactions that are essential for AI model training in reaction discovery and synthesis planning. For instance [21], illustrates the utilization of machine learning models, notably graph-convolutional neural networks (G-CNN), in forecasting the results of chemical interactions, thus facilitating the drug development process.

3.3 Materials Property Datasets

In materials chemistry, datasets provide critical information about material compositions, crystal structures, and properties such as band gaps, elastic moduli, and thermal conductivity. These datasets are used for designing new materials and optimizing existing ones. Notable examples include the Materials Project Database, which provides data on thousands of inorganic compounds and their computed properties like formation energies and band structures. Another important source is the Open Quantum Materials Database (OQMD), which focuses on density functional theory (DFT)-calculated material properties, particularly for crystalline materials. Such datasets allow AI to predict material properties and identify promising new materials with desired characteristics. The Materials Project Database and machine learning algorithms have been utilized in the development of novel solid-state battery materials. Utilizing datasets that encompass material compositions, crystal structures, and properties like band gaps and formation energies, researchers at Stanford University successfully trained an AI model to predict the ionic conductivity of diverse materials, a critical attribute of battery performance. The

researchers effectively identified lithium superionic conductors exhibiting high ionic conductivity, rendering them optimal candidates for next-generation solid-state batteries. This notably expedited the discovery process, as the conventional experimental method would have required substantially more time to evaluate all potential materials [27].

3.4 Molecular Dynamics Datasets

Molecular dynamics (MD) datasets are essential for studying time-dependent behaviors of molecules. These datasets provide information on molecular motion, interactions, and forces over time, enabling AI models to simulate dynamic molecular behavior and predict future states. An example is the MD17 dataset, which contains molecular trajectories of small molecules interacting over time, generated from ab initio molecular dynamics simulations. This type of dataset is valuable for training AI models that aim to predict molecular interactions and forces in real-time applications. A significant instance is researchers from ETH Zurich and TU Berlin who utilized deep neural networks (DNNs) on the MD17 dataset, comprising molecular trajectories and forces derived from ab initio molecular dynamics simulations of tiny molecules. Artificial intelligence models were developed using this information to forecast molecular forces and interactions, with the objective of minimizing the computational expenses linked to conventional quantum mechanical techniques like density functional theory (DFT). The result was the AI models' capacity to anticipate forces and energy with near quantum-chemical precision, attaining considerable accelerations relative to conventional quantum mechanical computations. This methodology facilitated real-time forecasting of molecular behavior, essential for applications including drug design, materials science, and molecular simulations [28].

3.5 Toxicity and Bioactivity Datasets

For AI models focused on drug discovery, datasets containing bioactivity, toxicity, and pharmacological properties of small molecules are critical. These datasets help predict a compound's interaction with biological targets or its potential toxicity. Examples include the ChEMBL Database, which contains bioactivity data for thousands of drug-like molecules, including their interactions with various biological targets. Another example is the Tox21 Dataset, a resource for predicting the toxic effects of small molecules, with toxicity screening data for thousands of compounds. In 2020, IBM Research employed AI models utilizing datasets such as ChEMBL and Tox21 to forecast therapeutic efficacy and possible adverse effects. Through the analysis of extensive datasets of chemical substances, the AI model recognized potential therapeutic candidates and indicated those likely to produce undesirable effects. This method accelerated the medication development process by early prediction of toxicity, hence decreasing both time and expenses. The AI models demonstrated exceptional accuracy in forecasting drug-target interactions and toxicity, hence improving the production of safer and more effective pharmaceuticals [29].

4 Sources of Datasets for AI Models in Computational Chemistry

The datasets used to train AI models in computational chemistry and materials science come from a variety of sources, each contributing uniquely to the advancement of AI-driven research. These sources range from publicly accessible databases to proprietary industrial collections, providing diverse data for a wide array of applications, including drug discovery, materials design, and reaction prediction.

4.1 Open-Source Databases

Open-source databases are crucial for advancing AI research in computational chemistry and materials science. These databases offer free and public access to large collections of data, often derived from both experimental and computational efforts. They are continuously updated with new entries, making them invaluable for training AI models that require large and diverse datasets. PubChem is one of the largest open-source chemical databases, providing information on the biological activities of small molecules. It includes chemical structures, properties, activities, and assay results, which are widely used for AI-driven drug discovery, toxicity prediction, and molecular property analysis. PubChem's comprehensive data coverage makes it a key resource for building AI models that predict interactions between small molecules and biological targets. The PubChem Database has been utilized to create AI models for drug development, toxicity assessment, and molecular property evaluation. The team utilized PubChem's extensive library of chemical structures, characteristics, and biological activities to train AI models for predicting the binding affinity of small compounds to biological targets and evaluating possible toxicity. This method facilitated enhanced virtual screening of drug candidates and increased prediction accuracy for molecular interactions, hence diminishing the necessity for experimental testing [30].

The **Materials Project** is another significant open-source platform, offering computational data on thousands of inorganic materials. This resource is vital for AI models tasked with predicting material properties, such as band

gaps, formation energies, and mechanical stability. The use of AI in materials discovery has grown significantly due to the Materials Project's accessible and detailed database, accelerating the design of new materials with optimized properties for applications like batteries, photovoltaics, and thermoelectrics. **ChEMBL**, a widely used open-source database, provides bioactivity data on drug-like molecules. AI models in drug discovery and pharmacology often rely on ChEMBL's high-quality data to predict the efficacy and safety of new compounds. Regularly updated with experimental data from scientific literature, ChEMBL serves as a foundation for building AI models that assist in virtual screening and lead optimization. Researchers at Lawrence Berkeley National Laboratory utilized AI models trained on data from the Materials Project to forecast essential material properties, including band gaps, formation energies, and mechanical stability. This extensive computational dataset facilitated the expedited discovery of materials tailored for applications such as batteries, photovoltaics, and thermoelectrics. The AI models facilitated the identification of novel, high-performance materials, diminishing dependence on conventional experimental techniques and accelerating materials innovation [31].

4.2 Experimental Data

Experimental data is critical for creating AI models that reflect real-world chemical and material behavior. These datasets, derived from lab experiments, include information such as crystal structures, spectroscopic results, and biological assays. Experimental data improves AI model accuracy by providing reliable ground truth for training and validation. The Protein Data Bank (PDB) is an essential resource for structural biology, offering detailed atomic structures of biological macromolecules like proteins, nucleic acids, and complexes. AI models in drug discovery and molecular biology use PDB to predict the interactions between proteins and small molecules or to model protein folding and dynamics. The use of PDB data has enabled AI-driven advancements in understanding biological mechanisms and developing novel therapeutics. Spectroscopic data, such as NMR, IR, and UV-Vis spectra, published in scientific journals, also serve as critical sources of experimental data. These datasets are particularly valuable for AI models focused on structural elucidation and functional group identification in organic and inorganic compounds. Incorporating spectroscopic data helps AI models make more precise predictions in chemical characterization and materials identification. The AlphaFold AI system, created utilizing data from the Protein Data Bank (PDB), attained unprecedented precision in forecasting protein shapes. Utilizing atomic-level data from the PDB, AlphaFold successfully modeled protein folding, greatly enhancing the comprehension of protein-ligand interactions. This advancement has transformed drug discovery by enabling more accurate predictions of protein structures, resulting in the creation of tailored treatments [32].

4.3 Simulated Data

Simulated data is becoming increasingly prevalent as computational chemistry tools such as Gaussian, VASP (Vienna Ab initio Simulation Package), and Quantum ESPRESSO generate high-quality datasets for quantum mechanical and molecular dynamics simulations. These datasets offer valuable information on electronic structures, molecular interactions, and energy states, which are difficult to obtain experimentally. Quantum mechanical simulations, such as those performed by VASP, provide data on the electronic structure and material properties of complex systems. AI models can be trained on these datasets to predict molecular reactivity, electronic behavior, and material stability. For example, by training on datasets from simulations, AI models can predict the outcomes of chemical reactions or the properties of new materials faster and more accurately than traditional methods. Similarly, molecular dynamics (MD) simulations using tools like LAMMPS or GROMACS generate large volumes of time-dependent data on the movement and interactions of atoms and molecules. These datasets are essential for training AI models that predict dynamic behaviors in materials and biomolecules, such as protein folding or the diffusion of ions in solid electrolytes. Simulated data from tools like as VASP (Vienna Ab initio Simulation Package) has been utilized to train AI models for predicting material qualities, encompassing electrical structures and stability. Utilizing VASP-generated datasets, AI models successfully predicted chemical reactivity and the electrical properties of intricate systems, thereby expediting the development of novel materials for energy storage and electronic applications. The incorporation of simulation data-enabled AI models to surpass conventional methods in both velocity and precision for forecasting chemical reactions and material characteristics [33].

4.4 Patent and Literature Databases

Patent and literature databases, such as USPTO, Reaxys, and SciFinder, provide vast amounts of chemical reaction data extracted from patents and published scientific articles. These datasets are rich in experimental results, including reaction conditions, pathways, and yields, which are invaluable for AI models focused on chemical synthesis and reaction prediction. The USPTO (United States Patent and Trademark Office) database contains millions of

chemical reactions and synthetic procedures, making it a critical resource for AI models in retrosynthesis and synthetic pathway planning. AI models trained on patent data can predict the feasibility of chemical reactions or propose novel synthetic routes, reducing the time and cost associated with experimental trial and error [34].

Reaxys, a comprehensive chemical reaction database, provides data on reaction mechanisms, compound properties, and experimental procedures. It is widely used by AI models in pharmaceutical and materials chemistry to optimize reaction conditions and discover new synthetic routes. SciFinder is another essential resource, offering access to a vast collection of chemical reactions and scientific literature. AI models leverage SciFinder data to predict reaction outcomes, identify reaction pathways, and assess the scalability of synthetic methods. Recent developments have enabled AI models to utilize Reaxys for the analysis of extensive reaction data, resulting in enhanced synthesis routes and reaction outcomes, particularly in intricate organic reactions [34].

4.5 Proprietary and Private Databases

Proprietary and private databases, maintained by companies and research institutions, offer unique datasets derived from years of experimental research and development. These datasets are typically not publicly available but are crucial for industrial applications of AI in fields like pharmaceuticals, materials science, and chemical engineering. Pharmaceutical companies, for example, often maintain proprietary datasets of compound libraries, bioactivity results, and clinical trial data. These datasets are used to train AI models that predict drug efficacy, safety, and toxicity. By leveraging proprietary data, companies can develop AI models that outperform those based solely on publicly available datasets, giving them a competitive edge in drug discovery. Pharmaceutical firms utilize private datasets, including compound libraries and bioactivity information, to train AI models for drug discovery. AstraZeneca utilized its own chemical and biological databases to create AI models that accurately anticipate medication toxicity and efficacy. These models were employed to evaluate prospective drug candidates more effectively, minimizing the duration and expense of the drug development process while enhancing safety forecasts relative to models constructed solely on public datasets [35]. Similarly, in materials science, companies may maintain proprietary databases of material compositions, properties, and performance data. These datasets enable the development of AI models that predict the performance of new materials in real-world applications, such as energy storage or structural engineering. Access to proprietary data allows companies to accelerate the discovery and optimization of materials for industrial use. Corporations such as BASF have created proprietary databases that encompass extensive information on material compositions, characteristics, and actual performance. These datasets were utilized to build AI models that forecast the efficacy of novel materials in applications including battery technology and structural engineering. Utilizing proprietary data enabled BASF to expedite material discovery and optimization, resulting in advancements in high-performance materials for industrial applications [36].

5 AI Model Validation in Computational and Material Chemistry

AI models would be beneficial if the results generated from them are reliable, hence the importance of validation. Models need to be validated before their application as useful tools in computational and/or materials chemistry. The model validation process ensures that AI models perform as expected, both in terms of meeting the designed objectives and meeting end-user requirements. In computational and material chemistry, validation techniques are particularly helpful in determining whether a model performs accurately. AI models are frequently used in the prediction of physical properties like band gaps, heat capacity, or catalytic activity of various materials. A prominent example is the use of ML algorithms in predicting the bad gaps in organic photovoltaic materials, where AI models trained based on large datasets have demonstrated the ability to outperform the traditional density functional theory (DFT) calculations with respect to the time of completion and accuracy [36]. For instance, ML models for predicting the dielectric breakdown strength of polymeric materials have been developed [16]. Data-driven approaches have been used in predicting polymer properties based on their molecular structures and in validating the model by benchmarking its predictions against data obtained from experiments, thereby achieving a high degree of accuracy. To the best of our knowledge, it is a crucial step in the development and deployment of AI systems that are robust and reliable in the field of computational chemistry. It improves model accuracy, reliability, and performance [37]. Furthermore, model validation is equally important due to the problems associated with fitting datasets. Often, the fitted datasets work perfectly for trained data and perform poorly on test data. Overfitting is the term used to describe this phenomenon. Additionally, there are situations in which model complexity is not properly accounted for. This also leads to poor results, and it is known as underfitting. However, it is important to examine a system's entire life cycle when selecting a validation method [38].

Presently, there are several methods for AI model validation, and the method selection largely depends on the nature of the datasets used and the problem being faced. Different validation methods have their pros and cons and some of these methods perform better than others. However, a lot of time should be allocated to the validation method selection since it has a great influence on the type of results. The chosen validation method should be robust and accurate when applied to test datasets. Given similar datasets and various validation techniques, the accuracy may vary. According to [39] and [40], the commonly used methods for AI model validation are holdout validation [41], K-fold cross-validation [22], and Leave-one-out cross-validation. Other methods include re-substitution, where the entire dataset is utilized for training the model and the error rate is estimated from the actual and predicted values within the datasets. These are only a few of the numerous techniques for validating AI models. It is crucial to validate the model using several different techniques and to present the results with the proper statistical significance and measurements of uncertainty [42].

6 Application of AI Models in Material Design

The prediction of material qualities, enhancing structures, and selecting appropriate precursors are just a few of the ways AI may be applied to the creation of novel materials [43]. To predict the properties of new materials, machine learning algorithms can be trained on data from existing materials [44]. New materials with certain qualities, like high thermal conductivity, stability and/or tensile strength, can be created this way [45, 46]. For instance, neural networks can be trained to predict a material's band gap, the density of states, or formation enthalpy. Figure 3 shows a stepwise illustration of the workflow for desegregating AI in computational and material chemistry. The overall process begins with the collection of data from relevant material databases; then data preprocessing and further data engineering which results in samples required for model building. Several ML techniques, including but not limited to regression, classification, and clustering, are often applied in model development. The resulting models are then evaluated by subjecting them to an evaluation method, which may be the cross-validation, and bootstrapping, with performance metrics including CA, MAPE, RMSE, and R^2. These processes result in scientific knowledge feedback, enhancing understanding of material properties and structure.

Computational models such as neural networks and linear regression have been reportedly used in designing polymer-based organic photovoltaics (OPVs) which have the potential to replace silicon-based solar cells. OPVs are conductive organic polymers that can produce electricity from sunlight [47]. They have the advantages of cost-effectiveness, low density, ease of processability, increased mechanical flexibility and toughness, and reduced energy consumption compared to conventional solar cells [48]. AI models also contributed to the design of organic light-emitting diodes (OLEDs). Multivalent calcium ion batteries (CIBs) have enlarged capacities and high operating voltages over monovalent lithium-ion batteries. A four-step computational workflow for the development of novel multivalent CIBs has been devised [49]. To achieve this, high-throughput DFT was used to screen 357 metal−calcium binary and ternary compounds to discover CIB anodes with higher properties. The tin electrochemical calciation reaction was conducted using density functional theory before being validated experimentally [49].

AI can also be used to optimize a material's structure such that it has a specific desired property. Propertiessuch as high conductivity, high strength, reactivity, surface area and the ideal atomic arrangements can be determined using techniques like genetic algorithms or Bayesian optimization. AI

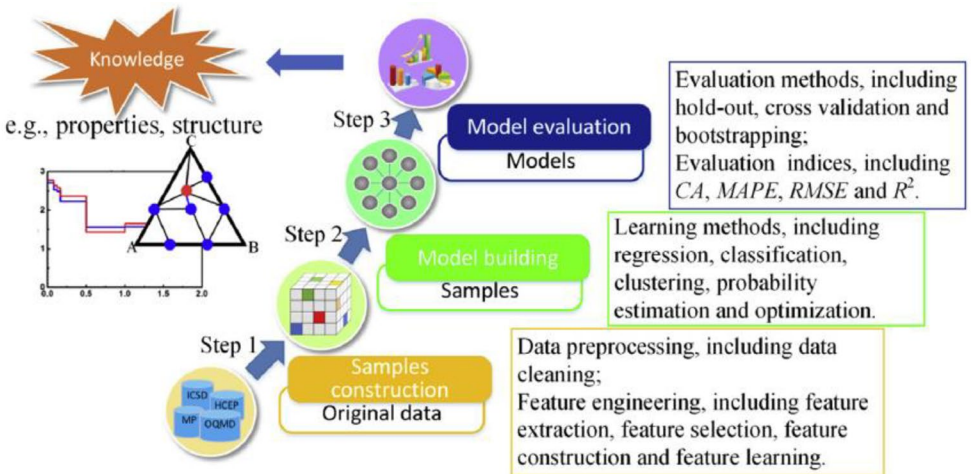

Fig. 3 AI-driven workflow for material property prediction - adapted with permission from [45]

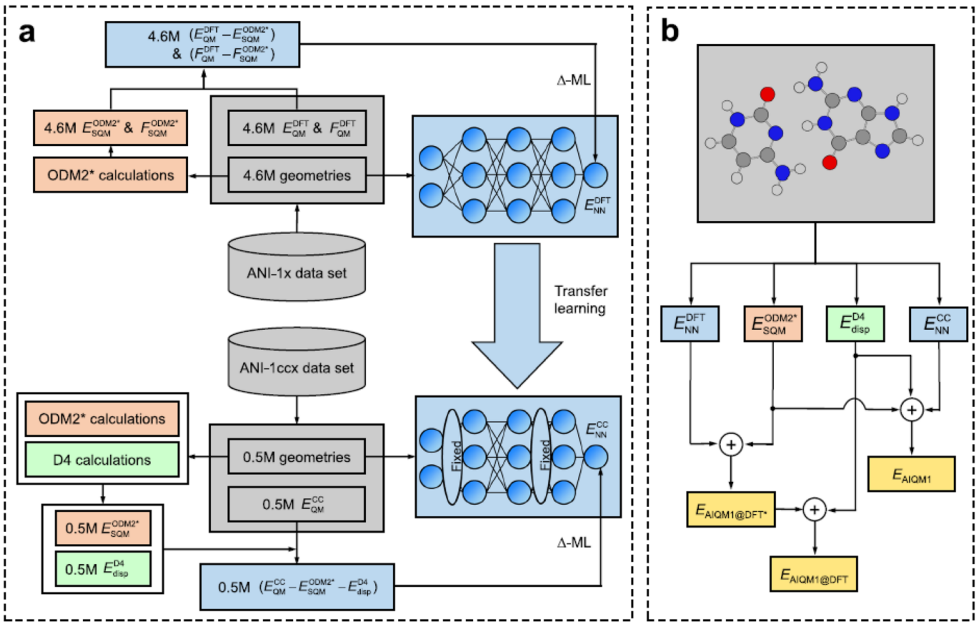

Fig. 4 The flowchart of a general-purpose and AI-enhanced quantum chemical calculation method. Adapted with permission from [57]

can also be employed to discover novel materials by searching through enormous databases of well-known chemicals and looking for patterns that point to a novel class of materials [50]. Techniques like unsupervised learning, which can find patterns in data without the need for - labeling, can be used to do this. Additionally, by predicting their qualities and then synthesizing them in a lab, AI can be used to create novel materials [51, 52]. The field of study in AI-based material design is expanding quickly, and it has the potential to completely change how materials are created and discovered. Scientists and engineers may swiftly and effectively find new materials with certain characteristics with the aid of AI-based material design, and they can also modify the structure of current materials to produce properties.

7 Application of AI Models in Quantum Chemistry

Research on the application of artificial intelligence (AI) in quantum chemistry has improved in recent years. Understanding chemical reactions and the behavior of materials depends on quantum chemistry, which applies quantum mechanics to chemical and biological systems, especially in the study of the wavefunction-based electronic structure, chemical kinetics, and characteristics of matter. This is achieved by utilizing methodical approximation calculations, which are computationally viable for the interpretation of molecular framework and projection of chemical reactivity [12]. Figure 4 summarizes the AI-enhanced quantum chemical calculation method. In quantum chemistry, one of the key applications of AI is to predict the characteristics of molecules [53]. Machine learning methods, including neural networks, have been used by researchers to forecast the energies and electron distributions of molecules [11, 54, 55]. This can assist in locating molecules in stable configurations and predicting the results of chemical events [56].

Simulating the kinetics of chemical reactions is another application of AI in quantum chemistry [58]. Researchers have studied the processes of chemical reactions and predicted the consequences of these events using AI-based techniques, such as neural networks [57, 59]. AI is also being used to enhance the effectiveness and precision of quantum simulations. To approximate the interactions between electrons and nuclei, for instance, researchers are employing quantum machine learning and neural network potentials, which can considerably lower the processing cost of quantum simulations [57]. Despite advancements, the complexity of quantum systems and the expense of replicating them continue to be a barrier to the application of AI in quantum chemistry. Similarly, AI models often struggle with larger, more complex molecules or materials due to the exponential increase in quantum interactions as the system grows. To address these challenges more effectively, researchers are still attempting to create reliable, understandable, and generalizable AI models [60, 61].

8 Application of AI Models in Medicine

Drug discovery is one of the most exciting applications of AI in computational chemistry [62]. The activity of certain targets, such as proteins and enzymes, and possible drug candidates against them has been predicted using AI-based

techniques [63–68]. For instance, researchers have been able to predict the activity of small molecules against a variety of targets using deep neural networks (DNN), a machine-learning method that utilizes vast datasets [69–74]. DNNs have also been used to forecast a compound's toxicity, which is crucial for developing drugs that are both safe and effective [35, 75–77]. Drug design has experienced a boost in ease of operations, particularly in the modeling of the interaction of drugs with various biological and non-biological species since the advent of computational chemistry in the late 1980s. Machine learning has been an integral part of modern drug design. Before experimental validations, several aspects of drug discovery such as databases search for hits, molecular descriptors optimization of hits to lead compounds, three-dimensional data of proteins and ligands, bioactivity predictions against desirable biological targets, toxicity predictions, and mechanisms by which pharmacological actions are elicited can be conducted by AI models [62]. Machine learning models, QSAR were used to identify inhibitors of kallikrein 5 protease [59] and Bayesian algorithms identified inhibitors of G-protein coupled receptors (cannabinoid receptors, CB_2) [78].

The creation of reliable and understandable models for drug discovery is still challenging, as many of the current AI-based techniques are not transparent in the sense that researchers do not yet properly understand their judgment or decision-making processes. An additional challenge faced by researchers in the creation of reliable and understandable models specifically for drug discovery is the fact that these techniques usually require a significantly large amount of data, which can sometimes be challenging to acquire, particularly for uncommon or highly specified objectives [17]. Researchers are investigating the use of additional AI techniques, such as rule-based systems and Bayesian networks [79], which offer greater transparency and data efficiency, to get around these restrictions [80, 81].

9 Integration of Physical Model in Computational Chemistry

An emerging field of study in computational chemistry is the integration of physical models with AI, to maximize the benefits of both physical models and AI-based approaches to enhance the comprehension and prediction of chemical systems [82]. Improving the interpretability and dependability of AI models is one of the key advantages of integrating physical models with AI. The fundamental understanding of the underlying chemical processes is provided by physical models, which can also be used to constrain the predictions of AI models and increase their physical significance. Additionally, this can aid in recognizing and overcoming the limitations of AI models. The ability to generalize AI models to new systems is another advantage of combining physical models with AI. Physical models can offer a prior knowledge about the chemical system, which can be used to direct AI models throughout their learning process and enhance their capacity to predict outcomes for novel systems reactions [83]. An instance of such was considered to be a learning-based AI model armed with model-based and model-free methods' integration of intuitive physics and having core ingredients that can crucially foster learning and cognition similar to those of humans [84]. Applicable expressions can be expected in many areas including embodied AI and deep learning for comprehending and interpreting image scenes rather than basic algorithms for object recognition [85].

In computational chemistry, there are numerous instances of how physical models and AI are combined. Researchers have used physical models of material properties to direct the design of novel materials using evolutionary algorithms and have used physical models of chemical reactions to constrain the predictions of neural network models. It has been reported that DeepMind's protein-folding deep learning AI (AlphaFold2) was utilized in 2020 to challenges associated with protein structure prediction to the width of an atom at the Critical Assessment of Protein Structure Prediction (CASP14) [86]. Computational structural chemists and biologists continue to explore AI tools in making advancements to solve the next pressing biological challenge which includes the prediction of the various forms that some proteins can take and the determination of the interactions between proteins and other molecules [87].

Despite the advancements made in this area, there are still problems that need to be solved, such as how to successfully combine physical models with AI models and how to make sure the resulting models are reliable, understandable, and generalizable. Future research in this field should concentrate on creating fresh techniques for fusing AI with physical models and using such techniques on a variety of chemical systems. In conclusion, computational chemistry research is actively pursuing the integration of physical models with AI, which has the potential to significantly enhance comprehension and prediction of chemical systems. It can deliver more trustworthy and comprehensible AI models and enhance their adaptability to new systems. Future growth in this field of study is anticipated, with computational chemistry becoming more dependent on the incorporation of physical models and AI.

10 Facilitation of Data Efficiency in Artificial Intelligence

Artificial intelligence (AI) has the potential to significantly increase data efficiency in computational chemistry. AI methods like active learning and transfer learning can be used to do this. Instead of being given a predetermined dataset, active learning involves the AI model actively choosing the data it wants to learn from. This can be especially helpful in computational chemistry, where it might be expensive to generate data from simulations or experiments. The AI model can reduce the overall number of samples needed for training by learning from a small subset of the data using active learning and then choosing the samples that are the most informative to learn from next [88].

Another strategy for improving data efficiency in computational chemistry is transfer learning. It entails applying an AI model that has already been trained to learn from a fresh task or dataset. When there is a data gap for a particular chemical system or operation, this can be especially helpful. The new model can learn from fewer samples and produce results like those of a pre-trained model by transferring knowledge from the trained model. AI models can also be used to make better use of the data that is already available in addition to active learning and transfer learning. These methods include data augmentation, which artificially increases the size of the dataset, and combining multiple datasets to increase the diversity of the data. Using AI to improve computational chemistry's data efficiency is a growing field of study that has the potential to drastically reduce the quantity of data needed to develop AI models and make predictions. To maximize the use of the data at hand and cut down on data waste, strategies including active learning, transfer learning, data augmentation, and merging several datasets can be used. Active learning, transfer learning, data augmentation, and the combination of numerous datasets can be used to make better use of the available data, hence decreasing the cost of experiments and simulations and allowing researchers to examine a broader range of chemical systems.

11 Scalability of the AI Models in Computational and Materials Chemistry

For large-scale simulations and predictions the application of artificial intelligence (AI) in computational chemistry necessitates extensive computer resources. The ability of AI models to handle big datasets and intricate simulations makes scalability a crucial factor that must be considered when applying AI to computational chemistry. Using distributed computing and cloud-based resources is one method to improve scalability. Due to the ability to distribute the computation across numerous computers, processing massive datasets and intricate simulations can be done in a shorter time. Another strategy is to use high-performance computing (HPC) systems, which can significantly speed up the development and prediction of AI models. Examples of these systems include tensor processing units (TPUs) and graphical processing units (GPUs) [89].

Employing lower precision data representation and computations, such as using 16-bit or 8-bit floating-point numbers rather than 32-bit or 64-bit integers, is an alternative strategy that can reduce memory usage and speed up processing [64]. Additionally, researchers are investigating the usage of edge computing, which enables the deployment of AI models on edge devices rather than in the cloud, such as smartphones and IoT devices. This can decrease latency, communication costs, and improve the security and privacy of the data. As a result, scalability is crucial when utilizing AI in computational chemistry because it enables the AI models to manage sizable datasets and intricate simulations. One strategy to improve the scalability of AI in computational chemistry is to leverage distributed computing and cloud-based resources, high-performance computing systems, low-precision data representation and processing, and edge computing.

12 Limitations of AI in Computational and Materials Chemistry

While computational chemistry's use of artificial intelligence (AI) has the potential to completely change how chemical systems research is carried out, there are also some limitations including the complexity of quantum systems, the difficulty of interpreting AI models, and the requirement for high-quality data. The complexity of quantum systems is one of the key AI in computational chemistry limits. Chemistry is built on quantum mechanics; however, this is a very difficult and abstract subject. Due to this complexity, modelling and predicting the behaviour of quantum systems using AI is challenging since the models must be able to faithfully represent the underlying physics. The issue of using AI to simulate complicated quantum systems persists despite recent advancements in the field. The inability of AI models to be understood is another drawback in computational chemistry. Numerous AI models, including neural networks, are regarded as "black boxes," making it challenging to comprehend how they produce predictions. Due to this, it may be challenging to comprehend the underlying chemical processes that the model is capturing as well as to spot any flaws or limits in the model. Finally, computational chemistry uses AI but needs high-quality data [90]. The accuracy

and generalizability of AI models may be constrained by the caliber and amount of accessible data. Particularly important are precise, comprehensive datasets that are typical of the chemical systems under investigation. It might be challenging to build robust and trustworthy AI models without high-quality data.

It is noteworthy that the necessity for high-quality data, the complexity of quantum systems, and the lack of interpretability of AI models are some of the primary issues that must be resolved to fully realize the potential of AI in computational chemistry. The complexity of quantum systems, the inability of AI models to be understood, and the requirement for high-quality data are the main limitations of AI in computational chemistry. To overcome these constraints, more work will be needed in the area, including the design of interpretable AI models, novel approaches to modeling and simulating quantum systems, and the gathering and curation of high-quality datasets. Despite these drawbacks, computational chemistry's application of AI holds the potential to significantly improve understanding of chemical processes and hasten the development of new materials and drugs.

In addition to the mentioned limitations, further challenges persist in the context of material chemistry. One significant limitation is the representation of complex molecular structures. While AI models have made significant strides in predicting molecular properties, representing intricate chemical structures accurately remains a challenge. Materials often exhibit a wide range of structural diversity, and capturing this diversity in AI models requires substantial data and computational resources. This limitation hinders the comprehensive study of complex materials with unique structures. Moreover, the transferability of AI models across different material systems is a noteworthy concern. AI models trained on specific materials may struggle to generalize their predictions to entirely different material classes. This limitation poses a challenge when attempting to apply AI-driven insights to novel materials with distinct characteristics. Developing AI models that can seamlessly adapt and generalize across diverse material chemistries is a pressing research area. Furthermore, the scalability of AI-driven simulations for material chemistry is constrained by computational resources. Simulating the behavior of materials at the atomic or molecular level demands extensive computational power. While AI has the potential to enhance the efficiency of simulations, the computational cost associated with detailed material modeling remains a bottleneck. Overcoming this limitation requires innovations in algorithmic efficiency and access to high-performance computing infrastructure.

Overall, the application of AI in material chemistry faces challenges related to accurately representing complex molecular structures, transferring AI models across material classes, and managing the computational demands of detailed simulations. Addressing these limitations will be pivotal in realizing the full potential of AI-driven advancements in material chemistry and accelerating the discovery of new materials with tailored properties.

13 Conclusion and Recommendations

Despite the potential of AI in computational chemistry, including its application in materials chemistry, several challenges and limitations still exist. These challenges encompass the complexity of quantum systems, the interpretability of AI models, and the necessity for high-quality data. In the realm of materials chemistry, these limitations can impact the discovery and optimization of materials with specific properties critical for various industries. Understanding and predicting the behavior of quantum systems are essential in materials chemistry. However, the intricate nature of quantum mechanics poses a significant challenge. Modeling and simulating these complex quantum systems using AI methods remain a formidable task, necessitating innovative approaches to bridge this gap effectively. AI models, such as neural networks, are often viewed as "black boxes." This lack of interpretability can hinder researchers from comprehending the underlying chemical processes captured by the model. In materials chemistry, where precise material properties are crucial, understanding how AI models arrive at predictions is essential. High-quality and comprehensive datasets are the backbone of robust AI models.

In materials chemistry, where variations in material properties are vast, the availability of precise, well-curated data is paramount. Limited access to such data can constrain the accuracy and generalizability of AI models. To address these limitations, several recommendations may be considered:

I. Enhance Model Transparency and Interpretability: Researchers should focus on developing AI models that are more transparent and interpretable. This would aid in understanding model predictions and building trust among practitioners and researchers.
II. Advanced Quantum Simulation Techniques: Addressing the complexities of quantum systems requires innovative approaches to quantum simulations. Researchers should explore new methodologies and algorithms that combine AI and quantum mechanics to enable more accurate and efficient predictions.
III. Curate High-Quality Datasets: High-quality and comprehensive datasets are essential for training robust AI models. Collaborative efforts to gather, curate, and share datasets specific to various chemical systems can greatly benefit the field.

IV. Explore Hybrid Models: Combining physical models with AI can lead to more accurate and reliable predictions. Researchers should continue to explore ways to effectively integrate these two approaches and maximize their benefits.
V. Optimize Data Efficiency Strategies: Active learning, transfer learning, and data augmentation hold promise for improving data efficiency. Researchers should continue to develop and refine these strategies to make the most of available data.
VI. Utilize Scalable Computing Resources: To handle large-scale simulations, researchers should leverage distributed computing, high-performance computing, and cloud-based resources. These technologies can significantly enhance scalability and computational efficiency.
VII. Collaborate Across Disciplines: The interdisciplinary nature of AI in computational chemistry calls for collaboration between chemists, physicists, computer scientists, and domain experts. Collaborative efforts can accelerate progress and address complex challenges.
VIII. Promote Ethical AI Practices: As AI technologies advance, ethical considerations become increasingly important. Researchers should prioritize ethical AI practices, transparency, and responsible data usage in their work.
IX. Educate and Train Researchers: Training programs and educational initiatives should be established to equip researchers with the skills needed to effectively apply AI in computational chemistry. This can foster innovation and ensure the responsible use of AI technologies.
X. Continued Research and Innovation: The field of AI in computational chemistry is still evolving. Researchers should continue to explore novel techniques, methodologies, and applications to push the boundaries of what AI can achieve in understanding and predicting chemical systems.

Declarations

Conflict of Interest The authors declare that there is no conflict of interest regarding the publication of this article.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Hiller SA, Golender VE, Rosenblit AB, Rastrigin LA, Glaz AB (1973) Cybernetic methods of drug design. I. Statement of the problem—the perceptron approach. Comput Biomed Res 6(5):411–421
2. Venkatasubramanian V, Chan K, Caruthers JM (1995) Evolutionary design of molecules with desired properties using the genetic algorithm. J Chem Inf Comput Sci 35(2):188–195
3. Chen D, Wang Z, Guo D, Orekhov V, Qu X (2020) Review and prospect: deep learning in nuclear magnetic resonance spectroscopy. Chemistry–A Eur J 26(46):10391–10401
4. Carleo G, Troyer M (2017) Solving the quantum many-body problem with artificial neural networks. Science 355(6325):602–606
5. Pettit RW, Fullem R, Cheng C, Amos CI (2021) Artificial intelligence, machine learning, and deep learning for clinical outcome prediction. Emerg Top Life Sci 5(6):729–745
6. Olawade DB, Ige AO, Olaremu AG, Ijiwade J, Adeola AO (2024 Sep) The synergy of artificial intelligence and nanotechnology towards advancing innovation and Sustainability-A Mini-Review. Nano Trends 29:100052
7. Jaiswal A, Babu AR, Zadeh MZ, Banerjee D, Makedon F (2020) A survey on contrastive self-supervised learning. Technologies 9(1):2
8. Glielmo A, Husic BE, Rodriguez A, Clementi C, Noé F, Laio A (2021) Unsupervised learning methods for molecular simulation data. Chem Rev 121(16):9722–9758
9. Van Engelen JE, Hoos HH (2020) A survey on semi-supervised learning. Mach Learn 109(2):373–440
10. Rivera-Lopez R, Canul-Reich J, Mezura-Montes E, Cruz-Chávez MA (2022) Induction of decision trees as classification models through metaheuristics. Swarm Evol Comput 69:101006
11. Mitikiri P, Jana G, Sural S, Chattaraj PK A machine learning technique toward generating minimum energy structures of small Boron clusters
12. Keith JA, Vassilev-Galindo V, Cheng B, Chmiela S, Gastegger M, Muller KR, Tkatchenko A (2021) Combining machine learning and computational chemistry for predictive insights into chemical systems. Chem Rev 121(16):9816–9872
13. Polanski J (2022) Unsupervised learning in drug design from self-organization to deep chemistry. Int J Mol Sci 23(5):2797
14. Schwaller P, Hoover B, Reymond JL, Strobelt H, Laino T (2021) Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. Sci Adv 7(15):eabe4166
15. James G, Witten D, Hastie T, Tibshirani R, Taylor J Linear regression. InAn introduction to statistical learning: with applications in python 2023 Jul 1 (pp. 69–134). Cham: Springer International Publishing
16. Montavon G, Samek W, Müller KR (2018) Methods for interpreting and Understanding deep neural networks. Digit Signal Proc 73:1–5
17. Raphael A, Dubinsky Z, Iluz D, Netanyahu NS (2020) Neural network recognition of marine benthos and corals. Diversity 12(1):29
18. Tayyab SM, Chatterton S, Pennacchi P (2022) Image-Processing-Based intelligent defect diagnosis of rolling element bearings using spectrogram images. Machines 10(10):908
19. Jordan J, Convolutional neural networks. https://www.jeremyjordan.me/convolutional-neural-networks/
20. Sharma A, Kumar R, Ranjta S, Varadwaj PK (2021) SMILES to smell: decoding the structure–odor relationship of chemical compounds using the deep neural network approach. J Chem Inf Model 61(2):676–688
21. Coley CW, Jin W, Rogers L, Jamison TF, Jaakkola TS, Green WH, Barzilay R, Jensen KF (2019) A graph-convolutional neural

network model for the prediction of chemical reactivity. Chem Sci 10(2):370–377
22. Wong TT (2015) Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recogn 48(9):2839–2846
23. Kim H, Park M, Kim CW, Shin D (2019) Source localization for hazardous material release in an outdoor chemical plant via a combination of LSTM-RNN and CFD simulation. Comput Chem Eng 125:476–489
24. Peng Y, Rios A, Kavuluru R, Lu Z (2018) Chemical-protein relation extraction with ensembles of SVM, CNN, and RNN models. ArXiv Preprint ArXiv:1802.01255. Feb 5
25. Dral PO (2024) AI in computational chemistry through the lens of a decade-long journey. Chem Commun 60(24):3240–3258
26. Guzman-Pando A, Ramirez-Alonso G, Arzate-Quintana C, Camarillo-Cisneros J Deep learning algorithms applied to computational chemistry. Mol Divers 2023 Dec 27:1–36
27. Sendek AD, Cubuk ED, Antoniuk ER, Cheon G, Cui Y, Reed EJ (2018) Machine learning-assisted discovery of solid Li-ion conducting materials. Chem Mater 31(2):342–352
28. Schütt KT, Arbabzadah F, Chmiela S, Müller KR, Tkatchenko A (2017) Quantum-chemical insights from deep tensor neural networks. Nat Commun 8(1):13890
29. Zhang L, Tan J, Han D, Zhu H (2017) From machine learning to deep learning: progress in machine intelligence for rational drug discovery. Drug Discovery Today 22(11):1680–1685
30. Jiménez-Luna J, Grisoni F, Schneider G (2020) Drug discovery with explainable artificial intelligence. Nat Mach Intell 2(10):573–.https://doi.org/10.1038/s42256-020-00236-4
31. Xie T, Grossman JC (2018) Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical review letters.;120(14):145301. Available from: https://doi.org/10.1103/PhysRevLett.120.145301
32. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A (2021) Highly accurate protein structure prediction with alphafold. Nature 596(7873):583–589
33. Lu Z (2021) Computational discovery of energy materials in the era of big data and machine learning: a critical review. Materials Reports: Energy.;1(3):100047. Available from: https://doi.org/10.1016/j.matre.2021.100047
34. Chen LY, Li YP (2024) AutoTemplate: enhancing chemical reaction datasets for machine learning applications in organic chemistry. Journal of Cheminformatics.;16(1):74. Available from: https://doi.org/10.1186/s13321-024-00869-2
35. Schneider P, Walters WP, Plowright AT, Sieroka N, Listgarten J, Goodnow RA Jr, Fisher J, Jansen JM, Duca JS, Rush TS, Zentgraf M (2020) Rethinking drug design in the artificial intelligence era. Nat Rev Drug Discovery 19(5):353–364
36. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A (2018) Machine learning for molecular and materials science. Nature.;559(7715):547–55. Available from: https://doi.org/10.1038/s41586-018-0337-2
37. Butler KT, Davies DW, Cartwright H, Isayev O, Walsh A (2018) Machine learning for molecular and materials science. Nature 559(7715):547–555
38. Hand DJ, Khan S (2020) Validating and verifying AI systems. Patterns.;1(3)
39. Heaton J, Goodfellow I, Bengio Y, Courville A (2018) Deep learning: the MIT press, ISBN: 0262035618, genetic programming and evolvable machines. 7–305
40. Raschka S, Mirjalili V (2019) Python machine learning: machine learning and deep learning with python, scikit-learn, and tensorflow 2. Packt publishing ltd. Dec 12
41. Géron A (2022) Hands-on machine learning with Scikit-Learn, keras, and tensorflow. O'Reilly Media, Inc. Oct 4
42. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow PM, Zietz M, Hoffman MM, Xie W (2018) Opportunities and Obstacles for deep learning in biology and medicine. J Royal Soc Interface 15(141):20170387
43. Gómez-Bombarelli R (2018) Reaction: the near future of artificial intelligence in materials discovery. Chem 4(6):1189–1190
44. Maleki R, Shams SM, Chellehbari YM, Rezvantalab S, Jahromi AM, Asadnia M, Abbassi R, Aminabhavi T, Razmjou A (2022) Materials discovery of ion-selective membranes using artificial intelligence. Commun Chem 5(1):132
45. Liu Y, Zhao T, Ju W, Shi S (2017) Materials discovery and design using machine learning. J Materiomics 3(3):159–177
46. Saal JE, Oliynyk AO, Meredig B (2020) Machine learning in materials discovery: confirmed predictions and their underlying approaches. Annu Rev Mater Sci 50:49–69
47. Hedley GJ, Ruseckas A, Samuel ID (2017) Light harvesting for organic photovoltaics. Chem Rev 117(2):796–837
48. Pollice R, dos Passos Gomes G, Aldeghi M, Hickman RJ, Krenn M, Lavigne C, Lindner-D'Addario M, Nigam A, Ser CT, Yao Z, Aspuru-Guzik A (2021) Data-driven strategies for accelerated materials design. Acc Chem Res 54(4):849–860
49. Yao Z, Hegde VI, Aspuru-Guzik A, Wolverton C (2019) Discovery of calcium-metal alloy anodes for reversible Ca-ion batteries. Adv Energy Mater 9(9):1802994
50. Zhai C, Li T, Shi H, Yeo J (2020) Discovery and design of soft polymeric bio-inspired materials with multiscale simulations and artificial intelligence. J Mater Chem B 8(31):6562–6587
51. Oliveira ON Jr, Oliveira MC (2022) Materials discovery with machine learning and knowledge discovery. Front Chem 10:930369
52. Sumita M, Yang X, Ishihara S, Tamura R, Tsuda K (2018) Hunting for organic molecules with artificial intelligence: molecules optimized for desired excitation energies. ACS Cent Sci 4(9):1126–1133
53. Ullah A, Dral PO (2022) Predicting the future of excitation energy transfer in light-harvesting complex with artificial intelligence-based quantum dynamics. Nat Commun 13(1):1930
54. Imbalzano G, Anelli A, Giofré D, Klees S, Behler J, Ceriotti M (2018) Automatic selection of atomic fingerprints and reference configurations for machine-learning potentials. J Chem Phys.;148(24)
55. Schlexer P, Winther KT, Garrido-Torres JA, Streibel V, Zhao M, Bajdich M, Abild-Pedersen F, Bligaard T (2019) Machine learning for computational heterogeneous catalysis, CatChem. 11 (16) 601–3581
56. Ang SJ, Wang W, Schwalbe-Koda D, Axelrod S, Gómez-Bombarelli R (2021) Active learning accelerates Ab initio molecular dynamics on reactive energy surfaces. Chem 7(3):738–751
57. Zheng P, Zubatyuk R, Wu W, Isayev O, Dral PO (2021) Artificial intelligence-enhanced quantum chemical method with broad applicability. Nat Commun 12(1):7022
58. Baylon JL, Cilfone NA, Gulcher JR, Chittenden TW (2019) Enhancing retrosynthetic reaction prediction with deep learning using multiscale reaction classification. J Chem Inf Model 59(2):673–688
59. Fang X, Bagui S, Bagui S (2017) Improving virtual screening predictive accuracy of human Kallikrein 5 inhibitors using machine learning models. Comput Biol Chem 69:110–119
60. Badu S, Melnik R, Singh S (2020) Mathematical and computational models of RNA nanoclusters and their applications in data-driven environments. Mol Simul 46(14):1094–1115
61. Smith Z, Ravindra P, Wang Y, Cooley R, Tiwary P (2020) Discovering protein conformational flexibility through artificial-intelligence-aided molecular dynamics. J Phys Chem B 124(38):8221–8229

62. Jordan AM (2018) Artificial intelligence in drug design—the storm before the calm? ACS Med Chem Lett 9(12):1150–1152
63. Dimitrov I, Bangov I, Flower DR, Doytchinova I (2014) AllerTOP v. 2—a server for in Silico prediction of allergens. J Mol Model 20:1–6
64. Gupta A, Müller AT, Huisman BJ, Fuchs JA, Schneider P, Schneider G (2018) Generative recurrent networks for de Novo drug design. Mol Inf 37(1–2):1700111
65. Merk D, Friedrich L, Grisoni F, Schneider G (2018) De Novo design of bioactive small molecules by artificial intelligence. Mol Inf 37(1–2):1700153
66. Thomford NE, Senthebane DA, Rowe A, Munro D, Seele P, Maroyi A, Dzobo K (2018) Natural products for drug discovery in the 21st century: innovations for novel drug discovery. Int J Mol Sci 19(6):1578
67. Siramshetty VB, Nguyen DT, Martinez NJ, Southall NT, Simeonov A, Zakharov AV (2020) Critical assessment of artificial intelligence methods for prediction of hERG channel Inhibition in the big data era. J Chem Inf Model 60(12):6007–6019
68. Moret M, Helmstädter M, Grisoni F, Schneider G, Merk D (2021) Beam search for automated design and scoring of novel ROR ligands with machine intelligence. Angew Chem Int Ed 60(35):19477–19482
69. Yoshimori A, Bajorath J (2020) The SAR matrix method and an artificially intelligent variant for the identification and structural organization of analog series, SAR analysis, and compound design. Mol Inf 39(12):2000045
70. Aronica PG, Reid LM, Desai N, Li J, Fox SJ, Yadahalli S, Essex JW, Verma CS (2021) Computational methods and tools in antimicrobial peptide research. J Chem Inf Model 61(7):3172–3196
71. Melo MC, Maasch JR, de la Fuente-Nunez C (2021) Accelerating antibiotic discovery through artificial intelligence. Commun Biology 4(1):1050
72. Wang B, Su Z, Wu Y (2021) Characterizing the function of domain linkers in regulating the dynamics of multi-domain fusion proteins by microsecond molecular dynamics simulations and artificial intelligence. Proteins Struct Funct Bioinform 89(7):884–895
73. Karthikeyan A, Priyakumar UD (2022) Artificial intelligence: machine learning for chemical sciences. J Chem Sci 134:1–20
74. Wigh DS, Goodman JM, Lapkin AA (2022) A review of molecular representation in the age of machine learning. Wiley Interdisciplinary Reviews: Comput Mol Sci 12(5):e1603
75. Panteleev J, Gao H, Jia L (2018) Recent applications of machine learning in medicinal chemistry. Bioorg Med Chem Lett 28(17):2807–2815
76. Miljković F, Rodríguez-Pérez R, Bajorath J (2021) Impact of artificial intelligence on compound discovery, design, and synthesis. ACS Omega 6(49):33293–33299
77. Mouchlis VD, Afantitis A, Serra A, Fratello M, Papadiamantis AG, Aidinis V, Lynch I, Greco D, Melagraki G (2021) Advances in de Novo drug design: from conventional to machine learning methods. Int J Mol Sci 22(4):1676
78. Renault N, Laurent X, Farce A, El Bakali J, Mansouri R, Gervois P, Millet R, Desreumaux P, Furman C, Chavatte P (2013) Virtual screening of CB2 receptor agonists from bayesian network and High-Throughput docking: structural insights into Agonist-Modulated GPCR features. Chem Biol Drug Des 81(4):442–454
79. Zhao C, Liu D, Teng B, He Z (2015) BagReg: protein inference through machine learning. Comput Biol Chem 57:12–20
80. Peña-Guerrero J, Nguewa PA, García-Sosa AT (2021) Machine learning, artificial intelligence, and data science breaking into drug design and neglected diseases. Wiley Interdisciplinary Reviews: Comput Mol Sci 11(5):e1513
81. Wang Y, Chen TY, Vlachos DG (2021) NEXTorch: a design and bayesian optimization toolkit for chemical sciences and engineering. J Chem Inf Model 61(11):5312–5319
82. Li T, Zhang C, Li X (2022) Machine learning for flow batteries: opportunities and challenges. Chem Sci 13(17):4740–4752
83. Meredig B, Agrawal A, Kirklin S, Saal JE, Doak JW, Thompson A, Zhang K, Choudhary A, Wolverton C (2014) Combinatorial screening for new materials in unconstrained composition space with machine learning. Phys Rev B 89(9):094104
84. Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ (2017) Building machines that learn and think like people. Behav Brain Sci 40:e253
85. Karpathy A, Fei-Fei L Deep visual-semantic alignments for generating image descriptions. InProceedings of the IEEE conference on computer vision and pattern recognition 2015 (pp. 3128–3137)
86. Alexander LT, Lepore R, Kryshtafovych A, Adamopoulos A, Alahuhta M, Arvin AM, Bomble YJ, Böttcher B, Breyton C, Chiarini V, Chinnam NB (2021) Target highlights in CASP14: analysis of models by structure providers. Proteins Struct Funct Bioinform 89(12):1647–1672
87. Callaway E (2023) Protein-folding contest seeks next big breakthrough. Nature
88. Bendell CJ, Liu S, Aumentado-Armstrong T, Istrate B, Cernek PT, Khan S, Picioreanu S, Zhao M, Murgita RA (2014) Transient protein-protein interface prediction: datasets, features, algorithms, and the RAD-T predictor. BMC Bioinformatics 15:1–2
89. Gentili PL (2013) Small steps towards the development of chemical artificial intelligent systems. RSC Adv 3(48):25523–25549
90. Von Lilienfeld OA (2018) Quantum machine learning in chemical compound space. Angew Chem Int Ed 57(16):4164–4169

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