International Journal of Innovation Scientific Research and Review
Vol. 06, Issue, 12, pp.7504-7513, December 2024
Available online at http://www.journalijisr.com
SJIF Impact Factor 2023: 6.599

IJISRR

ISSN: 2582-6131

Review Article

THE INTEGRATION OF MACHINE LEARNING AND PREDICTIVE MODELING: A NEW ERA IN COMPUTATIONAL CHEMISTRY AND MATERIALS SCIENCE

*** Ramachandran Dushanan and Rajendram Senthilnithy**

Department of Chemistry, The Open University of Sri Lanka, Sri Lanka.

Received 14th October 2024; Accepted 15th November 2024; Published online 30th December 2024

ABSTRACT

This review provides an in-depth analysis of the transformative integration of machine learning (ML) into computational chemistry and materials science, emphasizing its potential to revolutionize predictive modeling and accelerate discovery. Traditional computational techniques, such as density functional theory (DFT) and molecular dynamics, although effective, are constrained by high computational costs and scalability limitations. ML, with its ability to process complex, high-dimensional datasets, addresses these challenges by enabling rapid and accurate predictions of molecular properties, reaction mechanisms, and material behaviors. These advancements are driving progress in critical applications, such as drug discovery, where ML accelerates virtual screening and binding affinity predictions, and materials design, which benefits from faster identification of novel materials with tailored properties. The review also delves into the challenges impeding broader ML adoption, including data scarcity, bias in training datasets, over fitting, and the interpretability of complex ML models. Strategies to overcome these barriers, such as feature engineering, explainable AI, and the development of comprehensive, high-quality datasets, are explored. Furthermore, the importance of interdisciplinary collaboration among chemists, material scientists, and computer scientists is underscored, as such partnerships are vital for advancing ML-driven approaches. Future directions are discussed, including the integration of ML with multi-scale modeling, leveraging quantum computing for enhanced simulations, and improving explain ability to foster trust and adoption. This work highlights ML's potential to drive groundbreaking innovations in energy, healthcare, and sustainable materials, establishing it as a cornerstone for the future of computational science.

Keywords: Machine Learning, Explainable AI, Multi-scale Modeling, Quantum Computing.

INTRODUCTION

Computational chemistry and materials science have been central to scientific advances, using predictive modeling to simulate molecular and material behaviors [1] Traditionally, these models relied on first-principles calculations and statistical methods, which, while successful, had limitations in scalability, efficiency, and adaptability[1]. The introduction of Machine Learning (ML) has revolutionized the field, providing faster, more accurate, and scalable predictive capabilities [2].

Overview of Predictive Modeling in Computational Chemistry and Materials Science

Predictive modeling in computational chemistry and materials science has been essential for simulating chemical processes, predicting molecular properties, and designing materials[3]. Traditional methods like density functional theory (DFT) and molecular dynamics offer valuable insights but are computationally expensive and require expertise [4]. Integrating ML enhances predictive modeling using data-driven algorithms, overcoming traditional limitations and enabling more efficient exploration of chemical and material spaces, driving innovations previously impossible with conventional methods[5].

Evolution of Machine Learning in the Field of Computational Chemistry and Materials Science

The journey of ML in computational chemistry and materials science can be traced back to the early adoption of statistical learning methods to model structure-property relationships [6]. Initially, linear regression and principal component analysis (PCA) were employed to establish correlations between molecular descriptors and desired properties [7]. Over time, advancements in computational power and the availability of large-scale datasets catalyzed the adoption of more sophisticated algorithms, such as neural networks, support vector machines (SVMs), and ensemble methods [8].

Recent years have witnessed exponential growth in applying deep learning, a subset of ML characterized by multi-layered architectures capable of capturing complex, non-linear relationships[9]. Innovations such as graph neural networks (GNNs) for molecular structures and convolutional neural networks (CNNs) for imaging-based material characterization have further expanded the applicability of ML in the field [10]. Furthermore, developing hybrid models combining quantum mechanics with ML has opened new avenues for studying chemical and material systems with unprecedented precision [11]. The Figure 1 illustrate the historical development and increasing complexity of ML applications.

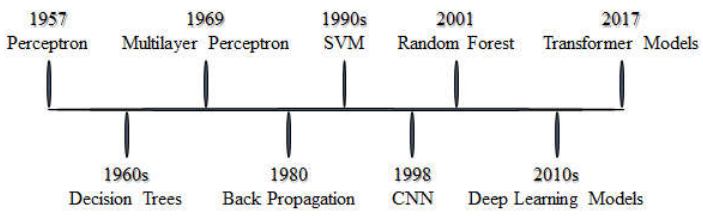

Figure 01. Evolution of machine learning techniques in computational chemistry

Importance and Potential of Integrating Machine Learning for Predictive Tasks

Integrating ML into predictive modeling represents a paradigm shift in chemistry and materials science. ML excels at identifying patterns in

***Corresponding Author: Ramachandran Dushanan,**
Department of Chemistry, The Open University of Sri Lanka, Sri Lanka.

large datasets and making accurate predictions, particularly valuable in areas like drug discovery, catalysis, and materials design, where experimental validation is resource-intensive [12]. It can predict molecular properties such as solubility, toxicity, reactivity, and screen material libraries for properties like thermal conductivity and mechanical strength. ML also aids in uncovering underlying mechanisms and enhancing understanding of systems [13].

Additionally, ML can improve traditional computational methods, such as approximating quantum mechanical calculations, reducing costs while maintaining accuracy [14]. The use of explainable AI (XAI) further increases the interpretability of predictions, fostering trust in the scientific community. In conclusion, the fusion of ML with predictive modeling is set to revolutionize computational chemistry and materials science, driving advancements in energy, healthcare, and materials innovation and becoming central to future research and development [15].

FUNDAMENTAL CONCEPTS

Integrating ML with predictive modeling in computational chemistry and materials science builds upon established principles and methodologies[16]. To understand this transformative approach, it is crucial to grasp the fundamentals of predictive modeling in these fields, along with the core machine learning paradigms and key algorithms driving modern ML applications.

Basics of Predictive Modeling in Computational Chemistry and Materials Science

Predictive modeling forecasts the properties and behaviors of chemical and material systems using theoretical frameworks and experimental data [17]. In computational chemistry, it predicts molecular properties like energy levels and binding affinities, while in materials science, it aids in designing materials and exploring performance under various conditions [18]. Traditional methods, including quantum mechanics, molecular dynamics, and statistical models, have been successful but are computationally intensive and limited in scalability for complex systems [19]. The integration of ML has revolutionized predictive modeling, offering data-driven approaches to overcome these challenges efficiently [20].

Overview of Machine Learning Techniques

ML involves algorithms that enable systems to learn patterns from data and make predictions or decisions without being explicitly programmed [21]. In computational chemistry and materials science, ML enhances predictive modeling by identifying correlations in high-dimensional datasets, which are often too complex for traditional approaches [22]. ML can be broadly categorized into three paradigms, as shown in Table 1.

Table 1. Computational Methods in Computational Chemistry and Materials Science

Learning Paradigm	Description	Applications in Chemistry	Applications in Materials Science
Supervised Learning	Maps input features (e.g., molecular descriptors) to labeled outputs (e.g., property values).	Predicting solubility, toxicity, and binding affinity.	Predicting thermal conductivity, elasticity, and bandgap energies.
Unsupervised Learning	Identifies patterns in unlabeled datasets for clustering, dimensionality reduction, and trend discovery.	Grouping molecules based on structures or properties.	Clustering materials by synthesis conditions or performance metrics.
Reinforcement Learning	Trains models via trial-and-error to maximize a reward function.	Optimizing reaction conditions for maximum yield.	Discovering optimal synthesis pathways for novel materials.

Key Algorithms

Several ML algorithms are crucial in predictive modeling within computational chemistry and materials science, each suited to specific tasks. Neural networks (NNs), including convolutional and GNN, are highly effective at modeling complex, non-linear relationships and are commonly used for predicting molecular properties and analyzing molecular and crystal structures [23]. SVMs are beneficial for classification and regression tasks, such as predicting molecular activity or material phase transitions. Decision trees organize data hierarchically and, combined with ensemble methods like random forests and gradient boosting, enhance accuracy in predicting chemical reaction reactivity and material properties [24]. Clustering algorithms, such as k-means and hierarchical clustering, are valuable for grouping similar entities in datasets, helping to identify functional groups in chemicals or categorize materials based on performance metrics [25]. Lastly, kernel-based techniques like Gaussian processes are employed for probabilistic predictions and uncertainty quantification, making them valuable for applications such as modeling energy surfaces in molecular simulations and predicting material property uncertainties[26].

MACHINE LEARNING APPLICATIONS IN COMPUTATIONAL CHEMISTRY

ML has revolutionized computational chemistry by enabling efficient and accurate predictive modeling for molecular properties, drug discovery, reaction mechanisms, and Quantitative Structure-Activity Relationship (QSAR) studies [27]. By leveraging advanced algorithms and vast datasets, ML addresses challenges inherent in traditional methods, such as high computational costs and limited scalability[28]. This section explores transformative ML applications in computational chemistry.

Predicting Molecular Properties

Understanding molecular properties is essential in drug design, catalysis, and materials science. Traditional methods like DFT and molecular dynamics are insightful but computationally intensive [28]. ML provides efficient alternatives, predicting properties like solubility, stability, and reactivity using datasets and advanced architectures like GNN [29]. Solubility predictions rely on molecular descriptors, while stability predictions use features like bond energies and electronic distribution [30]. ML also identifies reactive sites and optimizes reaction conditions using attention mechanisms [31]. Advancements like adapting AlphaFold for small molecules showcase ML's versatility and transformative potential in predictive chemistry [32].

Accelerating Drug Discovery through Virtual Screening and Binding Affinity

ML has transformed drug discovery by accelerating virtual screening and binding affinity predictions, addressing the limitations of traditional high-throughput screening (HTS), which is time-consuming and costly [33]. In virtual screening, algorithms like CNNs and SVMs

rapidly identify potential drug candidates by predicting binding potential and eliminating low-efficacy compounds [34]. ML models trained on experimental data quantify ligand-target interactions, with hybrid approaches combining molecular docking and ML for enhanced accuracy. Techniques like Deep Dock and transfer learning improve predictions using deep learning and existing datasets [35]. ML has significantly reduced drug discovery timelines, exemplified by Pfizer's ML-based screening for COVID-19 treatments, demonstrating efficiency in urgent healthcare needs [36].

Modeling Reaction Mechanisms and Pathways

Understanding reaction mechanisms is crucial for advancements in organic synthesis, catalysis, and environmental chemistry. Traditional quantum mechanics and kinetic simulations are resource-intensive[37]. ML provides faster, accurate alternatives by analyzing reactants, intermediates, and products to predict plausible pathways. Graph-based models excel in capturing atomic connectivity, while ML predicts reaction rates and equilibrium constants, optimizing catalytic processes [38]. Advanced models, such as RNNs for sequential step prediction and Bayesian models for uncertainty quantification, further enhance predictions. Platforms like IBM's RXN for Chemistry use ML to predict reaction mechanisms, enabling virtual design and testing synthetic pathways, lowering barriers to exploring and optimizing reactions [39].

Quantitative Structure-Activity Relationship Models Enhanced by ML

QSAR modeling links chemical structure to biological activity and has been revitalized by ML, enabling non-linear modeling and enhancing prediction accuracy [40]. ML-enhanced QSAR models predict biological activity by analyzing features like hydrophobicity, electronic distribution, and steric factors [41]. Deep learning methods like autoencoders capture latent features effectively. ML-driven QSAR also predicts toxicity, correlating structural features with toxic effects, minimizing risks in drug development and materials science [42]. Graph-based QSAR using GNN improves predictions by representing molecules as graphs, while multitask learning enables simultaneous property prediction, enhancing efficiency [27]. ML-driven QSAR expands structure-activity studies, impacting pharmacology, agrochemicals, and materials science [43].

MACHINE LEARNING IN MATERIALS SCIENCE

Materials science focuses on discovering, designing, and optimizing materials with specific properties for various applications. Traditionally, it relies on experimental methods and computational simulations, which can be time-consuming and costly [44]. ML has emerged as a transformative tool, accelerating material discovery, improving predictions, and optimizing performance with enhanced efficiency [45]. ML is particularly impactful in designing novel materials, predicting phase stability and thermodynamic properties, exploring material behaviors, and streamlining HTS processes [46].

Designing Novel Materials with Desired Properties

Designing materials with tailored properties has long been a goal in materials science, traditionally relying on intuition, trial-and-error experimentation, and computational modeling [47]. ML has revolutionized this process by identifying patterns in large datasets, enabling the prediction of material properties, and guiding the synthesis of novel materials [48]. ML is particularly valuable in designing energy materials, such as high-performance battery electrodes, catalysts for fuel cells, and thermoelectric materials [45]. It also aids in designing polymers with desired mechanical strength, thermal stability, and biodegradability, and optimizing composite materials by predicting the impact of different fillers and matrices[49]. ML helps predict mechanical properties like tensile strength and flexibility in advanced alloys, which are crucial for aerospace and automotive applications [50]. Notable techniques, such as generative models (e.g., VAEs and GANs) and inverse design frameworks, allow the generation and designing of materials with specific property profiles. By shifting from trial-and-error methods to data-driven approaches, ML accelerates material discovery, reducing the time and costs associated with material development [51]

Prediction of Phase Stability and Thermodynamic Properties

Understanding phase stability and thermodynamic properties is crucial for predicting how materials behave under various conditions, such as temperature, pressure, and chemical environments. Traditional methods like DFT and molecular dynamics are computationally intensive, especially for complex systems[52]. ML offers a more efficient alternative, leveraging existing datasets to predict phase stability and thermodynamic properties accurately[45]. Critical applications include predicting phase diagrams for multicomponent systems and thermodynamic quantities such as Gibbs free energy, entropy, and heat capacity[53]. Advanced models like Gaussian Process Regression (GPR) and Bayesian Optimization further enhance the reliability and efficiency of these predictions[26]. Notable examples, like the Materials Project and Open Quantum Materials Database (OQMD), use ML to accelerate phase stability and thermodynamic property predictions, providing valuable data for material selection and design[54].

Exploring Electronic, Optical, and Mechanical Properties Using ML

ML significantly enhances the prediction of materials' electronic, optical, and mechanical properties, crucial for electronics, photonics, and structural engineering applications [45]. For electronic properties, ML models predict bandgaps, essential for semiconductors and optoelectronics, and carrier mobility, vital for transistors and solar cells[55]. In optics, ML helps predict refractive indices absorption and design photonic bandgaps for waveguides and filters[56]. ML predicts strength, elasticity, and resistance to fracture and fatigue for mechanical properties, aiding in the design of load-bearing materials[57]. Techniques like GNN, CNNs, and RNNs model these properties, while multitask learning improves prediction accuracy by simultaneously forecasting multiple properties [58]. ML's ability to predict complex material behaviors accelerates the development of advanced materials for various applications [59].

Accelerating High-Throughput Materials Screening

ML has revolutionized HTS by enabling faster and more efficient virtual screening of materials, reducing the resource-intensive nature of traditional methods[60]. ML models prioritize materials based on fundamental properties like activity, stability, capacity, and efficiency in applications such as catalyst discovery, battery materials, and photovoltaics [61]. Techniques like active learning focus experimentation on the most promising candidates, while transfer learning accelerates discovery by applying knowledge from well-studied systems to new materials [62]. ML-driven HTS significantly reduces the time and cost of material discovery, fostering rapid advancements in renewable energy, environmental remediation, and manufacturing [48].

INTEGRATION OF MACHINE LEARNING AND QUANTUM CHEMISTRY

Quantum chemistry provides essential insights into molecular and material properties, but traditional quantum mechanical calculations like DFT are computationally demanding [63]. Integrating ML into quantum chemistry has enhanced these simulations' accuracy, efficiency, and predictive power. ML techniques are improving DFT calculations, enabling hybrid quantum-ML models, and transforming reaction prediction and energy profiling, offering a more efficient approach to studying complex systems [64].

Improving Density Functional Theory Calculations with ML

DFT is a widely used quantum mechanical method for studying molecular systems, offering accurate results for molecular energies, reaction profiles, and electronic structures [65]. However, its accuracy depends on the choice of exchange-correlation functional, and traditional functional often fail to capture complex phenomena like dispersion interactions or excited-state properties [66]. Integrating ML with DFT helps improve these limitations by learning from existing computational results and correcting errors in real time [67].

Critical applications include error correction, where ML models identify and correct systematic errors in DFT calculations by comparing DFT results with higher-level quantum methods, leading to more accurate predictions of molecular properties such as binding and reaction energies [68]. By training on diverse datasets, ML can also aid in developing more precise exchange-correlation functionals, particularly for van der Waals interactions and excited-state dynamics[69]. Furthermore, ML techniques help optimize DFT calculations by predicting which regions or settings will likely provide the most accurate results, reducing computational costs in large-scale simulations [59].

Notable techniques like Deep Neural Networks (DNNs) and Gaussian Process Regression (GPR) are used to improve error corrections and energy predictions for molecular systems [55]. The combination of ML and DFT enables the study of larger, more complex systems with greater accuracy, making it a powerful tool in fields such as drug discovery, material design, and reaction modeling [51].

Hybrid Approaches: Combining Quantum Mechanics with ML

Hybrid approaches combining quantum mechanical simulations with ML leverage the strengths of both paradigms, offering accurate electronic structure descriptions and scalable, efficient predictions[70]. These methods transform computational chemistry by enabling faster and more efficient simulations of complex molecular systems [71]. Critical applications include using ML-trained models based on quantum mechanical data, such as DFT, to predict molecular energies and reaction barriers accurately [64]. ML also accelerates identifying transition states and optimizing reaction pathways, facilitating the rapid study of chemical reactions [72]. Additionally, ML surrogate models effectively simulate solvation effects, which is critical for understanding molecular behavior in solution [73].

Notable techniques include quantum-ML hybrid models that approximate exchange-correlation functionals or molecular energies and neural network potentials to construct potential energy surfaces for exploring large systems [73]. The impact of these hybrid methods is significant, reducing computational costs and enabling the study of larger systems and complex processes, particularly in materials design and drug discovery [51].

Case Studies in Reaction Prediction and Energy Profiling

ML transforms quantum chemistry by enabling rapid and accurate prediction of reaction mechanisms and energy landscapes, significantly reducing computational costs and accelerating research[74]. Critical applications include reaction pathway prediction, where ML models trained on quantum-derived potential energy surfaces (PES) can quickly identify intermediates and transition states, enhancing the design of synthetic routes [75]. In catalytic reactions, hybrid QM/ML approaches predict energy profiles for new catalysts with minimal resources, facilitating efficient screening [76]. Similarly, in drug discovery and materials science, ML accelerates the prediction of molecule binding energies, streamlining the search for promising candidates [77].

Techniques such as active learning focus on the most informative molecular structures for simulations, while RNNs predict reaction steps by leveraging the temporal sequence of chemical reactions [78]. These advancements drive faster chemical discoveries, optimize catalysis, drug design, and materials development, and broaden the scope of accessible molecular systems for study.[51]

DATASETS AND FEATURE ENGINEERING

ML models in computational chemistry and materials science depend on high-quality datasets and practical feature engineering to represent molecular and material properties accurately. Techniques like molecular fingerprints, descriptors, and embeddings are crucial in transforming raw data into meaningful inputs, enhancing model performance and predictive reliability [18].

Importance of High-Quality Datasets for Model Training

The quality of data used to train ML models is critical for achieving accuracy, generalization, and robustness. In computational chemistry and materials science, datasets often include molecular properties, reaction energies, material characteristics, and experimental data, which must be reliable and comprehensive to enhance model predictive capabilities [35].

Critical considerations for high-quality datasets include data accuracy and precision, essential for preventing prediction errors. Inaccurate quantum mechanical calculations or experimental results, such as inconsistencies in DFT calculations, can introduce systematic biases and compromise model reliability [4]. Data diversity and representativeness are also crucial; a dataset with a wide range of chemical environments, material compositions, and property values ensures the model can generalize well to unseen data. For example, diverse molecular scaffolds improve drug discovery predictions for novel target proteins. Data volume is another important factor; larger datasets typically enhance model performance, particularly for deep learning, which often requires extensive data. In smaller datasets, transfer learning and data augmentation can help mitigate limitations[28]. Finally, data consistency and quality control are vital to maintaining dataset integrity. Errors in molecular structures, computational protocols, and the presence of outliers can distort learning, so proper preprocessing and curation are essential for ensuring high-quality datasets [27].

Sources of Datasets in Computational Chemistry and Materials Science

High-quality datasets are essential for integrating ML into computational chemistry and materials science. These datasets originate from various sources, including experimental databases,

quantum mechanical simulations, and HTS processes. The primary sources of high-quality datasets are given in Table 2.

Table 2. Primary sources of high-quality datasets in Computational Chemistry and Materials Science

Key Sources of High-Quality Datasets	Description
Computational Databases and Repositories	
The Materials Project	Offers data on structural, electronic, and thermodynamic properties of inorganic materials, aiding in material property prediction and discovery.
Cambridge Structural Database (CSD)	Contains extensive crystallographic data, crucial for understanding molecular properties and crystal structures.
Open Quantum Materials Database (OQMD)	Hosts DFT-derived properties for over 10 million materials, supporting HTS and material discovery.
Protein Data Bank (PDB)	Provides 3D biomolecular structures, essential for ML in drug discovery and biomolecular modeling.
Computational Chemistry and Simulation Tools	
Quantum chemistry software (Gaussian, VASP, QuantumESPRESSO)	Generates large datasets of molecular energies, electronic structures, and reaction dynamics, useful for ML models, especially in reaction prediction and energy profiling.
Experimental Data from Publications and Collaborations	
Peer-reviewed studies and research collaborations	Supply high-quality experimental data such as solubility and stability measurements, enhancing ML model accuracy when combined with computational results.
Synthetic and High-Throughput Screening Data	
Automated experimental platforms and virtual screening processes	Generate vast bioactivity, toxicity, and material properties datasets, which are valuable for ML applications in drug discovery and materials optimization.

Feature Selection and Representation: Fingerprints, Descriptors, and Embeddings

Feature engineering, transforming raw data into meaningful inputs for ML models, is critical in computational chemistry and materials science. Feature engineering ensures that ML models can accurately learn and predict behavior by effectively selecting and representing molecular or material characteristics [35]. Fundamental techniques include molecular fingerprints, descriptors, and embeddings.

Molecular Fingerprints

Molecular fingerprints are concise representations of molecules that enable quick comparison and classification. Key types include Extended-Connectivity Fingerprints (ECFP), which focus on local chemical environments, MACCS Keys that capture structural features like functional groups, and Daylight Fingerprints, which encode atom-pair connectivity. These fingerprints are essential in virtual screening, QSAR modeling, and predicting properties like solubility and binding affinity, making them crucial tools in drug discovery and materials science [27].

Descriptors

Descriptors are numerical values that quantify molecular properties such as size, shape, electronic structure, and reactivity. Topological descriptors reflect atomic connectivity and, geometrical descriptors capture 3D shapes like surface area and volume, and electronic descriptors represent electronic properties like dipole moment and orbital energies [40]. These descriptors are vital in QSAR modeling, materials property prediction, and molecular structure optimization, providing insights into how molecular features affect biological activity and material properties, which are essential for drug discovery and materials science [41], [42], [43].

Embeddings

Embeddings are advanced representations that reduce high-dimensional data into lower-dimensional spaces while preserving critical structural and chemical features. Techniques such as GNN represent molecules as graphs to capture atomic interactions, and chemoinformatic embeddings enable clustering[10], regression, and classification. These embeddings are widely used in drug discovery, materials design, and reaction prediction because they capture complex, non-linear relationships.

ADVANCEMENTS IN MACHINE LEARNING ARCHITECTURES

ML in computational chemistry and materials science has advanced significantly, evolving from traditional techniques like linear regression and decision trees to sophisticated architectures such as deep learning, GNN, transfer learning, and multitask learning[10]. These innovations enhance prediction accuracy and enable the exploration of new materials and molecular behaviors, addressing complex challenges in the field.

Deep Learning in Molecular and Materials Modeling

Deep learning (DL) is a powerful tool for handling complex, high-dimensional data in molecular and materials modeling. Deep neural networks (DNNs) excel at identifying intricate patterns, enabling accurate predictions of properties like solubility and stability. DL also advances materials discovery by predicting attributes like superconductivity and optimizing designs through reinforcement learning (RL) [18].

DL accelerates quantum mechanical calculations, reducing costs and enabling high-throughput simulations. In drug discovery, RNNs and CNNs aid in virtual screening and bioactivity prediction, optimizing molecular scaffolds for pharmacological activity [78].

Challenges in DL include interpretability, often perceived as a "black box." Efforts in XAI aim to improve transparency [79]. Additionally, DL's need for large datasets is addressed through techniques like regularization, data augmentation, and transfer learning, enhancing its applicability in computational chemistry [78].

Graph Neural Networks for Molecular and Crystal Structure Prediction

GNNs excel in tasks involving graph-structured data, such as molecular and crystal structure prediction. Representing molecules as graphs, GNNs model atomic relationships, predicting properties like toxicity, solubility, and binding affinity. In drug discovery, GNNs predict molecular docking scores and drug-target binding affinity [10]. In materials science, GNNs analyze atomic connectivity to predict crystal structures, phase transitions, and properties like conductivity and hardness [27]. They also model chemical reactions, aiding synthetic chemistry and materials processing. Challenges for GNNs include high computational demands and limited generalizability across diverse chemical spaces [27]. Research focuses on optimizing

GNN architectures and developing transfer learning techniques to improve scalability and applicability [10].

Transfer Learning and Multitask Learning in Predictive Modeling

Transfer learning and multitask learning enhance ML models by addressing data scarcity and improving generalization. Transfer learning applies knowledge from one task to another, allowing models trained on one molecule or material class to predict properties in others. Challenges include ensuring relatedness between tasks, with domain adaptation improving its effectiveness [75].

Multitask learning trains a model on multiple related tasks simultaneously, identifying shared patterns to enhance accuracy. It predicts multiple molecular properties or materials' attributes like strength and stability. Balancing performance across tasks remains a challenge, requiring advanced loss functions and model designs [7]. These techniques promise significant advancements in computational chemistry and materials science, with ongoing research refining their potential.

CHALLENGES AND LIMITATIONS

Integrating ML into computational chemistry and materials science has led to significant advancements, but challenges remain. These include data-related issues like scarcity and bias and technical challenges such as over fitting and balancing accuracy, scalability, and computational efficiency. This section discusses these challenges and potential solutions.

Data Scarcity and Bias Issues

A key challenge in ML applications is the scarcity of high-quality, labeled datasets, particularly in specialized areas like material design and novel chemical reactions. ML models require large volumes of data, but datasets for rare compounds or specific properties are often limited or incomplete [68]. Collecting data is costly and time-consuming, especially for high-level quantum mechanical calculations. Moreover, datasets often vary in format and lack uncertainty details, complicating training and validation. Bias also impacts model accuracy, with issues like sampling bias, where certain data types are underrepresented, and label bias, stemming from flawed labeling processes. These biases lead to poor generalization and inaccurate predictions.

To address these issues, techniques like data augmentation and generative models (e.g., VAEs, GANs) expand datasets, while transferring learning leverages knowledge from related tasks. Public databases like the Materials Project and ChemBL mitigate data scarcity, fostering collaboration and standardization [51].

Overfitting and Interpretability of ML Models

Overfitting is a major challenge, particularly with flexible models like deep neural networks (DNNs), which may memorize training data instead of generalizing. Sparse or noisy data exacerbates this issue, leading to poor performance in real-world applications [18].

Interpretability is another challenge. Deep learning models often act as "black boxes," making it hard to understand underlying principles or diagnose errors. This lack of transparency complicates validation and reduces trust in the predictions. Solutions include regularization techniques like L1/L2 regularization, dropout, and early stopping to prevent overfitting. XAI methods, such as saliency maps and SHAP values, enhance transparency, offering insights into critical features influencing predictions and aligning models with scientific principles[79].

Balancing Accuracy, Scalability, and Computational Efficiency

Advanced ML models demand extensive data, computational resources, and time, posing challenges in balancing accuracy, scalability, and efficiency [61]. Deep learning models, especially those for large datasets or complex simulations, require high-performance computing (HPC) infrastructure, which is costly. Additionally, scaling models efficiently for growing datasets while maintaining performance is difficult. Real-time predictions in areas like drug discovery or materials design further add complexity [35].

Efforts to address these challenges include developing lightweight architectures, pruning, and low-precision computations. Knowledge distillation enables smaller models to replicate larger ones. Parallel and distributed computing, including cloud platforms, helps scale models and reduce training times. Hybrid approaches combining ML and traditional methods optimize speed and accuracy, leveraging their respective strengths.

FUTURE DIRECTIONS AND OPPORTUNITIES

Integrating ML with predictive modeling in computational chemistry and materials science has already led to significant advancements, with even more tremendous potential on the horizon. As computational power and ML algorithms evolve, new opportunities are emerging to further revolutionize the prediction and engineering of molecular and material properties [23]. Future research will focus on areas such as integrating ML with multi-scale modeling, enhancing trust and adoption through explainable AI, advancing experimental collaboration, and exploring the impact of quantum computing on ML applications in computational chemistry [57].

Integration of ML with Multi-Scale Modeling

Multi-scale modeling in computational chemistry and materials science allows simulations at varying levels of complexity, from atomic interactions to macroscopic material properties. Traditionally, these models have relied on physics-based approaches, but integrating ML is increasingly enhancing efficiency and accuracy. ML can optimize quantum mechanical simulations, reducing the need for exhaustive calculations and enabling more accurate large-scale predictions. It also helps bridge different time and length scales by providing approximate solutions for macroscopic properties while maintaining atomic-level precision, improving the efficiency of simulations. In materials discovery and drug design, this hybrid approach accelerates the identification of new materials and molecules, especially when combined with high-throughput simulations. However, challenges such as calibrating and validating ML models, avoiding overfitting, and ensuring computational efficiency must be addressed. Future research will focus on improving scalability and developing novel algorithms or leveraging cloud computing to manage the complexity of integrated models.

Role of Explainable AI in Improving Trust and Adoption

As ML models are increasingly integrated into computational chemistry and materials science, the demand for XAI is rising [55]. A key barrier to adopting ML predictions in research is the lack of transparency in how models make decisions, which is especially critical in fields like drug development and materials engineering. XAI can build trust by explaining how models process data and generate predictions, helping validate results by identifying key features that

influence predictions. In drug discovery, XAI can clarify the molecular interactions driving binding affinity, guiding further optimization [79]. Despite challenges in applying XAI methods like LIME and SHAP to complex problems, future research should focus on developing specialized XAI techniques for these fields. Balancing model performance with interpretability, especially in complex models like deep learning, will be vital to advancing XAI in scientific research [79].

Advancing Experimental Collaboration Using Predictive Tools

The synergy between computational predictions and experimental work is essential for scientific progress, with ML playing a pivotal role in enhancing this collaboration. ML can accelerate experimental design by predicting outcomes and generating data-driven hypotheses, helping researchers design more efficient experiments. It also uncovers patterns in existing data that can lead to discoveries, such as novel materials or drug compounds. Continuous feedback between experimental data and ML models improves predictions and experimental designs [35]. However, challenges such as the gap between theoretical predictions and real-world results remain, requiring refinement of both methods. Developing platforms that integrate ML predictions with experimental data will be crucial for fostering dynamic, real-time scientific collaboration.

Potential Impacts of Quantum Computing on ML in Computational Chemistry

Quantum computing has the potential to transform computational chemistry and materials science by accelerating quantum mechanical simulations. Quantum ML (QML), which combines quantum computing with ML, could enable more accurate predictions of molecular properties and chemical reactions [15]. QML could help model complex quantum systems beyond classical computers' reach, leveraging ML to analyze large datasets. However, challenges remain, including the development of specialized quantum-ML algorithms and the availability of quantum hardware. Integrating ML and quantum computing will likely enhance predictive modeling in these fields as quantum technology advances. Figure 2 visually summarizes the key areas for future exploration and innovation in ML applications within computational chemistry and materials science [74].

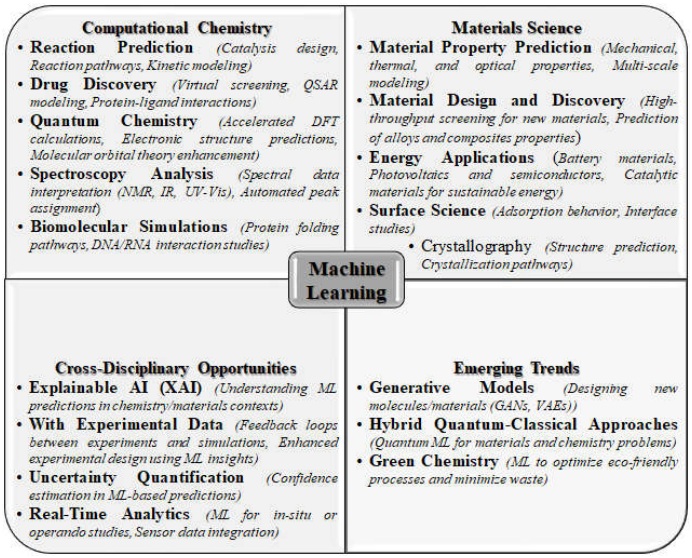

Figure 2. Conceptual mind map of future research directions in ML applications for computational chemistry and materials science.

CONCLUSION

Integrating ML with predictive modeling revolutionizes computational chemistry and materials science by enhancing predictions, accelerating drug discovery, and expanding modeling capabilities to more complex systems. ML speeds up simulations and improves the accuracy of predicting molecular properties, reactivity, and stability. It aids in virtual screening and drug candidate identification while broadening the scope of predictive modeling to include areas like reaction modeling, high-throughput materials screening, and multi-scale simulations. ML's ability to process large datasets and identify complex patterns has led to the discovery of novel materials and optimized molecules for specific applications, such as drug development and energy storage. Combining ML with quantum mechanical methods like DFT improves the efficiency of simulating chemical reactions and molecular interactions, overcoming computational bottlenecks. New ML architectures like deep learning and GNN have further advanced modeling accuracy, especially in large-scale systems. However, the success of ML depends on high-quality datasets and robust feature engineering to ensure accurate predictions.

Despite its potential, ML faces challenges such as data scarcity, model overfitting, and issues with result interpretability. Balancing models' accuracy, scalability, and efficiency for high-throughput applications remains a key hurdle. Future advances in multi-scale modeling, explainable AI, and quantum computing will improve the reliability of ML predictions and facilitate closer integration with experimental efforts. ML can revolutionize drug discovery and material design, offering faster, more accurate drug efficacy and toxicity predictions and enabling the design of materials with specific properties, such as those used in clean energy technologies. Realizing the full potential of ML in these fields requires interdisciplinary collaboration between chemists, material scientists, computer scientists, and engineers. Collaborative efforts will lead to the development of new algorithms, computational models, and databases tailored for ML applications. Addressing the challenges of data scarcity, model interpretability, and computational efficiency is crucial for the future of this integration. Ultimately, combining ML, computational methods, and experimental research will drive significant molecular design and materials science breakthroughs. Still, it will require ongoing collaboration and investment in computational resources.

Competing Interests

Authors have declared that no competing interests exist

Authors' Contributions

Dr. R. Dushanan: Designed the study, searches for literature, and wrote the first draft of the manuscript.

Prof. R. Senthilnithy: Literature searches and approved the final manuscript.

REFERENCES

[1] J. Westermayr, M. Gastegger, K. T. Schütt, and R. J. Maurer, "Perspective on integrating machine learning into computational chemistry and materials science," Journal of Chemical Physics, vol. 154, no. 23, Jun. 2021, doi: 10.1063/5.0047760/200193.

[2] J. Konečný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Scalable Machine Learning Algorithms for Big Data Analytics: Challenges and Opportunities," Journal of Artificial Intelligence Research, vol. 2, no. 2, pp. 124–141, Aug. 2022, Accessed: Dec. 03, 2024. [Online]. Available: https://www.thesciencebrigade.com/JAIR/article/view/327

[3] G. Pilania, "Machine learning in materials science: From explainable predictions to autonomous design," Comput Mater Sci, vol. 193, p. 110360, Jun. 2021, doi: 10.1016/J.COMMATSCI.2021.110360.

[4] K. Burke, "Perspective on density functional theory," Journal of Chemical Physics, vol. 136, no. 15, Apr. 2012, doi: 10.1063/1.4704546/941589.

[5] Nwakamma Ninduwezuor-Ehiobu et al., "Tracing the Evolution of AI and Machine Learning Applications in Advancing Materials Discovery and Production Processes," Engineering Science & Technology Journal, vol. 4, no. 3, pp. 66–83, Sep. 2023, doi: 10.51594/estj.v4i3.552.

[6] K. M. Jablonka, D. Ongari, S. M. Moosavi, and B. Smit, "Big-Data Science in Porous Materials: Materials Genomics and Machine Learning," Chem Rev, vol. 120, no. 16, pp. 8066–8129, Aug. 2020, doi: 10.1021/ACS.CHEMREV.0C00004/ASSET/IMAGES/MEDIUM/CR0C00004_0033.GIF.

[7] S. A. M. Stein, A. E. Loccisano, S. M. Firestine, and J. D. Evanseck, "Chapter 13 Principal Components Analysis: A Review of its Application on Molecular Dynamics Data," Annu Rep Comput Chem, vol. 2, no. C, pp. 233–261, Jan. 2006, doi: 10.1016/S1574-1400(06)02013-5.

[8] N. L. Rane, M. Paramesha, S. P. Choudhary, and J. Rane, "Machine Learning and Deep Learning for Big Data Analytics: A Review of Methods and Applications," Partners Universal International Innovation Journal, vol. 2, no. 3, pp. 172–197, Jun. 2024, doi: 10.5281/ZENODO.12271006.

[9] S. F. Ahmed et al., "Unveiling the frontiers of deep learning: innovations shaping diverse domains," Sep. 2023, Accessed: Dec. 03, 2024. [Online]. Available: https://arxiv.org/abs/2309.02712v1

[10] S. You et al., "Advancements and prospects of deep learning in biomaterials evolution," cell.comS You, Y Fan, Y Chen, X Jiang, W Liu, X Zhou, J Zhang, J Zheng, H Yang, X HouCell Reports Physical Science, 2024•cell.com, Accessed: Dec. 22, 2024. [Online]. Available: https://www.cell.com/cell-reports-physical-science/fulltext/S2666-3864(24)00394-1

[11] M. R. Pulicharla, "Hybrid Quantum-Classical Machine Learning Models: Powering the Future of AI," Journal of Science & Technology, vol. 4, no. 1, pp. 40–65, Jan. 2023, doi: 10.55662/JST.2023.4102.

[12] J. Benavides-Hernández and F. Dumeignil, "From Characterization to Discovery: Artificial Intelligence, Machine Learning and High-Throughput Experiments for Heterogeneous Catalyst Design," ACS Catal, vol. 14, no. 15, pp. 11749–11779, Aug. 2024, doi: 10.1021/ACSCATAL.3C06293/ASSET/IMAGES/MEDIUM/CS3C06293_0010.GIF.

[13] T. Le, V. C. Epa, F. R. Burden, and D. A. Winkler, "Quantitative structure-property relationship modeling of diverse materials properties," Chem Rev, vol. 112, no. 5, pp. 2889–2919, May 2012, doi: 10.1021/CR200066H/SUPPL_FILE/CR200066H_SI_001.PDF.

[14] Y. Han et al., "Machine learning accelerates quantum mechanics predictions of molecular crystals," Phys Rep, vol. 934, pp. 1–71, Nov. 2021, doi: 10.1016/J.PHYSREP.2021.08.002.

[15] I. Papadimitriou, I. Gialampoukidis, … S. V.-C. M., and undefined 2024, "AI methods in materials design, discovery and manufacturing: A review," ElsevierI Papadimitriou, I Gialampoukidis, S Vrochidis, I KompatsiarisComputational Materials Science, 2024•Elsevier, Accessed: Dec. 22, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025624000144

[16] J. A. Keith et al., "Combining Machine Learning and Computational Chemistry for Predictive Insights into Chemical Systems," Chem Rev, vol. 121, no. 16, pp. 9816–9872, Aug. 2021, doi: 10.1021/ACS.CHEMREV.1C00107/ASSET/IMAGES/LARGE/CR1C00107_0014.JPEG.

[17] L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," npj Computational Materials 2016 2:1, vol. 2, no. 1, pp. 1–7, Aug. 2016, doi: 10.1038/npjcompumats.2016.28.

[18] T. Cowen, K. Karim, and S. Piletsky, "Computational approaches in the design of synthetic receptors – A review," Anal Chim Acta, vol. 936, pp. 62–74, Sep. 2016, doi: 10.1016/J.ACA.2016.07.027.

[19] F. F. Abraham, "Computational statistical mechanics methodology, applications and supercomputing," Adv Phys, vol. 35, no. 1, pp. 1–111, Jan. 1986, doi: 10.1080/00018738600101851.

[20] W. Strielkowski, A. Vlasov, K. Selivanov, K. Muraviev, and V. Shakhnov, "Prospects and Challenges of the Machine Learning and Data-Driven Methods for the Predictive Analysis of Power Systems: A Review," Energies 2023, Vol. 16, Page 4025, vol. 16, no. 10, p. 4025, May 2023, doi: 10.3390/EN16104025.

[21] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science (1979), vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/SCIENCE.AAA8415.

[22] A. L. Ferguson, "Machine learning and data science in soft materials engineering," Journal of Physics: Condensed Matter, vol. 30, no. 4, p. 043002, Dec. 2017, doi: 10.1088/1361-648X/AA98BD.

[23] W. Chen et al., "Predicting Crystalline Material Properties with AI: Bridging Molecular to Particle Scales," Ind Eng Chem Res, Oct. 2024, doi: 10.1021/ACS.IECR.4C03224/ASSET/IMAGES/MEDIUM/IE4C03224_0012.GIF.

[24] H. S. Samuel, E. E. Etim, U. Nweke-Maraizu, and S. Yakubu, "Machine Learning in Chemical Kinetics: Predictions, Mechanistic Analysis, and Reaction Optimization.," Applied Journal of Environmental Engineering Science, vol. 10, no. 1, p. 36, May 2024, doi: 10.48422/IMIST.PRSM/AJEES-V10I1.47284.

[25] E. Swann, B. Sun, D. M. Cleland, and A. S. Barnard, "Representing molecular and materials data for unsupervised machine learning," Mol Simul, vol. 44, no. 11, pp. 905–920, Jul. 2018, doi: 10.1080/08927022.2018.1450982.

[26] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, and G. Csányi, "Gaussian Process Regression for Materials and Molecules," Chem Rev, vol. 121, no. 16, pp. 10073–10141, Aug. 2021, doi: 10.1021/ACS.CHEMREV.1C00022/ASSET/IMAGES/MEDIUM/CR1C00022_M072.GIF.

[27] S. K. Niazi and Z. Mariam, "Recent Advances in Machine-Learning-Based Chemoinformatics: A Comprehensive Review," Int J Mol Sci, vol. 24, no. 14, p. 11488, Jul. 2023, doi: 10.3390/IJMS241411488.

[28] N. Yao, X. Chen, Z. H. Fu, and Q. Zhang, "Applying Classical, Ab Initio, and Machine-Learning Molecular Dynamics Simulations to the Liquid Electrolyte for Rechargeable Batteries," Chem Rev, vol. 122, no. 12, pp. 10970–11021, Jun. 2022, doi: 10.1021/ACS.CHEMREV.1C00904/ASSET/IMAGES/MEDIUM/CR1C00904_0035.GIF.

[29] S. Shilpa, G. Kashyap, and R. B. Sunoj, "Recent Applications of Machine Learning in Molecular Property and Chemical Reaction Outcome Predictions," Journal of Physical Chemistry A, vol. 127, no. 40, pp. 8253–8271, Oct. 2023, doi: 10.1021/ACS.JPCA.3C04779/ASSET/IMAGES/MEDIUM/JP3C04779_0015.GIF.

[30] J. Zhou and M. Huang, "Navigating the landscape of enzyme design: from molecular simulations to machine learning," Chem Soc Rev, vol. 53, no. 16, pp. 8202–8239, Aug. 2024, doi: 10.1039/D4CS00196F.

[31] A. I. Visan and I. Negut, "Integrating Artificial Intelligence for Drug Discovery in the Context of Revolutionizing Drug Delivery," Life 2024, Vol. 14, Page 233, vol. 14, no. 2, p. 233, Feb. 2024, doi: 10.3390/LIFE14020233.

[32] G. C. Veríssimo, M. S. M. Serafim, T. Kronenberger, R. S. Ferreira, K. M. Honorio, and V. G. Maltarollo, "Designing drugs when there is low data availability: one-shot learning and other approaches to face the issues of a long-term concern," Expert Opin Drug Discov, vol. 17, no. 9, pp. 929–947, Sep. 2022, doi: 10.1080/17460441.2022.2114451.

[33] P. P. Parvatikar et al., "Artificial intelligence: Machine learning approach for screening large database and drug discovery," Antiviral Res, vol. 220, p. 105740, Dec. 2023, doi: 10.1016/J.ANTIVIRAL.2023.105740.

[34] R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. K. Ambasta, and P. Kumar, "Artificial intelligence to deep learning: machine intelligence approach for drug discovery," Molecular Diversity 2021 25:3, vol. 25, no. 3, pp. 1315–1360, Apr. 2021, doi: 10.1007/S11030-021-10217-3.

[35] D. Lee, W. Hwang, J. Byun, and B. Shin, "Turbocharging protein binding site prediction with geometric attention, inter-resolution transfer learning, and homology-based augmentation," BMC Bioinformatics, vol. 25, no. 1, pp. 1–26, Dec. 2024, doi: 10.1186/S12859-024-05923-2/TABLES/5.

[36] S. Kaur Thethi, "Chapter 8 Machine learning models for cost-effective healthcare delivery systems: A global perspective," Digital Transformation in Healthcare 5.0, pp. 199–244, Apr. 2024, doi: 10.1515/9783111327853-008/PDF.

[37] B. R. Taylor, N. Kumar, D. K. Mishra, B. A. Simmons, H. Choudhary, and K. L. Sale, "Computational Advances in Ionic Liquid Applications for Green Chemistry: A Critical Review of Lignin Processing and Machine Learning Approaches," Molecules 2024, Vol. 29, Page 5073, vol. 29, no. 21, p. 5073, Oct. 2024, doi: 10.3390/MOLECULES29215073.

[38] Y. Schuurman, L. Goulart de Araujo, L. Vilcocq, and P. Fongarland, "Recent Developments in the Use of Machine Learning in Catalysis Kinetics," 2024, doi: 10.2139/SSRN.5010723.

[39] W. Jin, C. W. Coley, R. Barzilay, and T. Jaakkola, "Predicting Organic Reaction Outcomes with Weisfeiler-Lehman Network," Adv Neural Inf Process Syst, vol. 30, 2017.

[40] R. P. Sheridan, A. Liaw, and M. Tudor, "Light Gradient Boosting Machine as a Regression Method for Quantitative Structure-Activity Relationships," Apr. 2021, Accessed: Dec. 20, 2024. [Online]. Available: http://arxiv.org/abs/2105.08626

[41] N. Mathur et al., "In Silico Docking: Protocols for Computational Exploration of Molecular Interactions," Unravelling Molecular Docking - From Theory to Practice [Working Title], Jul. 2024, doi: 10.5772/INTECHOPEN.1005527.

[42] R. Perkins, H. Fang, W. Tong, and W. J. Welsh, "Quantitative structure-activity relationship methods: Perspectives on drug discovery and toxicology," Environ Toxicol Chem, vol. 22, no. 8, pp. 1666–1679, Aug. 2003, doi: 10.1897/01-171.

[43] A. Cherkasov et al., "QSAR modeling: Where have you been? Where are you going to?," J Med Chem, vol. 57, no. 12, pp. 4977–5010, Jun. 2014, doi: 10.1021/JM4004285/ASSET/IMAGES/MEDIUM/JM-2013-004285_0009.GIF.

[44] J. Krzywanski, M. Sosnowski, K. Grabowska, A. Zylka, L. Lasek, and A. Kijo-Kleczkowska, "Advanced Computational Methods for Modeling, Prediction and Optimization—A Review," Materials 2024, Vol. 17, Page 3521, vol. 17, no. 14, p. 3521, Jul. 2024, doi: 10.3390/MA17143521.

[45] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, "Recent advances and applications of machine learning in solid-state materials science," nature.comJ Schmidt, MRG Marques, S Botti, MAL Marquesnpj computational materials, 2019•nature.com, doi: 10.1038/s41524-019-0221-0.

[46] D. Morgan and R. Jacobs, "Opportunities and Challenges for Machine Learning in Materials Science," Annu Rev Mater Res, vol. 50, no. Volume 50, 2020, pp. 71–103, Jul. 2020, doi: 10.1146/ANNUREV-MATSCI-070218-010015/1.

[47] J. Rondinelli, S. May, J. F.-M. bulletin, and undefined 2012, "Control of octahedral connectivity in perovskite oxide heterostructures: An emerging route to multifunctional materials discovery," cambridge.orgJM Rondinelli, SJ May, JW FreelandMRS bulletin, 2012•cambridge.org, 2012, doi: 10.1557/mrs.2012.49.

[48] K. Butler, D. Davies, H. Cartwright, O. Isayev, A. W.- Nature, and undefined 2018, "Machine learning for molecular and materials science," nature.comKT Butler, DW Davies, H Cartwright, O Isayev, A WalshNature, 2018•nature.com, Accessed: Dec. 20, 2024. [Online]. Available: https://www.nature.com/articles/s41586-018-0337-2

[49] A. Mannodi-Kanakkithodi, G. Pilania, T. H.-S. reports, and undefined 2016, "Machine learning strategy for accelerated design of polymer dielectrics," nature.comA Mannodi-Kanakkithodi, G Pilania, TD Huan, T Lookman, R RamprasadScientific reports, 2016•nature.com, Accessed: Dec. 20, 2024. [Online]. Available: https://www.nature.com/articles/srep20952/1000

[50] Z. Q. Chen, Y. H. Shang, X. D. Liu, and Y. Yang, "Accelerated discovery of eutectic compositionally complex alloys by generative machine learning," npj Computational Materials 2024 10:1, vol. 10, no. 1, pp. 1–12, Sep. 2024, doi: 10.1038/s41524-024-01385-5.

[51] B. Sanchez-Lengeling and A. Aspuru-Guzik, "Inverse molecular design using machine learning:Generative models for matter engineering," Science (1979), vol. 361, no. 6400, pp. 360–365, Jul. 2018, doi: 10.1126/SCIENCE.AAT2663.

[52] G. Kresse and J. Hafner, "Ab initio molecular dynamics for liquid metals," Phys Rev B, vol. 47, no. 1, pp. 558–561, 1993, doi: 10.1103/PHYSREVB.47.558.

[53] A. Jain et al., "Commentary: The materials project: A materials genome approach to accelerating materials," pdfs.semanticscholar.org, vol. 1, p. 11002, 2013, doi: 10.1063/1.4812323.

[54] S. Kirklin et al., "The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies," nature.comS Kirklin, JE Saal, B Meredig, A Thompson, JW Doak, M Aykol, S Rühl, C Wolvertonnpj Computational Materials, 2015 •nature.com, 2015, doi: 10.1038/npjcompumats.2015.10.

[55] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko, "Quantum-chemical insights from deep tensor neural networks," nature.com KT Schütt, F Arbabzadah, S Chmiela, KR Müller, A TkatchenkoNature communications, 2017•nature.com, 2017, doi: 10.1038/ncomms13890.

[56] J. Carrete, W. Li, N. Mingo, S. Wang, S. C.-P. R. X, and undefined 2014, "Finding unprecedentedly low-thermal-conductivity half-Heusler semiconductors via high-throughput materials modeling," APSJ Carrete, W Li, N Mingo, S Wang, S CurtaroloPhysical Review X, 2014•APS, vol. 4, no. 1, 2014, doi: 10.1103/PhysRevX.4.011019.

[57] S. Xin et al., "Bulk nanocrystalline boron-doped VNbMoTaW high entropy alloys with ultrahigh strength, hardness, and resistivity," J Alloys Compd, vol. 853:155995, 2021, Accessed: Dec. 22, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0925838820323598

[58] K. Choudhary *et al.,* "Recent advances and applications of deep learning methods in materials science," nature.comK Choudhary, B DeCost, C Chen, A Jain, F Tavazza, R Cohn, CW Park, A Choudharynpj Computational Materials, 2022•nature.com, doi: 10.1038/s41524-022-00734-6.

[59] T. Xie and J. C. Grossman, "Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties," Phys Rev Lett, vol. 120, no. 14, Apr. 2018, doi: 10.1103/PHYSREVLETT.120.145301.

[60] M. C. Sorkun, S. Astruc, J. M. V. A. Koelman, and S. Er, "An artificial intelligence-aided virtual screening recipe for two-dimensional materials discovery," npj Computational Materials 2020 6:1, vol. 6, no. 1, pp. 1–10, Jul. 2020, doi: 10.1038/s41524-020-00375-7.

[61] A. Toner-Rodgers et al., "Artificial Intelligence, Scientific Discovery, and Product Innovation," 2024, Accessed: Dec. 20, 2024. [Online]. Available: https://aidantr.github.io/files/AI_innovation.pdf

[62] A. Dunn, Q. Wang, A. Ganose, … D. D. C., and undefined 2020, "Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm," nature.com A Dunn, Q Wang, A Ganose, D Dopp, A Jainnpj Computational Materials, 2020•nature.com, doi: 10.1038/s41524-020-00406-3.

[63] J. S. Smith et al., "Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning," nature.com, doi: 10.1038/s41467-019-10827-4.

[64] K. T. Schütt, H. E. Sauceda, P.-J. Kindermans, A. Tkatchenko, and K.-R. Müller, "Schnet–a deep learning architecture for molecules and materials," pubs.aip.org, Accessed: Dec. 20, 2024. [Online]. Available: https://pubs.aip.org/aip/jcp/article/148/24/241722/962591

[65] P. Hohenberg and W. Kohn, "Inhomogeneous electron gas," Physical Review, vol. 136, no. 3B, 1964, doi: 10.1103/PHYSREV.136.B864.

[66] A. B.-T. J. of chemical physics and undefined 2014, "Perspective: Fifty years of density-functional theory in chemical physics," pubs.aip.org, vol. 140, pp. 18–301, 2014, doi: 10.1063/1.4869598.

[67] K. Ryczko, K. Mills, I. Luchak, … C. H.-C. M., and undefined 2018, "Convolutional neural networks for atomistic systems," ElsevierK Ryczko, K Mills, I Luchak, C Homenick, I TamblynComputational Materials Science, 2018•Elsevier, 2018, Accessed: Dec. 20, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025618301526

[68] M. Rupp, A. Tkatchenko, K. Müller, O. V. L.-P. review letters, and undefined 2012, "Fast and accurate modeling of molecular atomization energies with machine learning," APS, vol. 108, no. 5, Jan. 2012, doi: 10.1103/PhysRevLett.108.058301.

[69] G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, and A. Fazzio, "From DFT to machine learning: recent approaches to materials science–a review," Journal of Physics: Materials, vol. 2, no. 3, p. 032001, May 2019, doi: 10.1088/2515-7639/AB084B.

[70] J. Behler, "First Principles Neural Network Potentials for Reactive Simulations of Large Molecular and Condensed Systems," Angewandte Chemie - International Edition, vol. 56, no. 42, pp. 12828–12840, Oct. 2017, doi: 10.1002/ANIE.201703114.

[71] R. Ramprasad, R. Batra, G. Pilania, A. Mannodi-Kanakkithodi, and C. Kim, "Machine learning in materials informatics: recent applications and prospects," nature.com R Ramprasad, R Batra, G Pilania, A Mannodi-Kanakkithodi, C Kimnpj Computational Materials, 2017•nature.com, vol. 3, p. 54, 2017, doi: 10.1038/s41524-017-0056-5.

[72] P. Schwaller et al., "Machine intelligence for chemical reaction space," Wiley Interdiscip Rev Comput Mol Sci, vol. 12, no. 5, p. e1604, Sep. 2022, doi: 10.1002/WCMS.1604.

[73] A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé, and A. Laio, "Unsupervised learning methods for molecular simulation data," ACS Publications A Glielmo, BE Husic, A Rodriguez, C Clementi, F Noé, A LaioChemical Reviews, 2021•ACS Publications, vol. 121, no. 16, pp. 9722–9758, Aug. 2021, doi: 10.1021/acs.chemrev.0c01195.

[74] M. Sajjan et al., "Quantum machine learning for chemistry and physics," Chem Soc Rev, vol. 51, no. 15, pp. 6475–6573, Aug. 2022, doi: 10.1039/D2CS00203E.

[75] F. Strieth-Kalthoff, F. Sandfort, M. Kühnemund, F. R. Schäfer, H. Kuchen, and F. Glorius, "Machine Learning for Chemical Reactivity: The Importance of Failed Experiments," Angewandte Chemie International Edition, vol. 61, no. 29, p. e202204647, Jul. 2022, doi: 10.1002/ANIE.202204647.

[76] R. Jinnouchi, F. Karsai, and G. Kresse, "On-the-fly machine learning force field generation: Application to melting points," Phys Rev B, vol. 100, no. 1, Jul. 2019, doi: 10.1103/PHYSREVB.100.014105.

[77] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl, "Neural message passing for quantum chemistry," proceedings. mlr.pressJ Gilmer, SS Schoenholz, PF Riley, O Vinyals, GE DahlInternational conference on machine learning, 2017•proceedings.mlr.press, 2017, Accessed: Dec. 20, 2024. [Online]. Available: https://proceedings.mlr.press/v70/gilmer17a

[78] E. Podryabinkin, A. S.-C. M. Science, and undefined 2017, "Active learning of linearly parametrized interatomic potentials," Elsevier EV Podryabinkin, AV Shapeev Computational Materials Science, 2017•Elsevier, Accessed: Dec. 20, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025617304536

[79] V. Hassija et al., "Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence," Cognit Comput, vol. 16, no. 1, pp. 45–74, Jan. 2024, doi: 10.1007/S12559-023-10179-8/FIGURES/14.

****