Computation and machine learning for materials: Past, present, and future perspectives

Sobin Alosious, Meng Jiang, and Tengfei Luo*

Computational methods and machine learning (ML) are reshaping materials science by accelerating their discovery, design, and optimization. Traditional approaches such as density functional theory and molecular dynamics have been instrumental in studying materials at the atomic level. However, their high computational cost and, in certain cases, limited accuracy can restrict the scope of *in silico* exploration. ML promises to accelerate material property prediction and design. However, in many areas, the volume and fidelity of the data are critical barriers. Active learning can reduce the reliance on large data sets, and simulation has emerged as a critical tool for generating data on the fly. Despite these advances, challenges remain, particularly in data quality, model interpretability, and bridging the gap between computational predictions and experimental validation. Future research should develop automated frameworks capable of designing and testing materials for specific applications, and integrating ML with traditional simulations and experiments can contribute to this goal.

Introduction

The evolution of computational materials science has been shaped by several key milestones, creating today's data-driven trend. Initially, density functional theory (DFT) established itself as a tool for calculating electronic structures and materials properties, providing a foundation for quantitative prediction.[1,2] The introduction of high-throughput methodologies further expanded the field, with initiatives such as the Materials Project systematically generating large data sets and shifting research from isolated simulations to large-scale computational frameworks.[3,4] This progression from traditional simulations to data-driven discovery is illustrated in **Figure 1**, which outlines the historical development of computational materials science. The field progressed from early DFT and molecular dynamics (MD) simulations to high-throughput screening, accelerating materials discovery. The integration of machine learning (ML) has further transformed the landscape by enabling predictive modeling, surrogate simulations, and efficient exploration of complex material spaces. Current research has increasingly explored AI-driven discovery, generative models, and autonomous material design. As the volume and complexity of the data increased, traditional manual analysis methods became insufficient for uncovering patterns in large material data sets. This challenge led to the widespread adoption of ML techniques, which have been instrumental in extracting insights, identifying hidden correlations, and building surrogate models for rapid property prediction.[5,6] Materials cartography, which leverages these techniques, enables exploration, visualization, and analysis of high-dimensional material spaces by capturing structural and electronic relationships in an interpretable way.[7]

This article is organized into three main sections. We begin by exploring the evolution of computational materials science, tracing its progression from traditional atomistic simulations to the rise of ML-driven materials informatics. We then examine how ML has been integrated with established computational techniques, enhancing the efficiency and accuracy of property predictions, materials screening, and discovery. Finally, we discuss recent advancements,

Sobin Alosious, Lucy Family Institute for Data & Society, University of Notre Dame, Notre Dame, USA ; Department of Aerospace and Mechanical Engineering, University of Notre Dame, Notre Dame, USA; sobinalosious@gmail.com; salosiou@nd.edu
Meng Jiang, Lucy Family Institute for Data & Society, University of Notre Dame, Notre Dame, USA ; Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, USA; mjiang2@nd.edu
Tengfei Luo, Lucy Family Institute for Data & Society, University of Notre Dame, Notre Dame, USA ; Department of Aerospace and Mechanical Engineering, University of Notre Dame, Notre Dame, USA; Department of Chemical and Biomolecular Engineering, University of Notre Dame, Notre Dame, USA; Center for Sustainable Energy at Notre Dame (ND Energy), University of Notre Dame, Notre Dame, USA; tluo@nd.edu
*Corresponding author
doi:10.1557/s43577-025-00959-y

© The Author(s) 2025

persistent challenges, and future directions, focusing on the interplay between ML and high-throughput simulations, the potential of emerging technologies, and the critical role of interdisciplinary collaborations in advancing the field.

Historical perspectives: Evolution of computational materials science

Classical molecular simulations

The evolution of computational methods in materials science is deeply rooted in classical simulation techniques that have shaped modern studies. MD simulations, first introduced by Alder and Wainwright[8,9] and later expanded by Rahman,[10] provided a foundation for modeling atomic interactions in condensed matter. Similarly, Monte Carlo (MC) methods, pioneered by Metropolis et al.[11] revolutionized statistical sampling in computational physics. Building on these foundations, advancements in simulation techniques improved computational efficiency and statistical accuracy.

The advent of parallel computing significantly accelerated computational progress. Plimpton[12] developed efficient parallel algorithms for MD simulations, enabling larger-scale studies with improved performance across short-range and long-range interactions. These contributions established a robust computational framework, driving algorithm development advancements and high-performance materials science simulation. These advancements have enabled the evolution of atomistic simulations from modeling simple toy systems to simulating realistic materials.

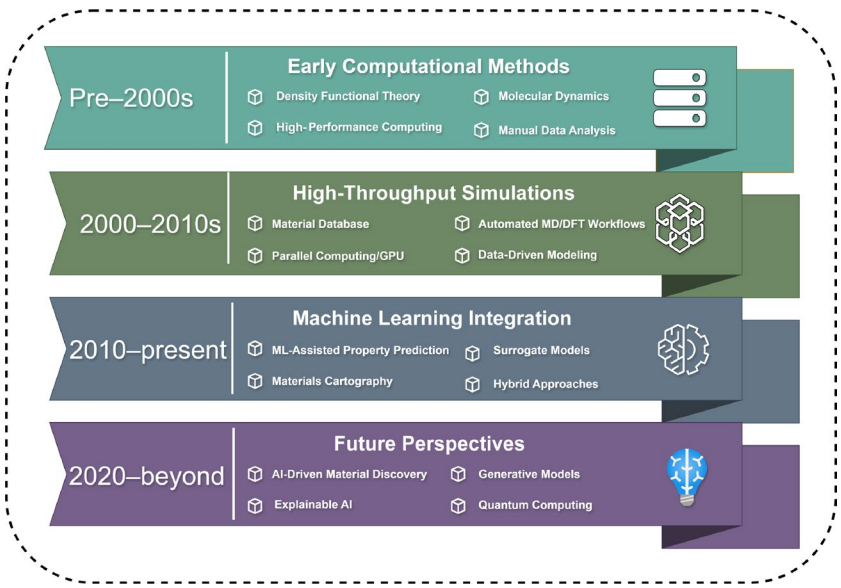


Figure 1. A schematic diagram illustrates the evolution of computational materials science, from early density functional theory (DFT) and molecular dynamics (MD) methods to the integration of high-throughput simulations and machine learning (ML). It also highlights artificial intelligence (AI)-driven discovery, generative models, and quantum computing as key future directions for accelerating materials innovation.

DFT

The development of DFT represented a breakthrough in computational materials science. Hohenberg and Kohn[13] established that the ground-state properties of an interacting electron system are uniquely determined by its electron density, forming the theoretical foundation of DFT. This concept was later made practical by Kohn and Sham,[2] who introduced self-consistent equations that account for exchange and correlation effects while remaining computationally efficient. These foundational works have shaped modern electronic-structure calculations. Over time, DFT has undergone continuous improvements, with Burke[14] offering a critical assessment of its strengths and limitations, particularly in addressing strongly correlated systems and excited state properties.

Beyond electronic-structure calculations, DFT has been integrated with MD, leading to the development of approaches such as Born-Oppenheimer molecular dynamics (BOMD) and Car-Parrinello molecular dynamics (CPMD). BOMD[15] computes interatomic forces at each time step using self-consistent DFT calculations, ensuring accuracy while balancing computational efficiency. In contrast, CPMD[16] introduces a fictitious mass term to propagate electronic degrees of freedom alongside nuclear motion, allowing for a more efficient treatment of electron–ion interactions while preserving DFT-level accuracy. Additionally, recent studies have explored using ML to improve the accuracy of DFT itself. For example, Snyder et al.[17] used ML to aid in developing new density functionals, enhancing predictive performance in electronic-structure calculations. These advancements have significantly broadened the scope of DFT, enabling the simulation of dynamic processes in complex systems and reinforcing its role as a fundamental tool in materials discovery.

Introduction of ML in materials research

In recent years, ML has become an emerging tool in materials research, complementing traditional simulation techniques. Butler et al.[18] provided a comprehensive review of how ML surrogate models can predict materials properties by learning from existing data. One of the most significant advancements in this area is the development of neural network potentials (NNPs), introduced by Behler and Parrinello,[19] which leverage DFT data to train NNPs to model interatomic interactions. Unlike conventional force fields, NNPs can capture complex many-body interactions with high accuracy, improving the reliability of MD simulations while enabling large-scale studies.

Generative models for inverse design have further expanded the role of ML in materials discovery. As demonstrated by Sanchez-Lengeling and Aspuru-Guzik,[20] these models can explore vast chemical spaces and propose novel materials with tailored properties that traditional methods may overlook. Integrating ML with high-throughput virtual screening (HTVS) has also played a key role by enabling the rapid identification of promising candidates for experimental validation. St. John et al.[21] presented a new database of approximately 91,000 candidate molecules for organic photovoltaic applications, including large structures and extrapolated properties for long polymer chains. Their work showed that message-passing neural networks could achieve accurate predictions without optimized 3D molecular geometries, making them well suited for scalable HTVS pipelines. Omar et al.[22] provided a detailed review of HTVS workflows in organic electronics, highlighting how systematic decisions throughout the screening pipeline, from candidate generation to property evaluation, can influence reliability and uncover valuable structure to property relationships beyond just top-performing materials. They further emphasized the importance of data set quality and size in determining prediction accuracy and described how HTVS continues to evolve alongside advances in data, methods, and applications.

Core computational techniques in materials science

MD simulations

MD simulations offer an atomistic perspective on the time evolution of materials by numerically solving Newton's equations of motion. For a system with N particles, each particle i with mass m_i follows:

$$m_i \ddot{\mathbf{r}}_i = \mathbf{F}_i = -\nabla_{\mathbf{r}_i} U(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N), \tag{1}$$

where $\mathbf{r}_i$ represents the position vector and U denotes the total potential energy. This equation governs the microscopic dynamics of molecular systems and serves as the foundation for MD simulations.[8,9]

A typical MD workflow consists of several key steps. First, an initial atomic configuration is generated, often using experimental crystallographic data or MC methods for disordered systems. Next, an appropriate interatomic potential or force field is selected to describe atomic interactions, such as the Lennard–Jones (L–J) potential[23] for noble gases or the embedded atom method (EAM) for metals.[24] Time evolution is handled using numerical integration schemes, with the Verlet algorithm being one of the most efficient and stable methods. The basic position update follows:

$$\mathbf{r}(t + \Delta t) = 2\mathbf{r}(t) - \mathbf{r}(t - \Delta t) + \frac{\Delta t^2}{m}\mathbf{F}(t), \tag{2}$$

which ensures time reversibility and energy conservation over long simulations. Alternative schemes, such as the leap-frog method, offer similar stability with minor differences in velocity calculations. Different ensemble techniques are available to model realistic thermodynamic conditions. For instance, the Nosé–Hoover thermostat regulates temperature by introducing auxiliary variables that couple the system to a heat bath, allowing for accurate sampling of the canonical (NVT) ensemble.[25] MD simulations have broad applications across materials science. In polymers, MD simulations provide insights into chain dynamics, entanglement, and viscoelastic properties.[26,27] In metals, MD has been used to study phenomena such as dislocation nucleation and grain-boundary motion, as well as mechanisms of phase transformation under shock.[28] Despite these advantages, MD is inherently limited by time scales typically spanning nanoseconds. Additionally, the accuracy of MD predictions depends heavily on the accuracy and transferability of the chosen force field.[29]

DFT

DFT provides a more rigorous quantum mechanical approach to determining the electronic structure of many-body systems. By reducing the complex problem of interacting electrons to an effective one-particle formulation based on the electron density $n(\mathbf{r})$, DFT simplifies calculations while maintaining accuracy. The foundation of this approach is established by the Hohenberg–Kohn theorems, which state that the ground-state properties of an electron system are uniquely determined by its density. In practical applications, the total energy of a system is expressed as a functional of the electron density:

$$E[n] = T_s[n] + \int V_{\text{ext}}(\mathbf{r})\, n(\mathbf{r})\, d\mathbf{r} + E_H[n] + E_{xc}[n], \tag{3}$$

where $T_s[n]$ represents the kinetic energy of noninteracting electron gas, $V_{\text{ext}}(\mathbf{r})$ is the external potential from atomic nuclei, $E_H[n]$ is the classical Coulomb (Hartree) energy, and $E_{xc}[n]$ accounts for many-body interactions beyond classical electrostatics.[30]

To make DFT computationally efficient, Kohn and Sham introduced an approach that reformulates the problem into a set of single-particle equations:[2]

$$\left[-\frac{\hbar^2}{2m}\nabla^2 + V_{\text{eff}}(\mathbf{r})\right]\psi_i(\mathbf{r}) = \epsilon_i \psi_i(\mathbf{r}), \tag{4}$$

where the effective potential is given by

$$V_{\text{eff}}(\mathbf{r}) = V_{\text{ext}}(\mathbf{r}) + V_H(\mathbf{r}) + V_{xc}(\mathbf{r}). \tag{5}$$

The electron density is reconstructed from the Kohn–Sham orbitals:

$$n(\mathbf{r}) = \sum_{i}^{\text{occ}} |\psi_i(\mathbf{r})|^2. \tag{6}$$

The self-consistent field (SCF) method, typically solved through iterative minimization techniques, ensures the convergence of electron density and total energy.

DFT has become an essential tool for predicting electronic band structures, densities of states, and charge distributions, providing crucial insights into the electronic properties of materials. Despite its success, DFT has limitations. Approximations to the exchange-correlation functional, such as the local density approximation (LDA) and the generalized gradient approximation (GGA), tend to underestimate bandgaps and struggle with strongly correlated systems.[31] These shortcomings have driven the development of hybrid functionals and many-body perturbation methods to improve accuracy. Nevertheless, DFT remains a powerful method for exploring a wide range of materials properties, from metal conductivity to semiconductor behavior and magnetism.

ML potentials

ML potentials have significantly improved atomistic simulations by offering a flexible, data-driven alternative to empirical interatomic potentials. Traditional models, such as L-J or EAM potentials, rely on fixed analytical expressions that often struggle to capture the full complexity of many-body interactions. In contrast, ML potentials learn these interactions directly from high-fidelity training data, typically obtained from first-principles DFT calculations, leading to notable improvements in accuracy and transferability. **Figure 2** outlines how ML-based interatomic potentials are developed. It starts with a structural database, where DFT-calculated energies and forces are assigned to atomic configurations. These structures are converted into descriptors, numerical representations of atomic environments that serve as input for ML models. The trained models learn the potential energy surface (PES), capturing complex many-body interactions beyond traditional force fields. Once developed, these ML potentials enable efficient MD and MC simulations, significantly accelerating large-scale atomistic modeling while maintaining high accuracy.

A widely used approach represents the total energy E of a system as the sum of atomic contributions:

$$E = \sum_{i=1}^{N} E_i, \tag{7}$$

where each atomic energy E_i is estimated based on a descriptor D_i that encodes the local atomic environment. NNPs, as discussed by Behler and Parrinello,[19] utilize high-dimensional descriptors and deep learning architectures to capture complex many-body effects that empirical potentials often fail to model. The mapping $E_i = \mathrm{NN}(D_i; \mathbf{w})$ is learned by optimizing the network weights $\mathbf{w}$ to minimize the difference between predicted and reference energies. Forces can be easily calculated using auto-differentiation of the neural network energy function.

Another widely used NNP-based approach is deep potential molecular dynamics (DPMD), developed by Zhang et al.[32] DPMD extends NNPs by employing hierarchical representations to improve accuracy and transferability across chemical environments. By leveraging deep neural networks, DPMD efficiently learns potential energy surfaces from DFT calculations, achieving near-first-principles accuracy while significantly reducing computational costs. However, while NNPs excel in accuracy, they often require large data sets and significant computational resources for training. To address this, alternative ML approaches, such as Gaussian process regression (GPR), offer probabilistic models that balance predictive power with built-in uncertainty quantification. In GPR, the energy of a configuration with descriptor x is given by

$$E(x) = \mathbf{k}(x, X)^{\top} \left(K + \sigma^2 I\right)^{-1} \mathbf{y}, \tag{8}$$

where X represents the training descriptors, $\mathbf{k}(x, X)$ is the kernel function, K is the covariance matrix computed from the training data, σ^2 is the noise variance, and $\mathbf{y}$ contains the reference energies.[33] This approach not only interpolates the PES but also identifies high-uncertainty regions, guiding further data acquisition for improved model performance.

Kernel methods, such as those incorporating the smooth overlap of atomic positions (SOAP) descriptor, further enhance ML potential accuracy by ensuring structural and symmetry-aware representations. Glielmo et al.[34] demonstrated that covariant kernels improve force predictions by enforcing physical symmetries. Similarly, the Gaussian approximation potential (GAP), introduced by Bartók et al.,[35] leverages kernel-based learning to construct interatomic force fields that surpass classical empirical models in accuracy. Another important kernel-based approach is moment tensor potentials (MTPs), introduced by Shapeev.[36] MTPs provide a systematically improvable framework by incorporating angular and radial features using moment tensor representations. Deringer et al.[37] demonstrate that ML potentials can achieve orders of magnitude speed increases compared to conventional DFT calculations while accurately modeling phase transitions, defect dynamics, and catalytic behavior. However, the predictive performance of these models is fundamentally limited by the quality and diversity of the training data. The model may fail to generalize reliably if relevant chemical environments are not represented.

Recent ML interatomic potentials (MLIPs) have increasingly adopted graph-based and symmetry-preserving architectures to improve accuracy and transferability. For instance, NequIP[38] utilizes E(3)-equivariant neural networks to learn force fields that respect physical symmetries, achieving state-of-the-art accuracy with significantly less training data. M3GNet[39] extends this concept to a general-purpose materials graph network trained on both structural and thermodynamic data, enabling property prediction and structure relaxation across a wide range of materials. MACE[40] combines equivariant message passing with atomic cluster expansions, delivering high data efficiency and predictive accuracy. These models outperform earlier MLIPs in many benchmarks but often come with higher training and inference costs, highlighting the tradeoff between physical fidelity and computational scalability.

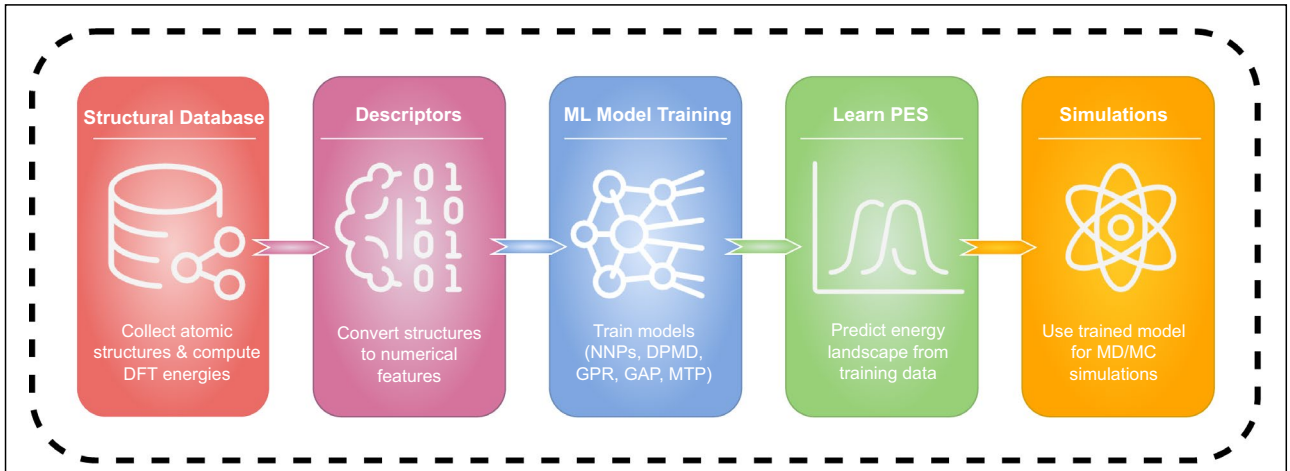


Figure 2. Workflow for constructing machine learning (ML) interatomic potentials. Atomic structures with density functional theory (DFT)-calculated energies are converted into descriptors, trained into ML models, and used for efficient molecular simulations. ML models include neural network potentials (NNPs), deep potential molecular dynamics (DPMD), Gaussian process regression (GPR), Gaussian approximation potentials (GAPs), and moment tensor potentials (MTPs). These models are used to learn the potential energy surface (PES) and accelerate molecular dynamics (MD) or Monte Carlo (MC) simulations.

To overcome the limitations of current ML potentials, recent efforts have focused on developing foundation models for materials simulations. One such approach is HIENet,[41] which introduces a hybrid architecture combining invariant and equivariant message-passing networks. This design allows the model to make consistent predictions of energies, forces, and stresses across a wide range of material systems while respecting key physical laws such as energy conservation and O(3) symmetry (i.e., invariance under 3D rotations and reflections). Such models aim to improve generalization and reduce the need for retraining in unseen chemical environments. Furthermore, because ML potentials use more complex ML models than simple empirical formulas in classical potentials, their computational efficiency is still lower than classical potentials. Nevertheless, the field is advancing rapidly, and ML potentials are increasingly becoming a practical tool for predictive and efficient atomistic modeling.

ML for materials property prediction, discovery, and inverse design

ML for polymeric materials

The integration of ML into polymer science has progressed from basic property prediction models to sophisticated high-throughput computational frameworks for accelerated materials discovery. One of the first notable contributions was the polymer genome, which introduced an ML-based approach for polymer property predictions.[47] This framework utilized curated data sets from experimental sources and DFT calculations, mapping hierarchical polymer fingerprints to properties such as the glass-transition temperature (T_g), dielectric constant, and bandgap.

Recognizing the limitations of small experimental data sets, Ma et al.[48] introduced PI1M, a benchmark database containing approximately one million generated polymer structures. The data set was built using a recurrent neural network (RNN) trained on polymers from the PoLyInfo database and is intended to support the development of polymer representations for ML. However, the PI1M database only contains polymer structures. To address this deficiency, Xu et al.[49] developed the second generation of PI1M, named POINT^2 (a Polymer Informatics Training and Testing Database), to include properties for each polymer. These properties were provided by different ML surrogate models trained on experimental or simulation data. The POINT^2 database also includes the uncertainty of the predicted properties, the interpretability of the ML models, and importantly, polymer synthesizability. Tao et al.[42] developed an ML framework to predict the T_g of polymers, leveraging molecular structure data for high-throughput screening. Their study demonstrated that integrating ML with molecular descriptors enables efficient identification of high-temperature polymers. **Figure 3**a shows the high-throughput screening of high-T_g polymers with a DNN fingerprint model. RadonPy, developed by Hayashi et al.,[50] automated high-throughput MD simulations to calculate thermal, mechanical, and electronic properties of amorphous polymers. The framework demonstrated how ML-assisted calibration could improve the accuracy of properties such as thermal conductivity (TC) and specific heat capacity. Additionally, RadonPy provided an automated pipeline for property computation.

Beyond static data set-based predictions, more adaptive frameworks such as active learning (AL) have emerged to improve discovery efficiency further. AL offers a different approach, one in which the ML surrogate model helps guide the learning process by dynamically suggesting or querying the "next-best experiment." Kim et al.[43] demonstrated how an AL framework can support the discovery of polymers with high T_g. The AL framework uses one of three decision-making strategies to choose the following experiment: exploitation, exploration, or a balanced combination of both. Their results showed that AL outperformed random selection in efficiently identifying

polymers that exceed a target T_g threshold. Figure 3b shows the relative performance of the four different strategies on changing the threshold temperature for consideration as a high-T_g polymer.

ML-driven high-throughput screening has been particularly impactful in the search for thermally conductive polymers. Ma et al.[44] combined high-throughput MD simulations and ML to identify polymers with relatively high TC (>0.3 W/mK). A set of 365 polymers was randomly selected from the PoLyInfo database, and their TCs were calculated using MD. These results were used to train a regression model capturing structure–TC relationships, which was then applied to screen the database. Finally, 121 polymers with MD-calculated TC above the threshold were identified. Selected polymers were reevaluated under different simulation conditions, and most retained TC above 0.3 W/mK. Figure 3c illustrates the workflow detailing monomer polymerization, chain replication, and TC computation via nonequilibrium MD (NEMD). Similarly, Xu et al.[51] used high-throughput MD simulations and AL to discover polymer blends with superior TC. They screened ~550,000 blends, identifying key correlations between TC, radius of gyration (R_g), and hydrogen bonding, which highlight effects from both inter- and intrachain interactions on TC.

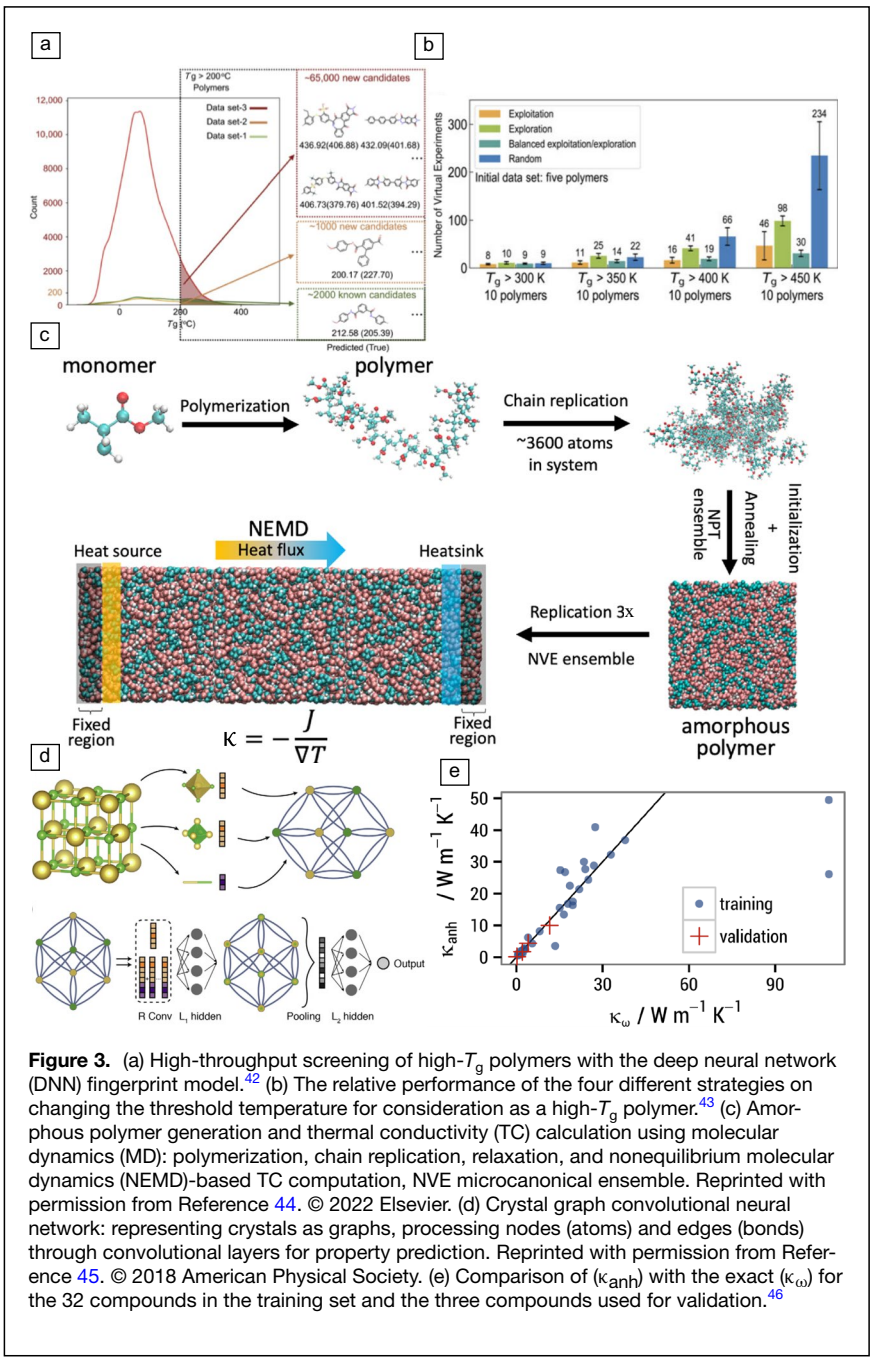


Figure 3. (a) High-throughput screening of high-T_g polymers with the deep neural network (DNN) fingerprint model.[42] (b) The relative performance of the four different strategies on changing the threshold temperature for consideration as a high-T_g polymer.[43] (c) Amorphous polymer generation and thermal conductivity (TC) calculation using molecular dynamics (MD): polymerization, chain replication, relaxation, and nonequilibrium molecular dynamics (NEMD)-based TC computation, NVE microcanonical ensemble. Reprinted with permission from Reference 44. © 2022 Elsevier. (d) Crystal graph convolutional neural network: representing crystals as graphs, processing nodes (atoms) and edges (bonds) through convolutional layers for property prediction. Reprinted with permission from Reference 45. © 2018 American Physical Society. (e) Comparison of (κ_{anh}) with the exact (κ_{ω}) for the 32 compounds in the training set and the three compounds used for validation.[46]

Beyond screening existing materials, ML is increasingly being applied in inverse design frameworks, enabling the generation of polymers that satisfy predefined property constraints. Kumar et al.[52] applied ML to predict and optimize the cloud point of poly(2-oxazoline) polymers. Their work integrated gradient boosting regression with particle swarm optimization, achieving accurate cloud-point tuning, thereby demonstrating the potential of inverse design in polymer formulation engineering. Ma et al.[53] employed a reinforcement learning approach to design amorphous polymers with high TC. They used an ML surrogate model trained on 469 TC values from high-throughput MD simulations as a surrogate predictor and an RNN trained on about one million virtual polymer structures as a polymer generator. The framework successfully generated polymers with predicted TC above 0.4 W/mK. Synthetic accessibility scores were used to assess synthesizability, and selected candidates were validated via MD simulations. The top-performing polymer reached a TC of 0.693 W/mK and was estimated to be easily synthesizable. Liu et al.[54,55] introduced the graph diffusion transformer (Graph DiT) for multi-conditional molecular generation, aiming to advance inverse design in materials and drug discovery. Graph DiT applies a novel graph-dependent noise model in the diffusion process and integrates a Transformer-based denoiser with an encoder that captures numerical and categorical property conditions. The model enables simultaneous control over multiple properties, such as synthetic accessibility and gas permeability. It

demonstrated superior performance across nine evaluation metrics. In a polymer inverse design task for gas separation, Graph DiT showed practical utility through expert feedback, highlighting its potential in real-world materials discovery.

While ML surrogate models require extensive training data, transfer learning (TL) has emerged as a promising approach to bridge gaps between low-fidelity and high-fidelity data sets, improving prediction reliability with minimal additional data. One early example of TL applied to polymer property prediction was demonstrated by Wu et al.,[56] in which a model trained on limited polymer data was used to identify candidates with high TC. Three polymers were synthesized from thousands of machine-generated candidates, achieving TCs up to 0.41 W/mK, comparable to those of state-of-the-art polymers in noncomposite thermoplastics. This study provided an early proof of concept for ML-guided polymer discovery and demonstrated the potential of TL in data-scarce scenarios. Liu et al.[57] developed semi-supervised learning techniques to address the label imbalance of polymer properties. The techniques were able to perform data augmentation in latent spaces, leveraging a large amount of unlabeled data.

In a comprehensive review, Xu et al.[58] surveyed ML techniques for designing polymeric membranes specifically for gas separation. They discussed that ML surrogate models, mainly supervised learning and TL, can accurately predict gas permeability and selectivity by leveraging structure–property relationships. Their review further highlighted how ML reduces experimental burden by identifying promising polymer candidates before synthesis, thereby streamlining material discovery for energy-efficient separations.

ML for inorganic crystalline materials

ML has also revolutionized inorganic crystalline materials, enabling rapid predictions of thermoelectric, mechanical, and electronic properties. Traditional computational methods such as DFT and MD have been widely used for studying thermoelectric properties, phonon transport, and materials stability, but their high computational cost limits scalability. ML addresses this limitation by offering a way to predict materials properties with much lower computational cost. While not a replacement for first-principles methods, ML surrogate models can deliver reasonable results for targeted problems, making them particularly useful for screening large materials spaces and accelerating the discovery process.

A major milestone in data-driven materials research was the introduction of the Materials Project.[59] Developed under the Materials Genome Initiative, this open-access database provides DFT-calculated properties for tens of thousands of inorganic compounds. It has enabled ML surrogate models to learn structure–property relationships, facilitating high-throughput screening for thermoelectrics, energy-storage, and electronics applications.

For novel material classes, ML accelerates the discovery of functional materials. Xie and Grossman[45] developed crystal graph convolutional neural networks (CGCNNs) to predict key properties, such as formation energy, bandgap, Fermi energy, and elastic moduli, directly from crystal structures. Their model was trained to reproduce DFT-calculated properties and achieved relatively small errors compared against the original DFT data, making it suitable for efficient property screening across large materials data sets. As shown in Figure 3d, CGCNN constructs a crystal graph in which nodes represent atoms and edges depict atomic connections within the unit cell.

ML has played a crucial role in thermoelectric materials research, where the goal is to optimize the thermoelectric figure of merit (ZT) by balancing the Seebeck coefficient, electrical conductivity, and TC. Carrete et al.[46] developed a high-throughput screening approach to identify half-Heusler compounds with low lattice TC (κ_L). By combining *ab initio* calculations, physical models, and ML techniques, they screened approximately 79,000 entries from the AFLOWLIB database. After filtering for mechanical and thermodynamic stability, they narrowed the list to 75 compounds and quantitatively estimated their κ_L values. Three semiconductors with $\kappa_L < 5$ W/mK were identified as promising candidates for experimental validation. Figure 3e shows a comparison between the simplified model predictions (κ_{anh}) and full anharmonic DFT calculations (κ_ω) for 32 training compounds and three validation cases. Wang et al.[60] provided a broader review of ML applications in thermoelectric research. It emphasized the role of feature selection, identifying atomic mass, bond length, and lattice symmetry as critical descriptors in thermoelectric property predictions.

Beyond thermoelectrics, ML has significantly advanced the study of phonon transport in inorganic crystalline materials. Accurate prediction of phonon dispersion and lattice TC (κ_L) is critical for designing thermal insulators (for thermoelectrics and thermal protection) and thermal conductors (for heat dissipation applications). Chen et al.[61] developed a GPR model trained on experimental data to predict κ_L. Their model identified bulk modulus, atomic mass, and bond length as key descriptors, achieving accuracy comparable to traditional semiempirical models while greatly reducing computational cost. Zhu et al.[62] expanded on this by integrating graph neural networks with random forest models to predict κ_L for all materials in the Inorganic Crystal Structure Database (ICSD). Their study introduced a structural chemistry map to classify materials based on TC characteristics and identified rare earth chalcogenides as promising thermoelectric candidates with ZT values exceeding 1.0. Liu et al. [63] demonstrated the effectiveness of TL in predicting phonon properties by leveraging more readily available electronic property data. They trained ML surrogate models on electronic bandgaps of 1245 semiconductors and transferred this knowledge to improve models trained on just 124 data points for phonon bandgap, group velocity, and heat capacity. TL significantly reduced the mean absolute error by 65%, 14%, and 54% for these three respective properties compared to direct learning. Liu et al.[64] employed TL to improve ML surrogate models for predicting the lattice TC of semiconductors. By using large low-fidelity data sets as

proxies for limited high-fidelity data, TL improved prediction accuracy by 23% in R^2 and reduced the average factor difference (AFD), a logarithmic metric that quantifies the typical order-of-magnitude error between predicted and reference values by 30 percent. Screening a semiconductor database, they identified candidates with TC > 350 W/mK, which were later validated by first-principles calculations. Luo et al.[65] reviewed ML applications in phonon transport, categorizing approaches into direct prediction models, which establish structure–property relationships without interatomic force constant (IFC) calculations and indirect methods that use ML interatomic potentials to accelerate phonon dispersion predictions. The review highlighted how ML significantly reduces computational costs while maintaining high accuracy, enabling large-scale materials screening.

Recent advances have also demonstrated ML's role in phase stability and crystal structure prediction. Zhang et al.[66] developed ML surrogate models using artificial neural networks (ANNs) and support vector machines (SVMs) to predict the phase formation of high-entropy ceramic carbides with 98.2% classification accuracy. These models utilize chemical descriptors to forecast single-phase stability in both equiatomic and nonequiatomic systems, offering targeted guidance for experimental synthesis.

Challenges, limitations, and opportunities

Despite the growing impact of ML in materials science, several key challenges persist, particularly in data quality, model interpretability, and extrapolation from computational predictions to experiments. Addressing these issues is essential for realizing the full potential of ML in accelerating materials discovery and optimization. One of the most significant challenges is data quality and standardization. Himanen et al.[68] point out that data sets from experiments and simulations often vary in formatting, accuracy, and completeness. The lack of standardized ontologies and effective data-sharing practices makes it hard to reproduce results reliably. Errors and biases in training data can also carry over into the models, reducing their overall predictive strength and generalizability. The accuracy of ML predictions for experimental synthesis mainly depends on the standardization, completeness, and diversity of data, but such data sets do not exist for most material classes.

A second major challenge is model interpretability and explainability. While traditional ML surrogate models, such as simple decision trees and linear regression, provide transparency in decision-making, more complex models, including ensemble methods such as random forests (RFs) and deep learning models such as convolutional and graph-based architectures, often function as black boxes, making their reasoning challenging to interpret. For instance, Ward et al.[6] showed that RFs can predict metallic glass formation accurately. However, while the model uses chemically meaningful features, the ensemble decision-tree structure is not easily interpretable, making it difficult to understand the physical reasons behind the predictions. This issue is even more pronounced in deep neural networks, which can learn complex patterns without clearly defined feature relationships. **Figure 4**a–b shows the experimental and predicted glass-forming ability in the Al–Ni–Zr ternary system, respectively. The model was not trained on any data from this system or its related binaries, yet it still produced predictions that matched experimental trends. This suggests that the model can generalize to new systems, but also highlights the difficulty in understanding why the predictions are reliable. To overcome these challenges, researchers recommend explainable artificial intelligence (XAI) techniques and hybrid approaches combining physics-based modeling with ML.

A third key challenge lies in extrapolating ML-trained computational models to experimental validation. Many ML-driven materials discovery efforts rely on data from DFT calculations and MD simulations, which may not fully capture experimental conditions. Peccati et al.[69] proposed incorporating uncertainty quantification by comparing results from multiple computational methods, such as Hartree-Fock and local density approximation, to provide confidence intervals in ML predictions. However, bridging the computational-to-experimental gap remains difficult even with improved computational accuracy. A closely related challenge is bridging the gap between computational screening and experimental synthesis. Although computational methods can identify thousands of potential materials, only a few prove experimentally viable. Chen et al.[67] illustrate this limitation in their ML-driven screening of 32 million candidates, where only 18 were successfully synthesized. For example, Figure 4c shows that electrochemical properties, such as the electrochemical stability window, are crucial for evaluating wide bandgap materials. Meanwhile, Figure 4d displays the self-diffusivity of lithium at various temperatures, categorized by anion type, further demonstrating the challenges in finding materials that can be realized in the laboratory.

Future directions and emerging trends

Advancing data-driven materials discovery

As we discussed, one of the most pressing challenges in materials informatics is the scarcity of high-quality, diverse data sets, which would preferably integrate experimental, computational, and textual data. Pyzer-Knapp et al.[72] reviewed how materials science is beginning to embrace multitask prompted training and instruction tuning approaches inspired by multimodal AI models such as Flamingo and LLaVA to develop foundational models tailored for materials discovery. Models such as Text-chemT5 and Regression Transformer have already demonstrated the potential of multimodal learning in improving materials property predictions across multiple data sources. Liu et al.[73] have developed multimodal large language models that use retrosynthetic planning to do inverse molecular or polymer design.

Improving interpretability in materials informatics

Despite ML's success in property prediction, model interpretability remains a critical challenge. Deep learning models such as GNNs have improved predictive accuracy but often still function as blackbox models, making it difficult to extract meaningful physical insights. To address this, future advancements will emphasize XAI, where models predict properties and provide human-understandable justifications for their decisions.

SHAP (SHapley Additive ExPlanations) has become a widely adopted tool for interpreting ML models by assigning feature importance values. Yang et al.[74] developed multitask deep learning models to predict gas permeabilities of polymer membranes and screened more than nine million hypothetical polymers for high-performance candidates. They used SHAP analysis to interpret the influence of chemical features on gas permeability in polymer membranes. They identified specific molecular descriptors and fingerprint substructures that positively or negatively affect permeability. The SHAP-derived insights aligned with established design heuristics, such as the role of methyl groups in increasing permeability and polar groups in enhancing selectivity. However, empirical and exact SHAP scores can produce misleading attributions, such as assigning high importance to irrelevant features and overlooking relevant ones. Recent analyses have uncovered fundamental theoretical and practical limitations in SHAP's ability to consistently reflect true feature relevance, particularly in complex models.[75] Additionally, attention mechanisms enhance interpretability by highlighting key features influencing predictions. For instance, Liu et al.[76,77] proposed a graph rationalization framework for molecular and polymer property prediction that identifies representative subgraph structures termed graph rationales that support GNN predictions. Their method introduces environment-based data augmentation to improve rationale extraction and avoids expensive graph decoding by learning representations in latent space. This approach improves the interpretability and generalization of graph-based models in materials science applications.

Multimodal learning also plays a crucial role in improving model interpretability. The Materials Informatics Transformer (MatInFormer) has demonstrated the ability to predict materials properties using textual descriptions of structures and properties.[78] Vu et al.[70] developed SCANN, an interpretable deep learning framework for materials informatics that not only predicts materials properties but also reveals the underlying structure–property relationships through attention mechanisms. In SCANN, local atomic representations are learned and combined into a global material representation using computed global attention (GA) scores. These scores quantify the importance of each local structure in determining the materials properties, providing physically meaningful insights into the data. **Figure 5**a shows GA scores estimated by the model for atomic sites in the $SmFe_{12}$, $SmFe_{11}Mo$, $SmFe_{11}Co$, and $SmFe_{11}Al$ crystal structures. In addition, the correlation between the ratio of GA scores of the substitution sites to the minimum GA scores among the Fe sites and the formation energy is illustrated in Figure 5b.

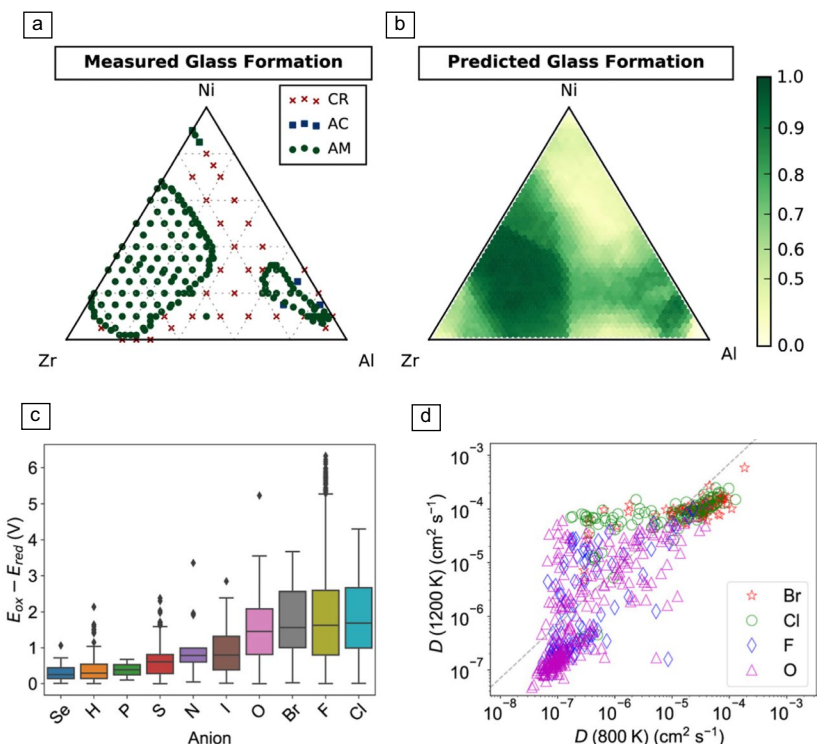


Figure 4. (a) Experimental measurements of metallic glass-forming ability in the Al–Ni–Zr ternary. (b) Predicted glass-forming ability from the machine learning surrogate model. Reprinted from Reference 6 under CC BY 4.0. (c) Electrochemical stability window distribution for 8795 candidates. (d) Lithium self-diffusivity at 800 and 1200 K for 583 materials, grouped by anion type. Reprinted with permission from Reference 67. © 2024 American Chemical Society.

Bridging computational predictions with experimental validation

A major challenge in ML-driven materials discovery is ensuring that computational predictions translate into experimentally realizable materials. While ML surrogate models trained on DFT and MD simulations have significantly accelerated computational screening, discrepancies between simulated and real-world conditions often limit their experimental applicability.

To bridge this gap, AL strategies are emerging as a powerful tool for improving model extrapolation. Ren et al.[79] demonstrated that integrating

AL with automation enhances experimental efficiency by prioritizing high-value experiments and mitigating uncertainties. Additionally, AI-driven autonomous labs, operating as modular cloud facilities, provide a scalable approach to optimizing synthesis and characterization workflows. Self-driving laboratories (SDLs) represent a transformative step forward in this direction. Beaucage et al.[80] highlight how these platforms leverage ML to automate synthesis and characterization, refining experimental conditions in real time.

The role of quantum computing in materials science

Quantum computing is poised to be a transformative tool in materials science, unlocking new possibilities for property prediction, material simulations, and inverse design. Integrating quantum algorithms with ML presents a promising solution, enhancing both accuracy and efficiency in materials discovery.

Melnikov et al.[81] explored the integration of quantum variational circuits into neural networks to improve materials property predictions. Their study demonstrated that hybrid quantum-classical models enhance generalization while reducing computational overhead. By leveraging quantum random walks, they identified structural patterns in material graphs, aiding in the prediction of crystal structures and material composition.

Beyond property prediction, quantum computing is also applied in materials design and optimization. Kitai et al.[82]

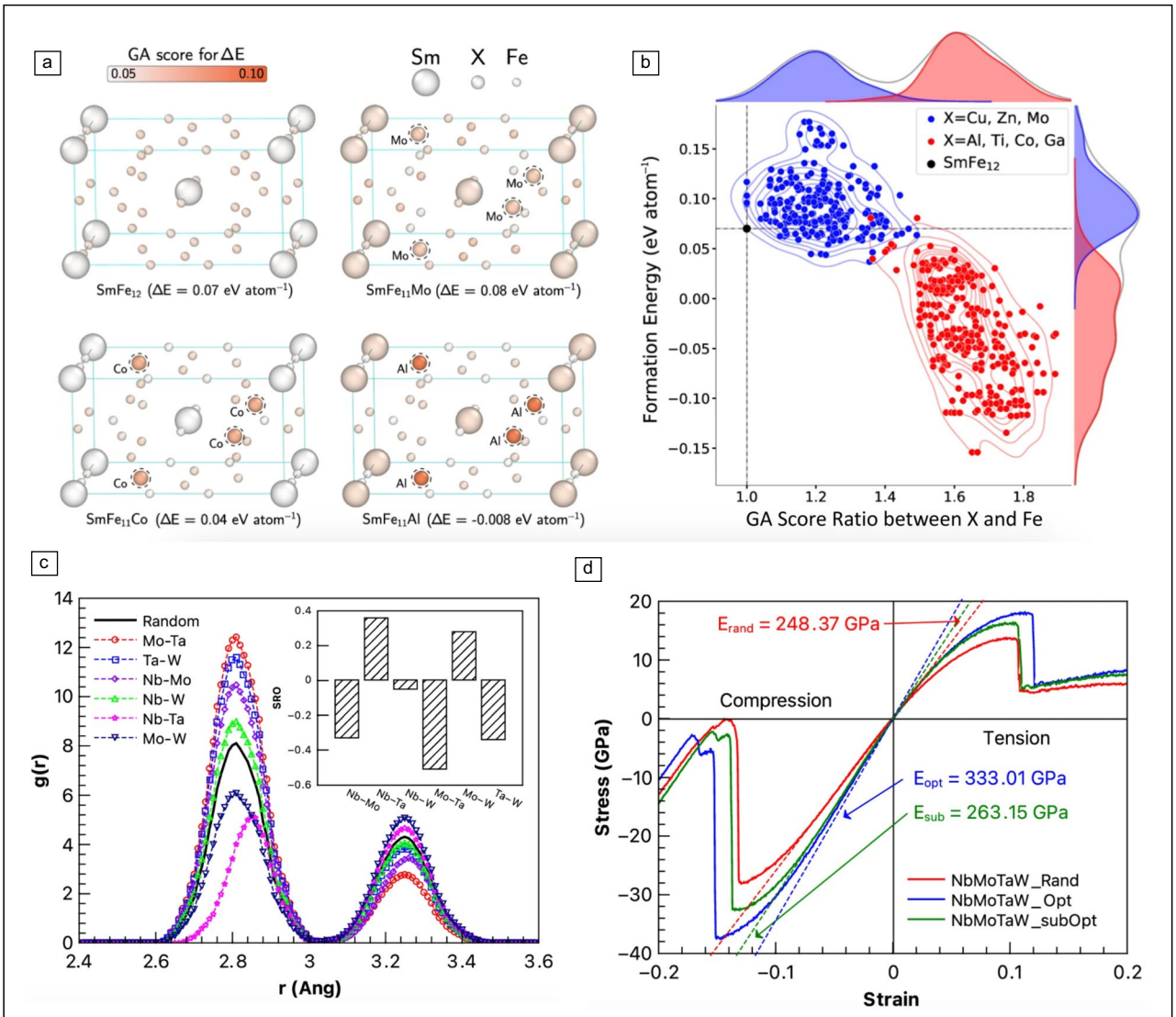


Figure 5. (a) Visualization of the global attention (GA) scores estimated by the model for atomic sites in the $SmFe_{12}$, $SmFe_{11}Mo$, $SmFe_{11}Co$, and $SmFe_{11}Al$ crystal structures. (b) Correlation between the ratio of GA scores of the substitution sites to the minimum GA scores among the Fe sites and the formation energy. Reprinted from Reference 70 under CC BY 4.0. (c) Partial radial distribution function (RDF) and the short-range orders (SRO) for optimized NbMoTaW lattice structure. (d) Uniaxial stress–strain behavior of NbMoTaW systems. Reprinted from Reference 71 under CC BY-NC-ND.

demonstrated the potential of quantum annealing for optimizing complex metamaterial architectures, such as wavelength-selective radiators. Their method, implemented on the D-Wave 2000Q quantum annealer, efficiently navigated high-dimensional design spaces, outperforming conventional inverse design approaches. Kim et al.[83–85] have applied quantum-assisted AL and optimization in multiple studies targeting advanced optical and energy materials. Their work includes the design of a visually transparent radiative cooler achieving 86.3 MJ/m^2 annual energy savings, titanium nitride metastructures with greater than 95% solar absorptance, and ultrathin (130 nm) optical diodes with 85% forward and 0.1% backward transmissivity. Xu et al.[71] introduced the quantum annealing assisted lattice optimization (QALO) framework for alloy design, combining ML models, quantum annealing, and energy calculations. Applied to NbMoTaW high-entropy alloys, QALO reproduced experimental trends such as Nb depletion and W enrichment, enhancing mechanical properties. As shown in Figure 5c–d, its predictions align with RDF analysis and stress–strain simulations, demonstrating its accuracy.

Despite its potential, quantum computing still faces challenges such as decoherence and hardware noise.[86] Another challenge is the limited number of qubits and their connectivities in any existing quantum computers. Future progress will depend on better error mitigation, more fully connected qubits, hybrid quantum-classical models, and integration with high-performance computing. As these technologies evolve, their synergy with ML and inverse materials design could unlock materials beyond the reach of classical computation.

Conclusions

Integrating computational methods and ML is transforming materials science, enabling faster, more accurate, and scalable approaches to material discovery and property prediction. Traditional techniques such as DFT and MD have provided the backbone of computational materials research, but their high computational costs often limit large-scale exploration. ML-based models now offer a powerful complement, allowing researchers to predict materials properties, optimize structures, and even perform inverse design to create materials tailored to specific applications with shorter development times.

ML enhances property prediction across various materials, from polymers to inorganic crystals. High-throughput ML frameworks make predicting mechanical, thermal, and electronic properties more efficient, and the increase in data volume and advancements in ML model have increasingly improved model accuracy. These in turn accelerate the identification of thermoelectric, electronic, and energy-storage materials among others. In the meantime, generative models open new pathways for inverse design, enabling *in silico* discovery before synthesis.

Despite these advances, challenges remain. Data quality, model interpretability, and bridging the gap between computational predictions and real-world experiments are ongoing concerns. Efforts in uncertainty quantification, XAI, and AL are helping improve simulation reliability and foster stronger collaboration between computational and experimental researchers. Emerging technologies such as self-driving laboratories and AI-assisted synthesis may further automate discovery, making experimental workflows more efficient.

Looking ahead, the convergence of ML, automation, and quantum computing has the potential to further reshape materials science. By addressing current limitations and fostering interdisciplinary collaboration, these innovations will accelerate the discovery of functional materials across various industries, from energy storage and electronics to biomaterials and sustainable materials development. The future of computational materials science is about fundamentally transforming how materials are designed and realized rapidly, and this may be achieved with the help of ML techniques.

Author contributions

S.A., M.J., and T.L. contributed equally to this work. All authors were involved in the conceptualization, literature search, original draft preparation, and review and editing of the manuscript.

Funding

Lucy Family Institute for Data & Society, Postdoctoral Fellowship and NSF Grant (Nos. 2332270 and 2102592).

Code availability

Not applicable.

Conflict of interest

The authors have no conflicts to disclose.

Ethical approval

Not applicable.

Consent for publication

Not applicable.

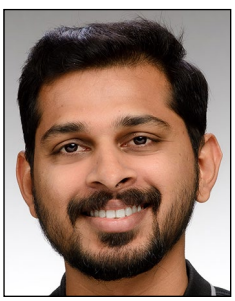
Open Access

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. W. Kohn, *Rev. Mod. Phys.* **71**, 1253 (1999)
2. W. Kohn, L.J. Sham, *Phys. Rev.* **140**, A1133 (1965)
3. S. Curtarolo, G.L.W. Hart, M.B. Nardelli, N. Mingo, S. Sanvito, O. Levy, *Nat. Mater.* **12**, 191 (2013)
4. S.P. Ong, W.D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V.L. Chevrier, K.A. Persson, G. Ceder, *Comput. Mater. Sci.* **68**, 314 (2013)
5. G. Pilania, C. Wang, X. Jiang, S. Rajasekaran, R. Ramprasad, *Sci. Rep.* **3**, 2810 (2013)
6. L. Ward, A. Agrawal, A. Choudhary, C. Wolverton, *NPJ Comput. Mater.* **2**, 16028 (2016)
7. S.B. Torrisi, M.Z. Bazant, A.E. Cohen, M.G. Cho, J.S. Hummelshøj, L. Hung, G. Kamat, A. Khajeh, A. Kolluru, X. Lei, H. Ling, J.H. Montoya, T. Mueller, A. Palizhati, B.A. Paren, B. Phan, J. Pietryga, E. Sandraz, D. Schweigert, Y. Shao-Horn, A. Trewartha, R. Zhu, D. Zhuang, S. Sun, *APL Mach. Learn.* **1**, 020901 (2023)
8. B.J. Alder, T.E. Wainwright, *J. Chem. Phys.* **27**, 1208 (1957)
9. B.J. Alder, T.E. Wainwright, *J. Chem. Phys.* **31**, 459 (1959)
10. A. Rahman, *Phys. Rev.* **136**, A405 (1964)
11. N. Metropolis, S. Ulam, *J. Am. Stat. Assoc.* **44**, 335 (1949)
12. S. Plimpton, *J. Comput. Phys.* **117**, 1 (1995)
13. P. Hohenberg, W. Kohn, *Phys. Rev.* **136**, B864 (1964)
14. K. Burke, *J. Chem. Phys.* **136**, 150901 (2012)
15. R.N. Barnett, U. Landman, *Phys. Rev. B* **48**, 2081 (1993)
16. R. Car, M. Parrinello, *Phys. Rev. Lett.* **55**, 2471 (1985)
17. J.C. Snyder, M. Rupp, K. Hansen, K.-R. Müller, K. Burke, *Phys. Rev. Lett.* **108**, 253002 (2012)
18. K.T. Butler, D.W. Davies, H. Cartwright, O. Isayev, A. Walsh, *Nature* **559**, 547 (2018)
19. J. Behler, M. Parrinello, *Phys. Rev. Lett.* **98**, 146401 (2007)
20. B. Sanchez-Lengeling, A. Aspuru-Guzik, *Science* **361**, 360 (2018)
21. P.C. St. John, C. Phillips, T.W. Kemper, A.N. Wilson, Y. Guan, M.F. Crowley, M.R. Nimlos, R.E. Larsen, *J. Chem. Phys.* **150**, 234111 (2019)
22. Ö.H. Omar, M. Del Cueto, T. Nematiaram, A. Troisi, *J. Mater. Chem. C* **9**, 13557 (2021)
23. J.E. Lennard, I. Jones, *Proc. Royal Soc. London. Series A, Contain. Papers Math. Phys. Charact.* **106**, 441 (1924)
24. M.S. Daw, M.I. Baskes, *Phys. Rev. B* **29**, 6443 (1984)
25. S. Nosé, *Mol. Phys.* **52**(2), 255 (1984)
26. X. Wei, T. Luo, *Phys. Chem. Chem. Phys.* **21**, 15523 (2019)
27. S. Sen, S.K. Kumar, P. Keblinski, *Macromolecules* **38**, 650 (2005)
28. V.I. Levitas, R. Ravelo, *Proc. Natl. Acad. Sci. U.S.A.* **109** (2012), pp. 13204–13207
29. A.R. Leach, *Molecular Modelling: Principles and Applications*, 2nd edn. (Pearson Education, Harlow, 2001)
30. R.M. Martin, *Electronic Structure: Basic Theory and Practical Methods*, 2nd edn. (Cambridge University Press, Cambridge, 2020)
31. A.J. Cohen, P. Mori-Sánchez, W. Yang, *Science* **321**, 792 (2008)
32. L. Zhang, J. Han, H. Wang, R. Car, E. Weinan, *Phys. Rev. Lett.* **120**, 143001 (2018)
33. C.E. Rasmussen, C.K.I. Williams, *Gaussian Processes for Machine Learning Adaptive Computation and Machine Learning* (MIT Press, Cambridge, 2006)
34. A. Glielmo, P. Sollich, A. De Vita, *Phys. Rev. B* **95**, 214302 (2017)
35. A.P. Bartók, M.C. Payne, R. Kondor, G. Csányi, *Phys. Rev. Lett.* **104**, 136403 (2010)
36. A.V. Shapeev, *Multiscale Model. Simul.* **14**, 1153 (2016)
37. V.L. Deringer, M.A. Caro, G. Csányi, *Adv. Mater.* **31**, 1902765 (2019)
38. S. Batzner, A. Musaelian, L. Sun, M. Geiger, J.P. Mailoa, M. Kornbluth, N. Molinari, T.E. Smidt, B. Kozinsky, *Nat. Commun.* **13**, 2453 (2022)
39. C. Chen, S.P. Ong, *Nat. Comput. Sci.* **2**, 718 (2022)
40. I. Batatia, D.P. Kovacs, G. Simm, C. Ortner, G. Csányi, *Adv. Neural Inf. Process. Syst.* **35**, 11423 (2022)
41. K. Yan, M. Bohde, A. Kryvenko, Z. Xiang, K. Zhao, S. Zhu, S. Kolachina, D. Sarıtürk, J. Xie, R. Arroyave, X. Qian, X. Qian, S. Ji, A materials foundation model via hybrid invariant-equivariant architectures (2025), Preprint, http://arxiv.org/abs/2503.05771
42. L. Tao, G. Chen, Y. Li, *Patterns* **2**, 100225 (2021)
43. C. Kim, A. Chandrasekaran, A. Jha, R. Ramprasad, *MRS Commun.* **9**(3), 860 (2019)
44. R. Ma, H. Zhang, J. Xu, L. Sun, Y. Hayashi, R. Yoshida, J. Shiomi, J.-X. Wang, T. Luo, *Mater. Today Phys.* **28**, 100850 (2022)
45. T. Xie, J.C. Grossman, *Phys. Rev. Lett.* **120**, 145301 (2018)
46. J. Carrete, W. Li, N. Mingo, S. Wang, S. Curtarolo, *Phys. Rev. X* **4**, 011019 (2014)
47. C. Kim, A. Chandrasekaran, T.D. Huan, D. Das, R. Ramprasad, *J. Phys. Chem. C* **122**, 17575 (2018)
48. R. Ma, T. Luo, *J. Chem. Inf. Model.* **60**, 4684 (2020)
49. J. Xu, G. Liu, R. Guo, M. Jiang, T. Luo, POINT^2: A polymer informatics training and testing database (2025), Preprint, http://arxiv.org/abs/2503.23491
50. Y. Hayashi, J. Shiomi, J. Morikawa, R. Yoshida, *NPJ Comput. Mater.* **8**, 222 (2022)
51. J. Xu, T. Luo, *NPJ Comput. Mater.* **10**, 74 (2024)
52. J.N. Kumar, Q. Li, K.Y.T. Tang, T. Buonassisi, A.L. Gonzalez-Oyarce, J. Ye, *NPJ Comput. Mater.* **5**, 73 (2019)
53. R. Ma, H. Zhang, T. Luo, *ACS Appl. Mater. Interfaces* **14**, 15587 (2022)
54. G. Liu, E. Inae, T. Zhao, J. Xu, T. Luo, M. Jiang, *Adv. Neural Inf. Process. Syst.* **36**, 21039 (2023)
55. G. Liu, J. Xu, T. Luo, M. Jiang, Graph diffusion transformers for multi-conditional molecular generation (2024), Preprint, http://arxiv.org/abs/2401.13858
56. S. Wu, Y. Kondo, M.-A. Kakimoto, B. Yang, H. Yamada, I. Kuwajima, G. Lambard, K. Hongo, Y. Xu, J. Shiomi, C. Schick, J. Morikawa, R. Yoshida, *NPJ Comput. Mater.* **5**, 66 (2019)
57. G. Liu, T. Zhao, E. Inae, T. Luo, M. Jiang, "Semi-Supervised Graph Imbalanced Regression," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, 2023), pp. 1453–1465
58. J. Xu, A. Suleiman, G. Liu, R. Zhang, M. Jiang, R. Guo, T. Luo, *Chem. Phys. Rev.* **5**, 041311 (2024)
59. A. Jain, S.P. Ong, G. Hautier, W. Chen, W.D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, K.A. Persson, *APL Mater.* **1**, 011002 (2013)
60. T. Wang, C. Zhang, H. Snoussi, G. Zhang, *Adv. Funct. Mater.* **30**, 1906041 (2020)
61. L. Chen, H. Tran, R. Batra, C. Kim, R. Ramprasad, *Comput. Mater. Sci.* **170**, 109155 (2019)
62. T. Zhu, R. He, S. Gong, T. Xie, P. Gorai, K. Nielsch, J.C. Grossman, *Energy Environ. Sci.* **14**, 3559 (2021)
63. Z. Liu, M. Jiang, T. Luo, *Sci. Adv.* **6**, eabd1356 (2020)
64. Z. Liu, M. Jiang, T. Luo, *Mater. Today Phys.* **28**, 100868 (2022)
65. Y. Luo, M. Li, H. Yuan, H. Liu, Y. Fang, *NPJ Comput. Mater.* **9**, 4 (2023)
66. J. Zhang, B. Xu, Y. Xiong, S. Ma, Z. Wang, Z. Wu, S. Zhao, *NPJ Comput. Mater.* **8**, 5 (2022)
67. C. Chen, D.T. Nguyen, S.J. Lee, N.A. Baker, A.S. Karakoti, L. Lauw, C. Owen, K.T. Mueller, B.A. Bilodeau, V. Murugesan, M. Troyer, *J. Am. Chem. Soc.* **146**, 20009 (2024)
68. L. Himanen, A. Geurts, A.S. Foster, P. Rinke, *Adv. Sci.* **6**, 1900808 (2019)
69. F. Peccati, R. Laplaza, J. Contreras-García, *J. Phys. Chem. C* **123**, 4767 (2019)
70. T.-S. Vu, M.-Q. Ha, D.-N. Nguyen, V.-C. Nguyen, Y. Abe, T. Tran, H. Tran, H. Kino, T. Miyake, K. Tsuda, H.-C. Dam, *NPJ Comput. Mater.* **9**, 215 (2023)
71. Z. Xu, W. Shang, S. Kim, E. Lee, T. Luo, *NPJ Comput. Mater.* **11**, 4 (2025)
72. E.O. Pyzer-Knapp, M. Manica, P. Staar, L. Morin, P. Ruch, T. Laino, J.R. Smith, A. Curioni, *NPJ Comput. Mater.* **11**, 61 (2025)
73. G. Liu, M. Sun, W. Matusik, M. Jiang, J. Chen, Multimodal large language models for inverse molecular design with retrosynthetic planning (2024), Preprint, http://arxiv.org/abs/2410.04223
74. J. Yang, L. Tao, J. He, J.R. McCutcheon, Y. Li, *Sci. Adv.* **8**, eabn9545 (2022)
75. X. Huang, J. Marques-Silva, *Int. J. Approx. Reason.* **171**, 109112 (2024)
76. G. Liu, T. Zhao, J. Xu, T. Luo, M. Jiang, "Graph Rationalization with Environment-based Augmentations", in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, 2022), pp. 1069–1078
77. G. Liu, E. Inae, T. Luo, M. Jiang, *ACM Trans. Knowl. Discov. Data* **18**(4), 86 (2024)
78. H. Huang, R. Magar, C. Xu, A.B. Farimani, Materials informatics transformer: A language model for interpretable materials properties prediction (2023), Preprint, http://arxiv.org/abs/2308.16259
79. Z. Ren, Z. Ren, Z. Zhang, T. Buonassisi, J. Li, *Nat. Rev. Mater.* **8**, 563 (2023)
80. P.A. Beaucage, D.R. Sutherland, T.B. Martin, *Macromolecules* **57**, 8661 (2024)
81. A. Melnikov, M. Kordzanganeh, A. Alodjants, R.-K. Lee, *Adv. Phys. X* **8**, 2165452 (2023)
82. K. Kitai, J. Guo, S. Ju, S. Tanaka, K. Tsuda, J. Shiomi, R. Tamura, *Phys. Rev. Res.* **2**, 013319 (2020)
83. S. Kim, W. Shang, S. Moon, T. Pastega, E. Lee, T. Luo, *ACS Energy Lett.* **7**, 4134 (2022)
84. S. Kim, S. Wu, R. Jian, G. Xiong, T. Luo, *ACS Appl. Mater. Interfaces* **15**, 40606 (2023)
85. S. Kim, S.-J. Park, S. Moon, Q. Zhang, S. Hwang, S.-K. Kim, T. Luo, E. Lee, *Nano Converg.* **11**, 16 (2024)
86. J. Preskill, *Quantum* **2**, 79 (2018) □

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



Sobin Alosious is a postdoctoral researcher at the Lucy Family Institute for Data & Society and the Department of Aerospace and Mechanical Engineering at the University of Notre Dame. His current research focuses on leveraging machine learning for materials discovery using molecular simulations. He earned his PhD degree in nanofluidics through a joint program between the Indian Institute of Technology Madras, India, and Swinburne University of Technology, Australia. His research interests include nanofluidics, molecular dynamics, nanoscale energy transport, and interfacial thermal transport. Alosious can be reached by email at sobinalosious@gmail.com or salosiou@nd.edu.



Meng Jiang is an associate professor in the Department of Computer Science and Engineering at the University of Notre Dame. He is the director of the Foundation Models Laboratory at the Lucy Family Institute for Data & Society. His research focuses on designing, developing, and evaluating advanced machine learning technologies that discover knowledge and patterns from graph and text data. He has published more than 150 peer-reviewed articles, most of which are accepted to top international conferences such as ICML, NeurIPS, ICLR, AAAI, KDD, and ACL. He is a senior member of the Association for Computing Machinery and IEEE. Jiang can be reached by email at mjiang2@nd.edu.



Tengfei Luo is the Dorini Family Professor and associate chair in the Department of Aerospace and Mechanical Engineering at the University of Notre Dame. He is a Fellow of the Lucy Family Institute for Data & Society. His research focuses on polymer physics and uses multiscale modeling techniques, from first-principles calculations and molecular simulations to mesoscale models. He has been working on polymer informatics with the focus of thermal-transport properties. He has published more than 180 journal articles. He is a Fellow of The American Society of Mechanical Engineers. Luo can be reached by email at tluo@nd.edu.