mater.scichina.com link.springer.com

Published online 12 March 2024 | https://doi.org/10.1007/s40843-023-2836-0
Sci China Mater 2024, 67(4): 1082–1100

SPECIAL TOPIC: Computation-assisted Materials Screening and Design

Recent advances in machine learning interatomic potentials for cross-scale computational simulation of materials

Nian Ran[1,2], Liang Yin[1,2,3], Wujie Qiu[1,2,4*] and Jianjun Liu[1,2,3*]

ABSTRACT **In recent years, machine learning interatomic potentials (ML-IPs) have attracted extensive attention in materials science, chemistry, biology, and various other fields, particularly for achieving higher precision and efficiency in conducting large-scale atomic simulations. This review, situated in the ML-IP applications in cross-scale computational models of materials, offers a comprehensive overview of structure sampling, structure descriptors, and fitting methodologies for ML-IPs. These methodologies empower ML-IPs to depict the dynamics and thermodynamics of molecules and crystals with remarkable accuracy and efficiency. More efficient and advanced techniques from interdisciplinary research field play an important role in opening a wide spectrum of applications spanning diverse temporal and spatial dimensions. Therefore, ML-IP method renders the stage for future research and innovation promising revolutionary opportunities across multiple domains.**

Keywords: machine learning interatomic potential, cross-scale computational simulation, structure sampling, encoding structure, fitting method

INTRODUCTION

Interactions between atoms play a pivotal role in many scientific and technological fields, as they significantly influence the behavior, properties, and dynamic characteristics of matters [1–3]. Understanding the principles governing atomic interactions is essential for investigating various aspects such as mechanical properties, thermodynamic stability, and electronic structures [4]. However, from the atomic scale to the macroscopic scale, phenomena in different scale ranges are intertwined with each other, forming a complex multi-scale system. In this context, the urgency to develop a methodology capable of accurately depicting atomic interactions while efficiently managing computations across multiple scales becomes evident.

Traditionally, the elucidation of atomic interactions involved the solution of the Schrödinger equation, which captures the quantum mechanical behavior of atoms and molecules. The solution of this equation provides insights into IPs and wavefunctions, forming the basis for predicting material properties and dynamics. However, the complexity of the Schrödinger equation arises from its treatment of the many-body problem, involving the positions, momenta, and intricate IP of numerous atoms [5–7]. This computational complexity presents challenges for investigating large-scale systems and extends the time required for research. To address these challenges, classical potential models have been widely adopted for atomic simulations [8]. These models have been widely adopted for atomic simulations due to their ability to represent the fundamental physics of interatomic interactions using simplified mathematical formulations. These models incorporate empirical parameters, which are often derived from experimental measurements or quantum mechanical computations. While such models abstract many-body complexities, they retain the ability to predict physical properties that are critical to material behavior, such as equilibrium bond lengths, bond angles, and cohesive energies. In systems where the interactions are well-understood and can be approximated by pairwise or simple many-body terms, classical potentials provide a computationally efficient means to simulate material properties. Despite their utility, classical potentials are limited by their empirical nature and often fail to capture complex quantum mechanical effects and precise interactions in systems with higher complexity, thus limiting their applications in material systems containing complex interactions such as defects, surfaces, phase transitions, and dipoles.

The introduction of machine learning interatomic potentials (ML-IPs) presents a novel approach to overcome these limitations [9–12]. This method utilizes ML algorithms, such as neural networks, to learn potential energy functions for atomic interactions from extensive datasets encompassing atomic structures and properties. By inputting atomic positions, types, and other relevant information S of the system, the model f can produce the corresponding potential energy E, expressed as $E = f(S)$ [9]. Through continuous refinement of model parameters, the ML-IPs gradually converge towards accurate approximations of quantum mechanical results. ML-IP methods offer distinct advantages. Primarily, they exhibit good transferability. Drawing insights from abundant data on atomic structures and properties, these methods accurately capture intricate interatomic interactions, demonstrating improved accuracy and adaptability

[1] State Key Laboratory of High Performance Ceramics and Superfine Microstructure, Shanghai Institute of Ceramics, Chinese Academy of Sciences, Shanghai 200050, China
[2] Center of Materials Science and Optoelectronics Engineering, University of Chinese Academy of Sciences, Beijing 100049, China
[3] School of Chemistry and Materials Science, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou 310024, China
[4] School of Mathematics, Physics and Statistics, Shanghai Polytechnic University, Shanghai 201209, China
* Corresponding authors (emails: wjqiu@sspu.edu.cn (Qiu W); jliu@mail.sic.ac.cn (Liu J))

across various material systems. Additionally, these methods offer significant advantages in computational efficiency. Conventional density functional theory (DFT) methods rely on iterative electronic density calculations, but they face limitations when applied to large systems due to their high computational demands. ML-IPs, through their data-driven paradigm, circumvent the necessity for explicit electron density computations, thus permitting more efficient simulations across diverse scales. By extrapolating knowledge from extensive datasets of quantum mechanical calculations, ML-IPs predict material behaviors across disparate scales, avoiding the computational complexity inherent in traditional methods. ML-IPs also offer significant advantages over conventional simulation techniques such as continuum mechanics, which often neglect atomic-scale phenomena crucial to material behavior, and Monte Carlo simulations, which become exceedingly time-consuming for intricate systems. What distinguishes ML-IP is their exceptional capacity to provide intricate atomic-scale insights while accommodating larger system sizes without proportionately increasing computational time. Coupled with their predictive accuracy and adaptability, ML-IPs emerge as an indispensable tool for computational simulations spanning different scales in materials science.

Over the last few years, significant progress has been made in the precision of potential energy predictions, thereby enabling simulations of complex phenomena such as phase transitions, surface reactions, and defect dynamics with high accuracy. This progress is attributed to the utilization of sophisticated ML architectures, such as graph neural networks and kernel methods, significantly enhancing the representation and predictability of atomic-scale interactions [13]. Notably, the evolution of descriptors has been pivotal in improving the accuracy of ML-IPs. Descriptors serve as critical tools for transforming information about atomic environments into formats amenable to ML algorithms. Recent research has expanded from simple symmetry functions and fixed geometric parameters to more flexible and intricate descriptors, such as high-dimensional Fourier series descriptors, many-body distribution functions, and local environment fingerprints [14–16]. These advanced descriptors can capture more detailed physical information and adapt to varying types of material systems and chemical environments. Furthermore, the advancements in sampling techniques, such as Metadynamics sampling, play an equally important role in the progress of ML-IPs [17]. These sampling methods are crucial for exploring the potential energy surfaces (PES) of complex systems and training ML-IPs to accurately predict rare events or states that occur infrequently in simulations. Metadynamics is a powerfully enhanced sampling technique, which facilitates the escape from metastable states and enables the exploration of the PES more thoroughly by iteratively adding a history-dependent potential to the PES. When combined with ML-IPs, Metadynamics can explore complex PES with improved efficiency and accuracy, as the ML models can predict the energy and forces in each new region explored, guiding the sampling process. Incorporating these advanced sampling techniques in ML-IP training ensures that models are not solely reliant on common configurations but also proficiently predict less frequent yet crucial events like chemical reactions, phase transitions, or material defects. Overall, these developments in model architecture, descriptors and sampling techniques contribute to the creation of more accurate and predictive ML-IPs, facilitating the development of cross-scale computational designs for materials.

In this review, we delved into the fundamental principles of ML-based potentials and their applications in addressing urgent needs in cross-scale computations. Specifically, we discussed methods for structure sampling, atomic environment descriptor construction, and structure-property fitting. Furthermore, we explored the applications of these methods in the design and discovery of materials across various domains. ML-IP methods strike a harmonious balance between computational efficiency and accuracy, thereby facilitating efficient material optimization, catalyst design, ion liquid research, drug discovery, and more. They promote the application of computational materials science in intricate structure design and performance prediction, providing fresh perspectives and methodologies for research in materials science and engineering.

NATURE OF CROSS-SCALE COMPUTATION BASED ON ML-IPs: $P = F[D(S)]$

In the realm of materials science, a recurring and fundamental question frequently emerges: how can we predict and understand the performance of materials, ultimately leading to the design of superior materials, starting from the intrinsic atomic structure of these materials? This question involves navigating the intricate relationship between the structural description of materials at the atomic level and their macroscopic performance [18]. To address this challenge, we have formulated the concept of "micro-to-macro mapping", which tightly links the structural (S) description (D) of a material to its performance (P) through a mapping function (F). Specifically, within the framework of ML-IPs, performance is typically described by quantity with quantum mechanical precision, such as energy, forces, electronic structures, and other physical properties. These parameters are of paramount interest when conducting computational simulations. For instance, accurate predictions of energy are required to determine the thermal stability or reactivity of a material. Likewise, understanding the interatomic interactions within the material, even including higher-order force derivatives, is crucial for simulating mechanical properties, diffusion dynamics, and heat transfer processes. Therefore, performance serves as the central focus, and ML-IP proves to be indispensable in accurately predicting these performance attributes.

In this process, the function (F) serves as a mapping or modeling tool responsible for linking the structural description ($D(S)$) of a material to its performance (P). This function can take the form of an ML model, such as a neural network, or other mathematical models like regression models [19,20]. The correlation between structure and performance is established through training this function, enabling accurate predictions and optimizations of material performance. Lastly, it is crucial to consider the structural description (D) of material. When crafting a structural description ($D(S)$), several key aspects typically need to be considered. First and foremost is the description of crystal structures, including lattice parameters and atomic coordinates, to precisely define the crystal structure of the material. In addition, it is essential to specify the types of atoms in the material and their relative quantities, i.e., the chemical composition. The symmetry of the crystal must also be described, including point group and space group symmetries. Studies of electronic properties often require a description of the electron density to illustrate the distribution of electrons within

the material [21]. Furthermore, the description of local environments is also critical, especially for electrocatalytic materials, which can be achieved through parameters such as coordination numbers, bond lengths, and bond angles to elucidate the interactions between atoms [22–24]. For amorphous or disordered materials, structural description may become more complex, involving additional information such as disorder, coordination diversity, and local preferences [25–27]. All these structural characters are considered to enable computational scale extension from microscopic into mesoscopic scale. It is well-known that mesoscopic simulations for domain structures usually are based on considering the bulk chemical energy, interfacial energy, elastic energy, and structural change energy from external fields. Therefore, ML-IP approach is significantly promising to realize cross-scale computation from microscopic to mesoscopic scales in material field.

To sum up, we tightly interconnect the above three key elements: by means of the function (F), we map the structural description of the material ($D(S)$) to its performance (P), as shown in Fig. 1. This relationship can be expressed by the following formula:

$$P = F[D(S)]. \quad (1)$$

This concise formula reveals a fundamental concept that the representation of material performance is not isolated but rather relies on the nonlinear mapping of structural description *via* a high-dimensional function F. This is also the essence of cross-scale computation: by mapping atomic-level information to macroscopic performance, we can provide powerful tools for material design and development. Furthermore, this approach can not only facilitate the discovery of new materials, but also enhance the performance of existing material.

In the subsequent sections, we will take an in-depth discussion of the three tightly coupled key elements of structure, performance, and function, along with an introduction to various methods, techniques, and their practical applications in material research. Through these discussions, we will gain a clearer understanding of how to leverage structure-property relationships to advance materials science, particularly in the realm of battery material research.

STRUCTURE SAMPLING—S

Achieving high-quality structure sampling on PESs is a crucial and fundamental task in the development of high-precision ML-IPs. It represents the pivotal and computationally resource-intensive step within the comprehensive framework of constructing ML-IPs. In recent years, a multitude of PES sampling techniques have flourished across virtually every scientific domain, spanning disciplines such as chemistry, condensed matter physics, materials science, and biology. These methodologies primarily include (Fig. 2): (1) static structure sampling: encompassing structural prediction approaches,

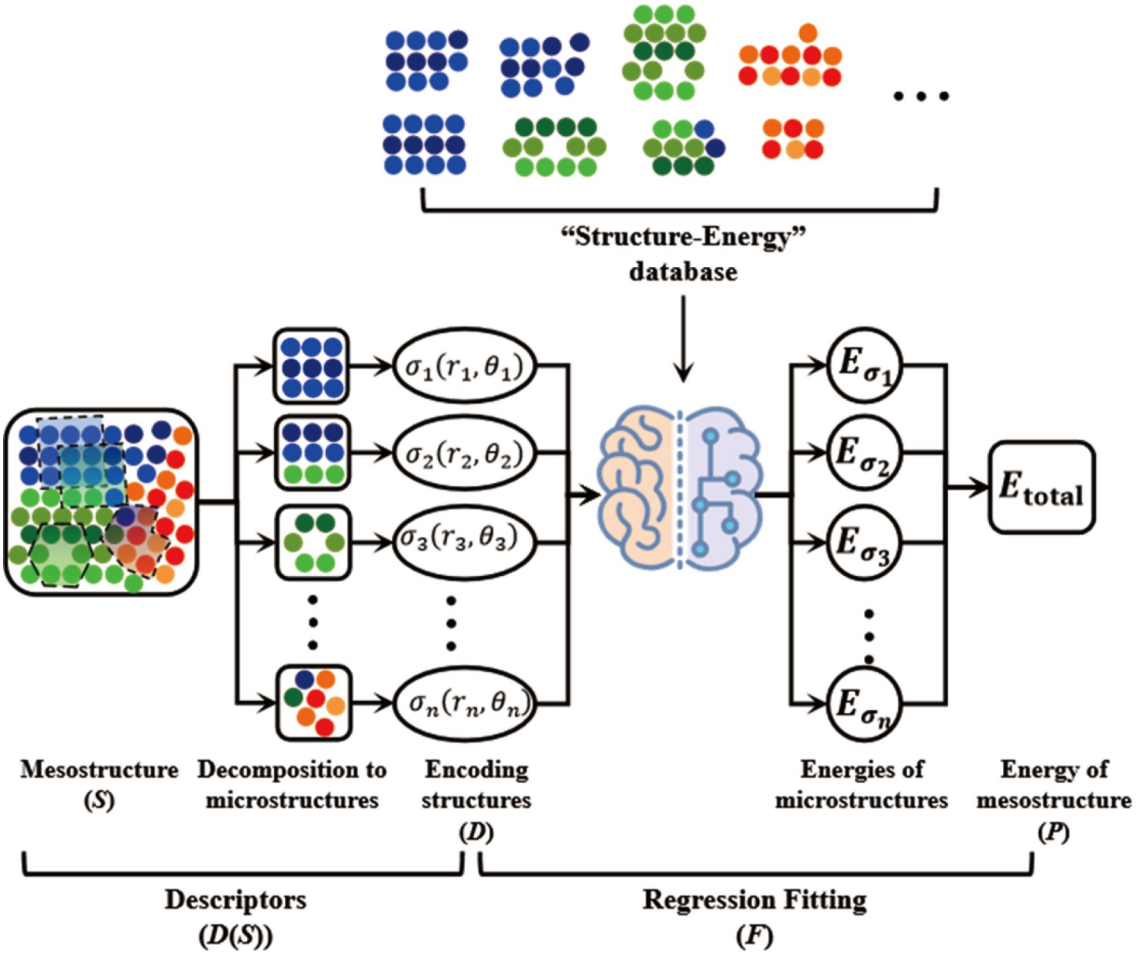

Figure 1 Cross-scale computational approach for microscale precision calculation of mesoscopic structures.

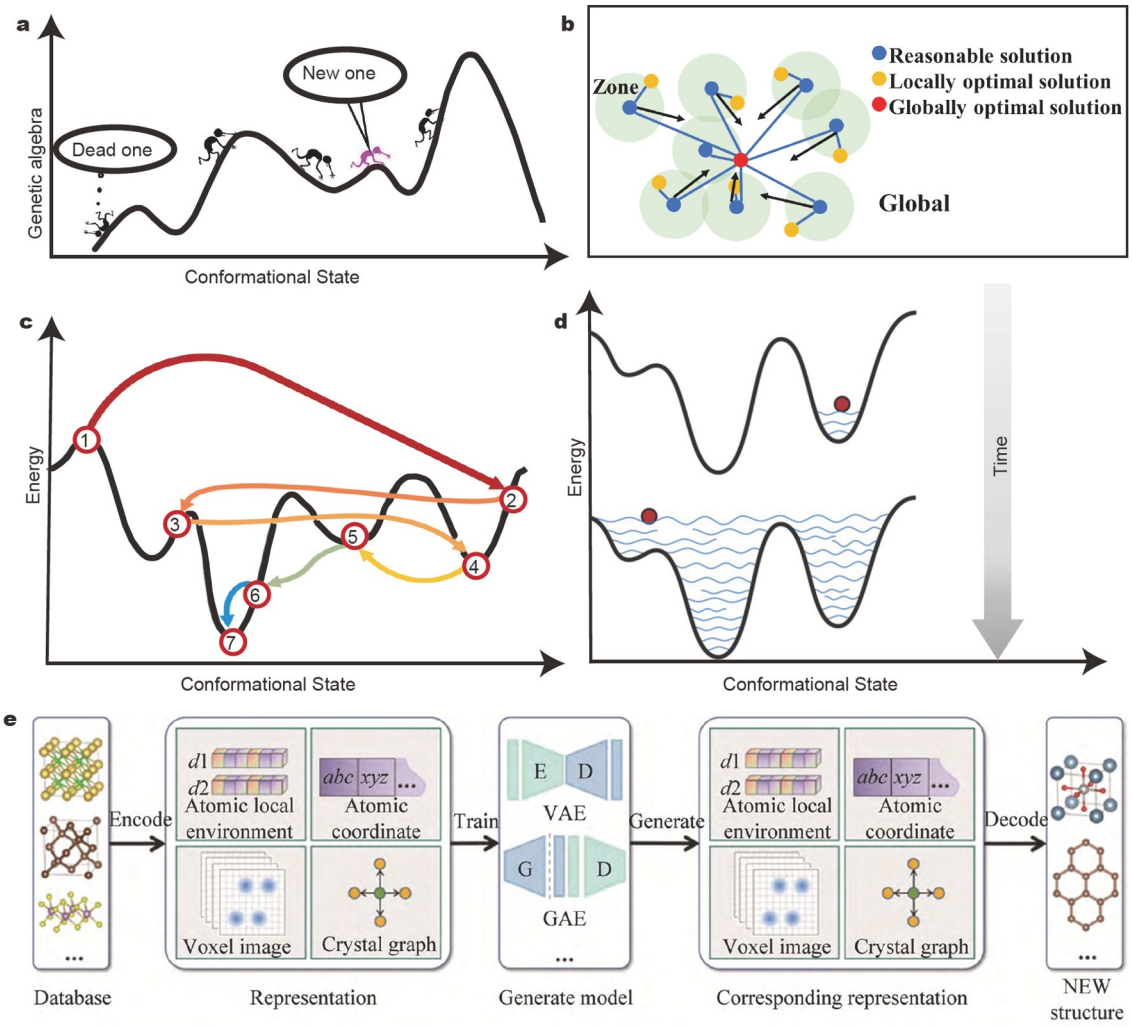

Figure 2 Structure sampling. (a) Static structure sampling of genetic algorithms. (b) Static structure sampling of particle swarm optimization. (c) Dynamic structure sampling of simulated annealing. (d) Dynamic structure sampling of Metadynamics. (e) Unsupervised learning sampling of generative models. Reprinted with permission from Ref. [39]. Copyright 2023, Chinese Ceramic Society.

including but not limited to basin hopping [28], genetic algorithms [29], and particle swarm optimization [30]. (2) Dynamic structure sampling: encompassing the realm of molecular dynamics (MD), exemplified by simulated annealing [9,31], as well as enhanced sampling methodologies like Metadynamics [32,33] and iMD-VR [34], which introduce external potentials to facilitate the escape of the system from local energy minima. (3) Unsupervised learning sampling: incorporating innovative techniques such as generative adversarial networks (GAN) [35], variational autoencoders (VAE) [36], flow models [37], and diffusion models [38]. These cutting-edge approaches have been intricately designed to overcome the formidable challenges associated with structure sampling, significantly advancing scientific investigations across diverse fields of study.

Static structure sampling

Static structure sampling is a crucial method for exploring stable configurations on PESs. This research area primarily focuses on seeking the lowest energy configurations and predicting crystal structures and molecular conformations. Basin hopping is a widely used static structure sampling technique, especially useful for finding local energy minima. This method operates on the concept of initiating from an initial structure and gradually descending to a local minimum on the PES through incremental structural adjustments. This process can be iteratively performed to explore diverse structures by searching for different local minima. Basin hopping has proven successful in structure prediction applications, including metal clusters [40], molecules [41], and crystals [42].

Genetic algorithms, depicted in Fig. 2a, draw inspiration from the biological evolution process and serve as another static structure sampling method. They generate a collection of structures with specific genetic characteristics and simulate evolutionary processes, including natural selection, crossover, and mutation, to iteratively optimize these structures in pursuit of the lowest-energy configuration. Genetic algorithms often demand substantial computational resources but excel in locating global minima for complex structures [43] and large systems [44].

As shown in Fig. 2b, particle swarm optimization is a static structure sampling method based on swarm intelligence. In this approach, a group of "particles" represent different structures, and they collectively search for the lowest-energy configuration on the PES. The particles interact with each other, adjusting

their positions by simulating the behavior of flocks of birds or schools of fish. This collaborative search method efficiently locates local minima on the energy surface. CALYPSO, as one of the representatives of particle swarm optimization, has been widely used in fields such as cluster structure prediction [45] and two-dimensional (2D) layer structure prediction [46].

In summary, these methods transform the PES by bypassing transition regions between local minima, enabling rapid global minimum searches. They typically derive the PES data from structural relaxation trajectories, potentially missing critical reaction pathways within transition regions.

Dynamic structure sampling

Dynamic structure sampling involves methods that explore the PES by simulating the dynamic behavior of a system. Unlike static structure sampling, which primarily focuses on finding the lowest energy configurations, dynamic structure sampling takes into account the evolution of structures over time. This method is particularly important for studying processes involving structural changes and transitions on the PES, especially for large conformational change involving chemical reactions. One of the prominent techniques in dynamic structure sampling is MD [9,31,47,48]. MD simulations (Fig. 2c) involve numerically solving the equations of motion for a system of atoms or molecules. By starting from an initial configuration and applying forces based on the PES, MD simulations allow researchers to track the trajectory of a system as it explores various configurations. This approach provides insights into the kinetics and thermodynamics of structural transitions. Simulated annealing, as a special variant of MD, has distinct advantages in finding the system's lowest energy configuration [49]. It starts from an initial high-temperature state and gradually cools the system, allowing it to explore different energy basins on the PES. As the temperature decreases, the system becomes trapped in local minima corresponding to stable configurations. The simulated annealing method that explores the PES by iteratively cycling between heating and cooling, has been widely applied in research related to crystal structure prediction, including the pressure-induced diamond to simple hexagonal phase transition, NbF_4, and *N*,*N*′-methylenebisacrylamide crystal structure predictions [50,51].

Compared with simulated annealing, which aims to find the system's lowest energy configuration, metadynamics (Fig. 2d) enhances sampling by introducing a memory mechanism, particularly in the study of chemical reactions and biological systems with energy barriers, making it of significant importance. Specifically, metadynamics introduces a history-dependent bias potential to drive the system out of local minima, thereby preventing the resampling of previously visited states and enabling the exploration of new regions on the PES. This method is particularly suitable for studying rare events and transitions that traditional static sampling might overlook, making it widely applied in research areas like water molecules, nicotine molecule, the liquid-liquid phase transition critical point of liquid phosphorus and the Claisen rearrangement of allyl vinyl ether to 4-pentenal [32,33,52]. In addition, other enhanced sampling methods derived from the same principles as metadynamics include coarse-grained MD [53], in the methods developed by Engkvist and Karlström [54] and directed dynamics methods such as adaptive biasing force (ABF) [55,56], and hyperdynamics [57].

In summary, enhanced sampling methods improve material sampling by introducing history-dependent bias potentials to help the system escape local minima. However, there is a "shortsightedness" issue with MD sampling used for data generation. At low temperatures, MD simulations have a significantly reduced probability of sampling chemical reactions due to the exponential increase in reaction barriers with decreasing temperature. Conversely, at high temperatures, simulated trajectories tend to favor structures with higher configurational entropy, resulting in insufficient sampling of stable structures. Consequently, the generated PES data are often highly redundant and limited to local regions around the input structures. In high-temperature conditions, enhanced sampling methods may also lead to structures being confined to local regions and exhibiting redundancy. Contributing factors to this sampling limitation encompasses the selection of unsuitable collective variables, bias potentials, or temperature ranges, as well as the vast structural space in multi-element systems. This inevitably leads to shortcomings in ML-IPs derived from such data when predicting unknown materials and reactions. Liu and colleagues [22] combined the main features of the metadynamics method and the Metropolis Monte Carlo (MC) method to propose a global optimization trajectory generation method called stochastic surface walking (SSW) for creating PES datasets [58,59]. Research results demonstrate the effectiveness of this method in material structure searches and predicting chemical reaction mechanisms [22,60–62], indicating that by effectively integrating global static structural material methods with enhanced sampling methods, it is possible to address deficiencies in stable structure and transition-state structural materials as well as redundancy issues.

In comparison to the traditional static structure sampling and enhanced sampling techniques, unsupervised learning generative models offer a fundamentally different approach to exploring PESs and generating new structures. These models have the capacity to either explicitly or implicitly model the probability distribution of a generated dataset. Subsequently, they can generate new data by sampling from this probability distribution. Prominent generative models in the field of ML include GAN [63], VAE [36], flow models [37], and diffusion models [38]. Presently, unsupervised learning generative models have showcased impressive capabilities across various domains, such as image generation, machine translation, speech synthesis, style transfer, and more. In recent years, unsupervised learning algorithms are integrated into the field of material design, particularly with the establishment of extensive databases containing information about material structures and properties. The fraction-based diffusion model stands as a leading example among deep generative models, skillfully generating new data samples by systematically removing noise from observed data. The diffusion model, introduced by Yang *et al.* [64], employs a graph-based representation to characterize crystals. Their approach incorporates a dual-based sampling method, expediting the diffusion process and enabling the generation of materials that exhibit both increased innovation and stability compared with earlier generative models. These unsupervised learning approaches have made substantial advancements in representation learning, PES exploration, and the generation of novel structures. Consequently, these emerging technologies are emerging as promising avenues to expedite the exploration of PESs in the context of structure prediction.

The fundamental architectures of generative models include GAN, comprising a generator, a discriminator, and VAE composed of an encoder and a decoder. Many innovative models also adhere to the GAN or VAE frameworks. Nouira andcolleagues [65] introduced CrystalGAN, which employs a 2D array representation of structures created by combining crystal lattice matrices and atomic fractional coordinates. They utilized GAN to sample ternary stable hydrides, starting from two stable binary hydrides. In a similar vein, Kim and colleagues [66] divided individual unit cells into voxel grids along directions parallel to the lattice *abc* axes, segmenting structures of varying sizes in the dataset into an equal number of voxels. They harnessed wGAN to learn the dataset's distribution and, with the assistance of an additional regression network, predicted lattice constants. This approach ultimately led to the generation of novel crystal images and was particularly successful in the design of various SiO_2 porous adsorption materials. Xie *et al.* [67,68] employed a multi-graph representation of crystal structures and developed a crystal diffusion variational autoencoder (CDVAE) network based on graph models and VAE principles. CDVAE learned the dataset's distribution and produced stable crystal structures through a diffusion model decoder. Its efficacy in generating structures was evidenced across various systems, encompassing perovskite structures, elemental carbon, and the structures in the materials project database [69]. Although flow models are relatively recent compared with GAN and VAE, they have not found extensive use in the realm of crystal structure generation due to their demanding requirements for reversibility in model design. Currently, diffusion models stand as the cutting-edge approach in the field of deep generative models [64], demonstrating ongoing rapid advancements.

In summary, when compared with static structure sampling and enhanced sampling methods, unsupervised sampling methods offer several distinct advantages in rapidly generating a substantial number of samples, particularly for systems characterized by intricate potential energy landscapes, thereby enhancing computational efficiency and resource utilization. Moreover, unsupervised learning models can autonomously acquire representations of the PES, eliminating the need for manual selection of sampling pathways or enhanced sampling parameters. However, it is crucial to acknowledge that unsupervised sampling methods come with their own set of limitations. Their performance is significantly contingent on the quality and diversity of input data. Furthermore, the structure generation process within existing models often exhibits a degree of stochasticity, posing challenges in precise control over factors like atomic numbers, ratios, and other structural intricacies. Consequently, the development of conditional generative models holds substantial promise for achieving controlled generation of crystal structures. By amalgamating unsupervised sampling methods with static structure sampling and enhanced sampling techniques, the efficient sampling of the global potential energy landscape, especially in contexts involving reactions, transition states, and more, becomes a tangible prospect. This endeavor has the potential to furnish high-quality PES datasets for ML-based atomic potentials, ushering in new horizons for research in materials science and chemistry.

ENCODING STRUCTURE—D

In contemporary material science research, gaining insights into and predicting the physical and chemical properties of materials stands as a paramount objective. To accomplish this goal, precise descriptions of material structures are indispensable [2,70]. Traditional methods of structural description often rely on manual selection of features or fixed parameter settings, a practice that can significantly curtail their adaptability and precision. Nevertheless, with the advancements in computer technology and algorithmic development, methods for structural description have rapidly evolved, especially those utilizing numerical encoding.

The objective of structural description is to convert the structural information of a material or molecule into numerical or vector formats, enabling its processing by computer programs. This conversion typically involves turning continuous spatial information or distributions into discrete numerical data, striving to retain crucial structural details. Properly encoding these structures not only provides accurate inputs for subsequent calculations or simulations, but also ensures the reliability and accuracy of the research outcomes. To better capture and numerically represent these structural details, researchers have developed a variety of descriptors. As shown in Fig. 3, these descriptors can be categorized into two main categories: global descriptors and local descriptors [70,71]. While global descriptors aim to offer a comprehensive overview of the entire structure, local descriptors focus on specific areas within a material or molecule, such as the environment around a particular atom or group of atoms. This section will present in detail the various approaches from these two categories of descriptors, along with their mathematical expressions, physical connotations, and scope of application. Therefore, this comprehensive analysis is favorable to gain a deep insight of how to select and apply the most suitable structural description to meet our specific models.

Global descriptors

Global descriptors offer a unified representation of the entire molecular or material structure. Unlike local descriptors, which focus on individual atoms or specific regions, global descriptors aim to capture characteristics of the structure as a whole. Therefore, the selection and design of global descriptors are critical. They not only need to convey information concisely, but also ensure that the described structural information is representative and distinguishable.

The Coulomb matrix is a commonly used global descriptor that relies on the Coulomb interactions between atoms to describe molecular structure [72]. This descriptor constructs a symmetric matrix by considering the nuclear charges of atoms and their distances from one another. The Coulomb matrix provides a simple and intuitive way to describe the electronic environment of a molecule and has been widely employed in various chemical applications [73,74]. The construction of the Coulomb matrix is relatively straightforward. For a molecule containing N atoms, the Coulomb matrix C is an $N \times N$ symmetric matrix, with each element (C_{ij}) defined as follows:

$$C_{ij} = \begin{cases} \dfrac{Z_i Z_j}{|R_i - R_j|}, i \neq j \\ 0.5 Z_i^{2.4}, i = j \end{cases}, \quad (2)$$

where Z_i and Z_j are the nuclear charges of the ith and jth atoms, respectively, and $|R_i - R_j|$ is the Euclidean distance between them. The diagonal Coulomb matrix element C_{ii} is an empirical

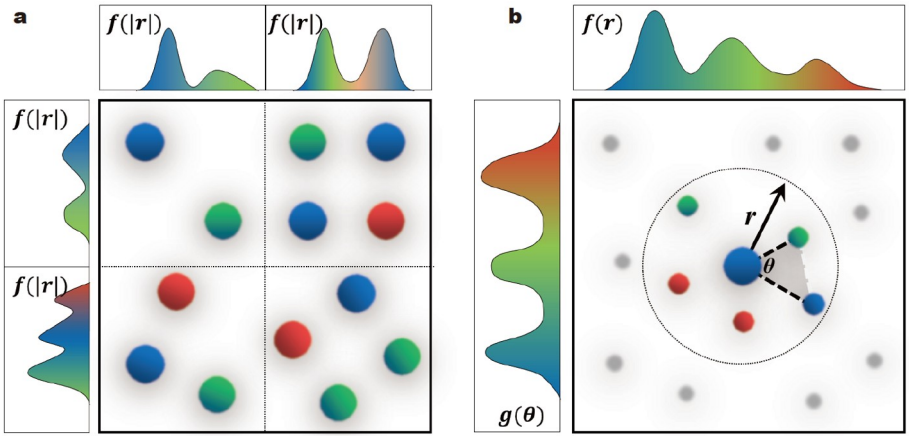

Figure 3 Structural descriptors: (a) global descriptors accounting for overall atomic environments; (b) local descriptors centered around selected atoms.

description of the self-interaction when $i = j$.

The primary physical insight behind the Coulomb matrix lies in its ability to describe the charge distribution within materials and the strength of interactions between atoms, which is crucial for understanding the atomic charge properties and performance of materials, particularly in molecular and lithium-ion battery materials. Elements on the diagonal provide information about the individual atomic charges, while off-diagonal elements describe the Coulomb interactions between two atoms. This is particularly relevant in processes such as Li-ion migration, which depends on the surrounding charge environment. The key advantages of the Coulomb matrix are its simplicity and universality. The Coulomb matrix can be directly calculated for any material without additional parameters or assumptions. In addition, the Coulomb matrix provides rich features for ML models, enabling effective learning and prediction of molecular properties. However, the Coulomb matrix approach also has some limitations. Since it is based on nuclear charges and atomic distances, this approach may be not comprehensive to capture all information for complex chemical environments, especially in intricate battery materials such as cathode materials and solid electrolytes, which involve ion channels and defects. Moreover, the size of the Coulomb matrix is proportional to the number of atoms, posing challenges for large-size materials, especially those with periodic structures [70].

A major issue with the Coulomb matrix approach is that it generates different representations for an identical structure with different atomic labellings. To address this issue, researchers have proposed the sorted Coulomb matrix [75]. In this approach, the ranks of the matrix are sorted according to the strength of atomic interactions, providing invariance to different atomic labels. The Coulomb matrix does not inherently possess rotational symmetry. However, this issue can be indirectly mitigated by combining the sorted Coulomb matrix with ML models, such as rotationally invariant neural networks. In addition, the direct use of Coulomb matrix for large structures may lead to computational and storage challenges. The randomly sampled Coulomb matrix approach reduces the dimensionality of the descriptors by randomly selecting atomic pairs while preserving the essential structural information [75]. Furthermore, in order to provide a richer description of materials, atomic properties such as atomic radius and electronegativity can also be incorporated as weights to form a weighted Coulomb matrix approach with multidimensional features [76,77]. These various variants not only retain the simple physical connotations, but also enhance its descriptive capability and application in diverse contexts.

The many-body tensor representation (MBTR) is a more complex global descriptor, which captures the effects of many-body atomic interactions, and it is widely employed to describe complex structures and properties [78,79]. In contrast to traditional two-body descriptors (e.g., Coulomb matrix approach), the MBTR considers three-body, four-body, and even higher-order atomic interactions, offering a more comprehensive description for complex systems [80]. The core idea of MBTR is to map many-body atomic interactions into a tensor form that encapsulates information about all atoms and captures various interaction patterns among them. This tensor can be regarded as a multidimensional array, where each dimension corresponds to a different atom or combinations of atoms, and the elements of the tensor represent the strength of interactions or other relevant information between these atoms. In this way, the MBTR transforms complex many-body interaction problems into high-dimensional tensor operations, providing a more flexible and comprehensive modeling tool.

In contrast to two-body descriptions like the Coulomb matrix, which typically utilizes a 2D matrix C, we can extend this concept and introduce an N-dimensional tensor T to describe N-body interactions in the MBTR. For three-body interactions, we can define a 3D tensor T_{ijk} mathematically:

$$T_{ijk} = V_{ijk}\left(r_{ij}, r_{jk}, r_{ki}, \theta_{ijk}\right), \tag{3}$$

where r_{ij} and θ_{ijk} are the respective distances between atomic pairs and the angle they formed. The key to the MBTR lies in its ability to capture complex phenomena such as angle dependence, electronic rearrangement and energy decreases due to many-body interactions. These ofen play crucial roles in systems such as organic molecules, polymers, solid-state physics, and biomacromolecules [2]. For example, in the study of protein folding and charge transfer systems, this approach can provide rich and accurate information [81,82]. Accurately describing interactions between neighboring sets of three atoms may significantly impact the overall stability and reactivity of certain organic molecules [83].

However, while the MBTR excels in capturing many-body interactions with high precision and rich physical insights, it also

has inevitable limitations. The first is computational complexity, particularly in large systems and high-order interactions. This complexity typically manifests as exponential growth, where an increase in system size or interaction order leads to an exponential increase in required computational resources. In addition, the MBTR approach faces the completeness concern of the basis set [84]. In quantum mechanics calculations, wave functions is required to be expanded in a set of basis functions, and an incomplete basis may lead to an inaccurate representation of many-body effects, especially in electron-correlated systems. This requirement places higher demands on the size and complexity of the basis set, further increasing the computational burden. Parameter selection and optimization are also issues of concern for MBTR. The MBTR approach involves numerous parameters to describe many-body potential functions, and the selection and optimization of these parameters can be complex and time-consuming, and may lead to overfitting or other numerical issues. Moreover, high-dimensional tensors often contain a large number of zero elements, particularly in sparse systems, which not only raise storage concerns, but may also result in waste of computational resources, as many calculations may be unnecessary.

Despite these limitations, the MBTR approach still demonstrates significant advantages, especially in complex systems where traditional two-body description methods face difficulties. Its high precision and wide applicability make it a promising tool in quantum chemistry, condensed matter physics, materials science, and the structural analysis of biomolecules. Therefore, overcoming its limitations through algorithm optimization and hardware acceleration may help broaden its applications in scientific research.

Local descriptors

Local descriptors are powerful and flexible tools in the study of structural encoding. In contrast to global descriptors, local descriptors focus on individual atoms or a small group of atoms and their roles in local environments. This approach offers significant advantages when dealing with large-scale or heterogeneous systems.

Atom-centered symmetry functions (ACSF) provide an effective mathematical framework for local descriptions of materials and molecular structures [85]. In this framework, each atom is described by its surroundings, which is crucial in the presence of many-body interactions and diverse atomic arrangements. ACSF is usually represented by a set of differentiable, rotationally invariant functions G_i, which are defined based on distances r_{ij} to neighboring atom j, angles θ_{ijk}, and higher-order geometric relationships. Mathematically, a commonly used radial part function can be expressed as

$$G_i^{\mathrm{R}} = \sum_j \exp\left[-\eta(r_{ij}-r_0)^2 f_{\mathrm{c}}(r_{ij})\right], \tag{4}$$

where r_{ij} is the distance between atoms i and j, η and r_0 are parameters, and $f_{\mathrm{c}}(r_{ij})$ is a cutoff function that ensures the function value is zero when r_{ij} exceeds a specified range [0, r_0]. The angular part is typically expressed as

$$G_i^{\mathrm{A}} = 2^{1-\zeta}\sum_{j,k}\left(1+\lambda\cos\theta_{ijk}\right)^{\zeta} \mathrm{e}^{-\eta\left(r_{ij}^2+r_{jk}^2+r_{ki}^2\right)} f_{\mathrm{c}}(r_{ij}) f_{\mathrm{c}}(r_{jk}) f_{\mathrm{c}}(r_{ki}), \tag{5}$$

where θ_{ijk} is the angle between atoms i, j, and k, and λ and ζ are parameters. This design enables the set of descriptors to have not only clear physical connotations, but also favorable mathematical properties, such as smoothness and differentiability, which is beneficial for subsequent ML models. It is noteworthy that the parameter η acts as a sensitivity factor, determining the range over which the descriptor is responsive to variations in interatomic distances. A larger value of η implies that the symmetry function is more sensitive to atomic pairs at closer distances, with contributions from farther distances diminishing rapidly. In contrast, a smaller η value makes the symmetry function more receptive to atomic pairs at extended distances. η is pre-determined, usually based on empirical or trial-and-error methods, and is not always directly related to the actual chemical properties of the system, which may lead to an inadequate capture of key features of the atomic environment. A fixed η value might not be appropriate for a variety of chemical environments, such as systems with different types of chemical bonds or complex coordination environments. Zhang *et al.* [86] proposed a physically inspired variation of ACSF by substituting the η parameter with atomic radii. This refinement links η more directly to specific types of chemical bonds in the target system, allowing the descriptor to reflect the actual physical and chemical environments more directly, which enhances the adaptability and predictability of ACSF for complex systems.

The ACSF inherently maintains permutational symmetry as it is constructed based on the local atomic environment. Rotational invariance is ensured by designing a series of functions based on distances and angles that remain unchanged under rotation. However, translational symmetry is typically not a requirement for local descriptors as they describe relative positions of atoms, not absolute ones. Such descriptors can reveal chemical interactions between atoms to a certain extent, such as covalent bonds and van der Waals interactions [87]. Since these descriptors are local, their calculations are highly scalable in large-scale or heterogeneous systems. As a result, this enables the ACSF to compete with more complex and computationally intensive first-principles methods in calculating energies, forces, and other atomic-level properties. The ACSF is therefore widely employed in the description of both periodic (e.g., crystals) and non-periodic (e.g., molecules or clusters) systems [88]. These descriptors have demonstrated high accuracy and reliability in a range of ML tasks, including MD simulations, and the screening and design of new materials [86,89].

Despite their many advantages, the ACSF approach also suffers from several limitations. Due to their highly nonlinear and high-dimensional nature, these descriptors require relatively large training sample sizes for effective fitting. In addition, a great deal of a priori knowledge and experimentation is often required to select the most suitable symmetry function form and parameters for a particular task.

Overall, as a local descriptor, the ACSF has emerged as a powerful and widely used tool in materials science and chemistry. It combines rich physical insights with good mathematical properties, providing a feasible approach to understanding and predicting many-body interactions and dynamic behavior in complex systems. However, in practical applications, careful selection and adjustment of their forms and parameters are still required to maximize their performance and applicability.

Smooth overlap of atomic position (SOAP) is a state-of-the-art technique for describing the local atomic environments of molecules and solid systems [90]. In contrast to ACSF, SOAP is primarily designed to provide a smooth and comprehensive

description of the local atomic environment and explicitly represent the relative arrangements of atoms. Due to its complexity and richness, SOAP is considered a powerful descriptor, especially in chemical environments with long-range interactions.

The central idea of SOAP is to smooth the distribution of local atomic environments and represent them in a rotation-invariant manner. For a specific atom, its SOAP descriptor can be written as

$$P_{ij}(r) = \sum_k f_c(r_{ik})\psi_i(r_{ik})\psi_j(r_{jk})Y_{lm}(\hat{r}ik)Y^*_{lm}(\hat{r}jk), \quad (6)$$

where r_{ik} represents the distance between atoms i and k, $f_c(r_{ik})$ is a cutoff function, ψ_i and ψ_j are radial basis functions, and Y_{lm} is a spherical harmonic function. This formula describes the two-body relationship between atoms i and j and the three-body relationship formed with other atoms k. The smoothing nature of the descriptor ensures that slight movements or perturbations of atoms do not lead to large variations in the descriptor. This is valuable in chemical and physical environments, where atoms often vibrate due to fluctuations of temperature, pressure, or other external factors. However, the cutoff function $f_c(r_{ik})$ in the traditional SOAP framework may lack smoothness at the cutoff radius r_c, leading to discontinuous changes in the contribution from atomic pairs. This can introduce artificial discontinuities on the PES, resulting in discontinuities in the forces at the boundaries. Tirelli *et al.* [91] proposed substituting the traditional cutoff function f_c with a piecewise-defined polynomial that smoothly vanishes at r_c, ensuring continuity at the cutoff boundaries. This modification allows for improved computational efficiency while maintaining physical accuracy and numerical stability. Furthermore, SOAP descriptors provide a rotation-invariant structural description by constructing its power spectrum, which can be achieved by calculating the autocorrelation function of SOAP descriptors [92].

The SOAP approach has been widely employed in various tasks of computational chemistry and materials science. For instance, SOAP can be used to describe atomic environments associated with specific properties, such as mechanical strength, in the screening and design of new materials [93,94]. Based on these descriptors, researchers can use ML models to predict material properties, allowing for rapid screening of potential material candidates without performing time-consuming first-principles calculations. In addition, SOAP has also been used for structural and dynamical studies of complex systems with high-pressure phase transition and grain boundary [95,96].

Similar to ACSF, the computational complexity of SOAP is also relatively high, especially when considering large-scale systems or high-order interaction calculations. This increases computational costs and may render it ineffective in certain applications. SOAP descriptor maintains rotational invariance through its power spectrum, constructed by calculating the autocorrelation function of the SOAP descriptors. SOAP provides a continuous description of the local atomic environment by smoothing out the atom distribution in 3D space, maintaining translational invariance for molecules or crystals. While SOAP descriptors do not directly consider permutational symmetry in their design, this symmetry can be implicitly achieved in feature space through appropriate kernel functions in ML applications. Furthermore, while SOAP provides detailed information about local environments, it may not be suitable for capturing long-range interactions or large-scale structural features.

Moment tensor potential (MTP) is a recent developed force-field method. Compared with traditional force field methods, MTP offers higher precision and flexibility, as this method do not only rely on pairwise distances between atoms but also higher-order, many-body interactions [97]. The emergence of this method addresses the current demand for high-precision calculations, especially in nonlinear materials and chemical systems. The MTP method is characterized by a series of moment tensors describing their atomic environments and many-body interactions, with the order of moment tensor being related to the number of participating atoms. For instance, a second-order tensor can describe a pairwise interaction, while a third-order tensor accounts for three-body interactions, and so forth. The energy expression in MTP can be written as follows:

$$E = \sum_i E_i = \sum_\alpha c_\alpha \Phi_\alpha(R), \quad (7)$$

where E_i represents the local energy of atom i. This local energy can be further expressed as a sum of products between a series of basis functions $\Phi_\alpha(R)$ and their corresponding coefficients c_α, where α represents the type and order of the moment tensor. The $\Phi_\alpha(R)$ here is a basis function associated with the atomic positions R and is defined by the contraction of moment polynomials $M_{\mu,\nu}$:

$$\Phi_\alpha(R) \overset{\text{def}}{=} \prod_{i=1}^{k} M_{\alpha_{ii},\alpha_i'}(R), \quad (8)$$

where α_i' represents the sum of off-diagonal elements in the matrix:

$$\alpha_i' = \sum_{j=1,j\neq i}^{k} \alpha_{ij}, \quad (9)$$

while $M_{\mu,\nu}$ is a ν-order tensor which contains a radial part and an angular part:

$$M_{\mu,\nu}(R_i) = \sum_j^{N_i} f_{\mu,\nu}(r_{ij}) r_{ij} \otimes \cdots \otimes r_{ij}, \quad (10)$$

where $\otimes$ is a tensor product, $r_{ij} \otimes \cdots \otimes r_{ij}$ is a tensor with rank ν, N_i is the number of atoms within the cutoff radius of the ith atom, and $f_{\mu,\nu}$ is a self-defined radial function similar to that in ACSF. Notably, the basis functions $\Phi_\alpha(R)$ utilized in MTP satisfy the intrinsic symmetries of translations, rotations, and permutations, which are essential for any IP aimed at improving physical accuracy and computational efficiency. The translational symmetry is achieved as the basis functions are defined based on relative positions and do not change when the entire system undergoes spatial shifts. Rotational invariance is incorporated by constructing the basis functions from scalar products of vectors and tensors that are invariant under rotation. Consequently, these functions produce the same set of values regardless of how the system is oriented in space. Permutational symmetry is sustained by formulating the basis functions as a summation over atoms or atom pairs, ensuring their independence from the order of atoms. This formulation guarantees that swapping any pair of atoms does not affect the potential's value. This comprehensive symmetry integration within the MTP

framework allows for the accurate prediction of physical properties under various configurations and orientations of the atomic system. In the MTP expression, there are no fitting parameters in $\Phi_\alpha(R)$, and the total energy E is linearly related to these coefficients $\{c_\alpha\}$ to be fitted. To determine these coefficients, a substantial amount of first-principles computational data is typically employed for training. This dataset encompasses precise energies and atomic forces for diverse geometric configurations, guiding the selection of coefficients to ensure that MTP predictions closely align with first-principles calculations. Selecting the appropriate tensor order and radial functions is paramount. If the order is too simplistic, it may struggle to accurately capture complex many-body effects. Conversely, overly complex orders can lead to overfitting, diminishing the accuracy of MTP predictions for unseen configurations. Consequently, a series of tests, often involving cross-validation and other model validation techniques, are conducted to identify the optimal tensor order and radial functions. From this, one of the limitations of MTP lies in its demand for a substantial number of parameters to accurately represent many-body effects. This requires an extensive training dataset for parameter fitting, potentially resulting in a complex and time-consuming fitting process. Furthermore, while MTP can effectively account for many-body effects, it may still exhibit reduced accuracy if the training data lack diversity or the structural sampling fails to include certain specific interactions.

Nevertheless, MTP has been employed to various systems, ranging from simple elements to binary systems [12,98]. In many cases, MTP predictions closely match first-principles calculations or experimental results while offering significantly faster computational speed [97,99]. For instance, MTP can provide in-depth insights into diffusion paths and coefficients without expensive quantum mechanical simulations in the study of Cu diffusion processes in high-temperature phase $Cu_{2-x}Se$ of classical thermoelectric materials [100].

Overall, the MTP method is a promising way to describe complex materials and molecular systems. Its ability to consider many-body effects renders it more accurate than traditional force field methods in many applications. Nevertheless, sufficient training data and elaborate parameter tuning are also required, as with most state-of-the-art methods, in order to achieve optimal performance of cross-scale calculations.

Comprehensive comparison of descriptors

In the realm of cross-scale calculations, achieving accurate and efficient descriptions of complex atomic structures is of paramount importance. Descriptors play a pivotal role in this process, each offering its unique advantages, limitations, and applications. Global and local descriptors serve as essential tools for capturing structural features from both a holistic and localized perspective, respectively. To select the most appropriate descriptor, it is required to deeply understand the characteristics of various descriptors and their performance in different applications.

Global descriptors, such as Coulomb matrices and MBTR, typically focus on the overall properties of a molecule or crystal, suggesting that they attempt to capture atomic interactions from a macroscopic viewpoint. For example, the Coulomb matrix approach relies on Coulomb energies to describe pairwise interactions throughout the molecular structure. This approach performs well in describing the total energy of systems and certain ground-state properties (e.g., melting point and statistical moment of spectra) [72–74], but may not be sensitive enough when local environments evolve. In contrast, MBTR takes more interaction details into account, such as many-body interactions between three atoms, providing a finer framework for capturing complex structural features.

Local descriptors, such as ACSF, SOAP, and MTP, primarily focus on the local environment of individual atoms or a small group of atoms. Their purpose is to describe and understand the characteristics of local atomic environments, such as coordination, chemical bonding, and local distortion. These descriptors therefore have advantages in describing the local properties of materials, such as atomic diffusion, point defect formation, and localized electrocatalytic activity. In particular, SOAP captures the local environment of atoms by considering the smooth overlap between atomic positions, rendering this approach especially effective in describing materials with similar structures but different properties. Whereas MTP provides an expandable framework for describing more complex atomic interactions, such as magnetism and spin-polarized interactions [101].

Thus, selecting the appropriate descriptor fundamentally depends on the research objectives and the physical effects to be considered. If the focus is primarily on macroscopic properties of materials, such as elastic modulus or atomization energies, global descriptors may be more appropriate. However, if the emphasis is on structure-property correlations at the microscopic level, such as enhancing ionic conductivity by introducing dopants or point defects, local descriptors should be considered. Therefore, there exists no definitive "optimal" descriptor, but rather descriptors that are most suitable for specific applications. Understanding the characteristics and limitations of various descriptors is key to selecting the right tool and successfully applying it to specific study. In future cross-scale calculations, novel descriptors or methods that combine multiple descriptors may emerge to provide more accurate and comprehensive structural encoding.

It is worth noting that descriptor transferability remains a central concern in cross-scale calculations, where performance on one dataset may not readily transfer to other chemical or physical environments. It is critical that training datasets should encompass a wide array of chemical elements, structural types, and external conditions, ensuring that descriptors capture universally applicable features rather than those specific to particular systems. Moreover, incorporating regularization techniques further reduces model complexity, which facilitates avoiding overfitting and enhancing the generalizability of model without compromising prediction accuracy. Regarding the descriptor training, it can be trained both independently or in conjunction with fitting models. While trained separately, descriptors are typically pre-trained on datasets independent of the final prediction task, followed by integration into neural networks. This approach offers the flexibility of independent training modules, but may require additional tuning to fit a specific system. In contrast, joint training optimizes both descriptor parameters and network weights. This method directly associates the model with energy, forces, or other properties throughout the training process, which may improve prediction accuracy of the cross-scale model. However, this method may also result in lower transferability of the descriptors, challenging their application in other systems.

FITTING METHODS FOR ML-IP—F

The selection of descriptors and the choice of regression models play a crucial role in shaping the functional form of ML-IPs. While theoretically, these two aspects can be considered independently, in practice, they are often closely linked, with the development of descriptors closely tied to specific regression models. However, it is important to note that this interdependence is not due to incompatibility between different methods but rather reflects the personal preferences of developers [102]. When it comes to fitting ML-IPs, the current fitting methods include linear regression [97,103], kernel methods [104,105], deep neural networks (DNN), and other methods.

Kernel methods

Kernel methods (Fig. 4a) involve mapping input data to a high-dimensional feature space using a kernel function, followed by model training and fitting within this feature space using algorithms such as linear regression, ridge regression, and support vector machines (SVM). Kernel methods typically require the kernel function to be square integrable and positive semidefinite, with their smoothness guaranteed by the smoothness of the kernel function itself. Currently, kernel methods utilize various types of kernel functions, including linear, polynomial, Gaussian, and Laplacian kernels. In the context of ML-IPs, Gaussian approximation potential (GAP) relies on Gaussian processes and utilizes kernels to approximate local atomic energies [104], defining the similarity between different atomic descriptors. This kernel-based ML-IP has been applied to predict the energy and forces of various materials, including metals, semiconductors, and amorphous solids [94,106–108]. Another popular kernel regression-based ML-IP is the adaptive generalized nearest-neighbor information (AGNI) potential, primarily used for simulating interactions within metals [105].

Kernel methods are known for their effectiveness in dealing with non-linear PESs in atomic potential fitting, making them suitable for modeling interactions among multiple atoms. In kernel methods, the cost of parameter training scales with the cube of the number of training points. In applications of the ML-IP, it is common for the number of training points to fall within the range of 10^6 to 10^7, demanding significant computational resources for parameter estimation. Additionally, kernel methods require the storage and factorization of a dense Gram matrix, which is of size with the square of the number of training points. This matrix impedes the hyperparameter tuning process, as it necessitates multiple-model runs to determine optimal hyperparameters. This training process becomes a bottleneck when dealing with multiple elements, such as in alloys, or when using active learning methods for training [110,111]. Another major drawback of kernel methods is that the cost of making predictions at new data points is proportional to the product of the number of input features and the number of training points, and the number of input features is typically on the order of 10^2. This results in slower execution times, especially in MD. To address the high computational costs in kernel methods, Dhaliwal and colleagues [112] proposed approximating local atomic energies as a linear combination of kernel-related random features, reducing the training time by 96% compared with the original kernel methods.

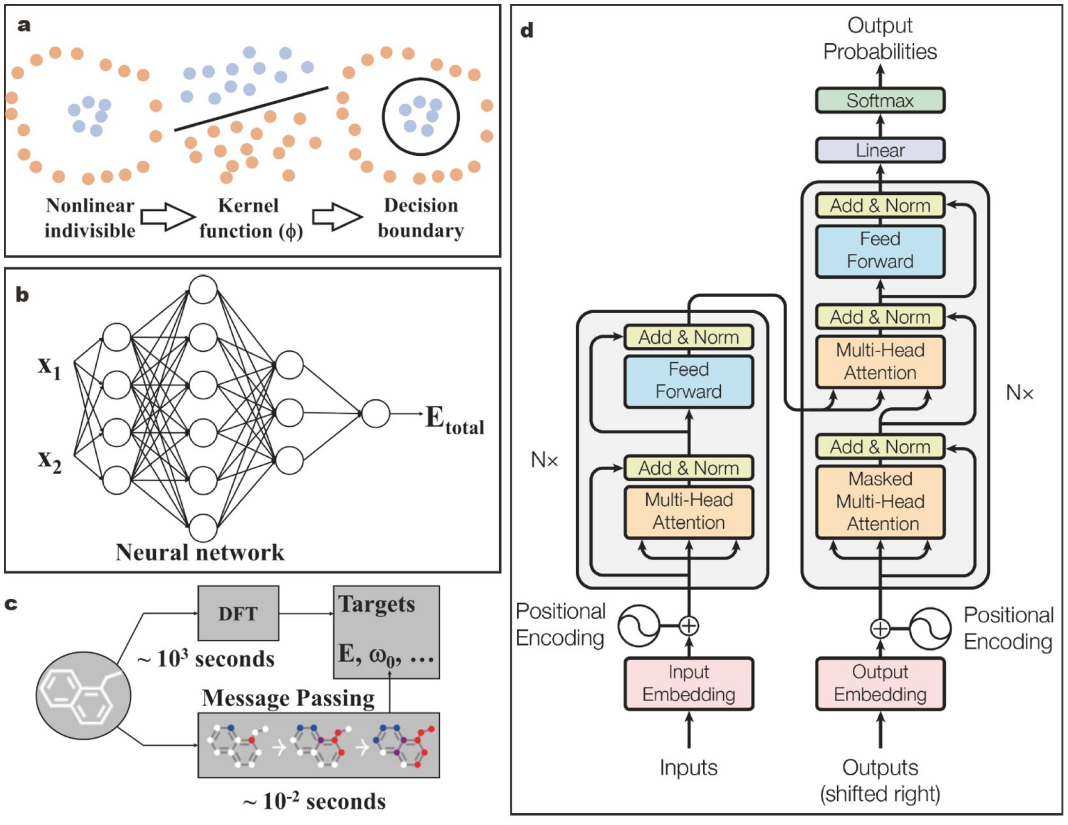

Figure 4 Fitting methods: (a) kernel function, (b) DNNs, (c) message passing neural networks, and (d) transformer. Reprinted with permission from Ref. [109], Copyright 2017, the Authors.

In the context of using kernel methods as fitting functions for ML-IPs, they have found extensive application due to their robust ability to perform nonlinear fitting, particularly excelling with smaller datasets. Nevertheless, they entail substantial computational costs. Kernel methods can also exhibit a susceptibility to overfitting in specific instances, especially when intricate kernel functions are employed, potentially leading to subpar generalization performance on novel data. Additionally, kernel methods lack the capability to autonomously acquire feature representations, necessitating manual intervention, thereby posing challenges in end-to-end learning scenarios.

DNN

In recent years, deep learning has outperformed traditional methods in various fields, including computer vision [113], natural language processing [114], and protein folding [115]. The cornerstone of deep learning lies in the impressive ability of DNNs to model high-dimensional nonlinear functions [116]. This technology has also been extensively researched in the realm of ML-IPs.

In 2018, Weinan and colleagues [117] introduced a groundbreaking MD approach called deep potential molecular dynamics (DeepMD). This innovative method is built upon the foundation of DNN. Within this framework (Fig. 4b), the process of fitting atomic potentials involves two crucial components: a feature network designed to describe chemical environments and a residual network dedicated to training the energies and forces of individual atoms. Through the application of this method, they accomplished an impressive dual achievement—dramatically enhanced simulation speed while concurrently preserving simulation results with an accuracy level comparable to the precise AIMD method. This method has been employed for a wide range of studies, including the elucidation of the thermal decomposition mechanism of the novel energetic material ICM-102 and dynamic nuclear magnetic resonance chemical shift calculations for paramagnetic battery materials [118,119]. Nitol and his fellow researchers [120] introduced an atomic potential approach grounded in artificial neural networks (ANN), with a specific emphasis on the element zinc. Remarkably, they effectively applied this method to replicate training data from DFT, attaining chemical precision, and precisely forecasting the *c*/*a* ratio of the hexagonal close-packed (HCP) ground state. These accomplishments highlight the vast potential of DNN and comparable deep learning techniques in the fields of materials science and molecular simulation. They enable researchers to furnish precise descriptions of atomic interactions and material properties.

However, with the proliferation of experiments like quantum chemistry calculations and MD simulations, a vast amount of data has been generated. Most classical ML techniques struggle to effectively harness this current data deluge. Yet, the symmetry inherent in atomic systems suggests that neural networks applicable to network graphs can also be applied to molecular models. Therefore, finding a more powerful model for fitting ML-IPs can be equivalent to discovering a model suitable for network graphs. Expanding on this concept, as shown in Fig. 4c, Gilmer *et al.* [121] introduced a framework for supervised learning based on graphs, incorporating both message passing and message reading neural networks (PMNN). Through this approach, he achieved the successful development of high-precision potential functions for 13 properties within the QM-9 dataset. Veronique and her colleagues [122] constructed ML-IPs for MOF materials based on equivariant PMNN. The results demonstrate that even for systems with multiple phases, accurate ML-IPs can be constructed with just around 1000 quantum mechanical evaluations. While PMNN demonstrates strong performance on small datasets with short-time simulations, the field nevertheless warrants further exploration. For example, there is currently a lack of clarity on how to assemble the most diverse training dataset for frameworks to ensure optimal material transferability. Additionally, it is crucial to investigate whether equivariant message-passing neural networks, such as neural equivariant interatomic potentials (NequIP), can maintain their precision across diverse materials design spaces. In such scenarios, integrating the message-passing architecture with more recent models may be necessary to offer a more accurate representation of long-range interactions.

In recent years, transformer (Fig. 4d), as a powerful class of deep learning methods, has been widely used in natural language processing and the field of proteins. AlphaFold2, developed by DeepMind [123], has achieved accuracy levels in predicting protein 3D structures from amino acid sequences that are comparable to experimentally determined structures. Building on this, Janson *et al.* [124] employed GAN to learn the 3D conformation distribution within conformational datasets. They introduced the idpGAN model, capable of generating 3D Cartesian coordinates for conformations with varying sequences and lengths. This model exhibits rapid sampling capabilities, enabling the generation of thousands of independent conformations in a short time, offering an efficient means of generating conformational ensembles. As the availability of large-scale DFT data is limited in materials science compared with the protein field, making transformer frameworks require substantial data challenging to apply, Liao and Smidt [125] harnessed the advantages of the transformer architecture and combined it with SE(3)/E(3)-equivariant graph neural networks (Equiformer) based on irreplaceable representations. They achieved results on QM-9, MD-17, and OC20 datasets that are comparable to previous models. In summary, transformer-based ML-IPs hold great promise for rapid development and application.

In conclusion, ML-IPs are fundamentally transforming the research paradigm in the field of molecular simulations. The wealth of data generated from first-principle calculations have significantly expanded the scope of these models. However, when faced with a new and complex system, the need to generate substantial new data for model training remains a challenge. Drawing inspiration from developments in other artificial intelligence domains, the question of whether we can harness the vast amount of existing data and reuse pre-trained models is a pressing issue for reducing the cost of model development. Zhang *et al.* [126] developed a significant pre-trained model according to a pivotal gated attention mechanism, known as DPA-1. This model was constructed using a comprehensive dataset that includes 56 distinct elements, with 2 million data points (comprising energy and force) randomly sampled from OC20. Their findings illustrate that the application of transfer learning with this pre-trained model substantially diminishes the reliance on new data across various datasets. Preferred networks, in collaboration with organizations including ENEOS, have developed a universal neural network atomic potential function known as PreFerred Potential (PFP) [127]. This achievement is

based on the TeaNet architecture from the Massachusetts Institute of Technology and the University of Tokyo and extensive training on a large-scale dataset. PFP covers all 72 elements present on the periodic table. In its pre-training phase, PFP utilizes an extensive dataset comprising approximately 150 million molecular conformations and 1 million protein pocket conformations. Zhou *et al.* [128] have developed a universal molecular representation learning framework based on 3D molecular structures, known as Uni-Mol. Built on the SE(3)-equivariant transformer architecture and pre-trained on a large-scale dataset comprising 210 million molecular conformations and 3.2 million protein pocket conformations, Uni-Mol demonstrates proficiency in handling diverse organic small molecules and protein pockets. Uni-Mol stands out in directly learning molecular representations from 3D molecular structures, while DPA-1 generates 3D molecular conformations from 1D or 2D molecular representations. Consequently, Uni-Mol and PFP effectively use 3D molecular information, whereas DPA-1 may necessitate more computational resources and time for the generation and optimization of molecular conformations. Both Uni-Mol and PFP take graph-based representations to characterize crystal structures, whereas DPA-1 utilizes a potential energy function called deep potential to assess the stability and activity of molecules. The strengths of Uni-Mol and PFP lie in their abilities to generate diverse molecular structures and consider the physical and chemical properties as well as pharmacological effects of molecules. On the other hand, DPA-1 excels in accurately simulating the dynamics and thermodynamics of molecules. These expansive pre-trained models alleviate the need for domain-specific DFT data, consequently reducing training costs and expediting the advancement of ML-IPs.

APPLICATIONS OF ML-IPs

In recent years, ML-IPs have found extensive applications in fields such as biology, chemistry, and materials science. These applications can be further explored from two perspectives: across different time scales and spatial scales.

Applications across time scales

When researching lithium-ion transport in lithium superionic conductors, the conventional method typically entails extrapolating room temperature ion diffusion properties directly from high-temperature (>600 K) *ab initio* AIMD using the Arrhenius assumption. However, this approach can introduce certain inaccuracies. As shown in Fig. 5a, Xu *et al.* [129] employing 1-μs ultra-long timescale low-temperature MLMD simulations, have unveiled the non-linear Arrhenius behavior of lithium ions within Li_3ErCl_6. This non-linearity stands as a primary factor contributing to the tendency of traditional AIMD simulations to overestimate its ionic conductivity. The photo-induced processes are fundamental in nature, but the precise simulation of their dynamics is severely constrained by the computational costs of basic quantum chemistry calculations, which hinders their application on long time scales. In light of this, Westermayr *et al.* [130] have developed an approach based on DNNs to learn the relationship between molecular geometry and its high-dimensional electronic properties, as shown in Fig. 5b, enabling accurate photodynamics on the nanosecond time scale. ML force fields have been employed to conduct *in-situ*, cross-scale, and 200-ps-long MD simulations of lithium dendrite morphology in an electrolyte environment (Fig. 5c). This approach has helped determine that surface energy and grain boundary energy are the primary driving forces behind the morphological evolution [131].

Applications across spatial scales

Jia *et al.* [132] employed the deep potential MD approach, enabling them to simulate trajectories exceeding 1 ns each day and simulating over 100 million atoms for more than 1 ns each day. In contrast to previous simulations, which had a maximum of 1 million silicon atoms (velocity = 4×10^{-3} s/step/atom), the fastest simulation speed reached 1.3×10^{-6} s/step/atom (for a system containing 9000 water molecules). The research successfully simulated 679 million water molecules and 127 million copper atoms (Fig. 5d), representing an improvement in speed by several orders of magnitude compared with previous simulations. Wang *et al.* [133] introduced AI_2BMD, as shown in Fig. 5e, a deep learning-based quantum-accurate protein dynamics simulation system. AI_2BMD incorporates novel protein segmentation techniques, an ML force field based on ViS-Net, and a self-developed dynamic simulation system. This system enables precise calculations for various proteins containing over 10,000 atoms and exhibits a wide range of applicability. Lai *et al.* [134] conducted large-scale MD simulations using a neural network potential with quantum mechanical accuracy at the lithium-copper interface (Fig. 5f). They investigated the dynamic behavior of lithium atom deposition on copper surfaces with different Miller indices and the arrangement characteristics of lithium atoms on copper surfaces. It was observed that the performance of Cu(100) and Cu(111) surfaces is significantly superior to that of the Cu(110) surface. These findings offer theoretical guidance for the manufacturing of commercial copper foils and the commercialization of anode-free lithium metal batteries. In addition, Milardovich *et al.* [135] accomplished high-precision simulations of 3024 Si_3N_4 atoms using ML-IPs developed through active learning combined with the GAP method. The computed neutron scattering structure factor of amorphous Si_3N_4 aligns well with experimental results.

In summary, the rapid advancement of ML-IPs has enabled scientists to conduct more precise and efficient MD simulations across both temporal and spatial scales in fields such as materials science, biology, and electrochemistry. This has unveiled numerous previously unobservable phenomena and introduced entirely new research methodologies. These innovative approaches and technologies provide powerful tools for gaining a deeper understanding of material behavior, materials design, and the functionality of biomolecules, and they hold significant promise for shaping the future of scientific research and applications.

CHALLENGES OF ML-IPs

Challenges for structure sampling

Based on my existing level of knowledge, I can summarize the challenges in the structural sampling section as follows: (1) the accuracy of ML-IPs largely depends on the accuracy and diversity of the training dataset. Training errors increase with increasing sampling temperature and with an increase in structural disorder [110,136,137]. Currently, active learning methods that select different data from MD trajectories and adaptively update active learning divergence thresholds have

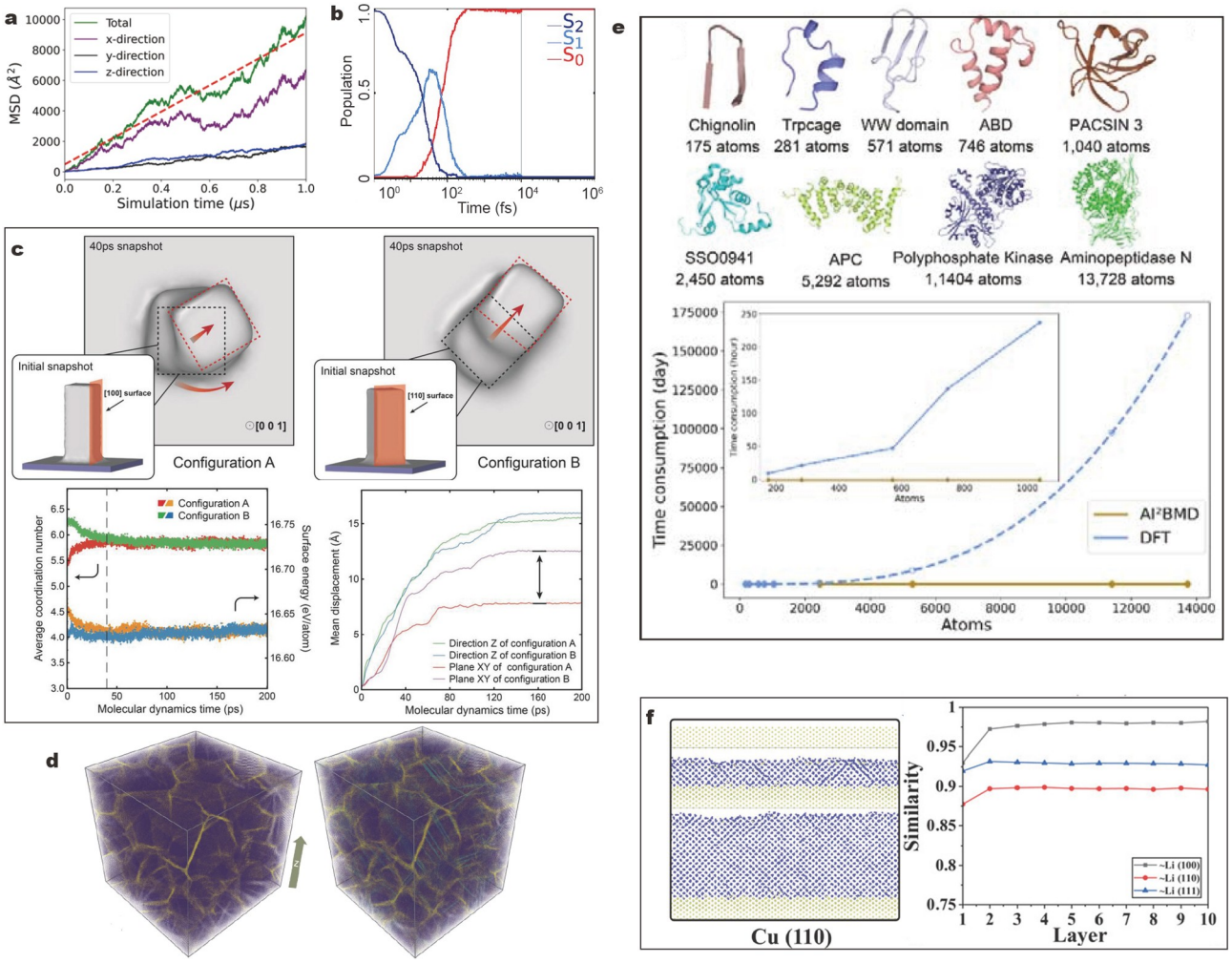

Figure 5 Applications of ML-IPs: (a) direction projected MSD curves of Li ion at 300 K of $Li_7P_3S_{11}$. Reprinted with permission from Ref. [129]. Copyright 2023, the Author(s). (b) Nonadiabatic MD simulations using DNNs for 1 ns. After excitation to the S2 state, ultrafast internal conversion to the S1 state takes place, followed by recovery of the S0 state within 300 fs. Until 10 ps, an ensemble of 200 trajectories is analyzed, followed by the population averaged from 2 trajectories. Reprinted with permission from Ref. [130]. Copyright 2019, the Royal Society of Chemistry. (c) Morphology changes of cuboid configurations with different exposed surfaces, top view of cuboid configuration with [98] exposed plane and [108] exposed plane, and statistics changes of surface atoms and mean displacement of configurations during MLFF-MD simulation. Reprinted with permission from Ref. [131]. Copyright 2022, Wiley-VCH GmbH. (d) 10,401,218-atom nanocrystalline copper consisting of 64 randomly oriented crystals with 15-nm averaged grain diameter, and the nanocrystalline copper after 10% tensile deformation along the *z* axis. Purple, yellow, and cyan denote the atoms in the grains, atoms in the grain boundaries, and atoms in the stacking faults. Reprinted with permission from Ref. [132]. Copyright 2020, IEEE. (e) Folded structures of 9 evaluated proteins. For these proteins, the number of atoms ranges from 175 to 13,728, and time consumption of energy calculation for 9 proteins. Reprinted with permission from Ref. [133]. Copyright 2023, the Authors. (f) Li homogeneous deposition on the Cu surfaces with different indices and the results of SSA, and snapshots of Li homogeneous deposition and the curve of SSA results of the first 10 Li monolayers on the Cu (110) surfaces. Reprinted with permission from Ref. [134]. Copyright 2022, Wiley-VCH GmbH.

significantly accelerated and improved structural sampling [99,138,139]. However, there are still challenges in balancing the proportion of transition state structures and equilibrium structures in chemical reactions to achieve global PES sampling. Developing an integrated approach combining active learning, crystal structure sampling, materials enhancement, and twin neural networks holds promise for rapidly sampling the global PESs in complex material systems and chemical reactions. (2) In accelerating PES exploration through unsupervised learning, research into designing unsupervised generative models for inorganic crystal structures is still in its exploratory stage. Further increasing the effective rate of crystal structure generation and enhancing structural diversity are important directions for the field. Additionally, existing models' structure generation processes are stochastic and often challenging to control in terms of the number of atoms and stoichiometry. Developing conditional generative models offers the potential for controlled crystal structure generation, providing new means for designing function-oriented materials. (3) Currently, there is a lack of systematic comparisons of ML methods for modeling atomic interactions in terms of both accuracy and efficiency. One of the main reasons is the absence of widely accepted and challenging efficiency. Existing datasets that are widely used, such as QM-9 [140], have relatively low difficulty levels and do not fully reflect the differences in the capabilities of different methods. The field urgently needs to establish datasets with a significant impact on the domain, similar to ImageNet [141]. In summary, the field of structural sampling needs to overcome challenges related to

datasets, accuracy, and diversity to enhance the efficiency and accuracy of materials design and PES exploration.

Challenges for descriptors

The current state of structural descriptors presents several critical challenges. (1) Most descriptors tend to focus on mathematically complete forms, leading to the neglect of key physical effects such as quantum fluctuations, electronic correlations, and spin-orbit interactions [71,84]. This issue may be solved by developing more advanced descriptor models that integrate additional physical principles, extending current atomic descriptors into "atom + electron" descriptors. (2) Since the training of potential functions relies on a large number of microscopic structural configurations, current descriptors face challenges in handling complex material systems with heterogeneous interfaces, nanoscale effects, or topological structures. Developing specific descriptors tailored to complex structures, such as novel descriptors combining short-range local environments with long-range interactions, could help address this challenge [1,142]. (3) Descriptors involving many-body interactions often suffer from computational inefficiency. Employing approximation algorithms, such as deep learning and graph neural networks, may facilitate the optimization of current descriptors with large to-be-fitted coefficients [47,68,76,143–145]. (4) Current structural descriptors might only reflect the static structural features, without considering the dynamic influence of external field conditions, such as temperature, pressure, and chemical environments on structural descriptors [19]. Therefore, integrating intrinsic structural parameters with experimental condition data is an important direction for constructing novel descriptors.

In general, the future of structural descriptors will encompass various developments, including better incorporation of physical effects, handling complex material systems, and improving computational efficiency. We can foresee that with continuous refinement of physical connotations, mathematical methods, and computational techniques, the emergence of accurate and efficient descriptors is on the horizon, providing powerful tools for cross-scale computational design of materials.

Challenges for fitting methods

When it comes to fitting methods for ML-IP, several primary challenges emerge: (1) harnessing the robust fitting capabilities of ML based on high-dimensional atomic environment descriptors and tensor features has led to a series of ML-IPs models using neural networks. These models approach chemical accuracy in small-scale benchmark tests. However, they have predominantly been tailored to specific material systems, and achieving a genuine universal model remains a challenge. While general potential function models like PFP and DPA-1 have been introduced, they still lack the vast datasets available in natural language processing. High-precision DFT datasets currently cover only a fraction of the entire materials landscape. Therefore, further research is needed to enhance the accuracy and generalization of such universal models. This can be accomplished by leveraging higher-precision quantum computations to provide training data and expanding the sample space through parallel high-throughput calculations, reinforcing support for rare cases. (2) Interpretability: in contrast to the high interpretability of DFT, ML-IP models exhibit relatively limited interpretability. While techniques such as SHAP and class activation maps offer some level of model interpretation, true interpretability, which encompasses rigorous physical foundations and the precise disclosure of atomic interactions, remains elusive. Developing model interpretation methods and integrating domain knowledge with ML model interpretations will enhance the quality and applicability of explanations.

In summary, the advancement of high-precision, fast, and universally applicable ML-IP simulation methods, coupled with substantial progress in model interpretability, will accelerate the widespread adoption of ML-IPs across various temporal and spatial scales in disciplines including biology, chemistry, and materials science.

CONCLUSIONS

The existing ML atomic potentials have significantly expanded the temporal and spatial scales of atomic simulations. This review provides an overview of three crucial aspects that are paramount to ML atomic potentials: (1) structure sampling, (2) structure descriptors, and (3) potential energy fitting methods. Atomic simulations based on ML atomic potentials could be broadly applied in fields such as materials science, chemistry, and biology. Leveraging the cross-temporal and cross-spatial scale characteristics of ML atomic potentials enables the prediction of new material structures and automatic exploration of chemical reaction mechanisms.

In the future, through further optimization of ML models and simulation methods, along with the construction of a large-scale, high-precision dataset, it is possible to develop a universal ML atomic potential model that spans the entire materials domain. This advancement will further expand the application scenarios of ML atomic potentials, accelerate the research cycle in theoretical simulations, and push over the boundaries of atomic simulations in terms of both time and spatial scales.

Received 1 October 2023; accepted 26 February 2024; published online 12 March 2024

1 Deringer VL, Caro MA, Csányi G. Machine learning interatomic potentials as emerging tools for materials science. Adv Mater, 2019, 31: 1902765
2 Musil F, Grisafi A, Bartók AP, *et al.* Physics-inspired structural representations for molecules and materials. Chem Rev, 2021, 121: 9759–9815
3 Kirkpatrick J, McMorrow B, Turban DHP, *et al.* Pushing the frontiers of density functionals by solving the fractional electron problem. Science, 2021, 374: 1385–1389
4 Behler J. Four generations of high-dimensional neural network potentials. Chem Rev, 2021, 121: 10037–10072
5 Kresse G, Furthmüller J. Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set. Phys Rev B, 1996, 54: 11169–11186
6 Kresse G, Furthmüller J. Efficiency of *ab-initio* total energy calculations for metals and semiconductors using a plane-wave basis set. Comput Mater Sci, 1996, 6: 15–50
7 White JA, Bird DM. Implementation of gradient-corrected exchange-correlation potentials in Car-Parrinello total-energy calculations. Phys Rev B, 1994, 50: 4954–4957
8 Car R, Parrinello M. Unified approach for molecular dynamics and density-functional theory. Phys Rev Lett, 1985, 55: 2471–2474
9 Behler J, Parrinello M. Generalized neural-network representation of high-dimensional potential-energy surfaces. Phys Rev Lett, 2007, 98: 146401
10 Botu V, Ramprasad R. Learning scheme to predict atomic forces and accelerate materials simulations. Phys Rev B, 2015, 92: 094306

11 Wen T, Zhang L, Wang H, *et al.* Deep potentials for materials science. Mater Futures, 2022, 1: 022601
12 Wang X, Sheng Y, Ning J, *et al.* A critical review of machine learning techniques on thermoelectric materials. J Phys Chem Lett, 2023, 14: 1808–1822
13 Tavakoli M, Mood A, Van Vranken D, *et al.* Quantum mechanics and machine learning synergies: Graph attention neural networks to predict chemical reactivity. J Chem Inf Model, 2022, 62: 2121–2132
14 von Lilienfeld OA, Ramakrishnan R, Rupp M, *et al.* Fourier series of atomic radial distribution functions: A molecular fingerprint for machine learning models of quantum chemical properties. Int J Quantum Chem, 2015, 115: 1084–1093
15 Xu P, Alkan M, Gordon MS. Many-body dispersion. Chem Rev, 2020, 120: 12343–12356
16 Jenke J, Subramanyam APA, Densow M, *et al.* Electronic structure based descriptor for characterizing local atomic environments. Phys Rev B, 2018, 98: 144102
17 Yoo D, Jung J, Jeong W, *et al.* Metadynamics sampling in atomic environment space for collecting training data for machine learning potentials. npj Comput Mater, 2021, 7: 131
18 Qiu W, Wang Y, Liu J. Multiscale computations and artificial intelligent models of electrochemical performance in Li-ion battery materials. WIREs Comput Mol Sci, 2022, 12: e1592
19 Guo ZY, Li CX, Gao M, *et al.* Mn–O covalency governs the intrinsic activity of Co–Mn spinel oxides for boosted peroxymonosulfate activation. Angew Chem Int Ed, 2021, 60: 2
20 Shi Z, Yang W, Deng X, *et al.* Machine-learning-assisted high-throughput computational screening of high performance metal–organic frameworks. Mol Syst Des Eng, 2020, 5: 725–742
21 Chandrasekaran A, Kamal D, Batra R, *et al.* Solving the electronic structure problem with machine learning. npj Comput Mater, 2019, 5: 22
22 Ma S, Huang SD, Liu ZP. Dynamic coordination of cations and catalytic selectivity on zinc–chromium oxide alloys during syngas conversion. Nat Catal, 2019, 2: 671–677
23 Liu G, Robertson AW, Li MMJ, *et al.* MoS_2 monolayer catalyst doped with isolated Co atoms for the hydrodeoxygenation reaction. Nat Chem, 2017, 9: 810–816
24 Ran N, Sun B, Qiu W, *et al.* Identifying metallic transition-metal dichalcogenides for hydrogen evolution through multilevel high-throughput calculations and machine learning. J Phys Chem Lett, 2021, 12: 2102–2111
25 Qiu W, Xi L, Wei P, *et al.* Part-crystalline part-liquid state and rattling-like thermal damping in materials with chemical-bond hierarchy. Proc Natl Acad Sci USA, 2014, 111: 15031–15035
26 Ding J, Patinet S, Falk ML, *et al.* Soft spots and their structural signature in a metallic glass. Proc Natl Acad Sci USA, 2014, 111: 14052–14056
27 Iwashita T, Nicholson DM, Egami T. Elementary excitations and crossover phenomenon in liquids. Phys Rev Lett, 2013, 110: 205504
28 Wales DJ, Doye JPK. Global optimization by basin-hopping and the lowest energy structures of Lennard-Jones clusters containing up to 110 atoms. J Phys Chem A, 1997, 101: 5111–5116
29 Jacobsen TL, Jørgensen MS, Hammer B. On-the-fly machine learning of atomic potential in density functional theory structure optimization. Phys Rev Lett, 2018, 120: 026102
30 Tong Q, Xue L, Lv J, *et al.* Accelerating CALYPSO structure prediction by data-driven learning of a potential energy surface. Faraday Discuss, 2018, 211: 31–43
31 Sosso GC, Miceli G, Caravati S, *et al.* Neural network interatomic potential for the phase change material GeTe. Phys Rev B, 2012, 85: 174103
32 Gastegger M, Marquetand P. High-dimensional neural network potentials for organic reactions and an improved training algorithm. J Chem Theor Comput, 2015, 11: 2187–2198
33 Herr JE, Yao K, McIntyre R, *et al.* Metadynamics for training neural network model chemistries: A competitive assessment. J Chem Phys, 2018, 148: 241710
34 Amabilino S, Bratholm LA, Bennie SJ, *et al.* Training neural nets to learn reactive potential energy surfaces using interactive quantum chemistry in virtual reality. J Phys Chem A, 2019, 123: 4486–4499
35 Goodfellow I, Pouget-Abadie J, Mirza M, *et al.* Generative adversarial networks. Commun ACM, 2020, 63: 139–144
36 Kingma DP, Welling M. Auto-encoding variational bayes. arXiv: 1312.6114
37 Rezende DJ, Mohamed S. Variational inference with normalizing flows. arXiv: 1505.05770
38 Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. arXiv: 2006.11239
39 Luo X, Wang Z, Gao P, *et al.* Review on machine learning accelerated crystal structure prediction. J Chin Ceram Soc, 2023, 51: 552–560
40 Ouyang R, Xie Y, Jiang DE. Global minimization of gold clusters by combining neural network potentials and the basin-hopping method. Nanoscale, 2015, 7: 14817–14821
41 Banerjee A, Jasrasaria D, Niblett SP, *et al.* Crystal structure prediction for benzene using basin-hopping global optimization. J Phys Chem A, 2021, 125: 3776–3784
42 Yang S, Day GM. Exploration and optimization in crystal structure prediction: Combining basin hopping with quasi-random sampling. J Chem Theor Comput, 2021, 17: 1988–1999
43 Wu SQ, Ji M, Wang CZ, *et al.* An adaptive genetic algorithm for crystal structure prediction. J Phys-Condens Matter, 2014, 26: 035402
44 Patra TK, Meenakshisundaram V, Hung JH, *et al.* Neural-network-biased genetic algorithms for materials design: Evolutionary algorithms that learn. ACS Comb Sci, 2017, 19: 96–107
45 Lv J, Wang Y, Zhu L, *et al.* Particle-swarm structure prediction on clusters. J Chem Phys, 2012, 137: 084104
46 Luo X, Yang J, Liu H, *et al.* Predicting two-dimensional boron–carbon compounds by the global optimization method. J Am Chem Soc, 2011, 133: 16285–16290
47 Schütt KT, Sauceda HE, Kindermans PJ, *et al.* SchNet—A deep learning architecture for molecules and materials. J Chem Phys, 2018, 148: 241722
48 Zhang L, Han J, Wang H, *et al.* Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. Phys Rev Lett, 2018, 120: 143001
49 Kirkpatrick S, Gelatt Jr. CD, Vecchi MP. Optimization by simulated annealing. Science, 1983, 220: 671–680
50 Martoňák R, Laio A, Parrinello M. Predicting crystal structures: The Parrinello-Rahman method revisited. Phys Rev Lett, 2003, 90: 075503
51 Pannetier J, Bassas-Alsina J, Rodriguez-Carvajal J, *et al.* Prediction of crystal structures from crystal chemistry rules by simulated annealing. Nature, 1990, 346: 343–345
52 Yang M, Karmakar T, Parrinello M. Liquid-liquid critical point in phosphorus. Phys Rev Lett, 2021, 127: 080603
53 Hummer G, Kevrekidis IG. Coarse molecular dynamics of a peptide fragment: Free energy, kinetics, and long-time dynamics computations. J Chem Phys, 2003, 118: 10762–10773
54 Engkvist O, Karlström G. A method to calculate the probability distribution for systems with large energy barriers. Chem Phys, 1996, 213: 63–76
55 Hénin J, Tajkhorshid E, Schulten K, *et al.* Diffusion of glycerol through *Escherichia coli* aquaglyceroporin GlpF. Biophys J, 2008, 94: 832–839
56 Comer J, Roux B, Chipot C. Achieving ergodic sampling using replica-exchange free-energy calculations. Mol Simul, 2014, 40: 218–228
57 Voter AF. Hyperdynamics: Accelerated molecular dynamics of infrequent events. Phys Rev Lett, 1997, 78: 3908–3911
58 Metropolis N, Rosenbluth AW, Rosenbluth MN, *et al.* Equation of state calculations by fast computing machines. J Chem Phys, 2004, 21: 1087–1092
59 Laio A, Parrinello M. Escaping free-energy minima. Proc Natl Acad Sci USA, 2002, 99: 12562–12566
60 Li YF, Liu ZP. Active site revealed for water oxidation on electrochemically induced δ-MnO_2: Role of spinel-to-layer phase transition. J Am Chem Soc, 2018, 140: 1783–1792
61 Huang SD, Shang C, Zhang XJ, *et al.* Material discovery by combining stochastic surface walking global optimization with a neural network.

Chem Sci, 2017, 8: 6327–6337
62 Huang SD, Shang C, Kang PL, *et al.* Atomic structure of boron resolved using machine learning and global sampling. Chem Sci, 2018, 9: 8644–8655
63 Goodfellow IJ, Pouget-Abadie J, Mirza M, *et al.* Generative adversarial networks. arXiv: 1406.2661
64 Yang MJ, Cho KH, Merchant A, *et al.* Scalable diffusion for materials generation. arXiv: 2311.09235
65 Nouira A, Sokolovska N, Crivello JC. Crystalgan: Learning to discover crystallographic structures with generative adversarial networks. arXiv: 1810.11203
66 Kim B, Lee S, Kim J. Inverse design of porous materials using artificial neural networks. Sci Adv, 2020, 6: eaax9324
67 Xie T, Fu X, Ganea OE, *et al.* Crystal diffusion variational autoencoder for periodic material generation. arXiv: 2110.06197
68 Xie T, Grossman JC. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Phys Rev Lett, 2018, 120: 145301
69 Jain A, Ong SP, Hautier G, *et al.* Commentary: The materials project: A materials genome approach to accelerating materials innovation. APL Mater, 2013, 1: 011002
70 Li S, Liu Y, Chen D, *et al.* Encoding the atomic structure for machine learning in materials science. WIREs Comput Mol Sci, 2021, 12: e1558
71 Himanen L, Jäger MOJ, Morooka EV, *et al.* DScribe: Library of descriptors for machine learning in materials science. Comput Phys Commun, 2020, 247: 106949
72 Rupp M, Tkatchenko A, Müller KR, *et al.* Fast and accurate modeling of molecular atomization energies with machine learning. Phys Rev Lett, 2012, 108: 058301
73 Low K, Kobayashi R, Izgorodina EI. The effect of descriptor choice in machine learning models for ionic liquid melting point prediction. J Chem Phys, 2020, 153: 104101
74 Vladyka A, Sahle CJ, Niskanen J. Towards structural reconstruction from X-ray spectra. Phys Chem Chem Phys, 2023, 25: 6707–6713
75 Çaylak O, Anatole von Lilienfeld O, Baumeier B. Wasserstein metric for improved quantum machine learning with adjacency matrix representations. Mach Learn-Sci Technol, 2020, 1: 03LT01
76 Lu S, Zhou Q, Guo Y, *et al.* On-the-fly interpretable machine learning for rapid discovery of two-dimensional ferromagnets with high Curie temperature. Chem, 2022, 8: 769–783
77 Lu S, Zhou Q, Guo Y, *et al.* Coupling a crystal graph multilayer descriptor to active learning for rapid discovery of 2D ferromagnetic semiconductors/half-metals/metals. Adv Mater, 2020, 32: e2002658
78 Huo H, Rupp M. Unified representation of molecules and crystals for machine learning. Mach Learn-Sci Technol, 2022, 3: 045017
79 Laakso J, Himanen L, Homm H, *et al.* Updates to the dscribe library: New descriptors and derivatives. J Chem Phys, 2023, 158: 234802
80 Laakso J, Todorović M, Li J, *et al.* Compositional engineering of perovskites with machine learning. Phys Rev Mater, 2022, 6: 113801
81 van der Vaart A, Bursulaya BD, Brooks CL, *et al.* Are many-body effects important in protein folding? J Phys Chem B, 2000, 104: 9554–9563
82 Merbis W, de Domenico M. Emergent information dynamics in many-body interconnected systems. Phys Rev E, 2023, 108: 014312
83 Pronobis W, Tkatchenko A, Müller KR. Many-body descriptors for predicting molecular properties with machine learning: Analysis of pairwise and three-body interactions in molecules. J Chem Theor Comput, 2018, 14: 2991–3003
84 Pozdnyakov SN, Willatt MJ, Bartók AP, *et al.* Incompleteness of atomic structure representations. Phys Rev Lett, 2020, 125: 166001
85 Behler J. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. J Chem Phys, 2011, 134: 074106
86 Zhang K, Yin L, Liu G. Physically inspired atom-centered symmetry functions for the construction of high dimensional neural network potential energy surfaces. Comput Mater Sci, 2021, 186: 110071
87 Tayfuroglu O, Kocak A, Zorlu Y. A neural network potential for the IRMOF series and its application for thermal and mechanical behaviors. Phys Chem Chem Phys, 2022, 24: 11882–11897
88 Guo Y, Wu X, Fu J. Revisiting the stable structures of gold clusters: Au_n (n = 16–25) by artificial neural network potential. J Phys D-Appl Phys, 2023, 56: 375302
89 Yanxon H, Zagaceta D, Tang B, *et al.* PyXtal_FF: A python library for automated force field generation. Mach Learn-Sci Technol, 2020, 2: 027001
90 Bartók AP, Kondor R, Csányi G. On representing chemical environments. Phys Rev B, 2013, 87: 184115
91 Tirelli A, Tenti G, Nakano K, *et al.* High-pressure hydrogen by machine learning and quantum Monte Carlo. Phys Rev B, 2022, 106: L041105
92 De S, Bartók AP, Csányi G, *et al.* Comparing molecules and solids across structural and alchemical space. Phys Chem Chem Phys, 2016, 18: 13754–13769
93 Ferreira AR. Chemical bonding in metallic glasses from machine learning and crystal orbital Hamilton population. Phys Rev Mater, 2020, 4: 113603
94 Rosenbrock CW, Gubaev K, Shapeev AV, *et al.* Machine-learned interatomic potentials for alloys and alloy phase diagrams. npj Comput Mater, 2021, 7: 24
95 Caruso C, Cardellini A, Crippa M, *et al.* *TimeSOAP*: Tracking high-dimensional fluctuations in complex molecular systems *via* time variations of SOAP spectra. J Chem Phys, 2023, 158: 214302
96 Song X, Deng C. Atomic energy in grain boundaries studied by machine learning. Phys Rev Mater, 2022, 6: 043601
97 Shapeev AV. Moment tensor potentials: A class of systematically improvable interatomic potentials. Multiscale Model Simul, 2016, 14: 1153–1173
98 Novoselov II, Yanilkin AV, Shapeev AV, *et al.* Moment tensor potentials as a promising tool to study diffusion processes. Comput Mater Sci, 2019, 164: 46–56
99 Yang H, Zhu Y, Dong E, *et al.* Dual adaptive sampling and machine learning interatomic potentials for modeling materials with chemical bond hierarchy. Phys Rev B, 2021, 104: 094310
100 Zhu Y, Dong E, Yang H, *et al.* Atomic potential energy uncertainty in machine-learning interatomic potentials and thermal transport in solids with atomic diffusion. Phys Rev B, 2023, 108: 014108
101 Novikov I, Grabowski B, Körmann F, *et al.* Magnetic moment tensor potentials for collinear spin-polarized materials reproduce different magnetic states of bcc Fe. npj Comput Mater, 2022, 8: 13
102 Behler J. Perspective: Machine learning potentials for atomistic simulations. J Chem Phys, 2016, 145: 170901
103 Thompson AP, Swiler LP, Trott CR, *et al.* Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. J Comput Phys, 2015, 285: 316–330
104 Bartók AP, Payne MC, Kondor R, *et al.* Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. Phys Rev Lett, 2010, 104: 136403
105 Botu V, Batra R, Chapman J, *et al.* Machine learning force fields: Construction, validation, and outlook. J Phys Chem C, 2017, 121: 511–522
106 Nyshadham C, Rupp M, Bekker B, *et al.* Machine-learned multi-system surrogate models for materials prediction. npj Comput Mater, 2019, 5: 51
107 Rowe P, Csányi G, Alfè D, *et al.* Development of a machine learning potential for graphene. Phys Rev B, 2018, 97: 054303
108 Deringer VL, Csányi G. Machine learning based interatomic potential for amorphous carbon. Phys Rev B, 2017, 95: 094203
109 Vaswani A, Shazeer N, Parmar N, *et al.* Attention is all you need. arXiv: 1706.03762
110 Podryabinkin EV, Shapeev AV. Active learning of linearly parametrized interatomic potentials. Comput Mater Sci, 2017, 140: 171–180
111 Li Z, Kermode JR, De Vita A. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. Phys Rev Lett, 2015, 114: 096405
112 Dhaliwal G, Nair PB, Singh CV. Machine learned interatomic potentials using random features. npj Comput Mater, 2022, 8: 7
113 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with

deep convolutional neural networks. Commun ACM, 2017, 60: 84–90
114 Wolf T, Debut L, Sanh V, *et al.* Huggingface's transformers: State-of-the-art natural language processing. arXiv: 1910.03771
115 Senior AW, Evans R, Jumper J, *et al.* Improved protein structure prediction using potentials from deep learning. Nature, 2020, 577: 706–710
116 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature, 2015, 521: 436–444
117 Wang H, Zhang L, Han J, *et al.* DeePMD-kit: A deep learning package for many-body potential energy representation and molecular dynamics. Comput Phys Commun, 2018, 228: 178–184
118 Chu Q, Luo KH, Chen D. Exploring complex reaction networks using neural network-based molecular dynamics simulation. J Phys Chem Lett, 2022, 13: 4052–4057
119 Lin M, Liu X, Xiang Y, *et al.* Unravelling the fast alkali-ion dynamics in paramagnetic battery materials combined with NMR and deep-potential molecular dynamics simulation. Angew Chem Int Ed, 2021, 60: 12547–12553
120 Nitol MS, Dickel DE, Barrett CD. Artificial neural network potential for pure zinc. Comput Mater Sci, 2021, 188: 110207
121 Gilmer J, Samuel SS, Patrick FR, *et al.* Neural message passing for quantum chemistry. arXiv: 1704.01212
122 Vandenhaute S, Cools-Ceuppens M, DeKeyser S, *et al.* Machine learning potentials for metal-organic frameworks using an incremental learning approach. npj Comput Mater, 2023, 9: 19
123 Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583–589
124 Janson G, Valdes-Garcia G, Heo L, *et al.* Direct generation of protein conformational ensembles *via* machine learning. Nat Commun, 2023, 14: 774
125 Liao YL, Smidt T. Equiformer: Equivariant graph attention transformer for 3D atomistic graphs. arXiv: 2206.11990
126 Zhang D, Bi H, Dai FZ, *et al.* DPA-1: Pretraining of attention-based deep potential model for molecular simulation. arXiv: 2208.08236
127 Takamoto S, Okanohara D, Li QJ, *et al.* Towards universal neural network interatomic potential. J Materiomics, 2023, 9: 447–454
128 Zhou G, Gao Z, Ding Q, *et al.* Uni-Mol: A universal 3D molecular representation learning framework. ChemRxiv, 2023, doi: 10.26434/chemrxiv-2022-jjm0j-v2
129 Xu Z, Duan H, Dou Z, *et al.* Machine learning molecular dynamics simulation identifying weakly negative effect of polyanion rotation on Li-ion migration. npj Comput Mater, 2023, 9: 105
130 Westermayr J, Gastegger M, Menger MFSJ, *et al.* Machine learning enables long time scale molecular photodynamics simulations. Chem Sci, 2019, 10: 8100–8107
131 Zhang W, Weng M, Zhang M, *et al.* Revealing morphology evolution of lithium dendrites by large-scale simulation based on machine learning force field. Adv Energy Mater, 2023, 13: 2202892
132 Jia W, Wang H, Chen M, *et al.* Pushing the limit of molecular dynamics with *ab initio* accuracy to 100 million atoms with machine learning. arXiv: 2005.00223
133 Wang T, He X, Li M, *et al.* AI$_2$BMD: Efficient characterization of protein dynamics with *ab initio* accuracy. bioRxiv, 2023, doi: 10.1101/2023.07.12.548519
134 Lai G, Jiao J, Fang C, *et al.* The mechanism of Li deposition on the Cu substrates in the anode-free Li metal batteries. Small, 2023, 19: 2205416
135 Milardovich D, Wilhelmer C, Waldhoer D, *et al.* Machine learning interatomic potential for silicon-nitride (Si_3N_4) by active learning. J Chem Phys, 2023, 158: 194802
136 Korotaev P, Novoselov I, Yanilkin A, *et al.* Accessing thermal conductivity of complex compounds by machine learning interatomic potentials. Phys Rev B, 2019, 100: 144308
137 Li R, Lee E, Luo T. A unified deep neural network potential capable of predicting thermal conductivity of silicon in different phases. Mater Today Phys, 2020, 12: 100181
138 Dickey JM, Paskin A. Computer simulation of the lattice dynamics of solids. Phys Rev, 1969, 188: 1407–1418
139 Ladd AJC, Moran B, Hoover WG. Lattice thermal conductivity: A comparison of molecular dynamics and anharmonic lattice dynamics. Phys Rev B, 1986, 34: 5058–5064
140 Ramakrishnan R, Dral PO, Rupp M, *et al.* Quantum chemistry structures and properties of 134 kilo molecules. Sci Data, 2014, 1: 140022
141 Deng J, Dong W, Socher R, *et al.* Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Miami, 2009. 248–255
142 Gao A, Remsing RC. Self-consistent determination of long-range electrostatics in neural network potentials. Nat Commun, 2022, 13: 1572
143 Zhou L, Zhu G, Wu Y, *et al.* A framework for metal surface energy prediction based on crystal graph convolutional neural network. J Chin Ceram Soc, 2022, 51: 389
144 Chen C, Ong SP. A universal graph deep learning interatomic potential for the periodic table. Nat Comput Sci, 2022, 2: 718–728
145 Schütt KT, Kessel P, Gastegger M, *et al.* SchNetPack: A deep learning toolbox for atomistic systems. J Chem Theor Comput, 2019, 15: 448–455

Acknowledgements This work was financially supported by the National Key R&D Program of China (2022YFB3807200), Shanghai Explorer Program (Batch I) (23TS1401500), the National Natural Science Foundation of China (22133005), the Project funded by China Postdoctoral Science Foundation (2022M723276 and GZB20230793), Shanghai Sailing Program (23YF1454900), and Shanghai Post-doctoral Excellence Program (2022660).

Author contributions Liu J designed the project. Ran N and Qiu W analyzed the data and wrote the manuscript. Yin L edited the figures. All authors contributed to the general discussion.

Conflict of interest The authors declare that they have no conflict of interest.

Nian Ran received her PhD degree in physical chemistry from the University of Chinese Academy of Sciences in 2022. She works as a postdoctoral researcher at Shanghai Institute of Ceramics, Chinese Academy of Sciences (SICCAS). Her research primarily focuses on utilizing artificial intelligence for materials design.

Wujie Qiu received his PhD degree in theoretical physics from the East China Normal University in 2016. He then worked as a postdoctoral fellow, an assistant research fellow, and later an associate research fellow at SICCAS. His research interests primarily focus on the development of computational electrochemical methods with artificial intelligence and the design of advanced materials.

Jianjun Liu received his PhD degree in physical chemistry from Jilin University in 2002, followed by enriching postdoctoral experiences at the Emory University and Southern Illinois University. In 2011, he joined SICCAS. His research interesting focuses on atom-level material design by various computational methods including classical, quantum, and machine-learning techniques.

机器学习原子间势在材料跨尺度计算模拟中的最新进展

冉念[1,2], 殷亮[1,2,3], 邱吴劼[1,2,4*], 刘建军[1,2,3*]

摘要 近年来，机器学习原子势(ML-IP)因其兼顾高精度和高效率的优势，在材料科学、化学、生物学等领域的大尺度原子模拟研究中引起了广泛关注. 本文聚焦于ML-IP在材料跨尺度计算模型中的应用，全面介绍了ML-IP的结构采样、结构描述符和拟合方法. 这些方法使ML-IP能够以高精度和高效率模拟分子和晶体的动力学和热力学特性. 跨学科研究领域中更高效、先进的技术在开拓覆盖不同时间和空间尺度的广泛应用方面发挥着重要作用. 因此，ML-IP方法为未来的研究和创新铺平了道路，为多个领域带来了革命性的机会.