npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01278-7

Pretraining of attention-based deep learning potential model for molecular simulation

Check for updates

Duo Zhang[1,2,3], Hangrui Bi[1,2], Fu-Zhi Dai[1], Wanrun Jiang[1], Xinzijian Liu[2], Linfeng Zhang[1,2] ✉ & Han Wang[4,5] ✉

Machine learning-assisted modeling of the inter-atomic potential energy surface (PES) is revolutionizing the field of molecular simulation. With the accumulation of high-quality electronic structure data, a model that can be pretrained on all available data and finetuned on downstream tasks with a small additional effort would bring the field to a new stage. Here we propose DPA-1, a Deep Potential model with a gated attention mechanism, which is highly effective for representing the conformation and chemical spaces of atomic systems and learning the PES. We tested DPA-1 on a number of systems and observed superior performance compared with existing benchmarks. When pretrained on large-scale datasets containing 56 elements, DPA-1 can be successfully applied to various downstream tasks with a great improvement of sample efficiency. Surprisingly, for different elements, the learned type embedding parameters form a *spiral* in the latent space and have a natural correspondence with their positions on the periodic table, showing interesting interpretability of the pretrained DPA-1 model.

Reliably representing the inter-atomic potential energy surface (PES) is core to the study of properties of molecules and materials in computational physics, chemistry, materials science, biology, etc. While electronic structure methods typically give accurate and transferable PES, they are prohibitively expensive for scaling to systems of more than thousands of atoms. On the other hand, empirical force fields are much more efficient but are inherently limited by their accuracy in many applications. By properly integrating machine learning (ML) methodologies and physical requirements like extensiveness and symmetries, various methods have emerged to address the accuracy *v.s.* efficiency dilemma in the realm of PES modeling[1–11]. Arguably, a new paradigm is forming: electronic structure methods are no longer used to generate the driving forces during molecular dynamics simulations but are used to generate data for training their alternatives, ML-based PES models.

Despite remarkable achievements of ML-based PES models[12–14], challenges still remain. For a domain expert who would like to apply such methodologies in their applications, a natural first question is on the efforts needed for obtaining a reliable PES model: Are there ready-to-use PES models? If not, what would be the amount of training data and time cost required? Can we take advantage of the ever-increasing publicly-available training data?

To address these issues, there have been several efforts. On one hand, general-purpose models for various systems, such as silicon[15], phosphorus[16], water[17], metals and alloys[18–22], etc., have been developed and are directly applicable to relevant studies. However, the range of applicability of such models is typically limited to small conformation or chemical space. For example, for alloys, the majority of general-purpose ML models are developed for systems with at most two element types. On the other hand, several efficient data generation protocols have been developed[23–26], of which a representative is DP-GEN[25,26], a concurrent learning procedure that iteratively explores the configuration space using models trained with existing data, and then labels only those configurations with high uncertainty level. Even with these protocols, the computational effort needed for complicated systems is still prohibitive. For example, to train a fairly general-purpose model for the AlMgCu alloy system, 100k density functional theory (DFT)[27,28] calculations were ultimately performed, resulting in the cost of ten million CPU core hours[18].

With the accumulation of high-quality electronic structure data covering almost all the elements on the periodic table, it is becoming possible to systematically develop pretraining schemes, which have been widely adopted in areas like computer vision (CV)[29,30] and natural language processing (NLP)[31,32]. In these schemes, one first trains a unified model on

[1]AI for Science Institute, Beijing 100080, China. [2]DP Technology, Beijing 100080, China. [3]Academy for Advanced Interdisciplinary Studies, Peking University, Beijing 100871, China. [4]Laboratory of Computational Physics, Institute of Applied Physics and Computational Mathematics, Beijing 100094, China. [5]HEDPS, CAPT, College of Engineering, Peking University, Beijing 100871, China. ✉e-mail: linfeng.zhang.zlf@gmail.com; wang_han@iapcm.ac.cn

large-scale datasets and then finetunes it for downstream tasks, expecting that a good representation can be learned in the first stage, and the amount of supervised data needed for the second stage will be significantly reduced. Recently, the pretraining-finetuning idea has been applied to organic molecules systems for energy and force predictions[33,34], and to tackle tasks beyond representing the PES[35–37]. Unfortunately, most ML-based PES models are premature for such schemes at scale in materials applications. Taking the widely used two versions of Deep Potential models[6,7] as examples, the ML parameters are element-type-dependent, making it highly inefficient when the training data contains many elements.

Constant efforts have been devoted to adapt the architecture of the ML-based PES models for large datasets. Among them, one class of models named equivariant graph neural networks (GNN)[38] that is built upon convolutions over atomic graphs of node and edge equivariant representations has shown promise of training on large datasets. SchNet[5], PaiNN[39], GemNet-OC[40], DimeNet++[41], PFP[42], SCN[43], SpinConv[44] and Equiformer/EquiformerV2[45,46] are trained on the OC20/OC2M[47] dataset containing about 133M/2M data frames of 56 elements. These models are benchmarked by the accuracy of energy, force and stable structure predictions. Very recently, it has been shown that introducing the attention architecture[45] in a GNN model improves the performance on the OC20/OC2M dataset[46]. Chen and Ong[48] proposed M3GNet, which was able to train on a subset of the Materials Project[49] that contains 187,687 configurations encompassing 89 elements and labeled at the generalized gradient approximation (GGA)[50] or GGA+U level. Takamota et. al.[42] introduced the PFP model, which was trained on a dataset composed of molecular and crystal configurations including approximately 9×10^6 frames of 45 elements. Choudhary et. al.[51] developed the ALIGNN model, and they were able to train the model on a subset of the JARVIS-DFT dataset[52] that is composed of 307,113 data frames of 89 elements. The M3GNet, PFP, and ALIGNN models are proposed as "universal" potential models, however, their accuracies are not on par with PES models trained for specific materials applications.

The equivariant GNN models are potential candidates for pretraining, several issues worth special attention before applying them in downstream real-world applications. First, the GNN approaches are not well-suited for massively parallel molecular dynamics simulations[53]. The update of each GNN layer requires communications between spatially decomposed subregions of the system. In each evaluation of the energy and forces, in total several to a dozen such updates are required, which may lead to a substantial communication overhead in massively parallel high-performance supercomputers. Second, some models, such as PaiNN, GemNet-OC, SCN, Equiformer/EquiformerV2, directly predict forces using rotationally equivariant networks[39,40,45,54] instead of energy gradients with respect to atomic coordinates. Therefore, the predicted force is not conservative, which serves as a basic assumption in guaranteeing the accuracy of molecular simulations[55]. The DimeNet++[41] Allegro[53] models are conservative. Last but not least, some models, such as GemNet-OC, SpinConv, M3GNet, and ALIGNN are not smooth, i.e. a sudden energy jump may happen as the positions of atoms infinitesimally varies. This leads to non-conserved energy in the Hamiltonian dynamics simulations, which is used in computing the dynamical properties like diffusion constant and viscosity.

By far, how much the downstream materials applications benefit from the ML models trained on the large-scale datasets are still not clear. To answer the question, in this article, we propose DPA-1, a Deep Potential model with a gated attention mechanism. Designed with a local descriptor, this model is exceptionally well-suited for parallel simulations on large-scale systems containing millions of atoms[56]. Notably, DPA-1 predicts conservative forces, ensures smoothness and demonstrates outstanding efficacy in learning interatomic interactions. Moreover, once pretrained, DPA-1 can significantly decrease the supplementary efforts needed for subsequent downstream tasks. We tested DPA-1 on various systems and observed superior performance compared with existing benchmarks. Then we took AlMgCu alloy systems[18] as an example, showing that after pretraining with single-element and binary samples, DPA-1 can save around 90% ternary samples compared with the DeepPot-SE model[7]. Finally, we pretrained DPA-1 using the OC20 dataset, which consists of 56 elements, and successfully applied it to various downstream tasks. We checked the interpretability of the pretrained model by looking into the learned embedding parameters for different element types, finding that the 56 elements are arranged on a *spiral* in the latent space, which has a natural correspondence with their physical properties on the periodic table. Above all, we believe that DPA-1 and the pretraining scheme will bring the field of molecular simulation to a new stage.

Results

We conducted a number of experiments to evaluate the performance of DPA-1, with its architecture illustrated in Fig. 1 and detailed in the Methods section. First, to test the model's ability to transfer among different compositions, we trained it from scratch against various systems and tested it under several challenging schemes. Then, we used an AlMgCu dataset to test its ability to transfer to ternary systems upon pretraining with single-element and binary data. Finally, we pretrained DPA-1 using the OC2M subset in OC20 dataset[47] and applied it to various downstream tasks. To illustrate the effectiveness of the type-embedding and attention schemes, we compared them against DeepPot-SE model[7] in all the experiments. In the following, we shall introduce first the datasets we used, and then the experiments we conducted.

Datasets

AlMgCu alloy systems[18]. This dataset is generated using DP-GEN[26], a concurrent learning scheme. After exploring 2.73 billion alloy configurations (derived from ~2000 bulk and surface systems), only a small portion (~100k configurations) of them are labeled and then compose the compact dataset. The exploration runs in the whole concentration space, i.e., $Al_xCu_yMg_z$ with $0 \le x, y, z \le 1$, $x + y + z = 1$, and x, y, z take discrete values permitted by the finite-size simulation boxes. We can divide the systems into ***single***, ***binary***, and ***ternary*** subsets, in the name of the number of non-zero *x, y, and z*. The configuration space covers a temperature range of around 50.0 K to 2579.8 K and a pressure range of around 1 bar to 50,000 bar.

Solid-state electrolyte (SSE) systems[57]. These systems contain $Li_{10}XP_2S_{12}$-type SSE materials, where X represents a single or combination of Ge/Si/Sn, and can be divided into three main parts: ***init***, ***mix*** and ***single***. The *init* part comes from the standard DP-GEN scheme starting from 590 structures that are generated via slightly perturbing DFT-relaxed crystal structures, $Li_{10}SiP_2S_{12}$ and $Li_{10}SnP_2S_{12}$ from Materials Project[49]. The exploration covers both ordered structures relaxed by DFT (i.e. structures downloaded from the Materials Project database, in which the position of Ge/Si/Sn/P atoms are fixed) and disordered structures whose 4d sites are randomly occupied by Ge/Si/Sn/P. Based on the *init* part, the *mix* part contains further exploration in binary and ternary mixture of Ge/Si/Sn, while the *single* part covers only a single X in Ge/Si/Sn with other changes in lattice and ratio of Li.

HEA systems. The High Entropy Alloy HEA dataset includes bulk TaNbWMoVAl alloy systems of various configurations and compositions. We employ DP-GEN to explore the composition space, starting from $Ta_3Nb_3W_3Mo_3V_3Al_1$, a 16-atom unit cell containing the former 5 elements as main components and Al as an additive. The dataset is divided into two subsets: ***interior*** and ***exterior***. The *interior* (higher entropy) subset includes composition variations near the starting point. It covers six-component, quinary, quaternary, and ternary alloys. The *exterior* (lower entropy) subset includes systems that are close to the corners and edges of the composition space. It includes systems where one or two elements dominate, binary alloys, and simple substance systems. For both subsets, the temperature range is around 50.0 K to 388.1 K and the pressure range is around 1 bar to 50000 bar.

OC20[47]. OC20 consists of single adsorbates (small molecules) physically binding to the surfaces of catalysts covering periodic bulk materials with 56 elements. Both the chemical diversity and system size are much more complex than other benchmark datasets, such as MD17[58], ANI-1x[24], or QM9[59]. OC2M is a subset including 2 million data points (energies and forces) randomly sampled from OC20, which is still challenging for model training and decent for pretraining. Johannes et al. recently provided several baselines on OC2M, taking months to converge[40].

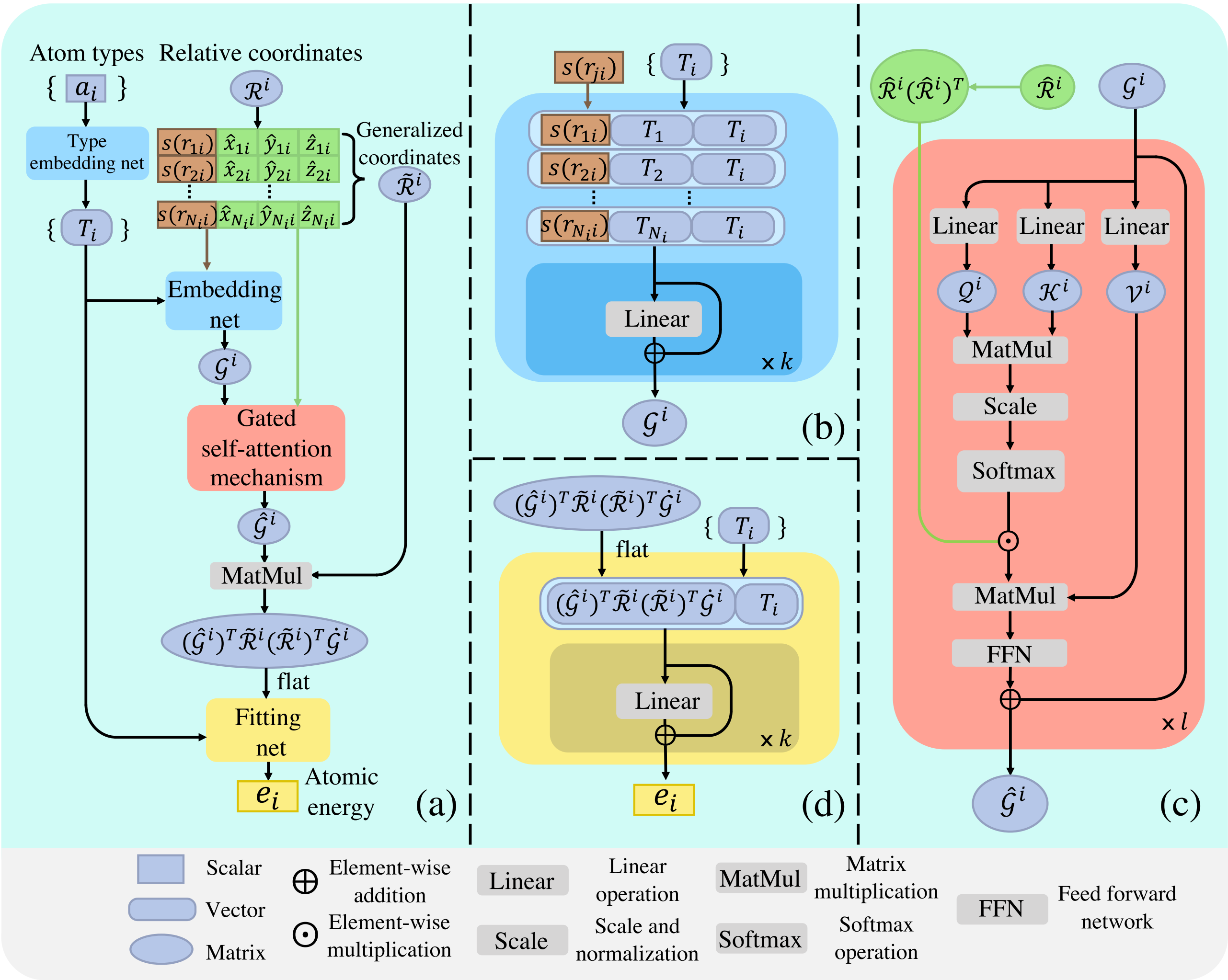

Fig. 1 | Schematic illustration of DPA-1. a Flowchart from $\mathcal{A}^i$ and $\mathcal{R}^i$ to the atomic energy e_i. **b** Structure of the Embedding net, which maps $s(r_{ji})$ and T_i, through multiple residual layers, to $\mathcal{G}^i$. **c** Self-attention mechanism on $\mathcal{G}^i$ through a standard scale-dot procedure gated by the angular information $\hat{\mathcal{R}}^i(\hat{\mathcal{R}}^i)^T$. **d** Fitting net structures, similar to Embedding net, from the descriptors $\mathcal{D}^i$ and T_i to final atomic energy e_i.

Accuracy on various datasets, trained from scratch

The majority of existing models usually focus on the ability to transfer among different configurations, in which case training and validation subsets consist of similar compositions (e.g. randomly sampled from the same dataset). However, to perform pretraining, the upstream and downstream datasets may differ violently. Thus, it's vital for models under the pretraining scheme to transfer among different compositions or even among different datasets, which has, as far as we know, rarely been discussed before. In this work, we mainly focus on a more general but challenging scheme to comprehensively test the generalization ability of the model.

We first designed several challenging tasks to test the model's ability to transfer among different compositions. For AlMgCu, SSE, and HEA systems, we divided them into subsets with different compositions for training and validation (See Datasets subsection for details). The results of DPA-1 and DeepPot-SE are shown in Table 1. With the training loss nearly the same (omitted in the table), the DPA-1 drastically outperforms DeepPot-SE in validation accuracy. For example, for AlMgCu systems, when trained only on single- and binary-element samples, the validation RMSE of DPA-1 on ternary samples can outperform DeepPot-SE by one order of magnitude (6.99 versus 65.1 meV/atom). This suggests that the DPA-1 model might have learned the latent interactions of ternary pairs Al-Mg-Cu from binary pairs Al-Mg, Al-Cu, Mg-Cu, and single-element interactions, possibly thanks to the type-embedding scheme and attention mechanism. We conducted an ablation study in Supplementary Note 1 on HEA systems to demonstrate the influence of each structural component.

To test the performance of DPA-1 in terms of predicting more physical quantities, we performed geometry relaxations on all AlMgCu ternary alloys available from the Materials Project to evaluate their accuracy in predicting formation energy and equilibrium volume (see details in Supplementary Note 2). We also used it to calculate the elastic moduli of AlMgCu systems, which requires accurately capturing the second-order information (see details in Supplementary Note 3). Additionally, we carried out molecular dynamics simulations on LiGePS systems to assess the diffusion coefficients in relation to temperature, comparing the results to ab initio molecular dynamics (AIMD) simulations and experimental studies (see details in Supplementary Note 4). In all tests, satisfactory agreement with the DFT and/or experimental references are obtained.

As a supplement, we also trained DPA-1 model on several simple systems to compare with other ML-based PES. Since these tasks are much easier than the above ones and out of our main focus, we place the results in Supplementary Note 8. Note that there may be relatively little room for improvement on these simple datasets, while DPA-1 can still outperform other methods with even less training samples.

Sample efficiency of pretrained models

As shown in Fig. 2, we use the learning curves to illustrate in terms of the amount of additional training data saved for downstream tasks thanks to model pretraining. In all the experiments, the learning curves were generated by an active learning procedure, in which a pool of data labeled by energy and force is prepared and three steps are repeated iteratively: using samples in the training pool to train the model; testing the model using the remaining samples; selecting 50 samples with the largest prediction errors on per-atom energies and adding them to the training pool. We use the term sample efficiency to denote the amount of training samples required by a model to achieve a given accuracy level for a certain task. The hyperparameter settings in these tests can be found in Supplementary Note 9.

We started with a relatively simple task to compare DeepPot-SE and DPA-1. In this task, both the two models were pretrained using single-element and binary subsets of the AlMgCu systems, and the learning curves were obtained using the AlMgCu ternary subset. As shown in Fig. 2a, DPA-1 exhibits a much better sample efficiency than DeepPot-SE, which should be expected.

Next, we used the OC2M dataset, which contains 56 elements, to pretrain DPA-1 and evaluated its performance on the HEA systems and the AlCu systems (Fig. 2b, c, respectively). As shown in Fig. 3c, the training cost of DeepPot-SE scales quadratically with the number of elements, making its pretraining computationally infeasible, while the number of elements has no effects on the training cost of DPA-1. It is observed that the sample efficiency of DPA-1 pretrained on OC2M is generally better than DPA-1 from scratch, while DeepPot-SE from scratch is the worst. Moreover, compared with AlCu systems, the improvement of pretraining is much more significant for HEA systems, possibly due to the fact that the number of elements of HEA is much larger than AlCu, and the local chemical environment is much more complicated.

Table 1 | Validation RMSE of DPA-1 and DeepPot-SE on energy (Δ E, meV/atom) and atomic forces (Δ F, meV/Å) with different settings of the training/validation sets (See Datasets Section for details)

Systems	Training	Validation	Validation RMSE			
			DPA-1		DeepPot-SE	
			Δ E	Δ F	Δ E	Δ F
AlMgCu	single + binary	ternary	**6.99**	**58**	65.1	92
	all (single + binary + ternary)	ternary	**2.26**	**35**	3.16	42
	all	all	**2.74**	**38**	3.67	45
SSE	init + single	mix	**0.56**	**60**	0.72	76
	init + mix	single	**3.72**	**69**	3.76	82
	all (init + single + mix)	all	**1.41**	**68**	2.92	85
HEA	interior	exterior	**31.2**	**158**	197	399
	exterior	interior	**6.88**	**117**	236	428
	all (interior + exterior)	all	**4.96**	**71**	28.7	141

The number of attention layers *l* in DPA-1 is set to 2 in the AlMgCu and SSE systems, and to 3 in the HEA systems. Bold numbers correspond to lower values.

The equivariant GNN models usually need thousands of GPU hours to be trained to a descent accuracy[40]. By contrast, the DPA-1 model only takes less than 200 GPU hours for training. The converged energy and force MAEs on the OC2M validation set are 0.681 eV and 0.076 eV/Å, respectively. This accuracy is comparable with the best energy-conserving GNN model DimeNet++, which achieves MAEs of 0.805 eV and 0.066 eV/Å, reported in ref. 40. A better performance of energy MAE 0.286 eV and force MAE 0.026 eV/Åis achieved by GemNet-OC at the cost of non-conservative forces and loss of smoothness[40].

In the potential energy model, the presence of non-conservative forces and unsmoothness introduce an artificial energy drift in MD simulations. While investigating static properties, this drift can be removed by incorporating a thermostat in the simulation. However, it is essential to carefully examine the potential impact on the accuracy of property estimation. To calculate the dynamical response of the system, such as the self-diffusion coefficient, viscosity, and heat conductivity, it is typically necessary to evaluate auto-correlation functions by using the Green-Kubo relations[60,61]. The estimations of auto-correlation functions usually require 10-100 ps long micro-canonical (NVE) simulations to achieve converged statistics and eliminate possible nonergodicity in the Hamiltonian dynamics[62]. In this context, energy conservation is critical; otherwise, the energy drift may lead the system to an undesired thermodynamic state or even cause a blow-up in the total energy. In Supplementary Note 5, we demonstrate the magnitude of the total energy drift during a 100-ps long NVE MD simulation for OC20 configurations. The drift observed in non-conservative models is approximately 10^{-2} eV/atom, which corresponds to a temperature of roughly 10^2 K. In contrast, the energy-conserving DPA-1 model, as anticipated, does not exhibit any energy drift.

As shown in Supplementary Note 6, it has been observed that, when trained with 1 million steps on the AlMgCu alloy dataset, the non-conservative models achieve relatively higher force accuracy but lower

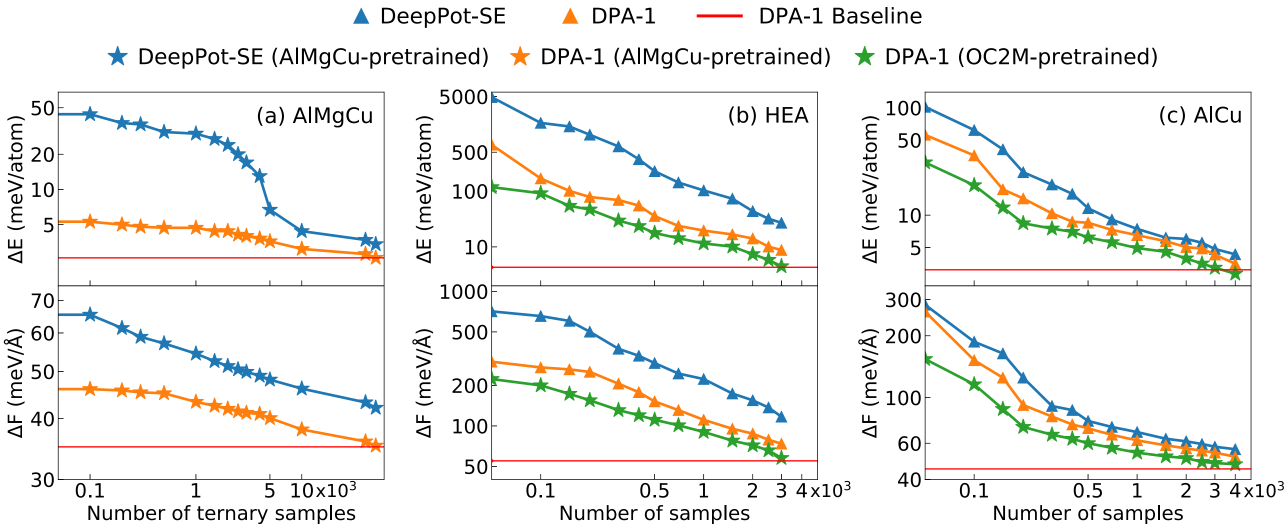

Fig. 2 | Learning curves of both energy and force with DeepPot-SE and DPA-1, under different setups and on different systems. a Learning curves on the AlMgCu ternary subset, with DeepPot-SE and DPA-1 models, pretrained on single-element and binary subsets; Learning curves on HEA (**b**) and AlCu (**c**), with DeepPot-SE (from scratch) and DPA-1 (both from scratch and pretrained on OC2M). The red line represents the full data training baseline with DPA-1.

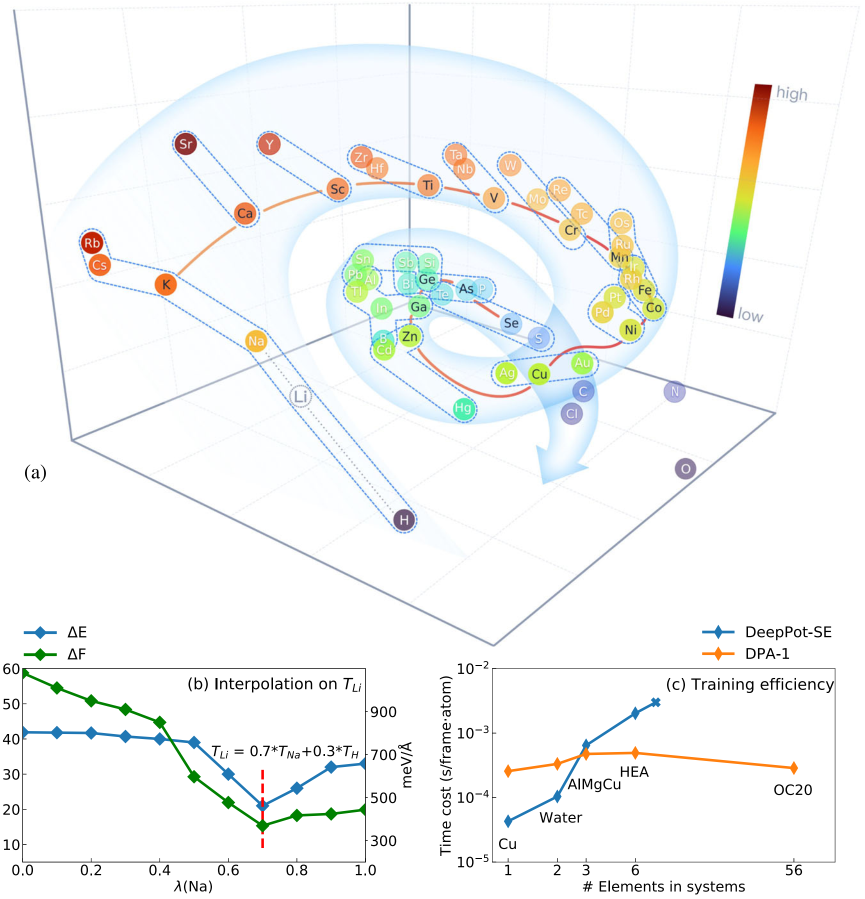

Fig. 3 | Interpretability of DPA-1 pretrained on OC2M and training efficiency comparison with DeepPot-SE. a 3-dimensional PCA visualization of the learned type embeddings of DPA-1 pretrained on OC2M. These 56 elements are roughly arranged on a ***spiral*** in the latent space. Elements in the fourth period are connected with the red line and elements belonging to the same family are grouped by the blue dot lines. Colors on the names of the elements represent the height in z-axis. We use dashed circle to denote the hypothetical position of Li, which is not contained in OC2M. See text for discussions. **b** RMSE of energy and force for SSE systems given by DPA-1 pretrained on OC2M, as functions of linear interpolation coefficient $\lambda(Na)$. Since Li is not contained in OC2M, we let $T_{Li} = \lambda(Na) * T_{Na} + (1 - \lambda(Na)) * T_H$ be the interpolated type embedding of Li. The OC2M-pretrained model with this interpolation and modified energy bias is directly tested on SSE systems without further training. **c** Training efficiency of DPA-1 and DeepPot-SE (considering type information of both two sides) with the growing number of element types in training systems. The maximum number of neighboring atoms to be considered is set to 120 in all the experiments.

energy accuracy compared to the conservative models. Furthermore, the accuracy of non-conservative models in predicting the equation of state (EOS), a fundamental material property, is lower than that of the conservative models. This may be attributed to the fact that non-conservative models predict energy and force separately, and thus accurate force prediction does not necessarily improve the shape of the energy landscape.

Interpretability of type embedding learned from pretraining

To see whether DPA-1 can learn physically meaningful information from pretraining, we investigated the 3-dimensional principal component analysis (PCA) visualization of the learned type embeddings in the OC2M-pretrained model. Interestingly, as shown in Fig. 3a, the arrangement of the elements generally follows the shape of a downward spiral. Elements belonging to the same period are lined up in the direction of the spiral; while elements belonging to the same family are listed in the direction orthogonal to the spiral. Even though some transition metal elements are almost bound together, this rule still roughly holds. It is observed that C, N, and O are outliers, possibly because in OC2M, C, N and O are mostly in organic molecules, which serve as adsorbates and have chemical environments that are very different from other elements.

In addition, we performed interpolation experiments for the type embedding of Li, an element unseen in OC2M. As shown in Fig. 3b, we let $T_{Li} = \lambda(Na) * T_{Na} + (1 - \lambda(Na)) * T_H$, since Li lies between H and Na in the same family. When tested on the SSE system, only the bias in the atomic energy is changed, since the setup of the electronic method used to label the SSE system is different from that for OC2M, which typically causes an energy shift. It is found that the RMSE of energy and force shows a sudden drop when $\lambda(Na) = 0.7$, which meets the chemical intuition and further confirms the interpretability of the pretrained DPA-1 model. Moreover, we conducted analogous interpolation experiments for Nb and Mo on the HEA systems, and reached similar conclusions as the Li interpolation (see detailed report in Supplementary Note 7).

Discussion

In this paper, we developed DPA-1, an attention-based Deep Potential model that allows for large-scale pretraining on atomistic datasets. We tested DPA-1 from different aspects, showing its excellent performance in terms of its accuracy on various datasets when trained from scratch, as well as its sample efficiency when pretrained with existing data. Further investigations on the type embedding parameters suggest the interpretability of DPA-1 pretrained on OC2M.

In the future, it will be of interest to extend the training dataset to cover the full periodic table, and, in particular, see a more converged "spiral" in the latent space; the embedding information of local chemical environments may be useful to characterize different conformations. Multi-task and unsupervised training schemes are worth exploring; and, for downstream tasks, just like what has happened in the fields of CV and NLP, schemes like model compression, distillation, and transfer, etc., are desperately needed. We leave these possibilities and more applications to future works.

Methods

Consider a system of N atoms, the elemental types are $\mathcal{A} = \{\alpha_1, \alpha_2, ..., \alpha_i, ..., \alpha_N\}$, and the atomic coordinates are $\mathcal{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_i, ..., \boldsymbol{r}_N\}$, with $\boldsymbol{r}_i$ being the three Cartesian coordinates of atom i. The PES of the system is denoted by E, a function of elemental types and coordinates, i.e. $E = E(\mathcal{A}, \mathcal{R})$. For each atom i, consider its neighbors $\{j | j \in \mathcal{N}_{r_c}(i)\}$, where $\mathcal{N}_{r_c}(i)$ denotes the set of atom indices j such that $r_{ji} < r_c$, with r_{ji} being the Euclidean distance between atoms i and j. E is represented as the summation of atomic energies $\{e_1, e_2, ..., e_i, ..., e_N\}$, where the atomic energy e_i only depends on the information of $\mathcal{N}_{r_c}(i)$. We define $N_i = |\mathcal{N}_{r_c}(i)|$, the cardinality of the set $\mathcal{N}_{r_c}(i)$. We use $\mathcal{A}^i$ to denote element types in $\mathcal{N}_{r_c}(i)$, and $\mathcal{R}^i \in \mathbb{R}^{N_i \times 3}$ their corresponding coordinates relative to i. The atomic energy e_i is thus a function of $\mathcal{A}^i$ and $\mathcal{R}^i$. The atomic force on atom i, $\mathcal{F}_i$, is defined as the negative gradient of the total energy with respect to i's coordinate:

$$\mathcal{F}_i = -\nabla_{\mathbf{r}_i} E. \tag{1}$$

We refer to ref. 7 for a detailed discussion of several requirements for PES modeling. In particular, the PES has to be invariant under translation, rotation, and permutation of the indices of atoms with the same element types.

The details of the model architecture are introduced below. We refer to Fig. 1 for the overall pipeline to predict the atomic energy e_i: from the embedded neighboring environment, through the self-attention scheme, to the symmetry-preserving descriptors, and finally to the fitting network.

Local embedding matrix with type information

We obtain the local embedding matrix with the following three steps. First, $\mathcal{R}^i$ is mapped to the generalized coordinates $\tilde{\mathcal{R}}^i \in \mathbb{R}^{N_i \times 4}$. In this mapping, each row of $\mathcal{R}^i$, $\{x_{ji}, y_{ji}, z_{ji}\}$, is transformed into a row of $\tilde{\mathcal{R}}^i$:

$$\{x_{ji}, y_{ji}, z_{ji}\} \mapsto \{s(r_{ji}), \hat{x}_{ji}, \hat{y}_{ji}, \hat{z}_{ji}\}, \tag{2}$$

where $\{x_{ji}, y_{ji}, z_{ji}\}$ denotes the Cartesian coordinates of $\boldsymbol{r}_{ji} = \boldsymbol{r}_j - \boldsymbol{r}_i$, $\hat{x}_{ji} = \frac{s(r_{ji})x_{ji}}{r_{ji}}, \hat{y}_{ji} = \frac{s(r_{ji})y_{ji}}{r_{ji}}, \hat{z}_{ji} = \frac{s(r_{ji})z_{ji}}{r_{ji}}$, and $s(r_{ji}) : \mathbb{R} \mapsto \mathbb{R}$ is a continuous and differentiable scalar weighting function applied to each component, defined as:

$$s(r_{ji}) = \begin{cases} \frac{1}{r_{ji}} & r_{ji} < r_{cs} \\ \frac{1}{r_{ji}}\left[u^3\left(-6u^2 + 15u - 10\right) + 1\right] & r_{cs} \le r_{ji} < r_c, \quad u = \frac{r_{ji} - r_{cs}}{r_c - r_{cs}}. \\ 0 & r_c \le r_{ji} \end{cases} \tag{3}$$

Here r_{cs} is a smooth cutoff parameter that allows the components in $\tilde{\mathcal{R}}^i$ to smoothly go to zero at the boundary of the local region defined by r_c.

Second, we add the atomic type embedding as supplemental information. For atom i, the type embedding map T_i is defined as:

$$T_i = \phi_T(\alpha_i), \tag{4}$$

where α_i is the atomic type of atom i and ϕ_T is a one-hot-like embedding network mapping from α_i to a length-fixed vector.

Then, given both $\tilde{\mathcal{R}}^i$ and type embeddings $\{T_i\} \cup \{T_j | j \in \mathcal{N}_{r_c}(i)\}$, we define the local embedding matrix $\mathcal{G}^i \in \mathbb{R}^{N_i \times M_1}$:

$$\left(\mathcal{G}^i\right)_j = G(s(r_{ji}), T_i, T_j), \tag{5}$$

where G is a neural network mapping from scalar weight $s(r_{ji})$ and type embeddings of both center and neighbor atoms, through multiple hidden layers, to M_1 outputs. Here we simply feed the concatenated inputs into G at once, as shown in Fig. 1b.

Attention method for building up trainable descriptors

The attention mechanism has achieved great success and played an increasingly important role in CV[63] and NLP[64]. It has become an excellent tool for modeling the importance or relevance of visual regions or text tokens, thus is potentially appropriate to reweight the interactions among neighbor atoms according to both distance and angular information.

In DPA-1, we follow the standard self-attention mechanism and obtain the queries $\mathcal{Q}^{i,l} \in \mathbb{R}^{N_i \times d_k}$, keys $\mathcal{K}^{i,l} \in \mathbb{R}^{N_i \times d_k}$, and values $\mathcal{V}^{i,l} \in \mathbb{R}^{N_i \times d_v}$:

$$\begin{aligned} \left(\mathcal{Q}^{i,l}\right)_j &= Q_l\left(\left(\mathcal{G}^{i,l-1}\right)_j\right), \\ \left(\mathcal{K}^{i,l}\right)_j &= K_l\left(\left(\mathcal{G}^{i,l-1}\right)_j\right), \\ \left(\mathcal{V}^{i,l}\right)_j &= V_l\left(\left(\mathcal{G}^{i,l-1}\right)_j\right), \end{aligned} \tag{6}$$

where Q_l, K_l, V_l represent three linear transformations which output the queries and keys of dimension d_k and values of dimension d_v, and l is the index of attention layer. Here we take $\mathcal{G}^{i,0} = \mathcal{G}^i$.

Then we adopt the scaled dot-product attention method[65] to mix the neighbor features after calculating the attention weights:

$$A(\mathcal{Q}^{i,l}, \mathcal{K}^{i,l}, \mathcal{V}^{i,l}, \mathcal{R}^{i,l}) = \varphi\left(\mathcal{Q}^{i,l}, \mathcal{K}^{i,l}, \mathcal{R}^{i,l}\right)\mathcal{V}^{i,l}, \tag{7}$$

where $\varphi\left(\mathcal{Q}^{i,l}, \mathcal{K}^{i,l}, \mathcal{R}^{i,l}\right) \in \mathbb{R}^{N_i \times N_i}$ is attention weights. In the original attention method, one typically has $\varphi\left(\mathcal{Q}^{i,l}, \mathcal{K}^{i,l}\right) = \mathrm{softmax}\left(\frac{\mathcal{Q}^{i,l}(\mathcal{K}^{i,l})^T}{\sqrt{d_k}}\right)$, with $\sqrt{d_k}$ being the normalization temperature. This is slightly modified to better incorporate the angular information:

$$\varphi\left(\mathcal{Q}^{i,l}, \mathcal{K}^{i,l}, \mathcal{R}^{i,l}\right) = \mathrm{softmax}\left(\frac{\mathcal{Q}^{i,l}(\mathcal{K}^{i,l})^T}{\sqrt{d_k}}\right) \odot \hat{\mathcal{R}}^i(\hat{\mathcal{R}}^i)^T, \tag{8}$$

where $\hat{\mathcal{R}}^i = \frac{\mathcal{R}^i}{\|\mathcal{R}^i\|_2} \in \mathbb{R}^{N_i \times 3}$ denotes normalized relative coordinates and $\odot$ means element-wise multiplication. Intuitively, in the neighborhood of center atom i, neighbor atom k may be highly correlated with j when both the relative distance attention $(\mathcal{Q}^{i,l})_j(\mathcal{K}^{i,l})_k^T$ and normalized product of relative coordinates $\frac{\mathbf{r}_{ji}(\mathbf{r}_{ki})^T}{r_{ji}r_{ki}}$ have high scores.

Then we add layer normalization in a residual way to finally obtain the self-attentioned local embedding matrix $\hat{\mathcal{G}}^i$ in one such attention layer:

$$\mathcal{G}^{i,l} = \mathcal{G}^{i,l-1} + \mathrm{LayerNorm}(A(\mathcal{Q}^{i,l}, \mathcal{K}^{i,l}, \mathcal{V}^{i,l}, \mathcal{R}^{i,l})). \tag{9}$$

We also tried other attention-related tricks such as pre-layer normalization, multi-head attention, etc., which brought little improvement. In practice, as shown in Fig. 1c, we repeated this procedure by $l (l \ge 2)$ times for a more complete representation. If not stated otherwise, we use $l = 2$ in the following sections of the work. Next, we define the encoded feature matrix $\mathcal{D}^i \in \mathbb{R}^{M_1 \times M_2}$ of atom i:

$$\mathcal{D}^i = (\hat{\mathcal{G}}^i)^T \tilde{\mathcal{R}}^i (\tilde{\mathcal{R}}^i)^T \dot{\mathcal{G}}^i, \tag{10}$$

where $\dot{\mathcal{G}}^i$ stands for a sub-matrix of $\hat{\mathcal{G}}^i$, which takes the first $M_2 (< M_1)$ columns of $\hat{\mathcal{G}}^i$. Here the feature matrix $\mathcal{D}^i$, i.e. the descriptor, preserves all the invariance mentioned above, of which the proof can be found in ref. 7. We then pass the reshaped $\mathcal{D}^i$, concatenated with the type embedding parameters of the center atom, through the multi-layer fitting network:

$$e_i = e\left(\mathcal{D}^i, T_i\right). \tag{11}$$

The total energy of the system is then given as the summation of e_i, and the atomic force $\mathcal{F}_i$ can be further computed via Eq. (1).

Model (pre-)training and finetuning

For model training or pretraining, we adopted the Adam stochastic gradient descent method[66] on all the trainable parameters $\boldsymbol{w}$ inside the model to minimize the loss:

$$\mathcal{L}_w(E^w, \mathcal{F}^w) = \frac{1}{|\mathcal{B}|}\sum_{t \in \mathcal{B}}\left(p_e|E_t - E_t^w|^2 + p_f|\mathcal{F}_t - \mathcal{F}_t^w|^2\right). \tag{12}$$

Here $\mathcal{B}$ represents a minibatch, $|\mathcal{B}|$ is the batch size, t denotes the index of the training sample. $E^w, \mathcal{F}^w$ denote the model outputs and $E, \mathcal{F}$ are the corresponding DFT results. We also adopted a scheduler to tune the prefactors p_e and p_f during the training process to make a better balance between energy and force labels. Virial errors, which are omitted here, can be added to the loss for training if available.

To finetune the pretrained model with a new dataset, we first change the energy bias in the last layer of the pretrained model with the new statistical results of the new dataset, and then we fix part of the parameters in the pretrained model and train the remaining. For the following experiments, we obtained the best performance when only the type embedding parameters are fixed.

Data availability

The dataset used for training OC2M-pretrained DPA-1 is available at: https://www.aissquare.com/datasets/detail?pageType=datasets&name=Open_Catalyst_2020(OC20_Dataset). Other datasets are available in their references or on reasonable request.

Code availability

The codes of DPA-1 are in the repository of DeePMD-kit: https://github.com/deepmodeling/deepmd-kit. The OC2M-pretrained model is available at: https://www.aissquare.com/models/detail?pageType=models&name=DPA_1_OC2M.

Received: 1 February 2023; Accepted: 19 April 2024;
Published online: 07 May 2024

References

1. Behler, J & Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
2. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. ábor Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
3. Thompson, A. P., Swiler, L. P., Trott, C. R., Foiles, S. M. & Tucker, G. J. Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **285**, 316–330 (2015).
4. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry. In *Proceedings of International Conference on Machine Learning*, 1263–1272 (PMLR, 2017).
5. Schütt, K. et al. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Proceedings of Advances in Neural Information Processing Systems* (2017).
6. Zhang, L., Han, J., Wang, H., Car, R. & Weinan, E. J. P. R. L. Deep potential molecular dynamics: a scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **120**, 143001 (2018).
7. Zhang, L. et al. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems. In *Proceedings of Advances in Neural Information Processing Systems* (2018).
8. Drautz, R. Atomic cluster expansion for accurate and transferable interatomic potentials. *Phys. Rev. B* **99**, 014104 (2019).
9. Gasteiger, J., Groß, J. & Günnemann, S. Directional message passing for molecular graphs. In *Proceedings of International Conference on Learning Representations* (2019).
10. Zhang, Y., Hu, C. & Jiang, B. Embedded atom neural network potentials: Efficient and accurate machine learning with a physically inspired representation. *J. Phys. Chem. Lett.* **10**, 4962–4967 (2019).
11. Gasteiger, J., Becker, F. & Günnemann, S. Gemnet: Universal directional graph neural networks for molecules. *Adv. Neural Inf. Process. Syst.* **34**, 6790–6802 (2021).
12. Deringer, V. L. et al. Gaussian process regression for materials and molecules. *Chem. Rev.* **121**, 10073–10141 (2021).
13. Unke, O. T. et al. Machine learning force fields. *Chem. Rev.* **121**, 10142–10186 (2021).
14. Wen, T., Zhang, L., Wang, H., Weinan, E. & Srolovitz, D. J. Deep potentials for materials science. *Mater. Futures* **1**, 022601 (2022).
15. Bartók, A. P., Kermode, J., Bernstein, N. & Csányi, G. ábor Machine learning a general-purpose interatomic potential for silicon. *Phys. Rev. X* **8**, 041048 (2018).
16. Deringer, V. L., Caro, M. A. & Csányi, G. ábor A general-purpose machine-learning force field for bulk and nanostructured phosphorus. *Nat. Commun.* **11**, 1–11 (2020).
17. Zhang, L., Wang, H., Car, R. & Weinan, E. Phase diagram of a deep potential water model. *Phys. Rev. Lett.* **126**, 236001 (2021).
18. Jiang, W., Zhang, Y., Zhang, L. & Wang, H. Accurate deep potential model for the Al–Cu–Mg alloy in the full concentration space. *Chin. Phys. B* **30**, 050706 (2021).
19. Szlachta, W. J., Bartók, A. P. & Csányi, G. ábor Accuracy and transferability of Gaussian approximation potential models for tungsten. *Phys. Rev. B* **90**, 104108 (2014).
20. Wang, X., Wang, Y., Zhang, L., Dai, F. & Wang, H. A tungsten deep neural-network potential for simulating mechanical property degradation under fusion service environment. *Nucl. Fusion* **62**, 126013 (2022).
21. Wang, Yi. Nan, Zhang, LinFeng, Xu, B., Wang, XiaoYang & Wang, H. A generalizable machine learning potential of Ag–Au nanoalloys and its application to surface reconstruction, segregation, and diffusion. *Model. Simul. Mater. Sci. Eng.* **30**, 025003 (2021).
22. Wen, T. et al. Specialising neural network potentials for accurate properties and application to the mechanical response of titanium. *npj Comput. Mater.* **7**, 206 (2021).
23. Podryabinkin, E. V. & Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Comput. Mater. Sci.* **140**, 171–180 (2017).
24. Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O. & Roitberg, A. E. Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **148**, 241733 (2018).
25. Zhang, L., Lin, De-Ye, Wang, H., Car, R. & Weinan, E. Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Mater.* **3**, 023804 (2019).
26. Zhang, Y. et al. Dp-gen: A concurrent learning platform for the generation of reliable deep learning based potential energy models. *Comput. Phys. Commun.* **253**, 107206 (2020).
27. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140**, A1133 (1965).
28. Car, R. & Parrinello, M. Unified approach for molecular dynamics and density-functional theory. *Phys. Rev. Lett.* **55**, 2471 (1985).
29. Russakovsky, O. et al. Imagenet large-scale visual recognition challenge. *Int. J. Comput. Vis.* **115**, 211–252 (2015).
30. Dosovitskiy, A. et al. An image is worth 16 × 16 words: Transformers for image recognition at scale. In *Proceedings of International Conference on Learning Representations* (2021).
31. Devlin, J., Chang, Ming-Wei, Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. Preprint at https://arxiv.org/abs/1810.04805 (2018).
32. Brown, T. et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **33**, 1877–1901 (2020).
33. Smith, J. S., Isayev, O. & Roitberg, A. E. Ani-1: an extensible neural network potential with DFT accuracy at force field computational cost. *Chem. Sci.* **8**, 3192–3203 (2017).
34. Smith, J. S. et al. Approaching coupled cluster accuracy with a general-purpose neural network potential through transfer learning. *Nat. Commun.* **10**, 1–8 (2019).
35. Liu, S. et al. Pre-training molecular graph representation with 3d geometry. In *Proceedings of International Conference on Learning Representations* (2022).
36. Stärk, H. et al. 3d infomax improves gnns for molecular property prediction. In *Proceedings of International Conference on Machine Learning*, 20479–20502 (PMLR, 2022).
37. Zhou, G. et al. Uni-mol: A universal 3d molecular representation learning framework. In *Proceedings of International Conference on Learning Representations* (2023).
38. Thomas, N. et al. Tensor field networks: Rotation-and translation-equivariant neural networks for 3D point clouds. Preprint at https://arxiv.org/abs/1802.08219 (2018).

39. Schütt, K., Unke, O. & Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *Proceedings of International Conference on Machine Learning*, 9377–9388 (PMLR, 2021).
40. Gasteiger, J. et al. Gemnet-oc: Developing graph neural networks for large and diverse molecular simulation datasets. In *Proceedings of Transactions on Machine Learning Research* (2022).
41. Gasteiger, J., Giri, S., Margraf, J. T. & Günnemann, S. Fast and uncertainty-aware directional message passing for non-equilibrium molecules. Preprint at https://arxiv.org/abs/2011.14115 (2022).
42. Takamoto, S. et al. Towards universal neural network potential for material discovery applicable to arbitrary combination of 45 elements. *Nat. Commun.* **13**, 2991 (2022).
43. Zitnick, L. et al. Spherical channels for modeling atomic interactions. *Adv. Neural Inf. Process. Syst.* **35**, 8054–8067 (2022).
44. Shuaibi, M. et al. Rotation invariant graph neural networks using spin convolutions. Preprint at https://arxiv.org/abs/2106.09575 (2021).
45. Liao, Yi-Lun & Smidt, T. Equiformer: Equivariant graph attention transformer for 3D atomistic graphs. In *Proceedings of International Conference on Learning Representations* (2023).
46. Liao, Y-L., Wood, B., Das, A. & Smidt, T. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations. In *Proceedings of International Conference on Learning Representations* (2024).
47. Chanussot, L. et al. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catal.* **11**, 6059–6072 (2021).
48. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728 (2022).
49. Jain, A. et al. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002 (2013).
50. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865 (1996).
51. Choudhary, K. et al. Unified graph neural network force-field for the periodic table: solid state applications. *Digit. Discov.* **2**, 346–355 (2023).
52. Choudhary, K. et al. The joint automated repository for various integrated simulations (Jarvis) for data-driven materials design. *npj Comput. Mater.* **6**, 173 (2020).
53. Musaelian, A. et al. Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.* **14**, 579 (2023).
54. Le, T., Noé, F. & Clevert, D.-A. Equivariant graph attention networks for molecular property prediction. Preprint at https://arxiv.org/abs/2202.09891 (2022).
55. Bond, S. D. & Leimkuhler, B. J. Molecular dynamics and the accuracy of numerically computed averages. *Acta Numer.* **16**, 1–65 (2007).
56. Jia, W. et al. Pushing the limit of molecular dynamics with ab initio accuracy to 100 million atoms with machine learning. In *Proceedings of SC20: International Conference For High Performance Computing, Networking, Storage And Analysis,* 1–14 (IEEE, 2020).
57. Huang, J. et al. Deep potential generation scheme and simulation protocol for the li10gep2s12-type superionic conductors. *J. Chem. Phys.* **154**, 094703 (2021).
58. Chmiela, S. et al. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
59. Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 1–7 (2014).
60. Green, M. S. Markoff random processes and the statistical mechanics of time-dependent phenomena. *J. Chem. Phys.* **22**, 398–413 (1954).
61. Kubo, R. Statistical-mechanical theory of irreversible processes. *J. Phys. Soc. Jpn.* **12**, 570–586 (1957).
62. Lee, H-S. & Tuckerman, M. E. Dynamical properties of liquid water from ab initio molecular dynamics performed in the complete basis set limit. *J. Chem. Phys.* **126**, 164501 (2007).
63. Guo, M.-H. et al. Attention mechanisms in computer vision: A survey. *Comput. Vis. Media* **8**, 331–368 (2022).
64. Galassi, A., Lippi, M. & Torroni, P. Attention in natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 4291–4308 (2020).
65. Vaswani, A. et al. Attention is all you need. In *Proceedings of Advances in Neural Information Processing Systems* (2017).
66. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. Preprint at https://arxiv.org/abs/1412.6980 (2014).

Acknowledgements

The work of H.W. and D.Z. is supported by the National Key R&D Program of China under Grant No. 2022YFA1004300, and the National Natural Science Foundation of China under Grant No. 12122103. We thank Y.L., Z.L., and G.K. for inspiring discussions. The computational resource was supported by the Bohrium Cloud Platform at DP technology.

Author contributions

D.Z., L.Z., H.W., and F.Z.D. conceived the idea of this work. D.Z., H.B., and H.W. designed the model structure. D.Z. implemented the model. D.Z., H.B., W.J., and X.L. performed the experiments on different systems. All authors contributed to the discussions and edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-024-01278-7.

Correspondence and requests for materials should be addressed to Linfeng Zhang or Han Wang.

Reprints and permissions information is available at http://www.nature.com/reprints

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