PCCP

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Check for updates

Cite this: *Phys. Chem. Chem. Phys.*, 2025, **27**, 8295

An active learning force field for the thermal transport properties of organometallic complex crystals†

Wenjie Zhang,[a] Weitang Li[b] and Zhigang Shuai*[ab]

The accurate prediction of lattice thermal conductivity in organometallic thermoelectric materials is crucial for advancing energy conversion technologies. Methods based on molecular dynamics simulations can solve this problem well, but require force fields with sufficiently high accuracy. Due to the complexity of chemical bonding in organometallic complex materials, the development of force fields with high predictivity has been a long standing challenge, particularly when thermal transport is concerned which requires even greater accuracy. In recent years, the rapid advancement of machine learning force fields has offered substantial potential for addressing these issues. However, there remain challenges for materials with large organometallic complexes in one unit cell and both inter- and intra-molecular interactions. In this work, we employ an active learning approach combined with deep neural networks to develop a force field taking copper phthalocyanine as an example. The model utilizes a local environment descriptor for representation without explicitly characterizing the metal–organic coordination. The nonlinear mapping capabilities of deep neural networks enable the model to effectively capture higher-order many-body interactions. Furthermore, we utilized the Green–Kubo method to calculate the thermal conductivity of copper phthalocyanine, revealing a value of 0.49 W m^{-1} K^{-1} at 300 K, consistent with experimental findings (0.39 W m^{-1} K^{-1}). This result significantly surpasses previous work with classical force fields. This work represents a significant advancement in demonstrating that machine-learning force fields can effectively characterize interactions in metal–organic complex systems and can significantly advance the development and discovery of organometallic thermoelectric materials.

Received 17th January 2025,
Accepted 12th March 2025

DOI: 10.1039/d5cp00232j

rsc.li/pccp

Introduction

The pursuit of efficient energy conversion technologies has intensified due to escalating energy demands and growing environmental concerns. Thermoelectric materials, which enable direct conversion between thermal and electrical energies, have garnered significant attention due to their potential applications in power generation and refrigeration without moving parts or emissions.[1] The efficiency of thermoelectric materials is quantified by the dimensionless thermoelectric figure of merit, $ZT = S^2\sigma T/\kappa$, where S is the Seebeck coefficient, σ is the electrical conductivity, T is the absolute temperature, and κ is the thermal conductivity.[2] Achieving high ZT values requires a delicate balance between these interdependent parameters, which poses a significant challenge in thermoelectric material design. Organometallic complexes have emerged as promising candidates in advancing thermoelectric technology due to their unique electronic structures and tunable properties.[3] These complexes consist of metal atoms bonded to organic ligands, creating a versatile platform for manipulating charge carrier transport and phonon scattering at the molecular level. The incorporation of heavy metal atoms can enhance the Seebeck coefficient and reduce thermal conductivity through increased phonon scattering, thereby improving the ZT value.[4]

Organometallic complexes typically have important characteristics such as enhanced electrical conductivity due to delocalized π-electron systems and metal-induced charge transfer, tunable electronic properties through modification of metal centers or ligands, and inherently low thermal conductivity from their organic components. The synergistic effects between the metal and organic ligands facilitate improved charge carrier mobility and reduced phonon transport, while their synthetic versatility enables molecular-level engineering to optimize thermoelectric performance.[5–10] Currently, organometallic complexes

---

[a] *MOE Key Laboratory of Organic OptoElectronics and Molecular Engineering, Department of Chemistry, Tsinghua University, Beijing 100084, China. E-mail: zgshuai@tsinghua.edu.cn*

[b] *School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, Guangdong 518172, China. E-mail: shuaizhigang@cuhk.edu.cn*

† Electronic supplementary information (ESI) available. See DOI: **https://doi.org/10.1039/d5cp00232j**

have very specific application systems in the thermoelectric field, demonstrating their potential to enhance thermoelectric performance through various mechanisms. For example, organometallic polymers, such as poly(3-hexylthiophene) doped with organometallic complexes, have shown increased electrical conductivity and improved thermoelectric efficiency.[11] Metal phthalocyanine complexes, including copper phthalocyanines, have been investigated for their thermoelectric properties due to their planar conjugated structures and thermal stability, leading to enhanced charge transport and Seebeck coefficients in thin-film devices.[12] Additionally, organometallic complexes have been incorporated into clathrates, coordination polymers, quantum dots, and skutterudites, all of which have contributed to the development of thermoelectric materials with optimized ZT values.[13–17]

The complex and often large unit cells of organometallic compounds, characterized by low symmetry and diverse bonding environments, present significant challenges in accurately predicting lattice thermal conductivity κ_L. Theoretical methods play a crucial role in understanding and forecasting the thermal transport properties of these materials. The two kinds of most reliable approaches for calculating κ_L are solving the phonon Boltzmann transport equation (BTE) combined with density functional theory (DFT) calculations[18–20] and molecular dynamics simulation.[21] For the method based on phonon BTE, first-principles calculation provides detailed insights into phonon dispersion relations and lifetimes by calculating interatomic force constants (IFCs). Studies have provided valuable insights into the thermal transport mechanisms of arsenic–phosphorus alloys, and IV–VI semiconductor compounds.[22,23] However, the computational cost associated with calculating IFCs for organometallic systems is substantial due to their structural complexity.[24] For MD-based methods, the wide usage includes non-equilibrium molecular dynamics based on Fourier's law of heat conduction and equilibrium molecular dynamics based on the Green–Kubo formalism.[25–27] MD simulations inherently include anharmonic effects and can handle large systems, making them suitable for studying materials with significant structural disorder or weak intermolecular interactions. Nevertheless, the accuracy of MD simulations heavily depends on the choice of the force field. Previous studies have attempted to model the copper phthalocyanine (CuPc) system using tailor-made force fields based on a hybrid-COMPASS force field.[28] While their approach provided valuable insights into the structural properties of CuPc, it exhibited significant limitations in accurately predicting its thermal transport properties, with discrepancies of an order of magnitude compared to experimental observations. Due to the poor description of higher-order many-body interactions, most of the classic force fields still lack the precision needed for systems in which vibrational properties have a strong influence on material properties. Developing accurate potentials for organometallic compounds is challenging, limiting the predictive power of MD in these systems.

In recent years, machine learning force fields (MLFFs) have emerged as a promising solution to these challenges.[29] MLFFs leverage large datasets generated from high-level quantum mechanical calculations to train models that can predict potential energy surfaces with near-quantum accuracy.[30] By learning directly from the data, MLFFs can capture complex interactions and coordination environments without relying on predefined functional forms or fixed parameters.[31] Many machine learning force field (MLFF) models have been developed, such as Gaussian approximation potential (GAP),[32] ANI,[33] deep potential (DP),[34,35] moment tensor potential (MTP),[36] neural equivariant interatomic potential (NequIP),[37] PhysNet,[31] TorchMD-NET,[38] NEP[39,40] and so on. These MLFFs have emerged as transformative approaches to molecular simulations, offering near *ab initio* accuracy at a fraction of the computational cost.

Furthermore, there have been significant advances in combining machine learning with thermal transport property calculations.[41] Baroni *et al.* investigated the viscosity and thermal conductivity of liquid water *via* molecular dynamics simulations based on the DP model.[42,43] Zojer *et al.* parameterized a machine-learned potential of the MTP model, achieving near-DFT precision in modeling the structural, thermal, and mechanical properties of metal–organic frameworks (MOFs).[44] Cheng *et al.* explored the effects of high pressure on the lattice dynamics and thermal transport properties of PbTe with the NEP model, revealing that pressure-induced changes enhance the lattice thermal conductivity by reducing phonon–phonon scattering.[45] These studies highlight the efficacy of MLFFs in exploring thermal transport phenomena, which are critical for the design of materials in thermoelectric applications and thermal management technologies.

This study focuses on solving the problem of calculating the lattice thermal conductivity of organometallic complex systems, which in turn leads to a better prediction of the thermoelectric properties. We choose copper phthalocyanine as an example and deploy a set of generalized workflows for implementing machine learning force field construction without relying on preconceived notions of fixed chemical bonds or prior knowledge of specific interactions, which are often inherent in classical force fields. This innovation effectively overcomes the traditional challenges associated with characterizing metal–organic coordination bonds in organometallic complexes. We propose utilizing an active learning approach[46–48] to construct a machine learning force field (MLFF) for the CuPc system that achieves accuracy comparable to DFT calculations. By accurately modeling the complex interactions within CuPc, our MLFF enables reliable molecular dynamics simulations for the prediction of thermal transport properties and enables accurate calculations of thermal conductivity in CuPc crystals. This advancement is critical for understanding the thermophysical behavior of CuPc and can also be used for other organometallic systems, which will significantly impact the design and optimization of organic thermoelectric materials.

Methods and implementation

In this section, the technical approach adopted in this study will be introduced. We deploy a workflow from the initial collection of the base dataset to the construction of the MLFF model and the subsequent calculation of the thermal transport properties. By utilizing an active learning framework, we systematically generate and organize datasets covering the potential energy surface in the solid state of copper phthalocyanine crystals.

Overall workflow

We employ a set of workflows capable of starting from the construction of the force field of metal–organic molecular complexes to the calculation of their thermal transport properties. The specific workflow is illustrated in Fig. 1 and can be broadly categorized into three distinct phases. Initially, we utilize an active learning framework to sample configurations of the CuPc system, generating a dataset suitable for force field development. This phase involves the systematic selection of representative structures, which are subsequently labeled using first-principles calculations performed with VASP.[49] This ensures that the dataset exhibits high accuracy, capturing the essential features of the system's potential energy landscape. Following the establishment of a dataset with first-principles accuracy, we employ deep neural networks (DNNs) to model the potential energy surfaces of CuPc. This step is crucial for constructing a robust mapping relationship between the structural configurations and the corresponding physical quantities, such as energy, force, and virial. The DNN fitting process leverages the rich dataset to enable precise predictions of interatomic interactions. With the high-precision machine learning force field in place, we proceed to compute the lattice thermal conductivity of the CuPc system. This is accomplished through equilibrium molecular dynamics simulations, combined with the Green–Kubo formalism to extract thermal transport properties. The resulting lattice thermal conductivity values provide insights into the thermal behavior of the system under investigation. A detailed description of each component of this workflow will be elaborated upon in the following sections.

Crystal information

Structurally, CuPc consists of a planar, cyclic tetrabenzoporphyrin framework with a central copper ion coordinated to the nitrogen atoms of the phthalocyanine ligand. We collect our initial dataset on the α and β forms of copper phthalocyanine, with specific crystal structures and data presented in Fig. 2 and Table 1. In its crystalline form, the α and β forms display different packing arrangements. Crystal structures are collected from experiments.[50,51]

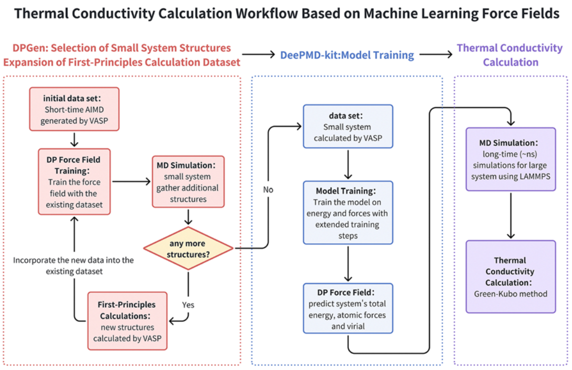

Fig. 1 Overall workflow.

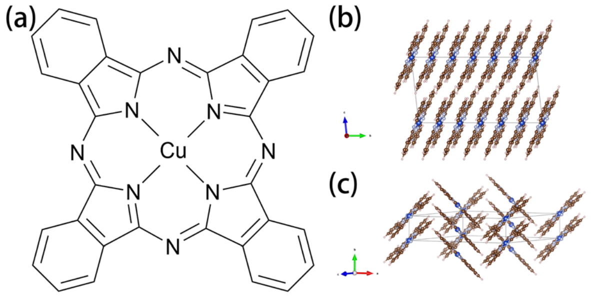

Fig. 2 Molecular structure of copper phthalocyanines.

Table 1 Unit cell parameters of CuPcs

	α-CuPC	β-CuPc
a (Å)	12.9	19.4
b (Å)	3.8	4.8
c (Å)	12.1	14.6
α (deg)	96.2	90.0
β (deg)	90.6	120.9
γ (deg)	90.3	90.0

Machine learning interatomic potential

The essential effect of machine learning interatomic potentials is to establish a connection between structural configurations and physical properties, enabling the efficient acquisition of a substantial amount of data that closely approximates the accuracy of first-principles calculations, while minimizing computational time costs. In this work, we utilize the deep potential framework,[34,35] which is based on deep neural networks (DNNs), to construct the force field model. Notably, this application does not require any explicit description of chemical bonds, which alleviates concerns regarding bond order when addressing complex coordination systems. The deep potential model has been extensively benchmarked not only against GAPs, MTPs, ACE, and MACE but also against other state-of-the-art models such as BPNN, SchNet, DimeNet++, GemNet-T, GemNet-dT, NequIP, Allegro, and SCN. These benchmarks have shown that DP performs well in both accuracy and speed. In addition, research on traditional thermoelectric materials, such as the work by Fan *et al.*,[39] further supports that the DP model offers higher accuracy and faster performance compared to MTP and GAP models, particularly in the context of heat transport properties.

In processing structural data, we employ the concept of the local environment to define descriptors for each atom. Specifically, the local environment of each atom is determined by the atom itself and its neighboring atoms within a specified truncation radius. Consequently, only those atoms within the cutoff radius influence the force field acting on the central atom. This approach ensures that atomic interactions depend on the arrangement of surrounding atoms in the local environment. This framework enhances the flexibility and accuracy of modeling interatomic interactions in complex systems, such as the copper phthalocyanine (CuPc) complex. In the framework of deep potential, the construction of a descriptor involves training a neural network to automatically determine the coefficients of different many-body interaction terms.

Consider a system containing N atoms, where the coordinates of the atoms are labeled as $r = \{r_1, r_2, \ldots, r_N\} \in \mathbb{R}^{3N}$. The total energy of the system E_{total} is represented as the sum of the local energies E_i of individual atoms:

$$E_{\text{total}} = \sum_i E_i = \mathcal{N}\Big(D\Big(x_i, \{x_j\}_{j\in n(i)}\Big)\Big),$$

where $\mathcal{N}$ is the fitting network and D is the descriptor. The input of the descriptor is $x_i = \{r_i, z_i\}$, while z_i represents the chemical species and $n(i)$ denotes the neighbouring atoms of atom i within the cutoff radius r_c. The descriptor D of atom i is given by:

$$D = \frac{1}{{N_c}^2} G^{\mathrm{T}} L (L)^{\mathrm{T}} G$$

where L is the local environment matrix of atom i, the row of which is given by:

$$L_j = \left[s(r_{ij}), \frac{s(r_{ij})x_{ij}}{r_{ij}}, \frac{s(r_{ij})y_{ij}}{r_{ij}}, \frac{s(r_{ij})z_{ij}}{r_{ij}} \right]$$

where r_{ij} is the relative distance between atom i and atom j. The $s(r)$ is the smooth version of the cutoff function, which guarantees that only contributions within distance r_c of atom i are considered and makes the second-order derivatives at the truncation still continuous.

The embedding matrix G is given by:

$$G = \mathcal{N}_{\text{embedding}}\big(s(r_{ij})\big)$$

where $\mathcal{N}_{\text{embedding}}$ is a three-layer MLP, the parameters of which are determined through training. The descriptor D preserves the translational, rotational, and permutational symmetries and is passed to another three-layer MLP to evaluate the potential energy.

After getting the potential energy E, the atomic force F_i and the virial tensor $\Xi = \Xi_{\alpha\beta}$ can be derived.

$$F_{i,\alpha} = -\frac{\partial E}{\partial r_{i,\alpha}},$$

$$\Xi_{\alpha\beta} = -\sum_{\gamma} -\frac{\partial E}{\partial h_{\gamma\alpha}} h_{\gamma\beta},$$

where α represents the αth component of r_i and F_i. $h_{\alpha\beta}$ is the βth component of the αth basis vector of the simulation region. The optimization of parameters within the model is achieved through the minimization of a loss function, as illustrated in the following equation:

$$\mathcal{L} = w_{\mathrm{E}}\left(E^{\text{pred}} - E^{\text{ref}}\right)^2 + w_{\mathrm{F}} \sum_i \left\| F_i^{\text{pred}} - F_i^{\text{ref}} \right\|^2 + w_{\Xi} \sum_{\alpha,\beta} \left(\Xi_{\alpha\beta}^{\text{pred}} - \Xi_{\alpha\beta}^{\text{ref}} \right)^2$$

where w_{E}, w_{F} and w_{Ξ} are weight coefficients that balance the contributions of different physical quantities. It is noteworthy that although the descriptors primarily rely on information from two-body and three-body interactions, the nonlinear mapping capabilities of deep neural networks enable the model to effectively capture higher-order many-body interactions. This allows the entire force field model to cover high-order interactions as comprehensively as possible. The hidden layers of the network are capable of learning complex relationships among the input features, which is difficult to achieve with traditional classical force fields.

Active learning framework

The generation of a comprehensive dataset is crucial for constructing an effective force field. It must balance sufficient data coverage of relevant phase space with minimal redundancy, as excessive redundancy can degrade model performance and increase computational costs. For the CuPc complex, high-quality and diverse datasets are essential to capture complex interactions and accurately represent configurational space, facilitating effective machine learning force field (MLFF) predictions.

The active learning framework serves as a strategic approach to efficiently generate and select informative data. It is an iterative process that efficiently explores the vast configurational space by selectively sampling the most informative structures, thereby ensuring the robustness and accuracy of the model.

As shown in Fig. 3, the active learning process was iteratively executed through the following steps:

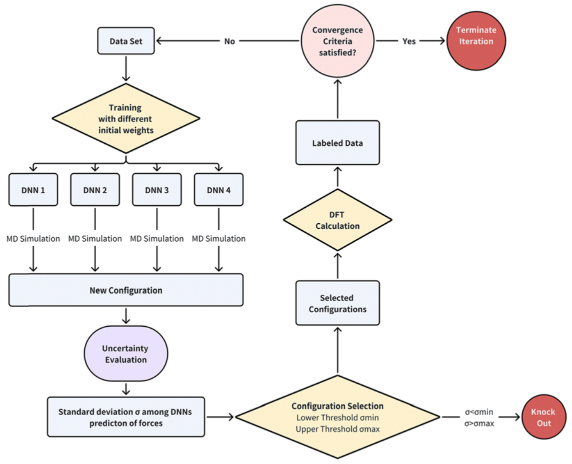

Fig. 3 Active learning iteration steps.

(1) Model training: four DNNs are trained using the current dataset with different initial weights, optimizing the neural network parameters to minimize the loss function defined by discrepancies between predicted and reference energies and forces.

(2) Candidate generation: new configurations are generated using molecular dynamics simulations. Lots of trajectories are performed under the current potential functions to explore new regions of the PES.

(3) Uncertainty evaluation: the ensemble of force field models predicts energies and forces for the candidate configurations. The max standard deviation (σ) among these predictions quantifies the uncertainty for each configuration.

$$\left(\sigma = \max\sqrt{\langle\|F_{m,i} - \langle F_{m,i}\rangle\|^2\rangle}\right)$$

where $F_{m,i}$ denotes the force on atom i predicted by model m. $\langle\ldots\rangle$ represents the average of model predictions.

(4) Configuration selection: configurations satisfying $\sigma_{\min} < \sigma < \sigma_{\max}$ are selected for inclusion in the training dataset. This ensures that we focus on regions where the model's uncertainty was informative but avoid regions likely associated with unphysical structures.

(5) High-fidelity calculations: selected configurations undergo density functional theory (DFT) calculations to obtain accurate reference energy, force and virial data, providing high-quality data for training.

(6) Dataset augmentation: The new data are incorporated into the training set, and the DNNs are retrained to include this additional information.

(7) Convergence assessment: the model's performance is evaluated against predefined criteria, such as root mean square error (RMSE) thresholds on a validation set. If the model meets the convergence criteria, the active learning loop will be terminated; otherwise, it will repeat from step 1.

Uncertainty evaluation is of importance in guiding the active learning process. We utilize an ensemble of neural network models trained on the same dataset but with different initial weights to estimate predictive uncertainty. The variance among the ensemble's predictions serves as an indicator of uncertainty for each configuration. The standard deviation of the predicted energies and forces is calculated across the ensemble models. This standard deviation provides a quantitative measure of the model's uncertainty for that configuration. A very small standard deviation indicates that the ensemble models agree closely in their predictions, suggesting that this region of the PES has been sufficiently explored and the DNN is confident in its predictions. A higher standard deviation represents regions that are not fully described or are more complex, indicating that additional data in these areas could improve the model's performance. However, if the standard deviation is excessively high, it may imply that the configuration corresponds to a physically incorrect or unstable structure, possibly due to numerical instabilities or artifacts in the candidate generation process.

Building on the above discussion, it is instructive to compare this ensemble-based approach to uncertainty quantification with the Bayesian approach employed in VASP's active learning force field. Bayesian approaches, such as Gaussian process regression (GPR), estimate uncertainty based on the posterior distribution of model parameters. These methods provide a probabilistic framework for uncertainty quantification, which is beneficial when detailed uncertainty analysis is required. They give not only a point estimate but also an associated variance, reflecting the model's confidence in its predictions. This approach is particularly advantageous for small datasets or situations where the relationship between the input and output is complex. In contrast, our method uses an ensemble of models to estimate uncertainty by comparing their predictions. This approach does not rely on a probabilistic framework and instead uses the variation in model outputs (forces or energies) to quantify uncertainty. The advantage of this method lies in its computational efficiency, as it avoids the heavy computational cost of calculating posterior distributions, making it well-suited for large systems or when computational resources are limited. Moreover, this approach is highly flexible and does not require assumptions about the prior distribution, which is particularly useful when prior knowledge is scarce.

Compared to Bayesian methods, the ensemble-based approach offers less detailed probabilistic uncertainty estimates, particularly in cases with sparse data or complex dependencies, and may overestimate uncertainty in regions where models agree closely. However, our ensemble-based approach is efficient, scalable, and flexible, and thus particularly well suited for high-throughput simulations in complex systems where specifying *a priori* distributions is impractical. It avoids intensive operations such as covariance matrix inversion. This approach provides a practical alternative to Bayesian error estimation in cases where computational efficiency is critical or prior information is limited.

When selecting configurations based on uncertainty, we consider both a lower and an upper bound on the standard deviation. Lower threshold ($\sigma_{\min}$), excluding configurations with very low uncertainty ($\sigma < \sigma_{\min}$), prevents the addition of data from regions where the MLFF is already well-trained. The upper threshold ($\sigma_{\max}$) excludes configurations with exceedingly high uncertainty, which could correspond to unphysical structures or numerical errors. To further ensure that selected configurations represented physically meaningful states of the CuPc complex, additional filters based on structural parameters are applied. Configurations exhibiting unrealistic bond lengths, angles, or atomic overlaps are excluded, even if they fall within the uncertainty range.

Furthermore, the $\sigma_{\min}$ and $\sigma_{\max}$ values are not fixed but dynamically adjusted based on the distribution of uncertainties in each iteration. This adaptive approach allows the selection criteria to evolve with the model's improving accuracy. A balance is maintained between exploring new regions of the PES (exploration) and refining the model's accuracy in already sampled regions (exploitation).

The idea of excluding configurations with excessively high uncertainty addresses the issue of potentially unphysical configurations that could arise in the active learning process, which can adversely affect the training of the model. In particular, large error estimates tend to occur in the early stages of active learning or when new structures are collected in regions of new temperature ranges. During these stages, the force field is not yet fully trained to describe the current potential energy surface, leading to the generation of unreasonable structures. For example, some atoms may be placed too close to each other, or other atoms may be positioned far from the system. If these structures were included in the dataset, they would degrade the predictive accuracy of the force field and introduce more unphysical configurations, exacerbating the problem. The potentials will exhibit greater variability in energy and force predictions, and the models struggle to generalize well to configurations that are outside the training set's scope. The inclusion of uncertain configurations leads to overfitting in certain regions of the configuration space, reducing the overall robustness of the model.

Thermal conductivity calculation

In this study, we employed equilibrium molecular dynamics (EMD) simulations in conjunction with the Green–Kubo formalism to calculate the thermal conductivity of the material under investigation. The Green–Kubo approach provides a rigorous theoretical framework that connects microscopic fluctuations in a system at thermal equilibrium to macroscopic transport coefficients, such as thermal conductivity.[25,26] By analyzing the time correlations of heat flux within the system, we can derive the thermal conductivity without imposing a temperature gradient, which is advantageous for simulating homogeneous systems and avoiding non-equilibrium effects.

The thermal conductivity tensor κ is given by the Green–Kubo relation:

$$\kappa = \frac{1}{k_{\mathrm{B}}T^2}\int_0^{\infty}\langle \mathbf{J}(0)\cdot\mathbf{J}(t)\rangle \mathrm{d}t,$$

where k_{B} is the Boltzmann constant, T is the absolute temperature and $\mathbf{J}(t)$ is the heat current vector at time t. $\mathbf{J}(0)\cdot\mathbf{J}(t)$ denotes the equilibrium ensemble average of the outer product of the heat flux at times 0 and t. The integral over time extends to infinity, but in practice, it is truncated at a time where the heat current autocorrelation function (HCACF) decays sufficiently. The microscopic heat flux $\mathbf{J}(t)$ is a vector quantity representing the flow of energy through the system at a given time. For classical systems, the heat flux can be expressed as:

$$\mathbf{J}(t) = \frac{1}{V}\left[\sum_i e_i\mathbf{v}_i - \sum_i \boldsymbol{r}_i\cdot\boldsymbol{F}_i\right],$$

where V is the volume of the simulation cell. e_i is the site energy of atom i, including kinetic and potential energy contributions. $\boldsymbol{v}_i$, $\boldsymbol{r}_i$ and $\boldsymbol{F}_i$ are the velocity, position vector and atomic force of atom i.

The molecular dynamics simulations are performed using the LAMMPS software package.[52] Periodic boundary conditions are applied in all three spatial dimensions to mimic an infinite bulk system and eliminate surface effects. Prior to data collection, the system is equilibrated under the NVT ensemble using a Nosé–Hoover thermostat. The equilibration phase ensures that the system reaches thermal equilibrium, with stable thermodynamic quantities such as temperature, pressure, and total energy. Production runs were conducted under the NVE ensemble to simulate the microcanonical ensemble. This setup is essential because the Green–Kubo formalism requires natural fluctuations of the system without external perturbations from thermostats or barostats. The production runs are sufficiently long, typically several nanoseconds, to ensure adequate sampling of heat flux fluctuations. A small timestep in the femtosecond range is used to accurately integrate the equations of motion and resolve atomic interactions.

Results and discussion

Dataset

In this work, we employ an active learning approach to generate molecular structures, followed by single-point calculations using VASP to derive first-principles energy, force, and virial data corresponding to these structures. DFT calculations are performed using the Vienna *ab initio* simulation package (VASP) with the projector augmented wave (PAW) method and the Perdew–Burke–Ernzerhof (PBE) exchange–correlation functional.[53,54] The PBE functional is augmented with the D3 van der Waals correction to accurately account for dispersion interactions. We set an energy cutoff of 800 eV for the PAW, a 2 × 6 × 3 *k*-point mesh, and a threshold of 10^{-5} eV for the electronic self-consistent loop. Detailed steps have already been described in the previous Methods section. This process is facilitated by the DPGEN[55] program.

Initially, we generate a preliminary dataset, from which we train an initial force field. This is iteratively expanded through molecular dynamics sampling. The initial dataset consists of structures selected from short-time *ab initio* molecular dynamics (AIMD) simulations conducted on α-CuPc. Subsequently, we conduct molecular dynamics simulations on different perturbed structures of α-CuPc and β-CuPc under the NPT ensemble across 3 temperature intervals ([50 K,100 K, 200 K], [300 K,400 K,500 K], and [600 K,700 K,800 K]) and five pressure levels [1 bar, 5 bar, 10 bar, 100 bar, and 1000 bar] for various configurations, resulting in a total of 5345 distinct structures. The distribution of these structures by temperature is presented in Table 2. Notably, the higher temperature regions yield a greater number of collected structures during the active learning iterations. When sampling in different temperature zones, we will keep selecting new structures based on uncertainty evaluation until the model accuracy converges. The detailed setting of thresholds is shown in ESI† Table S1. More structures collected in the temperature zone reflect the broader energy coverage that allows the exploration of a more extensive phase space.

Machine learning force field training

Utilizing the obtained dataset, we construct an efficient machine learning force field model based on high-accuracy

Table 2 Dataset distribution by temperature

	Structure counts	
Temperature (K)	α-CuPc	β-CuPc
[50,100,200]	226	489
[300,400,500]	1042	743
[600,700,800]	1411	1434

energy, force, and virial data extracted from first-principles calculations (DFT). The neural network is trained using the Adam optimization algorithm. We randomly select 80% of the dataset as the training set and reserve the remaining 20% for testing. In the setting of the training parameters, the cutoff radius is 6.50 Å. The embedding network has three layers with 25, 50, and 100 neurons, respectively, whereas the fitting network has three layers with 240 neurons. The loss function is optimized using the Adam stochastic gradient descent method. As shown in Fig. 4, we compare the force field accuracy before and after employing an active learning framework. Refining the force field through the active learning framework results in an order-of-magnitude improvement in accuracy on the same test set.

In the test set, the root mean square error (RMSE) for energy reaches 0.55 meV per atom, the RMSE for force is 57.9 meV Å^{-1}, and the RMSE for virial data per atom is 4.1 meV per atom. These figures indicate a high level of precision, with the model exhibiting similar performance across both the training and test sets. This close alignment further demonstrates that the dataset generated based on the active learning framework has a good coverage of the potential energy surface and that the deep potential model adequately obtains the required information from the dataset.

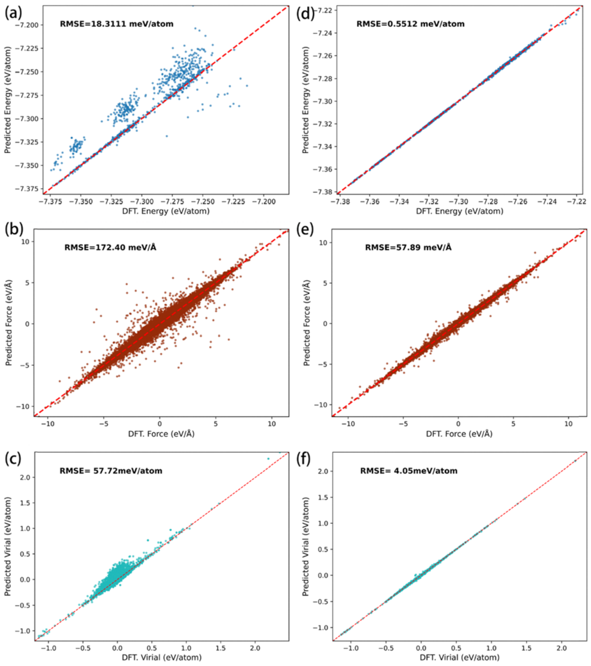

Fig. 4 (a) Energy, (b) force, and (c) virial data per atom before as calculated from the neural network before active learning compared with DFT calculations in training and test sets. (d) Energy, (e) force, and (f) virial data per atom are calculated from the neural network after active learning.

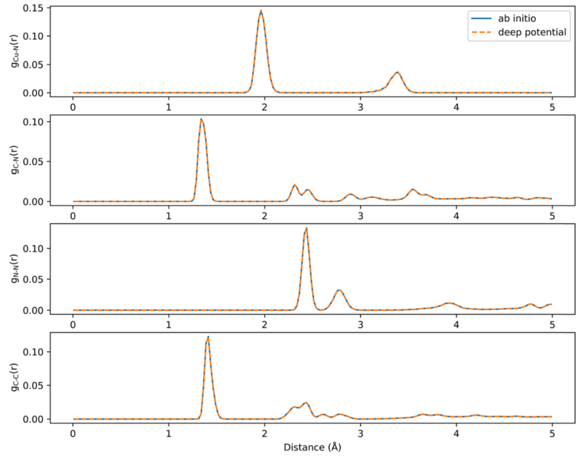

Fig. 5 Comparison of the radial distribution functions of β-CuPc (300 K) from *ab initio* (continuous blue line) and deep potential (dashed orange line) simulations, respectively.

However, further assessments are necessary to evaluate the model's physical performance, particularly whether the trained force field can yield results comparable to those obtained from DFT calculations. To test this, we compute the radial distribution function (RDF) for both the α and β phases of CuPc at 300 K using both the VASP and the trained model. The radial distribution function (RDF) $g(r)$ provides insights into the spatial distribution of particles within a system. As shown in Fig. 5, the RDF results from our force field model closely align with those from DFT calculations, reinforcing the reliability and accuracy of the trained force field model.

Convergence test

In this study, we employ equilibrium molecular dynamics (EMD) in conjunction with the Green–Kubo equation to compute the thermal conductivity of CuPc materials. While the EMD method does not strictly require the simulation cell size to exceed the phonon mean free path, size effects must be considered, necessitating a convergence test to mitigate their impact on the results. We conduct tests using various cell sizes, including 1 × 1 × 1, 2 × 2 × 2, . . . up to 10 × 10 × 10.

During the NVE simulations, we monitor fluctuations in pressure and temperature. Notably, the 1 × 1 × 1 cell exhibits substantial variations in physical parameters throughout the simulation, whereas larger cell sizes lead to increased stability, with convergence achieved at around the 5 × 5 × 5 cell size. Furthermore, despite our training set consisting solely of the unit cell and small supercell data, the trained force field model accurately captures variations in physical quantities as the cell size changes. This capability is attributed to the treatment of local environments within the neural network model.

Additionally, we perform convergence testing on the calculated thermal conductivity. For a given cell, we independently run 50 trajectories at 300 K to obtain ensemble averages, as shown

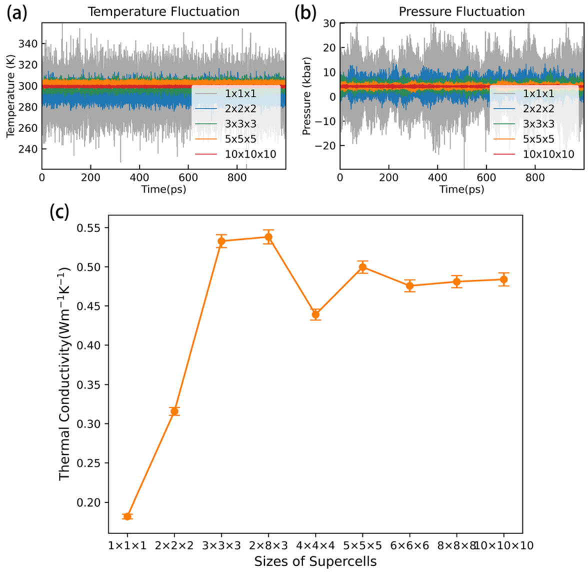

Fig. 6 (a) Temperature and (b) pressure fluctuations under NVE simulations for different sized crystal cells, (c) trend of thermal conductivity as a function of unit cell size. The data illustrate the convergence of thermal conductivity values with an increase in unit cell dimensions.

Table 3 Thermal conductivities from theoretical calculation and experiment

Thermal conductivity	Experiment[57] (W m^{-1} K^{-1})	Hybrid-compass[28] (W m^{-1} K^{-1})	DNN (W m^{-1} K^{-1})
β-CuPc	0.39	1.1 ± 0.2	0.499 ± 0.008

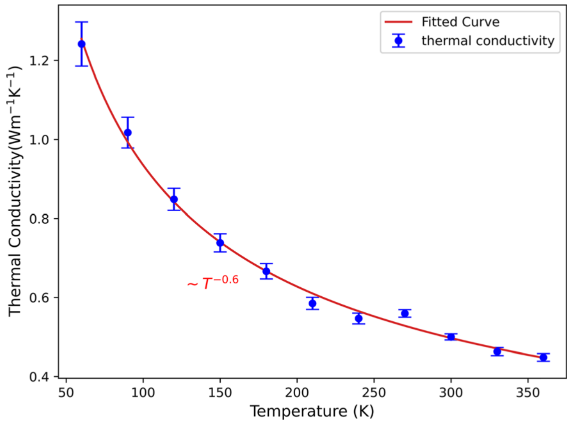

Fig. 8 Trend of thermal conductivity as a function of temperature.

in Fig. 6. The results indicate that convergence is reached with the 5 × 5 × 5 cell, aligning with the trends observed in the earlier physical parameter variations.

Thermal conductivity calculation

Following validation of the force field and convergence testing, we analyze the thermal transport properties of CuPc. Using a 5 × 5 × 5 cell, we execute fifty independent NVE simulation trajectories at the target temperature, employing the LAMMPS program. The heat flow computation module has been adapted from Tisi's work, correcting previous errors arising from LAMMPS' assumption of symmetry in atomic stress.[42,56]

Based on the heat flow data, we employ the Green–Kubo equation to derive the final thermal conductivity results. Taking the 300 K case as an example, the heat flow autocorrelation function is presented in Fig. 7, yielding a thermal conductivity of 0.49 W m^{-1} K^{-1}. In contrast, previous experimental results reported a value of 0.39 W m^{-1} K^{-1}, indicating an improved accuracy compared to the earlier hybrid-COMPASS-derived value of 1.2 W m^{-1} K^{-1} (Table 3). Given that our CuPc model is a perfect crystal in the molecular dynamics simulation, this overestimation of the calculated value is anticipated.

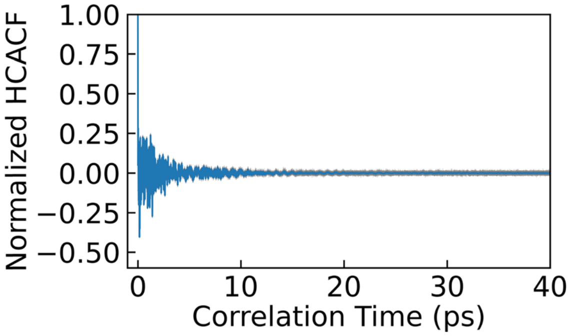

Fig. 7 Heat current autocorrelation function.

We also compute the thermal conductivity of CuPc across temperatures ranging from 60 K to 360 K, generating a temperature variation curve as shown in Fig. 8. At lower temperatures, diminished atomic random motion reduces phonon scattering, leading to increased thermal conductivity. The observed relationship of thermal conductivity with temperature, characterized by $\sim T^{-0.6}$, is consistent with the principles governing thermal transport in crystalline materials. In general, thermal conductivity arises from the collective motion of phonons, which are quantized lattice vibrations that carry thermal energy. At lower temperatures, phonon scattering is minimized, leading to higher thermal conductivity as phonons can propagate more freely through the crystal lattice. However, as temperature increases, several mechanisms contribute to increased phonon scattering. These include anharmonic interactions between phonons and the scattering caused by defects, grain boundaries, and other thermal excitations. Specifically, it implies that the phonon mean free path decreases with an increase in temperature, resulting in reduced thermal conductivity.

Conclusions

In conclusion, this research successfully demonstrates the efficacy of an active learning framework in generating a diverse dataset for training a machine learning force field tailored to CuPc. The nonlinear mapping capabilities of deep neural networks enable the model to effectively capture higher-order many-body interactions. The model utilizes a local environment matrix for representation without explicitly characterizing the metal–organic coordination. Nevertheless, the trained model exhibited impressive accuracy, as indicated by low RMSE values (energy: 0.55 meV per atom, force: 57.9 meV Å^{-1}, and virial data per atom: 4.1 meV per atom) and consistency with DFT results, confirming its potential for practical applications in predicting molecular interactions. Notably, our calculations reveal a thermal conductivity of 0.49 W m^{-1} K^{-1} at 300 K for the CuPc system, aligning well with experimental findings (0.39 W m^{-1} K^{-1}). In addition, our analysis of thermal conductivity elucidates the relationship between the temperature and thermal conductivity in CuPc. This work serves as an essential progress in demonstrating that the introduction of machine-learning force fields can effectively characterize the interactions of metal–organic complex systems and can significantly advance the development and discovery of organometallic thermoelectric materials.

Author contributions

Wenjie Zhang: conceptualization (equal); data curation; formal analysis; investigation; methodology (equal); visualization; software; writing – original draft (equal); and writing – review & editing (equal). Weitang Li: conceptualization (equal); methodology (equal); supervision (equal); and writing – review & editing (equal). Zhigang Shuai: conceptualization (equal); methodology (equal); project administration (equal); resources (equal); supervision (equal); and writing – review & editing (equal).

Data availability

Data for this article, including dataset, force field and MD results are available at **https://doi.org/10.5281/zenodo.13913359**. The code for the thermal conductivity calculation workflow can be found at **https://github.com/gdbhcxhmjk-z/ThermalConductivity-Workflow**. The version of the code employed for this study is version 0.1.2.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) (Grant No. T2350009 and 22433007), the Guangdong Provincial Natural Science Foundation (Grant No. 2024A1515011185), and the Shenzhen City "Pengcheng Peacock" Talent Program.

Notes and references

1 G. J. Snyder and E. S. Toberer, *Nat. Mater.*, 2008, **7**, 105–114.
2 L. E. Bell, *Science*, 2008, **321**, 1457–1461.
3 H. Liu, X. Shi, F. Xu, L. Zhang, W. Zhang, L. Chen, Q. Li, C. Uher, T. Day and G. J. Snyder, *Nat. Mater.*, 2012, **11**, 422–425.
4 J. P. Heremans, V. Jovovic, E. S. Toberer, A. Saramat, K. Kurosaki, A. Charoenphakdee, S. Yamanaka and G. J. Snyder, *Science*, 2008, **321**, 554–557.
5 M. Zebarjadi, K. Esfarjani, M. S. Dresselhaus, Z. F. Ren and G. Chen, *Energy Environ. Sci.*, 2012, **5**, 5147–5162.
6 Y. Sun, P. Sheng, C. Di, F. Jiao, W. Xu, D. Qiu and D. Zhu, *Adv. Mater.*, 2012, **24**, 932–937.
7 H. E. Katz, *J. Mater. Chem.*, 1997, **7**, 369–376.
8 O. Bubnova and X. Crispin, *Energy Environ. Sci.*, 2012, **5**, 9345–9362.
9 B. Russ, M. J. Robb, B. C. Popere, E. E. Perry, C.-K. Mai, S. L. Fronk, S. N. Patel, T. E. Mates, G. C. Bazan, J. J. Urban, M. L. Chabinyc, C. J. Hawker and R. A. Segalman, *Chem. Sci.*, 2016, **7**, 1914–1919.
10 D. Wang, W. Shi, J. Chen, J. Xi and Z. Shuai, *Phys. Chem. Chem. Phys.*, 2012, **14**, 16505–16520.
11 O. Bubnova, Z. U. Khan, H. Wang, S. Braun, D. R. Evans, M. Fabretto, P. Hojati-Talemi, D. Dagnelund, J.-B. Arlin, Y. H. Geerts, S. Desbief, D. W. Breiby, J. W. Andreasen, R. Lazzaroni, W. M. Chen, I. Zozoulenko, M. Fahlman, P. J. Murphy, M. Berggren and X. Crispin, *Nat. Mater.*, 2014, **13**, 190–194.
12 T. McAfee, B. C. Hoffman, X. You, J. M. Atkin, H. Ade and D. B. Dougherty, *J. Phys. Chem. C*, 2016, **120**, 18616–18621.
13 G. S. Nolas, J. L. Cohn, G. A. Slack and S. B. Schujman, *Appl. Phys. Lett.*, 1998, **73**, 178–180.
14 B. C. Sales, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 1997, **56**, 15081–15089.
15 Y. Lu and D. J. Young, *Dalton Trans.*, 2020, **49**, 7644–7657.
16 D. Yang, C. Lu, H. Yin and I. P. Herman, *Nanoscale*, 2013, **5**, 7290–7296.
17 X. Shi, J. Yang, J. R. Salvador, M. Chi, J. Y. Cho, H. Wang, S. Bai, J. Yang, W. Zhang and L. Chen, *J. Am. Chem. Soc.*, 2011, **133**, 7837–7846.
18 D. A. Broido, M. Malorny, G. Birner, N. Mingo and D. A. Stewart, *Appl. Phys. Lett.*, 2007, **91**, 231922.
19 W. Li, J. Carrete, N. A. Katcho and N. Mingo, *Comput. Phys. Commun.*, 2014, **185**, 1747–1758.
20 A. J. H. McGaughey and M. Kaviany, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 2004, **69**, 094303.
21 D. Wang, L. Tang, M. Long and Z. Shuai, *J. Phys. Chem. C*, 2011, **115**, 5940–5946.
22 Y. Sun, Z. Shuai and D. Wang, *Phys. Chem. Chem. Phys.*, 2018, **20**, 14024–14030.
23 R. Guo, X. Wang, Y. Kuang and B. Huang, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 2015, **92**, 115202.
24 A. Togo and I. Tanaka, *Scr. Mater.*, 2015, **108**, 1–5.
25 M. S. Green, *J. Chem. Phys.*, 1954, **22**, 398–413.
26 R. Kubo, *J. Phys. Soc. Jpn.*, 1957, **12**, 570–586.
27 F. Knoop, M. Scheffler and C. Carbogno, *Phys. Rev. B*, 2023, **107**, 224304.
28 C. Shao, Y. Jin, K. Pipe, M. Shtein and J. Kieffer, *J. Phys. Chem. C*, 2014, **118**, 9861–9870.
29 J. Behler, *Angew. Chem., Int. Ed.*, 2017, **56**, 12828–12840.
30 L. Zhang, J. Han, H. Wang, R. Car and W. E, *Phys. Rev. Lett.*, 2018, **120**, 143001.

31 O. T. Unke and M. Meuwly, *J. Chem. Theory Comput.*, 2019, **15**, 3678–3693.

32 A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, *Phys. Rev. Lett.*, 2010, **104**, 136403.

33 J. S. Smith, O. Isayev and A. E. Roitberg, *Chem. Sci.*, 2017, **8**, 3192–3203.

34 H. Wang, L. Zhang, J. Han and W. E, *Comput. Phys. Commun.*, 2018, **228**, 178–184.

35 J. Zeng, D. Zhang, D. Lu, P. Mo, Z. Li, Y. Chen, M. Rynik, L. Huang, Z. Li, S. Shi, Y. Wang, H. Ye, P. Tuo, J. Yang, Y. Ding, Y. Li, D. Tisi, Q. Zeng, H. Bao, Y. Xia, J. Huang, K. Muraoka, Y. Wang, J. Chang, F. Yuan, S. L. Bore, C. Cai, Y. Lin, B. Wang, J. Xu, J.-X. Zhu, C. Luo, Y. Zhang, R. E. A. Goodall, W. Liang, A. K. Singh, S. Yao, J. Zhang, R. Wentzcovitch, J. Han, J. Liu, W. Jia, D. M. York, W. E, R. Car, L. Zhang and H. Wang, *J. Chem. Phys.*, 2023, **159**, 054801.

36 A. V. Shapeev, *Multiscale Model. Simul.*, 2016, **14**, 1153–1173.

37 S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt and B. Kozinsky, *Nat. Commun.*, 2022, **13**, 2453.

38 R. P. Pelaez, G. Simeon, R. Galvelis, A. Mirarchi, P. Eastman, S. Doerr, P. Thölke, T. E. Markland and G. De Fabritiis, *J. Chem. Theory Comput.*, 2024, **20**, 4076–4087.

39 Z. Fan, Z. Zeng, C. Zhang, Y. Wang, K. Song, H. Dong, Y. Chen and T. Ala-Nissila, *Phys. Rev. B*, 2021, **104**, 104309.

40 Z. Fan, Y. Wang, P. Ying, K. Song, J. Wang, Y. Wang, Z. Zeng, K. Xu, E. Lindgren, J. M. Rahm, A. J. Gabourie, J. Liu, H. Dong, J. Wu, Y. Chen, Z. Zhong, J. Sun, P. Erhart, Y. Su and T. Ala-Nissila, *J. Chem. Phys.*, 2022, **157**, 114801.

41 M. F. Langer, F. Knoop, C. Carbogno, M. Scheffler and M. Rupp, *Phys. Rev. B*, 2023, **108**, L100302.

42 D. Tisi, L. Zhang, R. Bertossa, H. Wang, R. Car and S. Baroni, *Phys. Rev. B*, 2021, **104**, 224202.

43 C. Malosso, L. Zhang, R. Car, S. Baroni and D. Tisi, *npj Comput. Mater.*, 2022, **8**, 1–10.

44 S. Wieser and E. Zojer, *npj Comput. Mater.*, 2024, **10**, 1–18.

45 R. Cheng, X. Shen, S. Klotz, Z. Zeng, Z. Li, A. Ivanov, Y. Xiao, L.-D. Zhao, F. Weber and Y. Chen, *Phys. Rev. B*, 2023, **108**, 104306.

46 K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller and A. Tkatchenko, *Nat. Commun.*, 2017, **8**, 13890.

47 M. Rupp, A. Tkatchenko, K.-R. Müller and O. A. von Lilienfeld, *Phys. Rev. Lett.*, 2012, **108**, 058301.

48 S. J. Ang, W. Wang, D. Schwalbe-Koda, S. Axelrod and R. Gómez-Bombarelli, *Chem*, 2021, **7**, 738–751.

49 J. Hafner, *J. Comput. Chem.*, 2008, **29**, 2044–2078.

50 C. J. Brown, *J. Chem. Soc. A*, 1968, 2488–2493.

51 A. Hoshino, Y. Takenaka and H. Miyaji, *Acta Crystallogr., Sect. B*, 2003, **59**, 393–403.

52 A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott and S. J. Plimpton, *Comput. Phys. Commun.*, 2022, **271**, 108171.

53 J. P. Perdew and Y. Wang, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 1992, **45**, 13244–13249.

54 G. Kresse and J. Furthmüller, *Phys. Rev. B:Condens. Matter Mater. Phys.*, 1996, **54**, 11169–11186.

55 Y. Zhang, H. Wang, W. Chen, J. Zeng, L. Zhang, H. Wang and W. E, *Comput. Phys. Commun.*, 2020, **253**, 107206.

56 D. Surblys, H. Matsubara, G. Kikugawa and T. Ohara, *Phys. Rev. E*, 2019, **99**, 051301.

57 Y. Jin, A. Yadav, K. Sun, H. Sun, K. P. Pipe and M. Shtein, *Appl. Phys. Lett.*, 2011, **98**, 093305.