npj | computational materials

Article

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

https://doi.org/10.1038/s41524-024-01427-y

Quantum-accurate machine learning potentials for metal-organic frameworks using temperature driven active learning

Abhishek Sharma ✉ & Stefano Sanvito ✉

Understanding structural flexibility of metal-organic frameworks (MOFs) via molecular dynamics simulations is crucial to design better MOFs. Density functional theory (DFT) and quantum-chemistry methods provide highly accurate molecular dynamics, but the computational overheads limit their use in long time-dependent simulations. In contrast, classical force fields struggle with the description of coordination bonds. Here we develop a DFT-accurate machine-learning spectral neighbor analysis potentials for two representative MOFs. Their structural and vibrational properties are then studied and tightly compared with available experimental data. Most importantly, we demonstrate an active-learning algorithm, based on mapping the relevant internal coordinates, which drastically reduces the number of training data to be computed at the DFT level. Thus, the workflow presented here appears as an efficient strategy for the study of flexible MOFs with DFT accuracy, but at a fraction of the DFT computational cost.

Compounds presenting nanometer-size voids form a promising materials platform for various applications, including selective gas diffusion, adsorption and catalysis[1,2]. Flexible metal-organic frameworks (MOFs) have emerged as an intriguing class of nanoporous materials, which allow one to dynamically tune and control the structure and properties of such voids[3–5]. MOFs are crystalline materials made through reticular chemistry, where organic linkers are connected to metal units via coordination bonds. The flexibility of MOFs, in combination with external stimuli, affects the pores and pore channels and gives rise to interesting properties such as linker rotation, gate opening, swelling, negative thermal expansion, negative adsorption etc[3,6–8].

In order to study computationally the effects of an external stimulus, such as pressure and temperature, on the properties of flexible MOFs, a detailed analysis of the framework dynamics at an extended length and time scale is necessary[4,9–12]. This can be performed through molecular dynamics (MD) simulations. Ab-initio MD (AIMD), as implemented for instance with density functional theory (DFT), provides the most accurate estimation of the potential energy surface (PES). However, the computational overheads are significant and hence AIMD simulations are usually limited to few hundreds of atoms and pico-second time scales. Alternatively, one can use classical interatomic potential models or force-fields. These approximate the PES of a material with the help of parametric functions and may provide estimates of energy, forces, and virial-stress of thousand-atom atomic configurations in a short time. However, the use of classical force-fields for MOFs is hampered by their poor performance with atomic environments presenting coordination bonds[13]. Despite this limitation, a variety of classical force-fields have been used to study the properties of MOFs[14–23]. These force-fields are either transferrable (e.g., UFF[14], DREIDING[17], UFF4MOF[15,16] etc.) or developed for a specific MOF (e.g., QUICK-FF[19] and MOF-FF[21,23]).

A possible strategy to achieve DFT accuracy at a computational cost comparable to that of classical force fields is provided by machine-learning potentials (MLPs)[24–30]. In these, the atomic chemical environments are represented by mathematical descriptors at various levels of complexity[31], while the corresponding fitting parameters are obtained by training appropriate machine-learning models over the energy, forces, and virial-stress values of a large number of configurations. These are typically obtained by DFT. The accuracy of a MLP to predict the PES of a material depends on the chemical-environment descriptors, the number of parameters in the model, the size and diversity of the training set, and the training procedure.

Recently several computational works have used MLPs to study MOFs[12,32–42]. Most of these employ neural-network potentials (NNPs), which require thousands of training configurations and comprise hundred thousands of parameters to fit. In most of the earlier works, the training configurations were generated via picosecond-long AIMD simulations, which require a long computational time and significant computational resources. Furthermore, the fit of the many parameters is also computationally intensive and the model are little interpretable, namely it is not simple to define from the outset the boundary of the model's validity (e.g., the temperature and pressure range).

School of Physics, AMBER and CRANN Institute, Trinity College, Dublin 2, Ireland. ✉e-mail: asharma.ms.in@gmail.com; sanvitos@tcd.ie

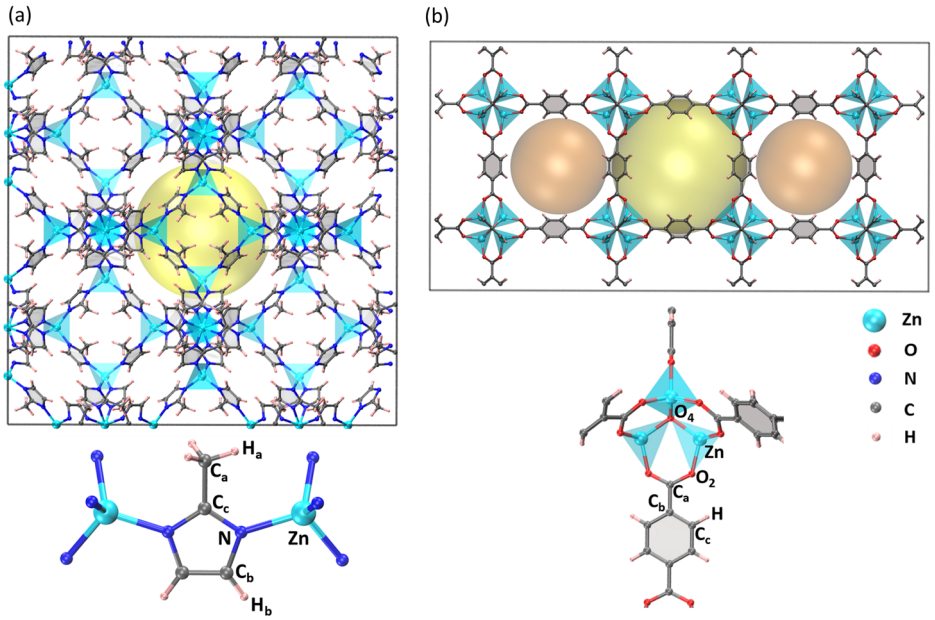

Fig. 1 | Atomic structures of the MOFs investigated in this work. a ZIF-8 and (**b**) MOF-5 (cyan: zinc, blue: nitrogen, red: oxygen, gray: carbon, pink: hydrogen). The basic building units with atom types (C_a, C_b, etc.) are shown below the main structures. In order to show clearly the pores (yellow and orange spheres), supercells of size 2 × 2 × 2 and 1 × 1 × 2 are shown for ZIF-8 and MOF-5, respectively. In all calculations described in this work, the unit cell of both ZIF-8 (containing 276 atoms) and MOF-5 (containing 424 atoms) are used.

In this work, we use the spectral neighbor analysis potential (SNAP)[43] as MLP model to study the structural and vibrational properties of MOFs at finite temperature and pressure. SNAP was previously shown to perform well for organic molecules and coordination complexes, and thus it appears as a natural choice for MOFs[44]. SNAP is based on many-body descriptors and linear models, hence, when compared to neural-network potentials, it requires only a few hundreds of parameters to obtain similarly accurate fits. For this reason, SNAP typically demands a much smaller training set than those needed by neural-networks, so that a limited number of DFT calculations is necessary. As a test bench, here we develop two SNAPs for the widely studied ZIF-8[45] and MOF-5[46] MOFs (their structures are shown in Fig. 1). These two particular MOFs have been selected for our study, since various experimental results are available, so that the validity of our approach can be thoroughly tested.

Firstly, we perform a very detailed analysis of the SNAP learning curves, which allows us to propose a protocol for generating correlation-free training sets. Then, we compute various structural and vibrational properties as a function of temperature, and we compare them with available experimental data, demonstrating an excellent agreement. Although applied here to ZIF-8 and MOF-5, our approach is completely general and widely applicable to any other MOFs, whose electronic structure is accessible by DFT (or other ab initio electronic structure methods). This allows one to develop predictive SNAPs for MOFs by using only a few hundreds DFT calculations and a simplified training procedure.

Results

Active learning algorithm

The construction of an adequate training set is crucial for the formulation of a MLP. In fact, MLPs are not physically informed, so that their knowledge of the energy and forces of a particular molecular structure is rooted in having been trained on structures that contain similar local environments. Importantly, in general, MLPs are not guaranteed to extrapolate to poorly known configurations, for which they can catastrophically fail. As such, an ideal training set needs to contain all the local environments that the system will experience when performing the inference, for instance the ones explored during MD simulations. Such training set should also be finely balanced, namely, even when complete, it should not be dominated by a particular pool of local environments. Finally, the size of the training set must be kept as limited as possible, so that the construction of the MLP itself will be numerically convenient in the computational economy of the workflow that one wants to pursue.

With all these requirements in mind we present here an active-learning strategy that allows us to construct a balanced training set, while performing a limited number of DFT calculations. This consists of two main tasks, namely (1) an algorithm that maps the diversity of the training set and ensures that all the relevant local environments are represented, and (2) a strategy to generate the molecular configurations containing those environments. The algorithm is then used for both ZIF-8 and MOF-5, although in two different ways. In fact, although in both cases the first step is identical, then for ZIF-8 the configurations are generated with AIMD, while for MOF-5 they originate from MD runs performed at increasingly large temperatures with subsequently more refined SNAPs. This is because we effectively use the construction of the ZIF-8 SNAP as a learning step to the construction of an efficient method, which is then used for MOF-5.

Selection of the configurations to include in the training set

The atomic configuration of a MOF structure can be defined through the knowledge of the unit cell parameters, the atomic bonds, the bond angles, and the dihedral angles [see Fig. 2a]. Classical force-fields use this information to compute the energy and forces of a configuration in terms of non-bonded and bonded interactions. In the case of non-bonded interactions, atoms are classified into different types, based on their chemical identity and connectivity, information which is then used to define the interaction parameters (e.g., electrostatic charges, Lennard-Jones parameters, etc.). Similarly, bond lengths, bond angles and dihedral angles are classified into different types and correspondingly interaction parameters of the bonded interaction are thus computed.

Inspired by this structure-informed approach, we have developed a simple algorithm to track the diversity and relevance of the local atomic environments included in a training set [see Fig. 2b for details]. As for the classical force fields, we would like to differentiate structures in terms of a limited number of structural descriptors, namely the cell-parameters, bonds, angles, and dihedrals (collectively called CBAD). The total number of structure descriptors, N_{CBAD}, depends on the MOF of interest and consists of 6 cell parameters, l bond types, m angle types, and n dihedral angle types [see Fig. 2a for details]. Then, we define the resolution of each descriptor, Δ, which is the minimum difference between two values of the descriptor that one can distinguish. This allows us to represent each value of the descriptors as an integer (representing bin index), namely as int(θ/Δ), where θ is the

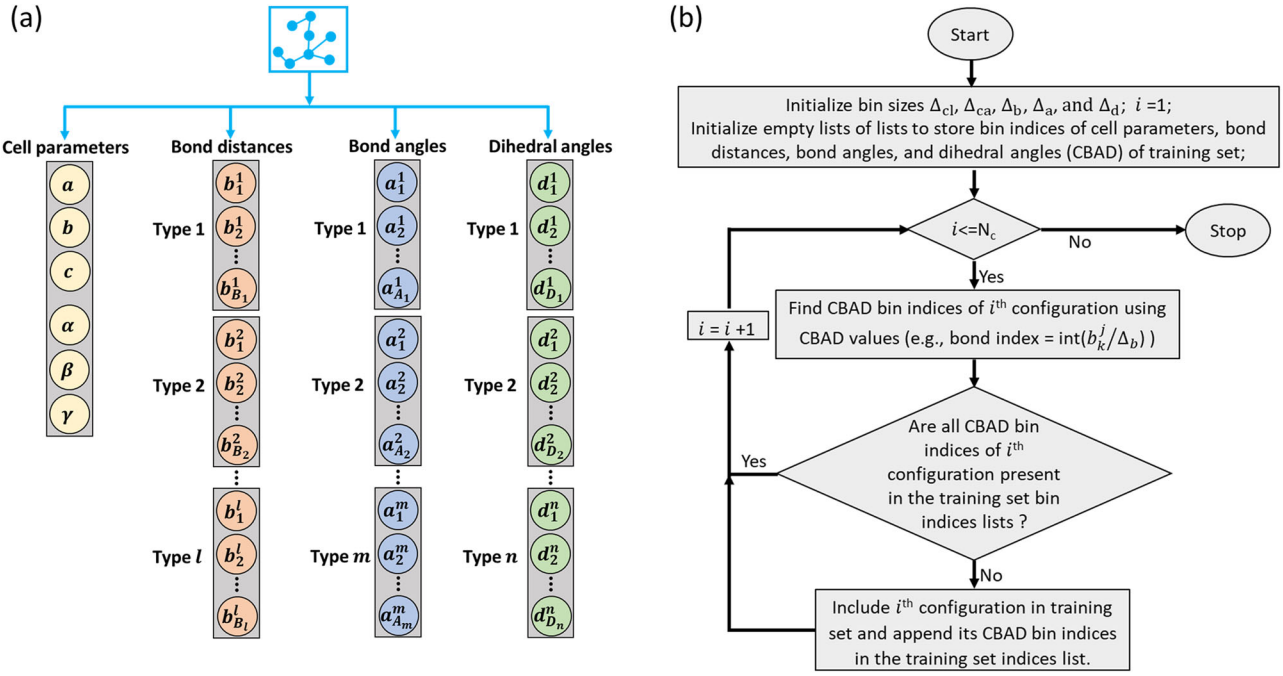

Fig. 2 | Outline of the active learning algorithm developed in this work. **a** Schematic showing the cell parameters (lattice parameters: a, b, c; cell angles: α, β, γ) and different types of bonds, angles, and dihedrals present in the atomic configuration of a MOF structure. This representation of an atomic configuration is termed here as CBAD (cell parameters, bonds, angles, and dihedrals). **b** A flowchart illustrating our CBAD-based algorithm to select training set configurations from a given set of N_c configurations. Here Δ_{cl}, Δ_{ca}, Δ_b, Δ_a, and Δ_d are the descriptor resolutions for the cell lengths (a, b, c), cell angles (α, β, γ), bond distances (b_k^j), angles (a_k^j), and dihedral angles (d_k^j), respectively. The Δ_i values are used to find CBAD bin indices (e.g., bond index = int(b_k^j/Δ_b)). The algorithm then compares the CBAD indices of a new configuration with the CBAD indices of the training set and will include the new configuration in training set, if at least one of the indices is not present in the configuration matrix.

value of the descriptor. Each MOF configuration has multiple values of each descriptor (except the 6 cell parameters, which have only one value). Therefore, for each MOF configuration we track N_{CBAD} lists of integers, where each list can have multiple integer values. If we wish to consider M possible values for each descriptor in the training set, we will have $M^{N_{CBAD}}$ possible configurations (or configuration matrix) to explore in a N_{CBAD} dimensional configuration space. The question is now how to populate such space with relevant configurations. In principle, one can generate MOF configurations where the descriptors are varied one at a time, but this will necessitate $M^{N_{CBAD}}$ DFT calculations, a number that can be prohibitively large. The strategy chosen here, instead, is that of mapping the entire space with a limited number of configurations.

In order to reduce the number of DFT calculations, we consider the multiplicity of descriptor values in an atomic configuration of MOF (within the N_{CBAD} lists of integers). If an atomic configuration results from a MD simulation (at certain temperature), then all values of a descriptor in that configuration would be different and belong to a distribution (of descriptor at certain temperature). In such configuration, the atoms in that configuration can have multiple local chemical environments. Thus, instead of tracking the $M^{N_{CBAD}}$ configuration matrix, we individually track N_{CBAD} descriptors and corresponding configuration matrix of size of $N_{CBAD} \times M$. With this simplification, our algorithm proceeds as follows. An initial configuration defines multiple value for each of the N_{CBAD} descriptors, we consider only unique values for each descriptor and populate the values in the N_{CBAD} descriptor lists (to sample the $N_{\mathrm{CBAD}} \times M$ configuration matrix). The next configuration will then define a new N_{CBAD} list of descriptors. If at least one of them is not already present in the corresponding descriptor lists, then such configuration will be accepted and it will be part of the training set. In this way we ensure that all the relevant values of each descriptor, within a precision Δ, are represented at least once in our training set. A schematic illustration of the algorithm is provided in Fig. 2b, while a detailed pseudocode is given in the Supplementary Fig. 1. It is to be noted that this CBAD algorithm selects configurations based on the order they are presented. Thus, if CBAD is applied individually to two different set of configurations which are in different order with same atomic structures (e.g., set1 = [C_1, C_2, C_3, C_4, C_5] and set2 = [C_4, C_2, C_5, C_1, C_3]), than it will select different type of training set configurations (e.g., [C_1, C_2, C_4] from set1 and [C_4, C_5, C_3] from set2). Furthermore, the CBAD algorithm can identify the diversity between a typical MD simulation in the NVT/NPT ensemble and a biased simulation (e.g., metadynamics), since the distributions of the CBAD values are different in these two cases.

Training the SNAP

Within the SNAP formalism[43] the total energy, E, of a molecule (or a solid) is expressed as the sum of individual atomic energies, E_i. These are, in turn, function of the local chemical environment of each individual atom, which is defined within a radial cut-off. SNAP then expands the local atomic-density distribution over four-dimensional spherical harmonics and constructs the associated bispectrum components, B_j, which form a rotationally invariant set of descriptors. Finally, the atomic energies are taken as a linear function of the bispectrum components, namely

$$E = \sum_i^{N_a} E_i = \sum_i^{N_a} \sum_l^{N_{2J}} \beta_i^l B_i^l$$

where, N_a is total number of atoms, N_{2J} is total number of bispectrum functions ($2J$ controls the order of the expansion), and β_i^l are the coefficients of the bispectrum components. The accuracy of the chemical-environment description can be tuned by tuning the number of bispectrum components. Earlier work[44] has shown that considering 56 bispectrum components (corresponding to $2J = 8$) per chemical species results in a reasonable accuracy, and thus we have used same value here. In both ZIF-8 and MOF-5 there are 7 different atom types (see Fig. 1 – note that C, O, and H atoms with

different coordination are considered as different atom types), so that our SNAP models are constructed over 392 bispectrum functions (and 392 β_i^I values). The SNAP training, namely the computation of the β_i^I values, is here performed over the energy, forces and the virial-stress of each of the configurations contained in the training set, with the reference values being computed with DFT (see Methods section for details). Thus, each configuration provides $3N_a + 7$ training data (1 energy, $3N_a$ forces, and 6 virial-stress components). The unit cells of ZIF-8 and MOF-5 contain 276 and 424 atoms, respectively. Therefore, if the training set comprises in the region of 600 configurations, we will have approximately 0.5 and 0.7 million of training data for ZIF-8 and MOF-5, respectively. We generate the bispectrum components of a given configuration by using the Large-scale Atomic/Molecular Massively Parallel Simulator (LAMMPS)[47,48] package and then obtain the bispectrum coefficients through ridge regression. We optimize the SNAP hyperparameters (atomic-species-dependent cutoff radius and chemical-species weights) by using the Scipy package and we drive the optimization by minimizing the error over energy, forces and stress tensor. Then, the SNAP training is performed at the optimal hyperparameters and the final model is used to perform MD simulations.

Generation of the training and test sets for ZIF-8

In order to establish a general SNAP-training protocol for MOFs, we have first developed the potential for ZIF-8. The unit cell of ZIF-8 contains four elements (C, H, N, and Zn), 7 atom types [see Fig. 1a] and 276 atoms. The different configurations to be included in the training set are generated by first following the computationally intensive approach used in earlier works, namely we perform AIMD simulations (details are given in Methods section). These have a duration of 1 ps (with a 0.5 fs timestep) at temperatures ranging from 100 K to 1000 K with a 100 K interval. From the generated 20,000 configurations, we then select those to include in the training set by using the simple algorithm described in the previous section. Note that for all configurations included in the training and test set, we run high-quality DFT calculations with the higher cut-off of 1000 Ry (details of all DFT calculations are given in Methods section).

In general, the number of configurations contained in the training set should be optimal, since a few configurations will result in poor a representation of the PES, while too many configurations are associated to a high computational cost and to a possible imbalance in the representation of the main structural characteristics of the MOF. Thus, finding the optimal number of configurations is an essential step for the development of the MLP. Our strategy to populate the configuration matrix mitigates the risk of oversampling, and we can systematically change the number of configurations by changing the resolution of the structural descriptors (how finely we sample each descriptor). In this way we create training sets ranging from 50 to 3000 configurations and fit a SNAP for each of these training sets. Then, the performance test is conducted over two different sets. The first, referred here as test set A, contains around 5000 configurations obtained from AIMD simulations (at different temperatures between 100 to 500 K), while the second (test set B) is generated by using classical MD simulations (using force-field proposed by Weng et al.[22]) of the ZIF-8 unit cell at 500 K. In this second case we select around 2000 configurations.

Depending on the training set size and diversity, a part of the PES (or the configuration space) can be represented more accurately (less error), less accurately (moderate but acceptable error), or unphysically (very high error). In order to study the effect of the training set size and diversity, we compute the SNAP learning curves for ZIF-8 (shown in Fig. 3). To obtain the learning curves, we create training sets of different sizes using different values of the descriptor resolutions (details are given in Supplementary Table 1). The learning curves are taken over the diverse test set and display both the root mean square error (RMSE) and the mean absolute error (MAE) as a function of the number of configurations in the training set. With a very small number of configurations, we observe high errors in the energy, forces and the virial-stress learning curves, signaling an unphysical representation of the PES. As the size of the training set increases, we observe a decrease in errors, indicating improvement in the representation of the PES. For training sets containing in excess of 600 configurations an error plateau is found, indicating that the PES is well described. This implies that 600 configurations are optimal for a good representation of the PES for a MOF like ZIF-8. Including more diverse configurations can further improve the representation of unexplored regions of the PES. Thus, after the plateau in the learning curves, the errors on the test set can decrease if the training set size is increased by adding configurations from a region close to the PES where test configurations are distributed. Interestingly, the errors can also increase if the additional configurations included are either far from the test set configurations (representing other portion of the PES) or have some level of correlation. Presence of correlation lowers the diversity in the training set and populates unevenly a particular region of the PES, a fact that may result in overfitting that region of the PES.

Here, for the test set A we observe a marginal error enhancement in energy (around 1 meV) when the configurations are increased beyond 600, a feature that may suggest minor overfitting. It is to be noted that, configurations are selected sequentially from a pool of 1 ps AIMD simulation trajectories, a selection that may be affected by correlation among the configurations (details are given in Supplementary Fig. 4) and this could be the possible reason of overfitting. For further MD simulations of ZIF-8 we use the current training set containing the selected 672 configurations; the associated parity plots, computed over energy, forces and viral stress, are displayed in Fig. 3. However, in order to check the cause of the overfitting, we have performed a new analysis, where we shuffle the order of the AIMD configurations (to disrupt correlation) and reselect the training set, using the CBAD algorithm. The learning curves for this new training process are shown in Supplementary Fig. 5, where we found that the marginal error enhancement in the energy of the test set A vanishes and a plateau in all energy, forces, and stress error values is observed beyond the 600 training configurations. In any case, the errors of the converged SNAP are extremely low, namely of the order of 0.5 meV/atom, 50 meV/Å and 25 MPa, respectively for energy, forces and stress tensor. This level of accuracy is certainly enough to perform reliable MD over a broad temperature range, as we will demonstrate later on.

Training and Test Set for MOF-5

In the construction of the ZIF-8 SNAP we did generate about 20,000 configurations, but then realised that 600 are ideal to fit a high-performing model. Now, for MOF-5 we wish to establish a method that allows us to compute only the 600 configurations needed without any redundancy. In a recent work, an incremental learning approach was used in combination with metadynamics to generate the training set configurations of a MLP[36]. In a metadynamics simulation, a few collective variables are defined and bias is added along their trajectories to explore a particular region of the phase space. Since increasing the temperature corresponds to enlarging the phase space explored for all structural descriptors (and not just the collective variables), here we develop a simple algorithm driven by temperature to generate the training set for MOFs (see Fig. 4 for details).

The proposed algorithm proceeds as following (see Fig. 4). Firstly, we take the experimental crystal structure and generate different configurations by introducing a small random perturbation to the atomic positions (Step 0 in Fig. 4). Among these configurations we select those to populate the defined configuration matrix according to the CBAD algorithm described before, and their electronic structure is computed by DFT. The DFT energies, forces and stress tensors are then used to train an initial SNAP (MLP_0). The following step (Step 1 in Fig. 4) performs 50 independent MD runs, starting from 50 inequivalent configurations obtained by random displacing atoms from the experimental structure. The MD is conducted, starting from different initial velocities, at the low temperature of 100 K (in the *NPT* ensemble) by using MLP_0 for approximately 2000 steps. We then select at most one configuration from each MD run to be included in the training set, according to the selection criterion discussed before, and for these we run again DFT simulations. Such expanded training set is then used to construct the next generation of SNAP (MLP_1). Step 1 is then repeated multiple times at a progressively higher MD temperature, which is here increased by 100 K

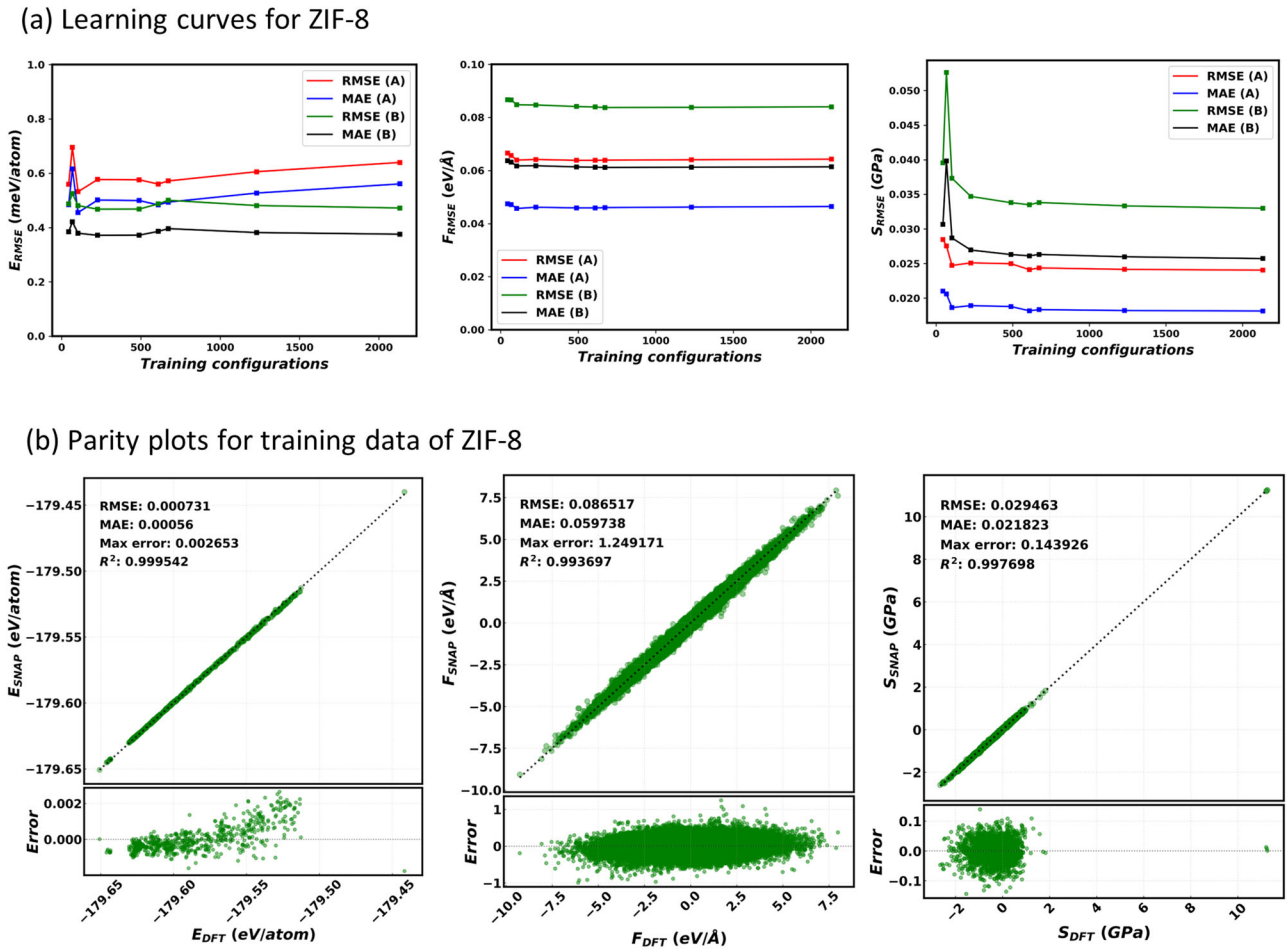

Fig. 3 | Performance of the trained SNAP model for ZIF-8. a Learning curves for the RMSE and MAE for energy (left-hand side panel), forces (middle panel) and virial-stress (right-hand side panel). Data are presented for test set A (composed of ~5000 configurations from AIMD simulations) and test set B (composed of ~2000 configurations from classical MD simulations) as a function of the number of configurations in the training set. Note that no significant change in the error is observed after the training set size reaches ~ 600 configurations. **b** Parity plots for energy, forces and virial-stress values comparing DFT and SNAP (trained over 672 configurations) values. The RMSE is 0.7 meV/atom, 86 meV/Å, and 29.5 MPa, respectively for energy, forces and virial-stress.

at each step. This process enhances the diversity of the training set and expands the range of temperature at which the SNAP can be used. For MOF-5 we performed iterations until the temperature reached 1000 K, obtaining a total of 487 training configurations. Further details about this approach are described in Supplementary Note 2.

Once the SNAP corresponding to the highest temperature (built over with 487 configurations) is constructed, we perform MD simulations with temperature now ramping between 10 K and 1000 K. In such MD simulations we use the same SNAP, trained over all the 487 configurations, across the entire temperature range and no other previous versions of SNAP are employed. Out of this last MD trajectory, we select the configurations to be used for the test set (1191 in total). Furthermore, we randomly select ~100 more configurations (from this MD simulation and from another at 400 K) to be included in the training set, so that the total number remains close to 600 (596 in our case). More details about the training set construction are given in Supplementary Table 2. The final SNAP is then trained on such data set (with 596 configurations) and used further for all analysis and MD simulations. Parity plots for final training set and test set are shown in Fig. 5. Again, we obtain a very high-quality potential with training-set energies accurate to sub meV/atom, and MAE on forces and stress components of 100 meV/Å and 19 MPa, respectively. Note that the errors on the test set are even lower than those on the training data. This is due to the fact that the test configurations are generated via MD simulations with a properly trained SNAP (where the temperature is ramped between 10 K to 1000 K), and therefore, they are less distorted when compared to the training configurations. As a consequence, the range of values for energy, forces and viral stresses in the test set is more limited than that of the training set. We now proceed to evaluate a number of structural and vibrational properties ZIF-8 and MOF-5 using MD simulations with trained SNAP.

Lattice constant

We begin by looking at the effect of temperature and pressure on the lattice constant. For this, the trained SNAPs are used to perform five independent 600ps-long MD simulations (with a timestep of 1 fs) in the *NPT* ensemble for both ZIF-8 and MOF-5. Then, the trajectories of the first 100 ps are considered as equilibration steps and the remaining 500 ps are used for property calculations. A window of 10 ps is used to estimate the average and the variance in the lattice parameters, with our results being summarised in Fig. 6. In general, we find an excellent agreement between the simulated lattice parameters and available experimental data. For example, our simulated lattice parameter (26.02 Å) of MOF-5 at 100 K is close to experimental single crystal X-ray diffraction data (25.89 Å) and to the simulation results of Eckhoff et al.[32] (26.082 Å) and Tayfuroglu et al.[37] (26.03 Å), obtained with neural-network potentials. In the case of ZIF-8 the lattice parameter increases with temperature (this is referred to as positive thermal expansion) and decreases with pressure. In contrast, MOF-5 has a negative thermal expansion. The computed linear thermal expansion coefficient at 300 K for ZIF-8 is $7.1 \times 10^{-6}\,K^{-1}$, which is within the experimental range determined by $11.9 \times 10^{-6}\,K^{-1}$ (Sapnik et al.[49]) and $6.5 \times 10^{-6}\,K^{-1}$ (Burtch et al.[50,51]). Similarly, we obtain a linear thermal

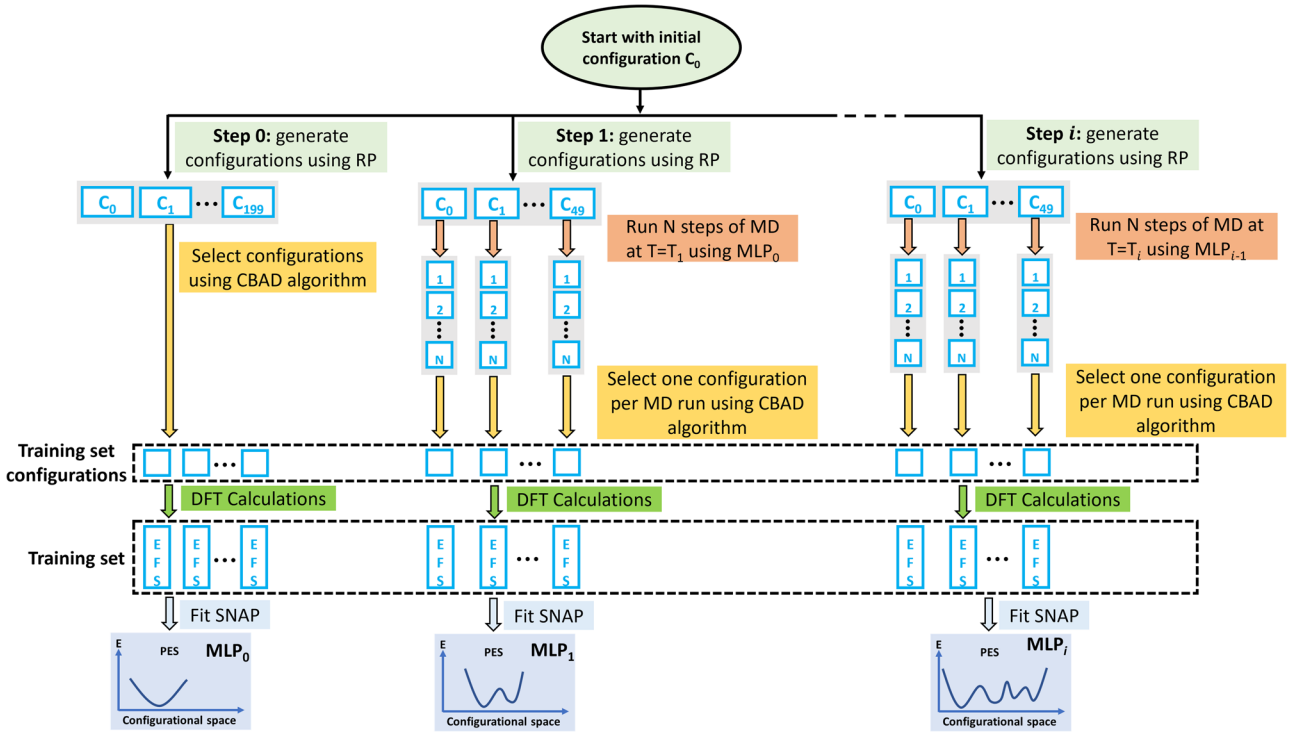

Fig. 4 | Overview of the computational workflow used to generate the training set for MOF-5. At each step a new SNAP is constructed by using the structure configurations obtained from the MD simulations performed with the SNAP trained at the previous step (see details in the text). 'RP' means random perturbation of the atomic positions.

expansion coefficient of $-13.3 \times 10^{-6}\,K^{-1}$ for MOF-5 at 300 K, which is close to experimental value[52] of $-13.1 \times 10^{-6}\,K^{-1}$ and to other simulation results from Eckhoff et al.[32] (-10.5 to $-8.3\,K^{-1}$) and Tayfuroglu et al.[37] (-13.17 to $-8.97\,K^{-1}$). Such excellent agreement indicates that our SNAPs are well capable of describing volumetric changes of the lattice parameters as a function of temperature. Note that the absolute value of the lattice parameters predicted by SNAP is slightly larger than that measured experimentally, by approximately 0.5% for both MOFs. This minor overestimation is due to the use of the DFT generalized-gradient approximation (GGA) to the exchange and correlation functional used for the construction of the training set. GGA sometime may slightly underbind and this feature is here transferred to the SNAP.

Vibrational density of states (VDOS)

Having investigated the temperature and pressure response of the MOFs we now move at analysing their vibrational properties. In particular, we compute the vibrational density of states (VDOS), which is here obtained as the Fourier transform of the mass-averaged velocity autocorrelation function along an MD trajectory. In this case, we perform an *NPT* simulation at 300 K for 1 ps, followed by a 500ps-long *NVE* simulation, from which we extract atomic configurations and velocities every 2 fs. Our computed VDOSs are shown in Fig. 7, while the partial VDOS (PVDOS) projected over each atom type are shown in Supplementary Figs. 12 and 14. Similar to results of Eckhoff et al.[32] for MOF-5, here we observe two main spectral regions (below 1700 cm^{-1} and after 2900 cm^{-1}) for both MOF-5 and ZIF-8. Then, we compare the PVDOS (Supplementary Figs. 12 and 14) of different atoms (see Fig. 1 for the definition of the atom types such as C_a, H_a, etc.) to identify the modes associated to the various peaks of the vibrational spectrum.

In general, the experimental[53] infrared (IR) and Raman spectrum of both ZIF-8 and MOF-5 agrees well with our simulated VDOS. Recently, in a detailed computational and experimental study of ZIF-8, Ahmad et al.[54] identified six regions defining the vibrational spectrum: (i) around 3200 cm^{-1} (stretching modes from C_b-H_b), (ii) around 3000 cm^{-1} (C_a-H_a methyl group's symmetric and asymmetric stretches), (iii) 1400–1500 cm^{-1} (C_a-H_a bending modes, H_b-C_b-C_b-H_b rocking modes, and ring deformation modes), (iv) 1310 cm^{-1} (rocking mode of C_b-H_b in the H_b-C_b-C_b-H_b moieties, and small deformation of the ring), (v) 1100–1200 cm^{-1} (combined scissoring and rocking motions of C_b-H_b in the H_b-C_b-C_b-H_b moieties of different rings in a unit cell, bending modes of C_b-H_b with respect to ring, breathing of entire ring, and minor C_a-H_a bending modes), and (vi) 990 cm^{-1} (C_a-H_a bending modes, in-plane C_b-H_b rocking in the H_b-C_b-C_b-H_b moieties, and small in-plane deformation of the ring). Consistently with their study, for the aromatic C_b-H_b dynamics, we observe VDOS spectral amplitude (Fig. 7 and Supplementary Figs. 11–12) in the 3200–3250 cm^{-1} range, which is also close to experimental Raman frequencies[55] of 3110 and 3131 cm^{-1} and the IR frequency[56] of 3135 cm^{-1}. The simulated VDOS for methyl C_a-H_a dynamics is observed in the window 2900–3150 cm^{-1}, which is also in agreement with range of Ahmad et al.[54] and to the experimental Raman[55], 2915 and 2931 cm^{-1}, and IR frequencies[56], 2927 and 2961 cm^{-1}. We also observe common PVDOS peaks around 1507 and 1140 cm^{-1} for H_b, C_b, C_c, and N atoms, which are associated with the dynamics of the entire ring (N-C_b, N-C_c, C_b-H_b). This value is close to the experimental Raman frequencies[55] at 1499 and 1508 cm^{-1}. In addition to these, we observe a common peak at 1390 cm^{-1} (for C_a, H_a, C_c, and N atoms), which corresponds to the coupled dynamics of the methyl group and the ring. For all C, H, and N atoms common peaks are observed near 1400–1450, 1310, 1200, 1000–1050, and 650–700 cm^{-1}, which correspond to vibrational dynamics of entire organic segment of ZIF-8. A peak around 600 cm^{-1} is common to H_b, C_b, and N atoms and corresponds to associated bending modes. Further analysis of PVDOS reveals Zn-N vibrational frequencies at around 180 cm^{-1}, 226 cm^{-1} and 286 cm^{-1}, which are close to the experimental Zn-N Raman frequencies[55] of 168 and 273 cm^{-1} and the far infrared (IR) frequencies[57] in the 265–325 cm^{-1} range. We also observe various peaks below 300 cm^{-1} which corresponds to collective atomic vibrations of ZIF-8.

Moving to MOF-5, we observe phonon bands up to 1650 cm^{-1} in the first spectral region and after 3000 cm^{-1} in the second spectral region. According to experimental IR/Raman spectra, Civalleri et al.[58] defined five

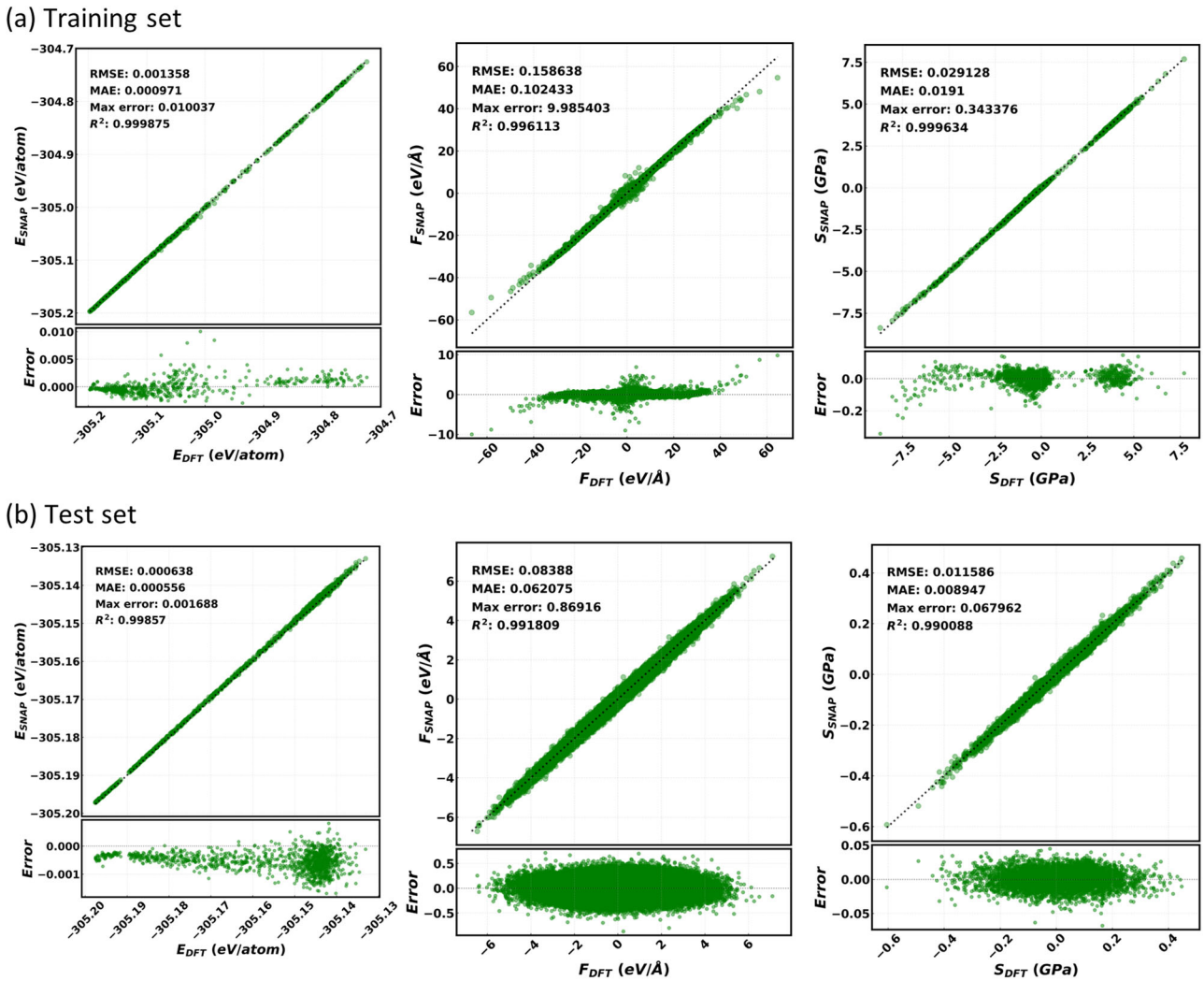

Fig. 5 | Performance of the trained SNAP model for MOF-5. Parity plots for energy (left-hand side panel), forces (middle panel) and virial-stress components (right-hand side panel) of the MOF-5 SNAP, computed over the training (596 configurations – upper panel) and test set (1191 configurations – lower panels). The RMSE for the test set are 0.6 meV/atom, 84 meV/Å and 11.6 MPa respectively for energy, forces and virial-stress components.

spectral regions for MOF-5: (i) 2900–3100 cm^{-1} (due to C-H stretching in phenylene), (ii) 1300–1650 cm^{-1} (carboxylate C=O and phenylene C=C stretching, and C-H bending vibrations), (iii) 600–1200 cm^{-1} (in-plane and out-of-plane deformation of phenylene ring including C-H groups), (iv) 200–600 cm^{-1} (due to Zn-O stretching and bending), (v) below 200 cm^{-1} (due to collective atomic vibrations and lattice modes). The PVDOS (Supplementary Figs. 13–14) reveals C_c-H stretching frequencies between 3050–3250 cm^{-1}, which are slightly outside the 2900–3100 cm^{-1} range, but are consistent with the frequencies obtained with the neural-network potentials simulations of Tayfuroglu et al.[37] (3126.4 and 3138.9 cm^{-1}) and Eckhoff et al. (3104/3148 cm^{-1}). In the PVDOS between 1250–1650 cm^{-1} we observe several peaks corresponding to carboxylate C=O, phenylene C=C, and C_c-H vibrations. These are in good agreement with experimental[59] C-O vibrational frequencies, 1377 and 1585 cm^{-1}, and previous simulation[37] results. Furthermore, we observe vibrational frequencies at 486–493 cm^{-1} and 556 cm^{-1} for O_4-Zn and frequencies in the 426–443 cm^{-1} range for O_2-Zn, which are close to the experimental[53,59] Zn-O IR frequency at 523 cm^{-1}. We also observe lower O_2-Zn frequencies at 263 cm^{-1} and 363 cm^{-1}, which are consistent with above mentioned spectral regions of the Zn-O stretching and bending modes. We further find various peaks below 200 cm^{-1}, which are attributed to collective atomic vibrations including both the metal nodes and organic linkers.

Free energy barrier for rotation of the MOF-5 phenylene rings

In order to unravel the internal dynamics of a MOF, it is essential to develop an understanding of the free energy barriers for the internal dynamics of different groups[4,60–63]. Such barriers can be studied with our MLP, which should be able to reliably map the atomic environments in the transition-state zone of the phase space. In MOF-5 the phenylene rings do not have a significant steric hinderance, however, significant interaction with the neighboring atoms creates a barrier to their rotational dynamics along the central axis. Therefore, in our MD simulations for MOF-5 at room temperature we did not observed rotation of any phenylene ring.

We have then analysed the distribution of the dihedral angles in MOF-5 (see Supplementary Fig. 9) in the training set configurations (generated with our temperature-driven active leaning algorithm at temperatures comprised between 100–1000 K). We have found that the training set contains configurations with all possible values of the C_c-C_b-C_a-O_2 dihedral angles (−180° to 180°). Thus, with our approach, the trained SNAP can reliably map atomic environments near the transition state corresponding to the rotational barrier. This motivates us to quantify the free-energy barrier for phenylene ring rotation in MOF-5. Earlier simulation work returned an energy barrier of 0.508/0.491 eV[32] and 0.58/0.65 eV[37] for the rotation of the phenylene ring in MOF-5, although these studies did not consider entropy effects. In order to evaluate the free-energy barrier (which includes both

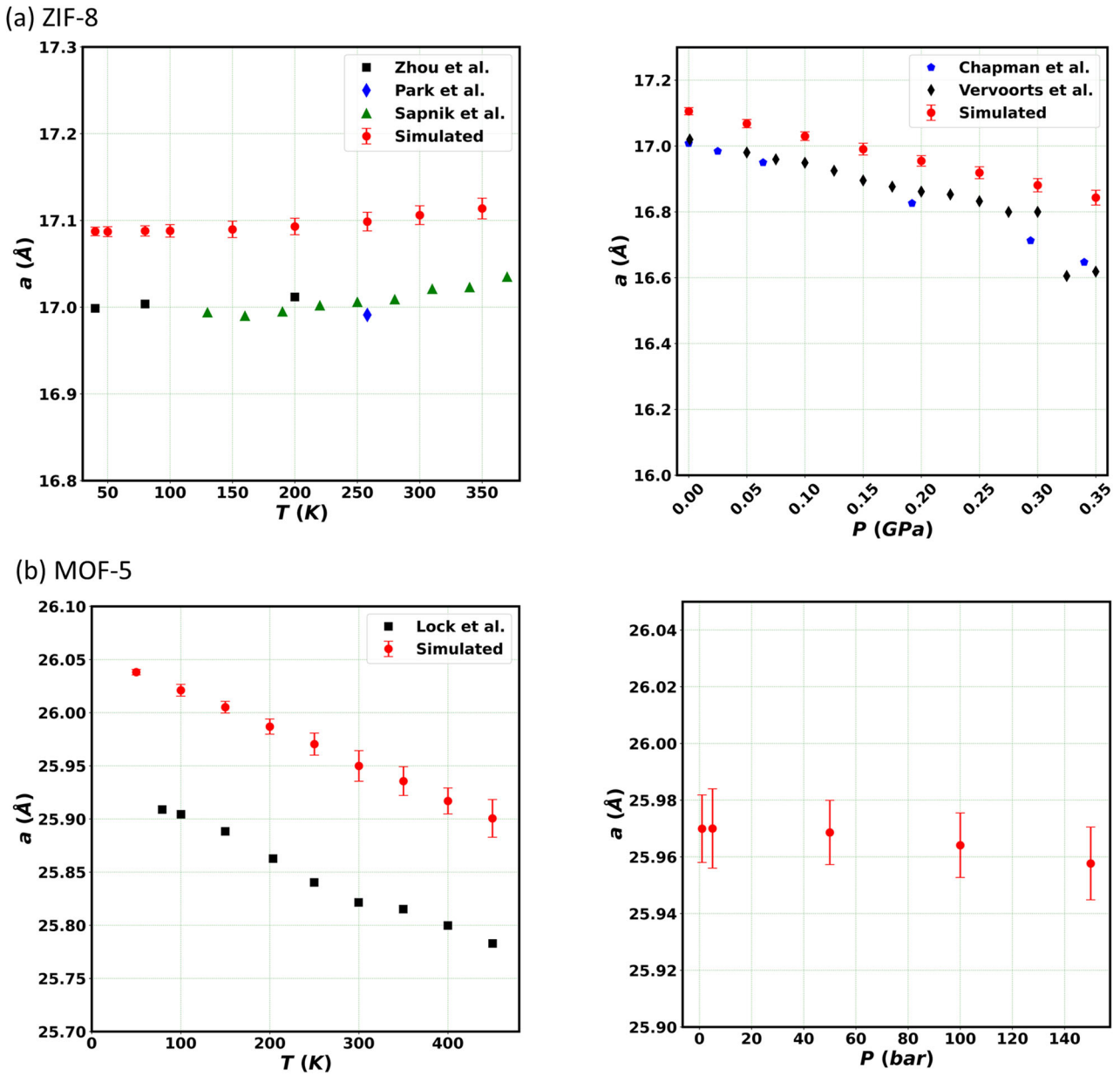

Fig. 6 | Simulated unit cell parameter for the two MOFs investigated as a function of temperature and pressure. Panel (**a**) is for ZIF-8 and (**b**) for MOF-5, with the results on the left-hand side panels concerning the temperature dependence and those on the right-hand side concerning the pressure. In the various panel we include a comparison with available experimental data (ZIF-8: Zhou et al.[83], Park et al.[45], Sapnik et al.[49], Chapman et al.[84], and Vervoorts et al.[85]; MOF-5: Lock et al.[52]). For ZIF-8 (MOF-5), the pressure is kept fixed at 1 bar during the temperature scan, while the temperature is kept fixed at 300 K (250 K) during the pressure scan. The error bars over the computed quantities correspond to the variance over the MD trajectory.

energy and entropy contributions) for the rotation of the phenylene ring, here we perform well-tempered metadynamics (WTM) simulations. In the WTM calculation, we consider two dihedral angles (ϕ_1 and ϕ_2) at each side of the phenylene ring as collective variables (see Fig. 8). All the WTM simulations are performed in the *NVT* ensemble, therefore, the resultant free energy corresponds to the Helmholtz free energy. Additional details about the WTM simulation are given in Methods section and Supplementary Note 3.

The free energy profile as a function of both collective variables is shown in Fig. 8a. The stable states in the rotation (blue regions) are separated by transition states (orange regions). If we consider the simplified case where the MOF-5 structure is rigid and only the rotational motion of the phenylene rings is allowed, one will have a negative correlation between the two collective variables and the only free-energy variation will be on the diagonal line of the 2D mesh of Fig. 8a. Along this diagonal, we compute a free energy barrier of more than 0.6 eV. Since the MOF-5 structure is flexible, the oxygen atoms wobble with respect to the Zn ones, a feature that relaxes the negative correlation between the collective variables and makes the free energy profile broader. This wobbling results in a transition path presenting a lower free-energy barrier than that along diagonal path, hence the wobbling of oxygen atoms helps in the rotation of phenylene rings of MOF-5 (see Supplementary Movie 1). The free-energy profile as a function of one dihedral angle, Fig. 8c, is finally obtained by integrating the effect of other angle[4] and a rotation free-energy barrier of 0.46 eV is thus computed. This value is close to the experimental one[63] of 0.49 eV, indicating once again the excellent quality of our interatomic potential.

Discussion

In the last few years, the application of MLPs to the study of MOFs has received a growing attention. In past studies, to select the training set

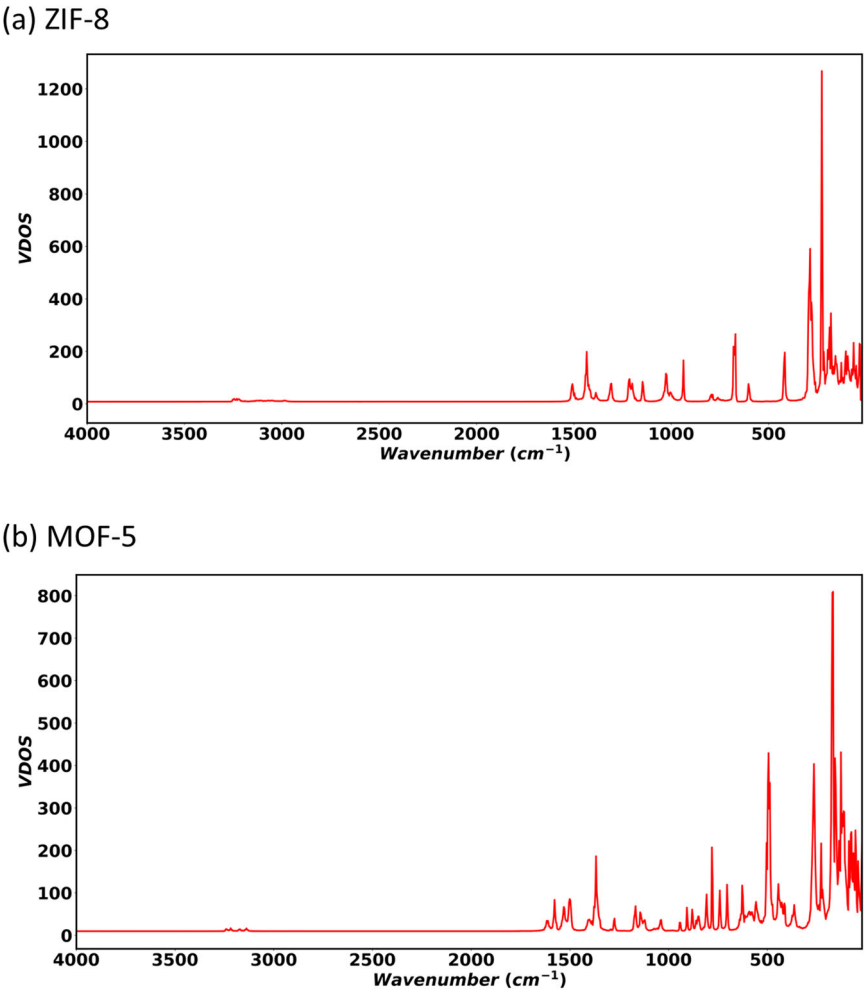

Fig. 7 | Simulated vibrational density of states (VDOS) for the two MOFs investigated. Panel (**a**) is for ZIF-8 and (**b**) for MOF-5. The corresponding partial VDOSs for each atom types (as defined in Fig. 1) are shown in Supplementary Figs. 11–14.

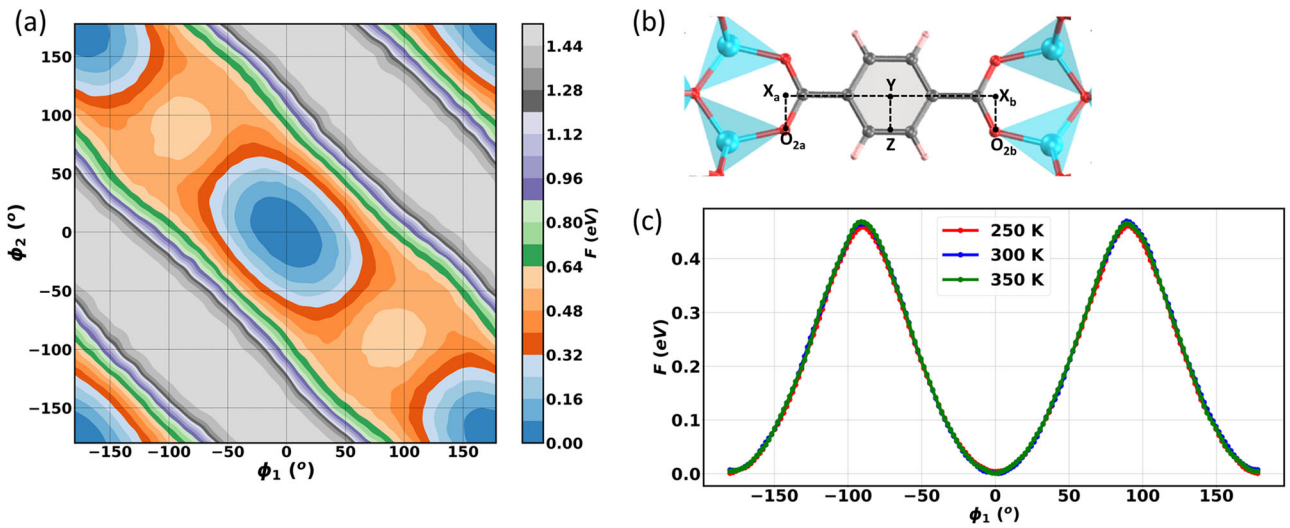

Fig. 8 | Investigation of the MOF-5 phenylene ring rotation performed with the trained SNAP. a Free energy profile for rotation of a phenylene ring in MOF-5 as a function of the two dihedral angle collective variables ϕ_1 and ϕ_2. The free energy minimum is shown in blue, and the barriers is shown in orange. **b** Schematic showing the collective variables, $\phi_1 = \phi_{O_{2a}X_aYZ}$ and $\phi_2 = \phi_{O_{2b}X_bYZ}$. **c** Free energy for the phenylene ring rotation as a function of one of the collective variables at different temperatures. A rotational barrier of 0.46 eV is observed for rotation of a phenylene ring in MOF-5.

configurations for the development of MLPs, two types of approaches have been reported. The first one involves the use of a configuration-selection metric, such as the model deviation (e.g., DP-GEN[64]) or the uncertainty quantification[26,65], to decide upon the inclusion of a configuration in the training set. Another approach uses biased simulation (e.g., metadynamics) and selects configurations separated by 1 ps simulation time to avoid correlation and to ensure diversity, without considering any particular configuration selection metric[36]. Here we focus primarily on bonded atomic systems (such as MOFs, small molecules, etc.). Therefore, our active learning algorithm relies on internal coordinates (bonds, angles, dihedrals), and cell parameters (for periodic configurations) and uses these as a configuration selection metric. The distribution of these values also gives an idea about the diversity of the training set. We avoid correlation and ensure diversity by selecting configurations from short (around 1 ps long) MD simulations (starting with different initial structure and velocities) at increasing temperatures. In our approach, we have not used any biased simulations, where the configuration space unravels along only a few collective variables. Instead, we vary the temperature, a strategy that allows the configuration space to expand in all possible directions. Geometries selected from this approach contribute to develop an effective MLP, which allows us to explore the configuration space in the direction of any feasible collective variables (as shown for the rotation of phenylene ring) and study the relevant transition states.

The complexity of the MLP training process can be understood by considering a typical MOF with N_a atoms in the unit cell. The DFT calculation of energy, forces and stress of N_c configurations will result in $(3*N_a+7)*N_c$ data points. This means that for a MOF with 400 atoms in the unit cell and 500–1000 configurations, there will be around 6–12 $\times 10^5$ data points. Then the data are used to fit the MLP model. Here, we have used SNAP, a MLP linear model constructed over only a few hundreds parameters (392 in our case). The training of a few hundreds parameters on such a large training set can be performed just on a laptop in a few minutes. This contrasts the training process of neural-network potential models, such as NequIP[36,66] and MACE[67,68], which requires the determination of a large number of parameters (of the order of 10^5–10^6) and it is usually performed on high-memory graphical-processing units in a time comprised between a few hours to a day. Having a large number of parameters, these deep-learning MLP models require more extended training data sets, but they are typically more accurate (e.g., force error ~ 30 meV/Å) than SNAP (force error ~ 60 meV/Å). The typical inference times are then more difficult to compare. In general, since linear models can be considered as a single-layer neural network, they are quicker to run. However, the final running time depends strongly on the time required to calculate the structure descriptors, which can vary widely depending on the specific implementations. In our MD simulations, we obtained speed of 0.35 s per MD step (with SNAP and D3 corrections) on a single core. Finally, as we have demonstrated here, our approach is general and can be widely deployed to construct high-performing MLPs at a low computational cost to accurately study the internal dynamics of MOFs. It is likely that the same SNAP is not able to describe bond-breaking events (at extreme temperature and pressure conditions) and other phenomena involving chemical reactions, such as adsorption and catalysis. This, however, is not an intrinsic limitation of SNAP, as of deep-learning MLPs, but rather depends on the specific configurations included in the training set. In the future, we will explore the use of SNAP for the study of such phenomena, including diffusion of gas molecules[34,42] in MOFs and chemical reactions.

Methods

Density functional theory (DFT) calculations

The QUICKSTEP[69] module of the CP2K[70] package is used for all the DFT calculations. Within this approach, the Kohn–Sham molecular orbitals are expanded over a linear combination of atom-centered Gaussian-type orbitals. All atoms are described using the MOLOPT basis set in combination with norm-conserving Goedecker–Teter–Hutter[71] (GTH) pseudopotentials. The Perdew–Burke–Ernzerhof (PBE)[72] exchange-correlation functional is employed throughout and the electron density is written over an auxiliary plane-wave basis set with appropriate energy cut-off (different for different types of calculations). The orbital transformation approach is used to find the solution of the Kohn–Sham equations and the self-consistent field (SCF) convergence of both the outer and inner loops is achieved with an accuracy of 10^{-7} Hartree.

Here, DFT is used for calculating energy, forces, and stress tensor values for a given atomic configuration, data that are used to construct the SNAP. In these calculations, DFT-D3[73] corrections are not included, and an energy cut-off of 1000 Ry is used. In addition, DFT is also employed to perform geometry and cell optimization of both ZIF-8 and MOF-5. In this case, dispersion corrections are included using the DFT-D3[73] approach with Becke-Johnson (BJ) damping[74] and an energy cut-off of 1000 Ry.

Ab-initio molecular dynamics simulations (AIMD) of ZIF-8

In order to generate the different atomic configurations of ZIF-8, AIMD simulations are performed. We first optimize the ZIF-8 unit cell (containing 276 atoms) using DFT with an energy cut-off of 600 Ry. Then, we perform CP2K AIMD simulations at different temperatures (100 to 1000 K in intervals of 100 K). The AIMD simulations are performed at constant temperature and pressure using a flexible cell. A timestep of 0.5 fs is used and all AIMD simulations are performed for 2000 steps (namely for 1 ps) at each temperature. To maintain the temperature, the Nose-Hoover thermostat is employed with time constant of 25 fs. The pressure is kept at 1 bar with the help of a barostat (as implemented in CP2K) with a time constant of 50 fs.

Molecular dynamics simulations

The trained SNAP[43] is used to perform molecular dynamics (MD) simulations of both ZIF-8 and MOF-5 using the LAMMPS[47] package. In addition to the SNAP, dispersion corrections are also included in these MD simulations using the Grimme's D3[73–75] approach with BJ damping. Namely, at each MD step, D3 corrections (to energy, forces, and virial-stress values) are added to the corresponding estimates from SNAP. Then, the atomic positions, velocities, and cell parameters are updated accordingly. A timestep of 1 fs is used in all LAMMPS MD simulations. For ZIF-8, an additional repulsive Ziegler-Biersack-Littmark (ZBL)[76] empirical potential is employed to create repulsion between the H_b-H_a, H_b-N, H_b-Zn, H_a-C_b, H_a-N, and H_a-Zn atom pairs. The inner and outer cut-off radius of the ZBL potential are chosen at 1.8 Å and 2.3 Å, respectively. For the atomic configurations contained in the training set of ZIF-8 the considered atomic pairs have a distance longer than 2.5 Å, resulting in zero ZBL contribution to the energy, forces, and stress. Therefore, the effect of ZBL is not subtracted from the training set data. In the case of MOF-5, no ZBL repulsion is considered.

The performance of SNAP against the lattice constants is estimated through MD simulations in the isothermal-isobaric ensemble, with constant number of particles, N, constant pressure, P, and temperature, T. In these MD simulations, a Nose-Hoover thermostat with 5 chains and time constant of 100 fs is used to maintain the temperature and a Nose-Hoover barostat (with time constant of 200 fs) balances the pressure. During these simulations, only the cell parameters are allowed to change, while the cell angles are kept constant.

Well-tempered metadynamics simulations

We perform well-tempered metadynamics (WTM)[77,78] simulation using the open-source community-developed Plumed[79,80] library patched with LAMMPS[47]. In the WTM simulations, a bias potential is added along the collective variables (CVs), to probe the free-energy landscape. In this work, we use two dihedral angles (described in main manuscript) as CVs. In the WTM simulations, we use a Gaussian width of 0.1 radian, an initial gaussian height of 0.05 eV, a biasfactor of 12, a grid spacing of 0.05 radian, and a bias deposition rate of 10^4 ns^{-1} (every 100 simulation steps). To maintain stability in the WTM simulations,

we apply upper walls on four Zn-O distances (near to the considered phenylene ring) at a value of 2.4 Å with force constant of 25 eV-Å^2. In addition to these, upper and lower walls are applied to two O_2-Zn-Zn-O_2 dihedral angles at a value of 0.85 radian with force constant of 25 eV-radian^2. We have performed WTM simulations at different temperatures in the isothermal ensemble.

Data availability

All datasets developed and used in this work are available via Zenodo[81].

Code availability

Our Python library and few examples of its use for training the SNAP and other different steps of temperature drive active learning are available at Github[82] in MOF_MLP_2024 repository.

Received: 14 May 2024; Accepted: 26 September 2024;
Published online: 08 October 2024

References

1. Kärger, J., Ruthven, D. M. & Theodorou, D. N. *Diffusion in Nanoporous Materials* (Wiley-VCH, 2012). https://doi.org/10.1002/9783527651276.
2. Roque-Malherbe, R. M. A. *Adsorption and Diffusion in Nanoporous Materials* (CRC Press, Taylor & Francis Group, 2007).
3. Schneemann, A. et al. Flexible metal-organic frameworks. *Chem. Soc. Rev.* **43**, 6062–6096 (**2014**).
4. Sharma, A., Dwarkanath, N. & Balasubramanian, S. Thermally activated dynamic gating underlies higher gas adsorption at higher temperatures in metal–organic frameworks. *J. Mater. Chem. A* **9**, 27398–27407 (2021).
5. Gu, C. et al. Design and control of gas diffusion process in a nanoporous soft crystal. *Science (80-.)* **363**, 387–391 (2019).
6. Dong, Q. et al. Tuning gate-opening of a flexible metal–organic framework for ternary gas sieving separation. *Angew. Chemie Int. Ed* **59**, 22756–22762 (2020).
7. Zhu, A.-X. et al. Tuning the gate-opening pressure in a switching Pcu coordination network, X-Pcu-5-Zn, by pillar-ligand substitution. *Angew. Chemie Int. Ed.* **131**, 18212–18217 (2019).
8. Coudert, F.-X. Responsive metal–organic frameworks and framework materials: under pressure, taking the heat, in the spotlight, with friends. *Chem. Mater.* **27**, 1905–1916 (2015).
9. Rogge, S. M. J., Waroquier, M. & Van Speybroeck, V. Unraveling the thermodynamic criteria for size-dependent spontaneous phase separation in soft porous crystals. *Nat. Commun.* **10**, 4842 (2019).
10. Vandenhaute, S., Rogge, S. M. J., Van Speybroeck, V. Large-scale molecular dynamics simulations reveal new insights into the phase transition mechanisms in MIL-53(Al). *Front. Chem*. 9 https://doi.org/10.3389/fchem.2021.718920 (2021).
11. Schaper, L., Keupp, J. & Schmid, R. Molecular dynamics simulations of the breathing phase transition of MOF nanocrystallites II: explicitly modeling the pressure medium. *Front. Chem*. 9 https://doi.org/10.3389/fchem.2021.757680 (2021).
12. Fan, D., Ozcan, A., Lyu, P. & Maurin, G. Unravelling negative in-plane stretchability of 2D MOF by large scale machine learning potential molecular dynamics. *arXiv*, No. arXiv:2307.15127. https://doi.org/10.48550/arXiv.2307.15127 (2023).
13. Li, P. & Merz, K. M. Jr. Metal ion modeling using classical mechanics. *Chem. Rev.* **117**, 1564–1686 (2017).
14. Rappe, A. K., Casewit, C. J., Colwell, K. S., Goddard, W. A. & Skiff, W. M. UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations. *J. Am. Chem. Soc.* **114**, 10024–10035 (1992).
15. Addicoat, M. A., Vankova, N., Akter, I. F. & Heine, T. Extension of the universal force field to metal–organic frameworks. *J. Chem. Theory Comput.* **10**, 880–891 (2014).
16. Coupry, D. E., Addicoat, M. A. & Heine, T. Extension of the universal force field for metal–organic frameworks. *J. Chem. Theory Comput.* **12**, 5215–5225 (2016).
17. Mayo, S. L., Olafson, B. D. & Goddard, W. A. DREIDING: a generic force field for molecular simulations. *J. Phys. Chem.* **94**, 8897–8909 (1990).
18. Bristow, J. K., Skelton, J. M., Svane, K. L., Walsh, A. & Gale, J. D. A general forcefield for accurate phonon properties of metal–organic frameworks. *Phys. Chem. Chem. Phys.* **18**, 29316–29329 (2016).
19. Vanduyfhuys, L. et al. Extension of the QuickFF force field protocol for an improved accuracy of structural, vibrational, mechanical and thermal properties of metal–organic frameworks. *J. Comput. Chem.* **39**, 999–1011 (2018).
20. Dubbeldam, D., Walton, K. S., Vlugt, T. J. H. & Calero, S. Design, parameterization, and implementation of atomic force fields for adsorption in nanoporous materials. *Adv. Theory Simulations* **2**, 1900135 (2019).
21. Dürholt, J. P., Fraux, G., Coudert, F.-X. & Schmid, R. Ab initio derived force fields for zeolitic imidazolate frameworks: MOF-FF for ZIFs. *J. Chem. Theory Comput.* **15**, 2420–2432 (2019).
22. Weng, T. & Schmidt, J. R. Flexible and transferable ab initio force field for zeolitic imidazolate frameworks: ZIF-FF. *J. Phys. Chem. A* **123**, 3000–3012 (2019).
23. Bureekaew, S. et al. Flexible first-principles derived force field for metal-organic frameworks. *Phys. status solidi* **250**, 1128–1141 (2013).
24. Behler, J. Perspective: machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901 (2016).
25. Behler, J. Four generations of high-dimensional neural network potentials. *Chem. Rev.* **121**, 10037–10072 (2021).
26. Kulichenko, M. et al. Uncertainty-driven dynamics for active learning of interatomic potentials. *Nat. Comput. Sci.* **3**, 230–239 (2023).
27. Friederich, P., Häse, F., Proppe, J. & Aspuru-Guzik, A. Machine-learned potentials for next-generation matter simulations. *Nat. Mater.* **20**, 750–761 (2021).
28. Domina, M., Patil, U., Cobelli, M. & Sanvito, S. Cluster expansion constructed over Jacobi-Legendre polynomials for accurate force fields. *Phys. Rev. B* **108**, 94102 (2023).
29. Behler, J. & Csányi, G. Machine learning potentials for extended systems: a perspective. *Eur. Phys. J. B* **94**, 142 (2021).
30. Mishin, Y. Machine-learning interatomic potentials for materials science. *Acta Mater* **214**, 116980 (2021).
31. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
32. Eckhoff, M. & Behler, J. From molecular fragments to the bulk: development of a neural network potential for MOF-5. *J. Chem. Theory Comput.* **15**, 3793–3809 (2019).
33. Achar, S. K., Wardzala, J. J., Bernasconi, L., Zhang, L. & Johnson, J. K. Combined deep learning and classical potential approach for modeling diffusion in UiO-66. *J. Chem. Theory Comput.* **18**, 3593–3606 (2022).
34. Zheng, B. et al. Quantum informed machine-learning potentials for molecular dynamics simulations of CO2's chemisorption and diffusion in Mg-MOF-74. *ACS Nano* **17**, 5579–5587 (2023).
35. Yu, Y., Zhang, W. & Mei, D. Artificial neural network potential for encapsulated platinum clusters in MOF-808. *J. Phys. Chem. C* **126**, 1204–1214 (2022).
36. Vandenhaute, S., Cools-Ceuppens, M., DeKeyser, S., Verstraelen, T. & Van Speybroeck, V. Machine learning potentials for metal-organic frameworks using an incremental learning approach. *npj Comput. Mater*. **9**, https://doi.org/10.1038/s41524-023-00969-x (2023).
37. Tayfuroglu, O., Kocak, A. & Zorlu, Y. A neural network potential for the IRMOF series and its application for thermal and mechanical behaviors. *Phys. Chem. Chem. Phys.* **24**, 11882–11897 (2022).

38. Shaidu, Y., Smith, A., Taw, E. & Neaton, J. B. Carbon capture phenomena in metal-organic frameworks with neural network potentials. *PRX Energy* **2**, 023005 (2023).
39. Ying, P. et al. Sub-micrometer phonon mean free paths in metal–organic frameworks revealed by machine learning molecular dynamics simulations. *ACS Appl. Mater. Interfaces* **15**, 36412–36422 (2023).
40. Liu, S. et al. Machine learning potential for modelling H2 adsorption/diffusion in MOF with open metal sites. *arXiv*, No. arXiv:2307.15528. https://doi.org/10.48550/arXiv.2307.15528 (2023).
41. Wieser, S. & Zojer, E. Machine learned force-fields for an Ab-Initio quality description of metal-organic frameworks. *arXiv*. arXiv:2308.01278. https://doi.org/10.48550/arXiv.2308.01278 (2023).
42. Zheng, B. et al. Simulating CO2 diffusivity in rigid and flexible Mg-MOF-74 with machine-learning force fields. *APL Mach. Learn.* **2**, 26115 (2024).
43. 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).
44. Lunghi, A. & Sanvito, S. A unified picture of the covalent bond within quantum-accurate force fields: from organic molecules to metallic complexes' reactivity. *Sci. Adv.* **5**, 1–8 (2019).
45. Park, K. S. et al. Exceptional chemical and thermal stability of zeolitic imidazolate frameworks. *Proc. Natl. Acad. Sci. USA* **103**, 10186–10191 (2006).
46. Li, H., Eddaoudi, M., O'Keeffe, M. & Yaghi, O. M. Design and synthesis of an exceptionally stable and highly porous metal-organic framework. *Nature* **402**, 276–279 (1999).
47. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.* **117**, 1–19 (1995).
48. Thompson, A. P. et al. LAMMPS - a flexible simulation tool for particle-based materials modeling at the atomic, meso, and continuum scales. *Comput. Phys. Commun.* **271**, 108171 (2022).
49. Sapnik, A. F., Geddes, H. S., Reynolds, E. M., Yeung, H. H.-M. & Goodwin, A. L. Compositional inhomogeneity and tuneable thermal expansion in mixed-Metal ZIF-8 analogues. *Chem. Commun.* **54**, 9651–9654 (2018).
50. Burtch, N. C. Engineering precisely controlled negative and zero thermal expansion be- haviors in metal-organic frameworks. *United States Sandia Natl. Lab. Rep*. https://doi.org/10.2172/156144 (2019).
51. Burtch, N. C. et al. Negative thermal expansion design strategies in a diverse series of metal–organic frameworks. *Adv. Funct. Mater.* **29**, 1904669 (2019).
52. Lock, N. et al. Elucidating negative thermal expansion in MOF-5. *J. Phys. Chem. C* **114**, 16181–16186 (2010).
53. Hadjiivanov, K. I. et al. Power of infrared and Raman spectroscopies to characterize metal-organic frameworks and investigate their interaction with guest molecules. *Chem. Rev.* **121**, 1286–1424 (2021).
54. Ahmad, M. et al. ZIF-8 vibrational spectra: peak assignments and defect signals. *ACS Appl. Mater. Interfaces* **16**, 27887–27897 (2024).
55. Kumari, G., Jayaramulu, K., Maji, T. K. & Narayana, C. Temperature induced structural transformations and gas adsorption in the zeolitic imidazolate framework ZIF-8: a Raman study. *J. Phys. Chem. A* **117**, 11006–11012 (2013).
56. Xu, B. et al. Monitoring thermally induced structural deformation and framework decomposition of ZIF-8 through in situ temperature dependent measurements. *Phys. Chem. Chem. Phys.* **19**, 27178–27183 (2017).
57. Ryder, M. R. et al. Identifying the role of terahertz vibrations in metal-organic frameworks: from gate-opening phenomenon to shear-driven structural destabilization. *Phys. Rev. Lett.* **113**, 215502 (2014).
58. Civalleri, B., Napoli, F., Noël, Y., Roetti, C. & Dovesi, R. Ab-initio prediction of materials properties with CRYSTAL: MOF-5 as a case study. *CrystEngComm* **8**, 364–371 (2006).
59. Tzitzios, V. et al. Solvothermal synthesis, nanostructural characterization and gas cryo-adsorption studies in a metal–organic framework (IRMOF-1) material. *Int. J. Hydrogen Energy* **42**, 23899–23907 (2017).
60. Pakhira, S. Rotational dynamics of the organic bridging linkers in metal–organic frameworks and their substituent effects on the rotational energy barrier. *RSC Adv.* **9**, 38137–38147 (2019).
61. Tafipolsky, M., Amirjalayer, S. & Schmid, R. Ab initio parametrized MM3 force field for the metal-organic framework MOF-5. *J. Comput. Chem.* **28**, 1169–1176 (2007).
62. Vogelsberg, C. S. et al. Ultrafast rotation in an amphidynamic crystalline metal organic framework. *Proc. Natl. Acad. Sci. USA* **114**, 13613–13618 (2017).
63. Gould, S. L., Tranchemontagne, D., Yaghi, O. M. & Garcia-Garibay, M. A. Amphidynamic character of crystalline MOF-5: rotational dynamics of terephthalate phenylenes in a free-volume, sterically unhindered environment. *J. Am. Chem. Soc.* **130**, 3246–3247 (2008).
64. Zhang, L., Lin, D.-Y., Wang, H., Car, R. & E, W. Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Mater.* **3**, 23804 (2019).
65. Briganti, V. & Lunghi, A. Efficient generation of stable linear machine-learning force fields with uncertainty-aware active learning. *Mach. Learn. Sci. Technol.* **4**, 35005 (2023).
66. Batzner, S. et al. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453 (2022).
67. Batatia, I., Kovács, D. P., Simm, G. N. C., Ortner, C. & Csányi, G. MACE: higher order equivariant message passing neural networks for fast and accurate force fields. *arXiv* https://doi.org/10.48550/arXiv.2206.07697 (2023).
68. Kovács, D. P., Batatia, I., Arany, E. S. & Csányi, G. Evaluation of the MACE force field architecture: from medicinal chemistry to materials science. *J. Chem. Phys.* **159**, 44118 (2023).
69. Vandevondele, J. et al. Quickstep: fast and accurate density functional calculations using a mixed Gaussian and plane waves approach. *Comput. Phys. Commun.* **167**, 103–128 (2005).
70. Kühne, T. D. et al. CP2K: an electronic structure and molecular dynamics software package -quickstep: efficient and accurate electronic structure calculations. *J. Chem. Phys* **152**, 194103 (2020).
71. Goedecker, S. & Teter, M. Separable dual-space Gaussian pseudopotentials. *Phys. Rev. B - Condens. Matter Mater. Phys.* **54**, 1703–1710 (1996).
72. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
73. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate Ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **132**, 154104 (2010).
74. Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.* **32**, 1456–1465 (2011).
75. DFT-D3 https://www.chemie.uni-bonn.de/pctc/mulliken-center/software/dft-d3, https://github.com/loriab/dftd3.
76. Ziegler, J. F., Biersack, J. P. & Littmark, U. The stopping and range of ions in matter. *Pergamon* **1**, https://en.wikipedia.org/wiki/Stopping_and_Range_of_Ions_in_Matter (1985).
77. Laio, A. & Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. USA.* **99**, 12562–12566 (2002).
78. Barducci, A., Bussi, G. & Parrinello, M. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Phys. Rev. Lett.* **100**, 020603 (2008).
79. Bonomi, M. et al. Promoting transparency and reproducibility in enhanced molecular simulations. *Nat. Methods* **16**, 670–673 (2019).
80. Tribello, G. A., Bonomi, M., Branduardi, D., Camilloni, C. & Bussi, G. PLUMED 2: new feathers for an old bird. *Comput. Phys. Commun.* **185**, 604–613 (2014).

81. Sharma, A. & Sanvito, S. Quantum-accurate machine learning potentials for metal-organic frameworks using temperature driven active learning. *Zenodo* https://doi.org/10.5281/zenodo.11176257 (2024).
82. MOF_MLP_2024 https://github.com/asharma-ms/MOF_MLP_2024.
83. Zhou, W., Wu, H., Udovic, T. J., Rush, J. J. & Yildirim, T. Quasi-free methyl rotation in zeolitic imidazolate framework-8. *J. Phys. Chem. A* **112**, 12602–12606 (2008).
84. Chapman, K. W., Halder, G. J. & Chupas, P. J. Pressure-induced amorphization and porosity modification in a metal−organic framework. *J. Am. Chem. Soc.* **131**, 17546–17547 (2009).
85. Vervoorts, P., Burger, S. & Hemmer, K. K. G. Revisiting the high-pressure properties of the metal-organic frameworks ZIF-8 and ZIF-67. *ChemRxiv* (2020).

Acknowledgements

The authors thank the Trinity Center for High Performance Computing (TCHPC) and the Irish Center for High-End Computing (ICHEC) for providing the computational resources. This work has been supported by Science Foundation Ireland through the Advanced Materials and BioEngineering Research (AMBER) (Grant: 12/RC/2278 − P2) and by the Qatar National Research Fund (Award: NPRP12C-0821-190017).

Author contributions

All calculations, including the creation of the density-functional-theory dataset, the training of the machine-learning model and the molecular dynamics simulations, have been performed by A.S.; The project was designed by A.S. and S.S.; S.S. provided supervision and financial support to the project. All authors wrote and reviewed the manuscript.

Competing interests

S.S. is an Associated Editor at npj Computational Materials. The authors declare no additional competing interest.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41524-024-01427-y.

Correspondence and requests for materials should be addressed to Abhishek Sharma or Stefano Sanvito.