ROYAL SOCIETY
OF CHEMISTRY

Check for updates

A universal machine learning model for the electronic density of states

Wei Bin How, [ID] Pol Febrer, [ID] Sanggyu Chong, [ID] Arslan Mazitov, [ID] Filippo Bigi, [ID] Matthias Kellner, [ID] Sergey Pozdnyakov [ID] and Michele Ceriotti [ID] *

In the last few years several "universal" interatomic potentials have appeared, using machine-learning approaches to predict energy and forces of atomic configurations with arbitrary composition and structure, with an accuracy often comparable with that of the electronic-structure calculations they are trained on. Here we demonstrate that these generally-applicable models can also be built to predict explicitly the electronic structure of materials and molecules. We focus on the electronic density of states (DOS), and develop PET-MAD-DOS, a rotationally unconstrained transformer model built on the Point Edge Transformer (PET) architecture, and trained on the Massive Atomic Diversity (MAD) dataset. We demonstrate our model's predictive abilities on samples from diverse external datasets, showing also that the DOS can be further manipulated to obtain accurate bandgap predictions. A fast evaluation of the DOS is especially useful in combination with molecular simulations probing matter in finite-temperature thermodynamic conditions. To assess the accuracy of PET-MAD-DOS in this context, we evaluate the ensemble-averaged DOS and the electronic heat capacity of three technologically relevant systems: lithium thiophosphate (LPS), gallium arsenide (GaAs), and a high entropy alloy (HEA). By comparing with bespoke models, trained exclusively on system-specific datasets, we show that our universal model achieves semi-quantitative agreement for all these tasks. Furthermore, we demonstrate that fine-tuning can be performed using a small fraction of the bespoke data, yielding models that are comparable to, and sometimes better than, fully-trained bespoke models.

1 Introduction

Machine learning (ML) methods are rapidly gaining popularity in the field of computational materials science due to their ability to predict material properties at a fraction of the cost of traditional *ab initio* methods, while maintaining comparable levels of accuracy.[1–3] ML models typically scale linearly with the system size, in contrast to *ab initio* methods that are usually more costly and exhibit poorer scaling behaviour,[4] which limits their usability for large or complex systems.

Early efforts in this domain were focused on highly specialized models, designed for specific properties in narrow regions of the chemical space. Examples of such early developments include interatomic potentials (IPs)[5,6] as well as models designed to predict bandgaps,[7–10] charge densities,[11] Hamiltonians,[12,13] nuclear magnetic resonance (NMR) spectra[14,15] or electronic density of states (DOS).[16,17] In recent years, there has been a shift towards developing universal models, *i.e.* models that are capable of generalizing well across a large fraction of the periodic table, spanning both molecules and extended

materials.[18–20] However, these efforts have been largely focused on constructing universal ML interatomic potentials (MLIPs) to enable stable molecular dynamics simulations across diverse chemistries. Lately, there has been growing interest in building universal ML models to predict other material properties beyond energies and forces, such as bandgaps,[21–24] Hamiltonians,[25,26] and the density of states.[27–29]

Recently, a new universal MLIP, PET-MAD,[30] has been introduced, reaching similar accuracies as existing state-of-the-art MLIPs for inorganic bulk systems while remaining reliable for molecules, organic materials and surfaces. The PET-MAD model employs the Point Edge Transformer (PET) architecture,[31] a transformer-based graph neural network that does not enforce rotational symmetry constraints, but learns to be equivariant to a high level of accuracy through data augmentation. PET-MAD was trained on the small (containing fewer than 100 000 structures) but extremely diverse Massive Atomic Diversity (MAD) dataset.[32] It encompasses both organic and inorganic systems, ranging from discrete molecules to bulk crystals. The dataset also includes randomized and heavily distorted structures to increase stability when performing complex atomistic simulations. Inspired by the success of the highly expressive PET architecture and highly diverse MAD dataset, we decided to apply this same combination to the

Laboratory of Computational Science and Modeling, Institut des Matériaux, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland. E-mail: michele. ceriotti@epfl.ch

prediction of the electronic density of states (DOS), a useful quantity for understanding the electronic properties of materials.

The DOS quantifies the distribution of available electronic states at each energy level and underlies many useful optoelectronic properties of a material, such as its conductivity, bandgap and optical absorption spectra.[33,34] These properties are highly relevant for applications like semiconductors and photovoltaic devices. Hence, the ability to easily obtain the DOS of a material can be instrumental for material discovery, paving the way for the development of better semiconductors or more efficient photovoltaics.[27] Furthermore, the DOS can also enhance MLIPs by accounting for finite temperature effects, such as the temperature dependent electronic free energy[35] or electronic heat capacity,[36] thereby broadening their utility.

In this work, we present PET-MAD-DOS, a universal machine-learning model for predicting the DOS, based on the PET architecture and MAD dataset. Uncertainty quantification (UQ) was also performed based on existing UQ methods[37,38] to provide a measure for the accuracy of the DOS predictions at different energies. We evaluate the performance of PET-MAD-DOS on atomistic benchmarks and ensemble quantities for a diverse set of scientifically interesting material systems, namely gallium arsenide (GaAs), lithium thiophosphate (LiPS), and high entropy alloys (HEA). We compare the ensemble quantities obtained using PET-MAD-DOS against bespoke models, i.e. PET models trained solely on those materials, and fine-tuned versions of PET-MAD-DOS for each material class. These bespoke models have roughly half the test-set error of PET-MAD-DOS. The fact that a model specialized for a single material is only twice as accurate as our universal predictor is

a testament to the robustness of PET-MAD-DOS. At the same time, having access to more accurate bespoke models trained on an entirely different specialized dataset allows us to assess the reliability of PET-MAD-DOS when using it in more complicated simulation workflows, whose validation by explicit electronic structure calculations would be prohibitively expensive.

2 Results

This section covers the performance of PET-MAD-DOS, our foundation DOS model based on the PET architecture and trained on the MAD dataset. We report the details of the model and its training in the Methods (Section 4). We first showcase the performance and generalizability of PET-MAD-DOS by evaluating the DOS predictions on different subsets of the MAD dataset and several public datasets. Afterwards, we show that the predicted DOS can be used to obtain accurate predictions of the bandgap. Finally, we demonstrate the utility of our model on three case-study materials by evaluating ensemble quantities derived from molecular dynamics (MD) trajectories. For these, we compared the performance of PET-MAD-DOS against that of (1) PET models trained solely on those systems and (2) the corresponding fine-tuned PET-MAD-DOS models.

2.1 Model performance

We evaluate the performance of PET-MAD-DOS both on the MAD test set and on samples from other popular atomistic datasets, covering a broad spectrum of systems from bulk inorganic systems to drug molecules. The MAD dataset was originally developed as a compact dataset to train universal
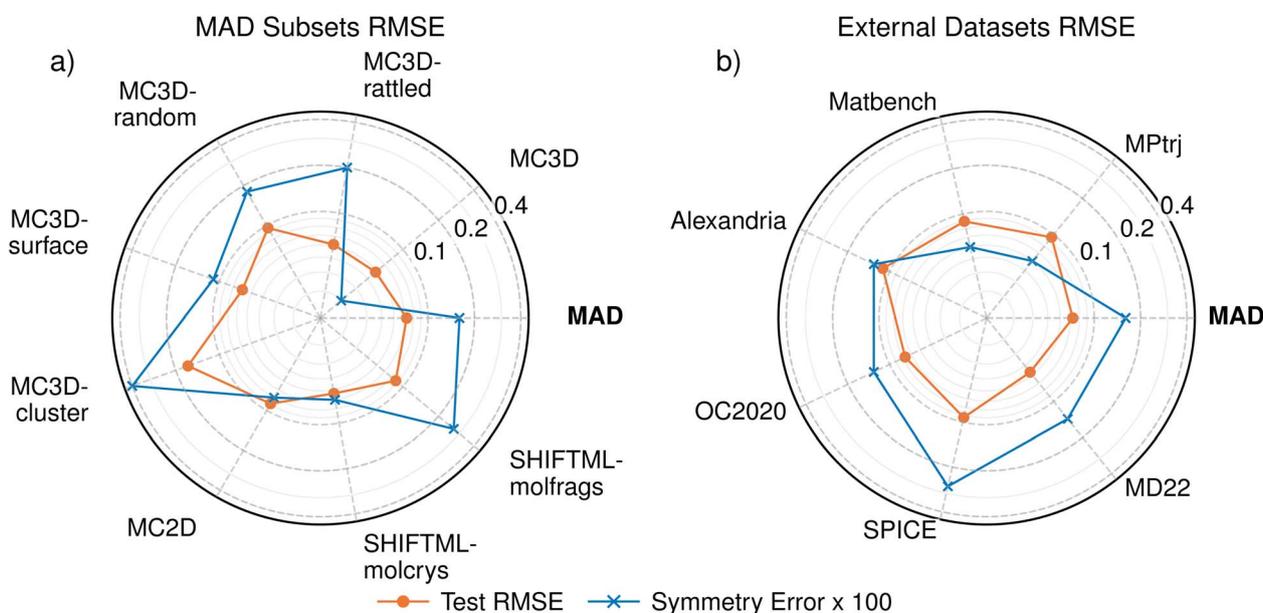

Fig. 1 Root mean square error (RMSE) of the DOS predictions (orange-line) on the test set of the MAD dataset across the different subsets (a) and the external datasets (b). The blue line shows the rotational discrepancy, arising from the fact that PET is rotationally unconstrained. The symmetry error is multiplied by 100 to plot it on the same scale as the test RMSE, which is two orders of magnitude higher. Both the RMSE and the symmetry error are scaled based on the number of electrons in the system and have units of $eV^{-0.5}$ electrons^{-1} state.

MLIPs, and is described in detail in ref. 32. It is divided into eight distinct subsets, which we summarize here:

MC3D & MC2D: materials Cloud 3D (33 596 structures) and 2D (2676 structures) crystal database respectively.[39,40]

MC3D-rattled: structures generated by adding Gaussian noise to the atomic positions of MC3D structures (30 044 structures).

MC3D-random: structures formed by randomizing the elemental composition of a subset of MC3D structures (2800 structures).

MC3D-surface: surfaces obtained by cleaving a subset of MC3D structures along random crystallographic planes with low Miller index (5589 structures).

MC3D-cluster: clusters formed by randomly subselecting two to eight atoms from some MC3D structures (9071 structures).

SHIFTML-molcrys & SHIFTML-molfrags: molecular crystals (8578 structures) and neutral molecular fragments (3241 structures) respectively from the SHIFTML dataset that is sampled from the Cambridge Structural Database.[41,42]

The samples from external datasets are recomputed using MAD DFT settings to maintain consistency between training and evaluation data. They come from six sources:

MPtrj: relaxation trajectories of bulk inorganic crystals dataset.[43]

Matbench: bulk inorganic crystals from the Materials Project Database.[44]

Alexandria: relaxation trajectories of bulk inorganic crystals as well as 2D and 1D systems.[45]

SPICE: drug-like molecules and peptides.[46]

MD22: molecular dynamics trajectories of peptides, DNA molecules, carbohydrates and fatty acids.[47]

OC2020 (S2EF): molecular relaxation trajectories on catalytically active surfaces.[48]

The errors of PET-MAD-DOS in these datasets are shown in Fig. 1, with further details of the error distributions in MAD illustrated in Fig. 2 which also provides a few representative example of DOS predictions, helping to relate the integrated
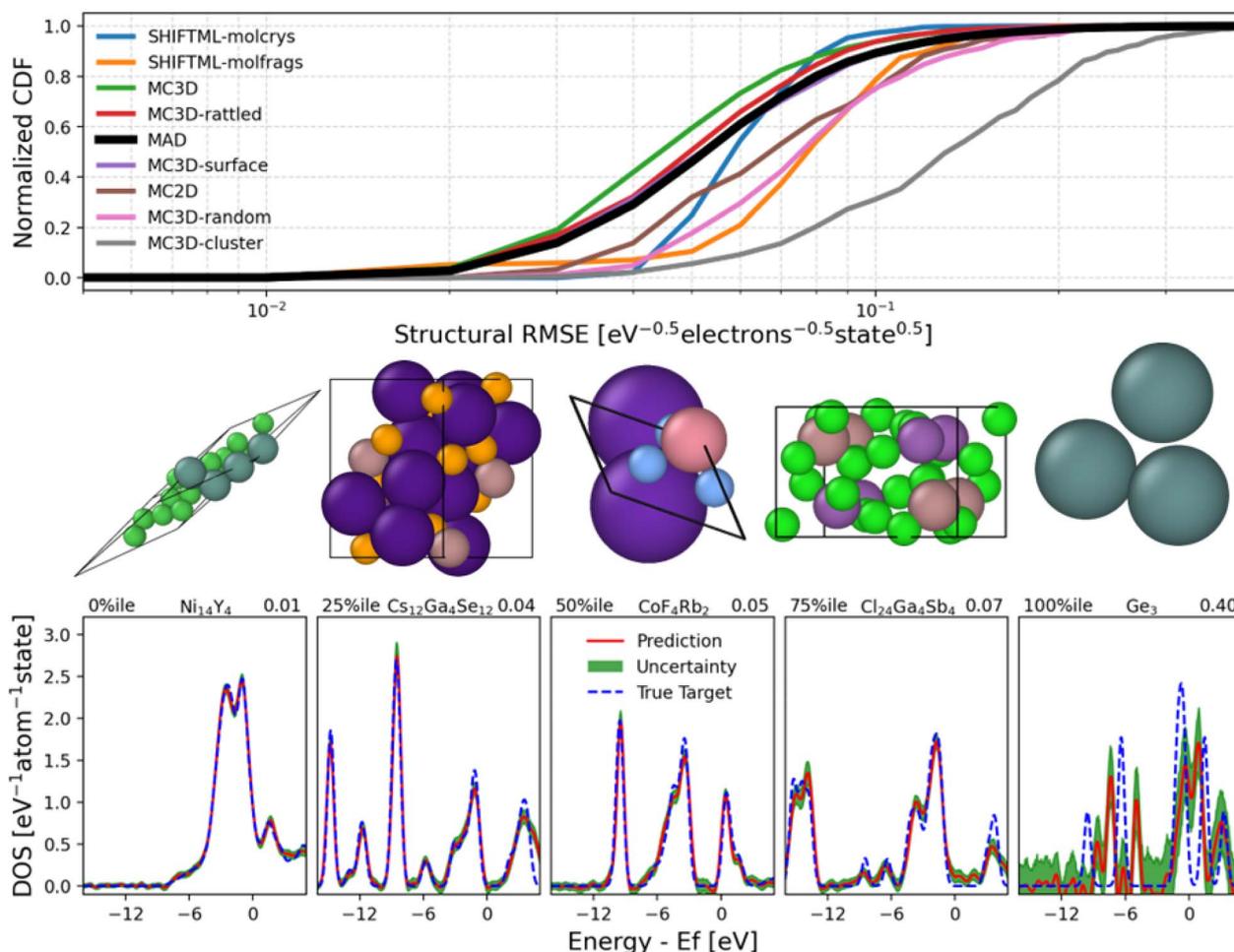

Fig. 2 Error distributions in the MAD test set. The top panel shows the normalized cumulative distribution function (CDF) of the RMSEs of each structure in each subset, represented by different colors, and the CDF of the entire MAD test subset in black. The bottom panel shows selected true DOS (blue dashed)/predicted DOS (red solid) comparisons from different parts of the MAD error distribution, for visual reference. Above each subplot, the atomic configuration of the corresponding structure is displayed. The green areas represent the uncertainty associated with the DOS prediction as estimated by the calibrated last-layer prediction rigidity (LLPR) ensembles. The routine computes the standard deviation σ associated with the prediction at each energy channel. The range of the x axis has been truncated to ease visualization of the DOS predictions and its corresponding uncertainties. The RMSE corresponding to each subplot in the bottom panel is at the top right corner.

errors to the visual quality of the predictions. Overall, the general performance trends of PET-MAD-DOS across the different datasets are similar to those of PET-MAD. For the MAD subsets, both models perform worst on the MC3D-random and MC3D-cluster subsets, likely due to the high chemical diversity in the subsets and the presence of several extreme cases of far-from-equilibrium configurations. The accuracy is especially poor for clusters, which have sharply-peaked DOS and often a highly nontrivial electronic structure. As shown in Fig. 2, the error distribution has a long tail, with a few high-error structures, but most of the structures having errors below 0.07 $eV^{-0.5}$ electrons^{-1} state. Considering the external datasets, Fig. 1b shows that PET-MAD-DOS performs best on MD22 and SPICE, which is consistent with the fact that the model performs better on the molecular part of the MAD dataset (SHIFTML subsets). Additionally, the performance of PET-MAD-DOS on the MAD dataset is comparable to that of the external datasets, highlighting both the chemical diversity of MAD and the ability of PET-MAD-DOS to capture the structure–property relationship in the extrapolative regime. Since the PET architecture does not impose any rotational constraints on the predictions, a rotated structure will not necessarily give the same prediction as the original structure despite the physical DOS being invariant to rotations. However, Fig. 1 shows that the rotational discrepancy is two orders of magnitude smaller than the RMSE of the DOS. Furthermore, recent works have shown that rotational discrepancies from rotationally unconstrained models have negligible impact on a model's performance in practical applications.[49] Therefore, the lack of rotational constraint for PET-MAD-DOS does not impact the reliability of the model.

In Fig. 2, we also provide the uncertainties that have been quantified at each energy channel using the standard deviation of the calibrated last-layer prediction rigidity (LLPR) ensembles.[37] Information regarding the construction of the LLPR ensemble can be found in Section 4.8 of the Methods. The quantified uncertainties correspond well with the error made by the model for the structures shown on the bottom of Fig. 2. Our LLPR-based uncertainty quantification (UQ) module is crucial for ensuring reliability in the model predictions, which is especially relevant for general-purpose models like PET-MAD-DOS as they are utilized in the "edge" cases where performance may deteriorate without warning. In particular for the DOS, the model's performance is inconsistent across energy channels, and thus our UQ module can be useful for identifying the model's confidence across different energy regions of the prediction.

2.2 Predicting the bandgaps

The bandgap plays a fundamental role in the optical and electronic properties of a material. Its magnitude provides insight into the electrical conductivity at different temperatures, as well as the wavelength of light that the material can absorb. Hence, predicting the bandgap can be very useful for material design in applications such as electronics, catalysis and photonics.

In this work, we define the bandgap as the difference between the valence band maximum (VBM) and the conduction

band minimum (CBM). To determine the bandgap from the DOS, one would normally first determine the Fermi level by finding the energy where the integrated DOS equals the total number of electrons in the system. The positions of the VBM and CBM can then be estimated to determine the bandgap. However, the application of this method to predicted DOS spectra poses several challenges. Although the DOS inside the bandgap should be zero, due to the use of Gaussian smearing to construct the target DOS, along with small prediction errors from the model, the DOS within the bandgap is often a small non-zero value. This introduces ambiguity in the choice of a threshold below which the DOS should be treated as zero. Another challenge is the determination of the Fermi level, which depends on the integrated DOS and therefore is very sensitive to accumulated errors. All these challenges are illustrated in Fig. 3 for $MgCl_2$, an insulator in the test set of MAD. The calculated Fermi level on the raw predicted DOS (red lines) is offset to the right of the gap by around 0.5 eV due to a slight underestimation of the integrated DOS. Since the Fermi level falls into a region with non-zero DOS, the physical interpretation is that $MgCl_2$ is a metal with no bandgap, which is qualitatively wrong. Even if the Fermi level was correctly determined, the oscillations in the predicted DOS (the most prominent one around −9 eV) would complicate the assessment of the gap's magnitude.

Given these issues, one may wonder whether the predicted DOS can be used to achieve the goals that motivated us to develop a DOS model in the first place. To this end, we
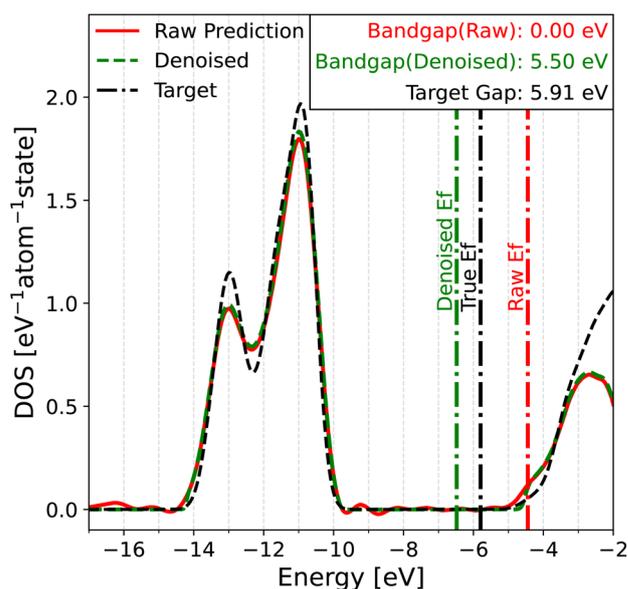
Fig. 3 Evaluation of the bandgap in $MgCl_2$, an insulator in the test set of MAD. The raw prediction of PET-MAD-DOS (solid red) is compared against that of the denoised prediction (dashed green) and true DOS (dash-dotted black). The colored vertical lines represent the Fermi level determined *via* integration of the corresponding DOS spectra. The target gap of 5.91 eV represents the HOMO−LUMO gap obtained from the underlying DFT calculation while the other bandgaps are obtained from the corresponding DOS spectra, using a threshold of 0.1 eV^{-1} atom^{-1} state.
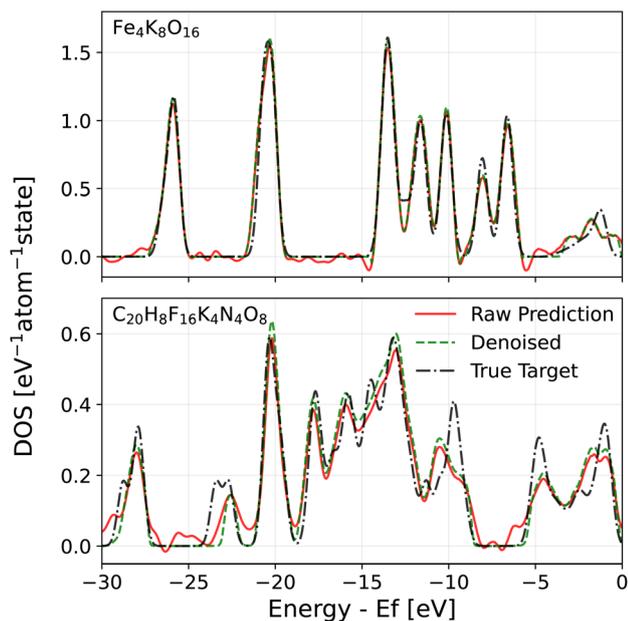
Fig. 4 Demonstration of the effects of denoising on two sample predictions on the MAD test set. The raw prediction of PET-MAD-DOS (solid red) is compared against that of the denoised prediction (dashed green) and true DOS (dash-dotted black). The x-axis is truncated to enhance visualization of the differences between each DOS.

developed two solutions. The first solution involves passing the raw DOS prediction through a denoising filter to eliminate model noise in gap regions. The denoised DOS is also scaled such that the DOS integrates to the correct number of electrons up to the Fermi level, which is predicted by a convolutional neural network (CNN) (see Section 4.6 in the Methods for

details). A demonstration of the denoising algorithm can be seen in both Fig. 3 and 4. Both figures show that the denoised prediction (green dashed line) exhibits virtually no oscillations in the gap regions, unlike the raw prediction (red solid line). For the case of $MgCl_2$ in Fig. 3, the bandgap obtained from the denoised DOS is much better thanks to the improved Fermi level determination and higher quality DOS predictions in the gap. The second solution relies on a fully data-driven approach: the raw predicted DOS is passed through a CNN to predict the bandgap directly. The idea behind this solution is that a trained CNN should be able to find a way of dealing with noise that outperforms our handcrafted denoising algorithm, at the cost of being less elegant. For both approaches, the point that the CNN is applied is crucial. PET-MAD-DOS predicts atomic contributions that are summed over the atom indices to produce the total DOS. It is at this point where the CNN, which introduces non-linearities, should be applied. Applying it at the level of individual atomic environments would amount to making the assumption that a global quantity such as the bandgap and position of the Fermi level can be written as a sum of atomic contributions. For the same reasons, the denoising filter is applied to the total DOS and not to individual atomic contributions.

The performance of each method's bandgap predictions is displayed in Fig. 5, accompanied by Tables I and II in the SI. For the MAD test set, the CNN method achieves MAE errors that are roughly 4× lower than the raw predictions and 2× lower than the denoised predictions. In general, using the CNN method achieves better accuracies for secondary quantities. For instance, when estimating the DOS at the Fermi level, the MAEs of the raw predictions, denoised DOS and CNN method are 0.15, 0.13 and 0.10 eV^{-1} $atom^{-1}$ state respectively. The results suggest
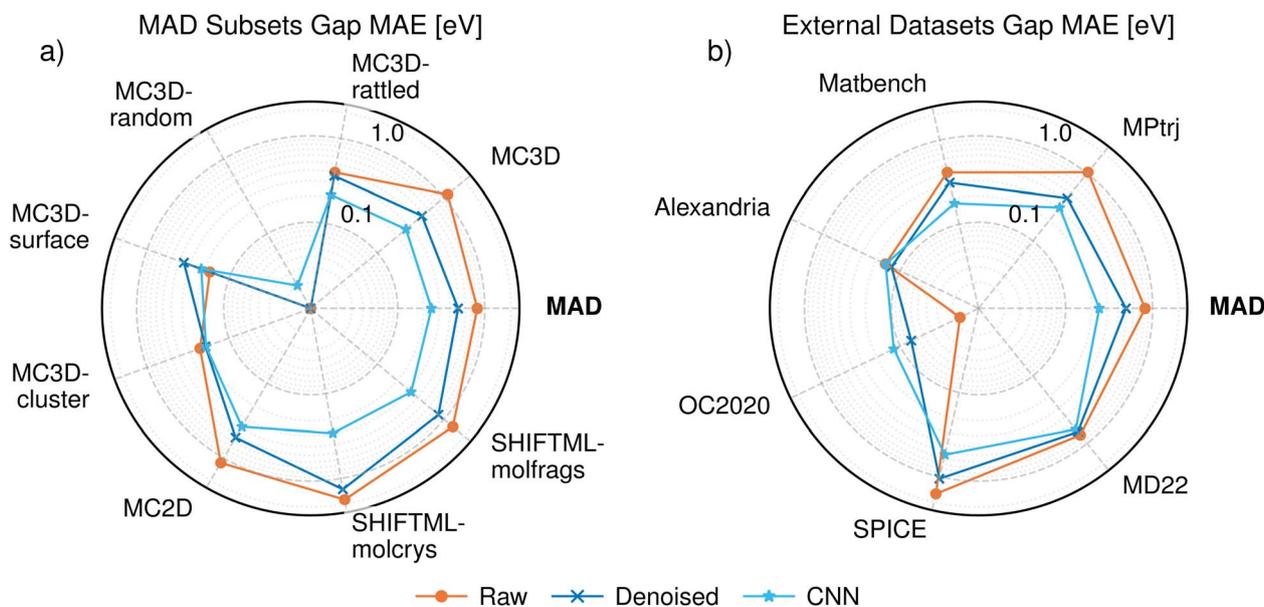

Fig. 5 Comparison of the mean absolute error (MAE) of the bandgap predictions determined by physically interpreting the raw DOS prediction (orange), physically interpreting the denoised DOS prediction (dark blue), and applying a CNN on the raw DOS (light blue). The results are reported on the test set of the MAD dataset across the different subsets (a) and the external datasets (b). The MAE have units of eV. The values plotted in this figure are listed in Tables I and II in the SI.

that the CNN method yields superior performance, although the denoised DOS offers a reasonable alternative while keeping the workflow physically sound. Physical interpretability can be an advantage since it allows the derivation of additional properties from the same DOS without having to train more models. For example, we use the denoised DOS in Section 2.3.2 to compute the electronic heat capacity.

The bandgap performance on the different MAD subsets and the external samples can also be seen in Fig. 5. The performance on bandgap does not necessarily follow the same trend as that of the DOS. Amongst the MAD subsets, the bandgap performance is best on the MC3D-random subset, where PET-MAD-DOS struggles to get good DOS predictions. A similar observation can be made for the Alexandria external dataset. On the other hand, the bandgap performance is poor on the SPICE and MD22 datasets, where PET-MAD-DOS performs well. This can be attributed to the distribution of bandgaps in those subsets. For instance, the MC3D-random test subset consists entirely of conductors with no bandgap, and are thus easier to predict especially when using the raw or denoised DOS which tend to underestimate the bandgap. A similar argument can be made for Alexandria, SPICE and MD22, where the error in each task correlates with its mean and standard deviation. In most cases the bandgap is predicted with an error around 100 meV, which is comparable to the Gaussian smoothing we apply to construct the DOS, and smaller than typical DFT errors.

As a point of reference for the bandgap performance, we refer to the Matbench mp_gap leaderboards, as of December 2025. Based on the CNN approach, with a MAE and RMSE of 0.1900 and 0.3875 eV, it would be ranked 5th and 1st respectively. However, we emphasize that this is only to give a point of reference regarding the performance of the model, and not to make a direct comparison with the models on the Matbench leaderboards. Firstly, the models on the Matbench leaderboards are trained on the Matbench dataset while our model is trained on the MAD dataset. Secondly, our evaluation is only done on a small sample of 140 structures, recomputed with MAD DFT settings while the Matbench leaderboard is based on the entire test subset, which we cannot use directly because it is computed with incompatible DFT settings.

2.3 Application to finite-temperature material simulations

In addition to benchmarking PET-MAD-DOS on atomistic datasets, we demonstrate it in realistic applications by using it out-of-the-box as a general purpose model or as a foundation model to be fine-tuned. Towards that end, we used PET-MAD-DOS to predict the finite-temperature thermal-averaged DOS of two technologically relevant systems, namely gallium arsenide (GaAs) and lithium thiophosphates (LPS), and to predict the electronic heat capacity of a high entropy alloy (HEA). Specific details with regards to the material simulations can be found in Section III of the SI.

GaAs is a semiconductor with excellent physical and optoelectronic properties, making it well suited for photovoltaic devices for a wide range of applications.[50] The Ga–As system forms a simple binary phase diagram with metallic and semiconducting liquid and solid phases, making it an interesting system to use as a benchmark.

LPS have garnered great interest in the scientific community for their potential as electrolytes for solid-state batteries.[51] Li_3PS_4, one of the most popular LPS, has been extensively studied and modelled computationally.[52,53] Li_3PS_4 has three main polymorphs, α, β, and γ. The system is most stable in the γ polymorph at room temperature but it transforms into the metastable β polymorph at 573 K and then into the α polymorph at 746K.[54] Although the γ polymorph is a poor ionic conductor at room temperature, the β polymorph has high ionic conductivity for Li^+, making it a promising candidate for a solid electrolyte.

HEAs refer to systems composed of 5 or more metals in approximately equimolar proportions. These materials typically have desirable mechanical and catalytic properties.[55–58] However, building machine learning models to study HEAs and explore their composition space is often challenging due to the inherently high chemical diversity in these systems. They are often used in high-temperature applications, where thermal electronic excitations become relevant.

For all systems, we built a bespoke model, *i.e.* a PET model that is trained solely on the GaAs dataset from Imbalzano and Ceriotti,[59] or the LPS dataset from Gigli *et. al.*,[53] or a subset of the HEA25S dataset from Mazitov *et. al.*[60] All the datasets are recomputed with MAD DFT settings. Additionally, we also built a set of fine-tuned models by using the low-rank adaptive (LoRA) technique on the PET-MAD-DOS model on those datasets. Details on the fine-tuning procedure are discussed in Section 4.7. The bespoke and fine-tuned models have typically half the test-set errors, and serve as an assessment of the accuracy of the zero-shot PET-MAD-DOS in these complex simulations that would be prohibitively expensive with DFT.

2.3.1 Test set performance. To evaluate the performance of PET-MAD-DOS, the bespoke model and the LoRA fine-tuned model, we compare their accuracy on the test subset of those datasets in Table 1. PET-MAD-DOS performs reasonably well out-of-the-box, achieving errors that are comparable with those computed on the MAD subsets. The first thing to note is that PET-MAD-DOS errors are roughly twice as high as the errors of bespoke models in these systems. This is a common fact observed in other foundation models like MACE[19] and PET-MAD[30] and does not diminish the utility of PET-MAD-DOS as a fast and inexpensive tool for qualitative DOS predictions for material systems across the periodic table.

Table 1 Comparison of test root mean squared error (RMSE) performance for bespoke, low rank adaptation (LoRA), and PET-MAD-DOS models on different systems. The best performing model for each material in indicated in bold

RMSE on test subset [$eV^{-0.5}$ electrons^{-1} state]

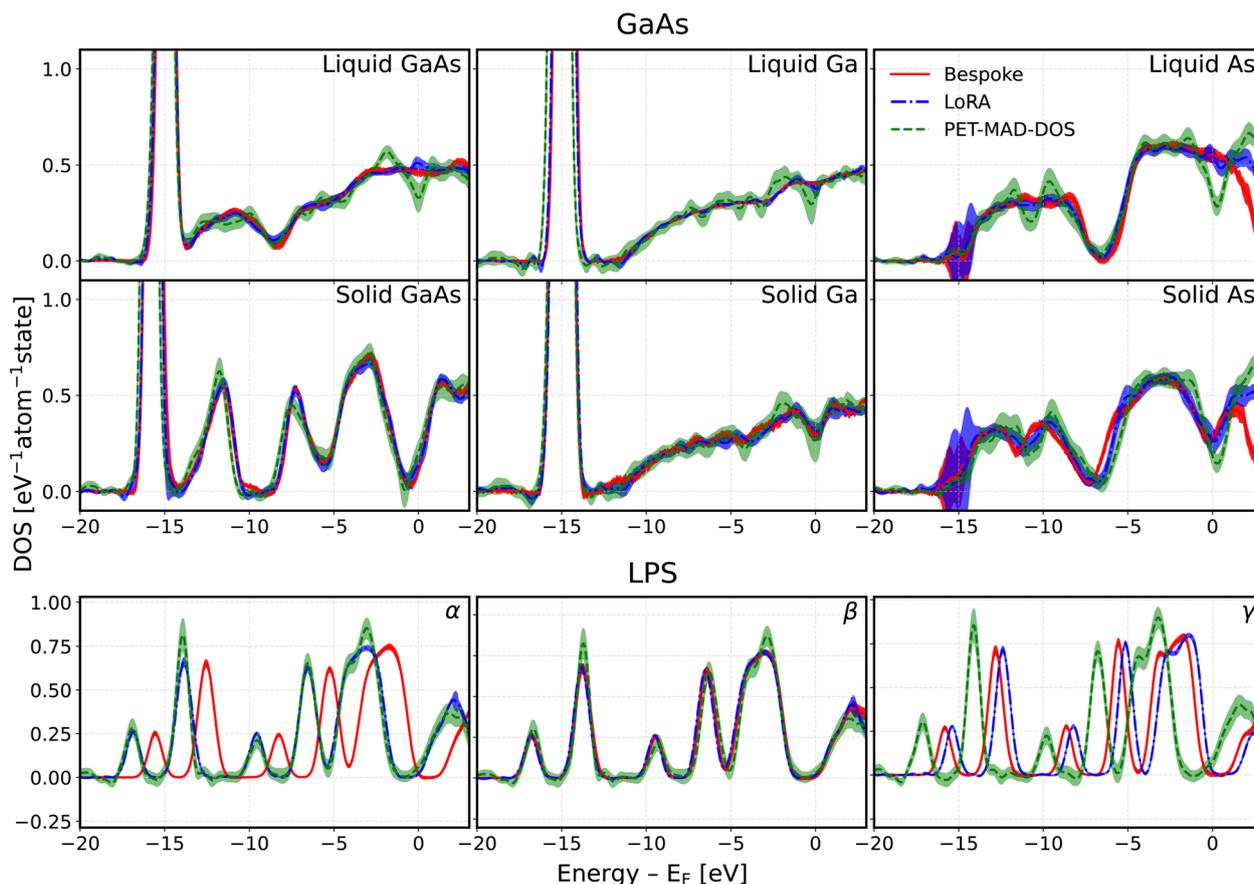| PET-MAD-DOS	Bespoke model	LoRA model
0.036	**0.016**	0.018
0.064	**0.027**	0.030
0.056	0.032	**0.029**

Fig. 6 Thermal-average DOS predictions of the MD trajectories of GaAs (top 2 rows) and LPS (bottom row) at different phases. The red solid lines represent the prediction of the bespoke model, the blue dash-dotted lines represent the prediction of the low rank adaptation (LoRA) model, and the green dotted line represents the prediction of PET-MAD-DOS. The colored areas represent the uncertainty associated with the DOS predictions of the corresponding model, obtained by propagating the uncertainties from each individual snapshot in the MD trajectory. In this procedure, the thermal-average DOS is computed for each member in the calibrated last-layer prediction rigidity (LLPR) ensemble, and the standard deviation across the ensemble members is taken as the uncertainty. Each system's phase is labelled at the top right corner of each subplot. The MD trajectories are obtained using a bespoke PET-MAD model. The energy axis shared between all systems is truncated to focus on the model's performance near the Fermi level, hiding the core and high energy states. A plot of the model predictions that includes the core states can be seen in Fig. 5 of the SI. For all subplots, the DOS is normalized with respect to the number of atoms in the system and the energy reference is set to the Fermi level determined based on each respective DOS prediction.

Once PET-MAD-DOS is finetuned, it offers a performance similar to or even better than that of the bespoke models. The fine-tuned models are able to achieve bespoke accuracies without significant impact to their performance on the MAD dataset (Table III of the SI). Furthermore, based on the learning curves in Section IV of the SI, the fine-tuned models have good performance even in the low-data regime, where they clearly outperform the bespoke ones. For the LPS and HEA datasets, the fine-tuned models are able to achieve bespoke accuracies using only 20% of the training data.

2.3.2 Thermal-average DOS. In addition to evaluating the models on their test set performance, we also compare each model's ability to compute the thermal-average DOS along molecular dynamics (MD) trajectories of GaAs and LPS in different phases. Studying phase transitions or interfaces requires atomistic models of thousands or more atoms, for which computing thermal-averages of the DOS is beyond the capabilities of conventional electronic structure methods.

Deringer *et al.*[61] have previously combined MLIPs with ML models for the DOS to reveal electronic properties in large amorphous silicon systems up to 100k atoms, proving the potential of the approach to reach unprecedented system sizes. However, their study relied on bespoke models. In this section, we demonstrate that similar results can also be obtained using only universal models, eliminating the need to train bespoke models, which can be computationally expensive during both the training and data generation phase.

For GaAs, we used NVT MD trajectories of Ga, GaAs, and As in both solid and liquid phases generated with the bespoke interatomic potential in ref. 30. For the solid systems, the MD simulations were performed at 150 K, 750 K and 550 K for Ga, GaAs, and As respectively. Meanwhile, for the liquid systems, the temperatures are 450 K, 2250 K, and 1650 K for the Ga, GaAs, and As systems. For both solids and liquids, the temperatures are chosen to be well into the solid or liquid phases, so as to avoid spurious phase transitions due to the limitations of the

reference DFT energetics. The simulations were performed for 4 ns, using a timestep of 4 fs.

For LPS, we used the MD trajectory generated by the bespoke interatomic potential in ref. 30. The trajectories for the three LPS phases were performed in the NpT ensemble at 400 K for a quasi-cubic cell containing 768 atoms at a pressure of zero bar. The trajectories were run for 3 ns, sampled every 20 fs.

Fig. 6 shows that PET-MAD-DOS is generally able to qualitatively predict the same DOS profile as the bespoke model, up to roughly 3 eV above the Fermi level. The LLPR module acts as a good estimate of the model confidence, as evidenced by the good overlap between the uncertainties of all three models. In this case, the profiles are a thermal average of model predictions across a MD trajectory, so we need to propagate uncertainty. To do so, we first compute the thermal-average predicted by each LLPR ensemble member. We then take the mean over LLPR ensemble members to get the final prediction, and use the standard deviation as a measure of uncertainty. It is crucial to note that the decay of the DOS above the Fermi level for the bespoke model is likely not physical as it arises due to the limited number of eigenstates in the DFT calculations used for the training set. For LPS, the predictions are observed to be offset relative to one another when aligned at the Fermi level. This is attributed to the difficulty in determining the Fermi level for a predicted DOS spectra as highlighted in Section 2.2. However, the shape of the DOS profiles still closely matches that of the LoRA and bespoke models. Along with the overlapping uncertainties, this highlights the fact that PET-MAD-DOS is able to yield good qualitative results out of the box in practical applications.

2.3.3 Electronic heat capacity. For HEAs, we evaluate the quality of the thermal-averaged DOS by using it to obtain the electronic heat capacity of a prototypical CoCrFeMnNi alloy. The electronic heat capacity can be particularly relevant at high temperatures, making it important for HEAs used in high temperature applications.

In this work, we calculated the electronic heat capacity from the HEA MD trajectories obtained using PET-MAD in ref. 30. The trajectories were obtained using a combination of replica-exchange molecular dynamics run with Monte-Carlo atom swaps. The simulation was performed with 16 replicas for 200 ps in the NPT ensemble using a 2 fs timestep at zero pressure and using a logarithmic temperatures grid ranging from 500 K to 1200 K.

To derive the heat capacity from the DOS, we used the denoised DOS as described in Section 2.2 instead of the raw DOS predictions, due to its higher physical interpretability. First, the thermal-averaged DOS was computed as the average of the denoised predictions along the MD trajectory. Then, the electronic contribution to the internal energy, U^{el}, was computed under the rigid band approximation as highlighted in ref. 62. The electronic heat capacity was then calculated as the derivative of U^{el} with respect to temperature using a finite difference scheme. Further details on the computation of the electronic heat capacity can be found in Section III of the SI. The uncertainties for the heat capacities are propagated by computing the heat capacity for each member in the calibrated
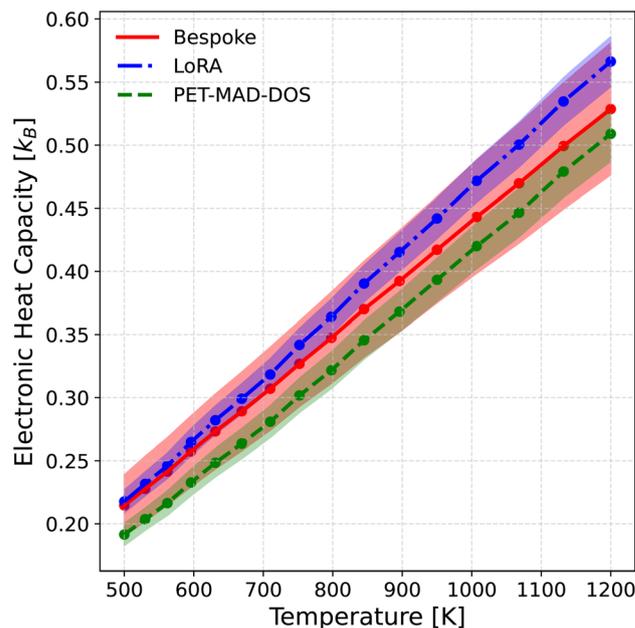

Fig. 7 Constant pressure electronic heat capacity derived from the thermal-average DOS of the HEA system at 16 different temperatures from 500 K to 1200 K. The red solid line represent the prediction of the bespoke model, the blue dash-dotted line represent the prediction of the low rank adaptation (LoRA) model, and the green dotted line represents the prediction of PET-MAD-DOS. The colored areas represent the uncertainty associated with the DOS predictions of the corresponding model, obtained by propagating the uncertainties from each individual snapshot in the MD trajectory. In this procedure, the heat capacity is computed for the denoised prediction of each member in the calibrated last-layer prediction rigidity (LLPR) ensemble, and the standard deviation across the ensemble members is taken as the uncertainty.

LLPR ensemble, taking the mean as the predicted heat capacity and the standard deviation as the uncertainty. The results are shown in Fig. 7, where it can be observed once again that PET-MAD-DOS performs well, being able to capture semi quantitatively the trend between heat capacity and temperature. Furthermore, the overlapping uncertainties reflect good agreement between all 3 models.

3 Discussion

PET-MAD-DOS consistently achieves semiquantitative predictions of the DOS and properties that can be extracted from it. Despite being trained on a small dataset and having a moderate number of parameters, it performs well across a broad spectrum of material classes, even on structures from external datasets. The generalizability of PET-MAD-DOS exceeds that of other universal DOS models[27,28] which are trained on datasets consisting solely of inorganic systems. Its performance out-of-the-box is only a factor of two worse than that of bespoke models trained on a medium-sized dataset for a specific class of material. This allows PET-MAD-DOS to yield results that are close to those of the bespoke models even in practical applications, highlighting the efficacy of PET-MAD-DOS as a general

purpose tool for DOS predictions. Furthermore, with the uncertainty quantification module based on an LLPR ensemble, it is also possible to have a reliable estimate of the model's error for both the DOS and the derived quantities at a relatively low cost. If the projected error is unsatisfactory, the performance can also be enhanced for particular applications by using PET-MAD-DOS as a foundation model to be fine-tuned for enhanced accuracies. The performance of these fine-tuned models is close to the bespoke models, sometimes outperforming them on their own validation domain. Learning curves show that fine-tuning works well with only about 100 additional structures, requiring far less data than bespoke models, and retaining stable predictions for the more general datasets. Transfer learning on a small number of reference calculations at a higher level of theory could be a practical strategy to obtain more accurate models for the many systems for which PBESol is not expected to provide a good description of electronic excitations.

Even though the residual noise in the prediction in the gap makes it somewhat difficult to determine its value from a direct analysis of the predicted DOS, a simple denoising filter or the application of a dedicated CNN to the raw prediction makes it possible to extract electronic properties with good accuracy. Although the PET architecture employed does not enforce rotational constraints, PET-MAD-DOS is still able to predict the DOS with a high level of rotational invariance, with the rotational variability being 2 orders of magnitude smaller than the accuracy of the model. PET-MAD-DOS is built and integrated within the metatensor[63] ecosystem, allowing the model to be easily accessible and for the training procedure to be easily replicated. Based on the accessibility, versatility and utility of PET-MAD-DOS, we believe that it can serve as a useful tool for materials discovery, especially in applications that require explicit information on the electronic structure.

4 Methods

In this section, we introduce details with regard to dataset construction, model architecture, loss functions for training and model evaluation, model training procedure, bandgap model architecture, and model fine-tuning, and uncertainty quantification. Further details regarding the MD simulations and hyper-parameters of the model can be found in Sections III and V of the SI.

4.1 Dataset construction

As the MAD dataset was primarily constructed to fit MLIPs, it was computed using a minimal number of energy bands. The energy range in which the DOS is well defined, based on the eigenvalues calculated, varies widely across the dataset. To increase data representation at energies above the Fermi level, a small subset of 850 structures was recalculated using four times the number of valence bands in the system. These structures are 750 monoelemental systems from the MC3D and MC3D-rattled subsets, together with the 100 structures that possess the lowest energy cutoff in the entire MAD dataset. Including these recomputed structures improves the DOS

predictions in the high energy range, as displayed in Section VI of the SI. Additionally, for bandgap benchmarking purposes, a small random subset comprised of 140 structures was taken from the Matbench dataset and recomputed with the same DFT settings outlined in ref. 30.

The calculations above were performed using the Quantum Espresso v7.2 package,[64] under a non-magnetic setting with the PBESol exchange–correlation functional. The pseudopotentials used were obtained from the standard solid-state pseudo-potentials library (SSSP) v1.2 (efficiency set),[65] using the highest settings for the plane-wave and charge density cutoffs across all 85 elements present in the MAD dataset (110 Ry and 1320 Ry respectively). The Marzari–Vanderbilt–deVita–Payne cold smearing[66] was used, with a spread of 0.01 Ry. For structures with periodicity, a fine k-point spacing of 0.125 π Å^{-1} was used in every periodic dimension while only one k point was used for the non-periodic dimensions. See ref. 32 for a detailed discussion of the makeup of the MAD dataset. Note that the PBESol exchange–correlation functional often predicts band gaps and electronic excitations which are not in good agreement with experiment, and that in these cases also PET-MAD-DOS should not be expected to produce reliable predictions. In such a case, the most straightforward solution would be to fine-tune PET-MAD-DOS using data computed at a satisfactory level of theory.

The target DOS for a structure, $DOS_A^Q(E)$, is then built *via* Gaussian smearing of the eigenvalues at each k-point and projecting it on a uniform energy grid as follows:

$$DOS_A^Q(E) = \sum_{n \in \text{bands}} \sum_{\mathbf{k}} w_k g(E - \varepsilon_n(\mathbf{k})) \tag{1}$$

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}}, \tag{2}$$

where N_A represents the number of atoms in the structure. $\varepsilon_n(\mathbf{k})$ represents the eigenvalues at each k point, with the energy reference set to the Fermi level determined by Quantum Espresso in the quantum chemical calculation. w_k represents the weight of the k-point in the Brillouin zone integral. σ is a Gaussian smearing parameter which is set to 0.3 eV, determined by comparing the constructed DOS of a sample structure against that of the same structure computed with a finer k-grid. E is the energy grid, which is a uniform grid containing 4806 points from -149.65 eV to 80.65 eV, representing 1.5 eV below and above the lowest and highest eigenvalue cutoff in the original MAD dataset (excluding recalculated structures). The lowest eigenvalue cutoff is the lowest eigenvalue in the dataset while the highest eigenvalue cutoff is the minimum energy of the highest energy band in the dataset.

4.2 PET model

The Point Edge Transformer (PET)[31] architecture combines both transformers and graph neural networks by using transformers in the message-passing layer. For every system, a directed graph is built by defining atoms as nodes and directed edges connect atoms within a specified cutoff radius. Feature vectors f_{ij}^t are then built on each directed edge between atoms i and j. These feature vectors serve as the messages that will be passed in the message-

passing layer, l. The dimensionality of f_{ij}^l is fixed and is defined by a hyperparameter of the architecture, d_{PET}. In each message-passing layer, a transformer is used to perform a permutation-covariant sequence-to-sequence transformation. The transformer takes as input all feature vectors f_{ij}^l, for a given central atom i and layer l, and outputs the corresponding feature vectors $\{f_{ij}^{l+1}\}_j$ for the next layer $l + 1$. This step also incorporates structural and chemical information regarding the central atom, such as the 3D positions of the neighbors and chemical species. After going through all the message-passing layers, all feature vectors f_{ij}^l are then used as inputs for a final feed-forward network. The output of the final feed-forward network is summed across bonds ij and layers l and represents the final target property, an array with size 4806 depicting the DOS in this case. To obtain better expressivity, the PET architecture does not impose any rotational constraints, allowing a single layer to theoretically access virtually unlimited body orders and angular resolution. To address the lack of rotational symmetry constraints, data augmentation is employed for the model to learn the rotational behaviour of the target, *i.e.* invariant for the case of the DOS. In this work, the only change made to the original PET architecture is at the last layer of the final feed-forward network, which is modified to give 4806 outputs, representing the size of the DOS array, instead of 1. For a more detailed description of the architecture and specific operations, the reader can refer to the original PET publication.[31] PET-MAD-DOS predicts the DOS of the system from a sum of atomic contributions, that are then summed to yield the total DOS, which is the physical target the model is trained on. The atomic contributions shall not be interpreted physically (*e.g.* as a projected DOS). Even though one could imagine fitting the atomic contributions to some projected DOS, the projections are not associated with well-defined physical observables, and are not uniquely defined. Constraining the model to be consistent with a given choice of atomic projections might have a negative impact on the accuracy of the predicted DOS without providing substantial advantages.

4.3 Training and evaluation functions

A simple mean squared error loss function is unable to properly reflect the underlying physical constraints of the DOS as a machine learning target, especially in a highly chemically diverse dataset where each calculation has a different energy cutoff in the eigenvalues. To account for the lack of an absolute energy reference in bulk systems,[67] we use a loss function that is agnostic to the energy reference of the prediction and the target. For this, we compute the loss only on the energy reference that minimizes the prediction error. We define the self-aligning loss, AL, for a single structure A as such:

$$\mathrm{MSE}(y(E), \hat{y}(E)) = \int_{E_{\min}}^{E_{\max}} \mathrm{d}E \ (y(E) - \hat{y}(E))^2 + \int_{G_{\min}}^{E_{\min}} \mathrm{d}E \ y(E)^2$$

(3)

$$\mathrm{AL_A}(\mathbf{W}) = \min_{\Delta \in 0, 1, \dots, \chi} \left[\mathrm{MSE}\big(\mathrm{DOS_A^W}(E + (\Delta \times e)), \mathrm{DOS_A^Q}(E)\big) \right].$$

(4)

$E_{\min}$ and $E_{\max}$ denote the energy minimum and maximum of the evaluation window. $\mathrm{DOS_A^Q}(E)$ represents the true DOS for structure A while $\mathrm{DOS_A^W}(E)$ represents the predicted DOS for structure A given model parameters $\mathbf{W}$. χ is an integer that denotes the maximum number of grid points the energy reference can shift by and e represents the energy grid interval. $G_{\min}$ refers to the minimum energy of the prediction grid and the second term in the eqn (3) essentially fits the DOS predictions below $E_{\min}$ to zero to reflect that there are no states below the minimum eigenvalue. This arises due to the fact that this minimization procedure requires the model to predict the DOS in a wider energy grid, resulting in $G_{\min} \leq E_{\min}$. The optimization algorithm then searches for the continuous subset within the prediction, corresponding to the size of the target, that minimizes the MSE. Based on preliminary testing, we have set χ to 200, corresponding to the prediction grid being 10 eV wider. This is similar to the adaptive energy reference used in ref. 68, with the exception that the loss is now fully minimized at every epoch instead of being optimized simultaneously with the model weights, but the energy reference can only shift in integer multiples of the energy grid interval. By restricting the search space to only integer multiples, it circumvents the need to compute derivatives or build splines of the DOS during the minimization procedure. Additionally, we were able to exploit full vectorization to evaluate the loss for all values of Δ simultaneously, ensuring that the minimization procedure obtains the global minima. We demonstrate the significance of the shift agnostic training procedure in Section VI of the SI.

Although every system, in principle, has an infinite number of eigenvalues at every k-point, electronic structure calculations consider only a finite number of them. Due to this restriction, calculating the DOS based on the method outlined in Section 4.1 will result in a sharp unphysical drop in the DOS to zero, past the maximum computed eigenvalue. This impacts the reliability of the DOS targets computed near the highest computed eigenvalue. To account for this during model evaluation and training, we set $E_{\max}$ in (4) for each structure to 0.9 eV, corresponding to 3× the smearing value, below the minimum energy of the highest energy band across every k-point. Since MAD was computed with a minimal number of energy bands, a large number of structures have a low $E_{\max}$, with some $E_{\max}$ values being lower than the Fermi level. Hence, it is not feasible to simply set the $E_{\max}$ of all structures to the minimum $E_{\max}$ in the dataset. Additionally, due to the wide range of $E_{\max}$ in the dataset, there is an uneven distribution of data across the energy grid. This results in highly oscillatory predictions at higher energy levels due to insufficient data in those regions. These oscillations can contaminate predictions during deployment if the structure contains atomic environments that comes from two training structures with very different $E_{\max}$ (Section VII of SI). To address these oscillations, we introduce a gradient loss, GL, that imposes a mean squared penalty on the gradient of the predictions, determined *via* finite differences, outside $E_{\max}$. The gradient loss for a single structure, A, is:

$$\mathrm{GL_A(\mathbf{W})} = \int_{E_{\max}}^{G_{\max}} \mathrm{d}E \left(\frac{\mathrm{dDOS_A^W}\left(E + \left(\Delta_{\mathrm{opt}} \times e\right)\right)}{\mathrm{d}E} \right)^2, \qquad (5)$$

where $G_{\max}$ represents the maximum energy of the prediction grid and Δ_{opt} is the optimal shift determined via (4).

In addition, we also include the loss on the cumulative DOS, CL, similar to ref. 28 and 69. The loss on the cumulative DOS for a single structure, A, is expressed as:

$$\mathrm{CL_A(\mathbf{W})} = \int_{E_{\min}}^{E_{\max}} \mathrm{d}E \left(\mathrm{cDOS_A^W}\left(E + \left(\Delta_{\mathrm{opt}} \times e\right)\right) - \mathrm{cDOS_A^Q}(E)\right)^2 \qquad (6)$$

where cDOS represents the cumulative DOS function. The final loss that the model is trained on is as follows:

$$L(\mathbf{W}) = \frac{1}{N}\sum_A \frac{1}{N_A}\left(\mathrm{AL_A(\mathbf{W})} + \alpha\mathrm{GL_A(\mathbf{W})} + \beta\mathrm{CL_A(\mathbf{W})}\right), \qquad (7)$$

where N refers to the number of structures in the training set and N_A denotes the number of atoms in structure A. The loss is normalized with respect to the number of atoms in each structure to make the loss independent of structure size. α and β are hyperparameters used to scale GL and CL respectively. In this work, α and β are set to 10^{-4} and 2 based on preliminary tests.

For evaluation, the RMSE is also normalized to account for the difference in the number of electrons represented by the DOS in the dataset:

$$n_A = \int_{E_{\min}}^{E_{\max}} \mathrm{d}E\ \mathrm{DOS_A^Q}(E) \qquad (8)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_A \frac{\mathrm{AL_A(\mathbf{W})}}{n_A^2}} \qquad (9)$$

where N represents the number of structures in the evaluation set. n_A represents the number of electrons represented in the target DOS.

We evaluate the symmetry error as the standard deviation of the DOS predictions of 38 rotated copies of the each structure, based on a Lebedev angular grid with a degree of 8. The standard deviations are only computed up to the point where the DOS target is defined so that it can be compared to the RMSE of the DOS predictions. The formula for the symmetry error, σ_A^{rot}, is as follows:

$$\sigma_A^{\mathrm{rot}} = \sqrt{\frac{1}{38}\sum_{i=1}^{38}\frac{1}{n_A}\int_{E_{\min}}^{E_{\max}} \mathrm{d}E\ \left(\mathrm{DOS_A^i}(E) - \mathrm{DOS_A^\mu}(E)\right)^2}, \qquad (10)$$

where i represents the index of the rotated copies, A represents the structure, $\mathrm{DOS_A^i}$ represents the prediction on the ith rotated copy of structure A and $\mathrm{DOS_A^\mu}(E)$ represents the mean prediction of structure A across all rotations. The symmetry error is normalized by the number of electrons so that it can be meaningfully compared against the RMSE in (9).

4.4 Training of PET-MAD-DOS

The structures in each one of the eight subsets in the MAD dataset were randomly split into training, validation, and test sets in a 8 : 1 : 1 ratio. We perform a hyperparameter search over the five points on the Pareto-front of PET-MAD[30] and select the hyperparameters that yield the best balance of performance and accuracy. The results are detailed in Section VIII of the SI, where we also report the computational cost of PET-MAD-DOS. The resulting optimal hyperparameters are the same as those in PET-MAD, with a cutoff radius of 4.5 Å, 2 message-passing layers, each comprising of two transformer layers with a token size of 256 and 8 heads in the multi-head attention layer. The output multi-layer perceptron contains 512 neurons, which are fed to a linear layer to give 4806 outputs, corresponding to the DOS at each energy channel. This results in a total of 8 625 226 parameters in the model. Model training was performed using the PyTorch framework and the metatrain package[63] on 1 NVIDIA H100 GPU with a batch size of 16 structures for a total of 760 epochs, taking roughly 72 hours. For model training, the Adam[70] optimizer was used, with an initial learning rate (LR) of 10^{-4}, using a warmup of 100 epochs that increases the LR linearly from 0 to 10^{-4}. Afterwards, a LR scheduler was employed to half the LR every 250 epochs.

4.5 CNN model specifications

For the CNN models used to predict secondary quantities like the bandgap, Fermi level and $\mathrm{DOS}(E_F)$ model, we utilize a simple 1D convolutional neural network (CNN) for univariate sequential input. The model takes the raw PET-MAD-DOS prediction of a structure as input and is composed of four sequential convolutional blocks followed by two fully connected layers. Each convolutional block contains a convolutional layer with 64 output channels and a SiLU activation function, and a 1D max pooling layer with a kernel size of 4. The kernel size of the convolution layer in the first, second, third and fourth block is 32, 16, 8 and 8 respectively. The two fully connected layers contains 1024 neurons each, with the SiLU activation function to produce a scalar output representing either the target. The model is trained on the mean squared error (MSE) against the DFT targets, using the Adam optimizer with an initial LR of 10^{-4} and 100 warmup epochs that increases the LR linearly from 0 to 10^{-4}. Early stopping is implemented to stop model training if the MSE on the validation set does not decrease after 50 epochs. The model is trained using the Pytorch framework on 1 NVIDIA H100 GPU with a batch size of 16 for roughly 150 epochs, taking around 30 minutes. During evaluation, the ReLU activation function is applied to the predictions of the bandgap model to remove unphysical negative bandgap values.

4.6 Prediction denoising

As highlighted in 2.2, relying on a physical interpretation of the raw predicted DOS for the Fermi level and bandgap requires extremely high DOS accuracies and minimal noise in the gap. As this is difficult to achieve under the current training approach,

an additional prediction denoising step was applied on the DOS predictions to obtain a DOS that can be physically interpreted.

Firstly, a CNN model was trained, as described in 4.5, to predict the position of the Fermi level of a structure based on the raw predicted DOS. Then, a 1-D Gaussian filter, with a standard deviation, σ, of 0.3 eV was applied on the raw predicted DOS as follows,

$$\text{DOS}_G(E) = \int_{G_{\min}}^{G_{\max}} \text{DOS}_{\text{pred}}(\tau) G(E - \tau) \mathrm{d}\tau \tag{11}$$

$$G(E) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{E^2}{2\sigma^2}\right), \tag{12}$$

where DOS_G represents the filtered DOS and DOS_{pred} represents the raw DOS prediction. Next, the filtered DOS is passed through a modified sigmoid function,

$$f(x) = \frac{1}{1 + \mathrm{e}^{-a(x-b)}}, \tag{13}$$

where the additional constants a and b determine the inflection point and slope of the sigmoid function. In this work, we chose a to be 0.1 and b to be 100. The output of the modified sigmoid function, β, is then used as a multiplier on the DOS output to obtain a thresholded DOS.

$$\text{DOS}_{\text{thresh}}(E) = \text{DOS}_{\text{pred}}(E) \times f(\text{DOS}_G(E)) \tag{14}$$

In the last step, the thresholded DOS is then scaled such that the physical Fermi level of the DOS lie on the same point as that predicted by the Fermi level CNN, described in the first step.

$$n = \int_{G_{\min}}^{\varepsilon_F^{\text{CNN}}} \text{DOS}_{\text{thresh}}(E) \tag{15}$$

$$\text{DOS}_{\text{clean}} = \frac{n_{\text{elec}}}{n} \text{DOS}_{\text{thresh}}(E) \tag{16}$$

where $\text{DOS}_{\text{clean}}$ represents the final denoised DOS output, n_{elec} refers to the number of electrons in the neutral system (excluding the ones in the pseudopotential), and $\varepsilon_F^{\text{CNN}}$ refers to the Fermi level of the system predicted by the CNN model described in the first step.

4.7 Fine-tuning

The popular low-rank adaption (LoRA) method[71] was employed to fine-tune the pre-trained PET-MAD-DOS models for specific applications. LoRA was selected for its efficiency and ability to reduce the impacts of catastrophic forgetting, which refers to a fine tuned model losing its predictive capabilities on its base dataset. Instead of fine tuning all the model weights as in conventional fine-tuning, LoRA instead trains an additional set of parameters while leaving the original model weights untouched. These parameters are comprised of two low-rank matrices which are added to each attention block of the model, scaled by a regularization factor that controls the influence of the matrices on the model's weights. Through tuning the rank of the matrices and the regularization factor, a model can be fine tuned to achieve better performance in

specific applications without compromising the generalizability of the model. In this work, we use the same LoRA parameters as PET-MAD, namely a rank of 8 and the regularization factor set to 0.5.

LoRA-fine-tuned models retain varying degree of accuracy (see the Table III of the SI for details) on the generic structures from the MAD dataset, while providing performance comparable to that of a bespoke model, even in the low data regime for certain systems. Hence, we recommend the use of LoRA when fine-tuning PET-MAD-DOS for a specific application.

4.8 Uncertainty quantification

To perform uncertainty quantification (UQ) for the PET-MAD-DOS model, we employed the last-layer prediction rigidity (LLPR) method by Bigi et al.,[37] which computes uncertainties as the inverse of the prediction rigidity.[72,73] The fact that the DOS is a vectorial prediction target presents limitations in the originally proposed UQ approach: the last-layer features of each structure used for DOS prediction is fixed for all energy channels, and calibration factors are obtained "globally" across the entire dataset, resulting in a fixed uncertainty profile for all structures, only scaled differently based on the relative magnitude of the prediction rigidity. We therefore initialize a last-layer ensemble of 128 models with the weights sampled following eqn (25) of ref. 37. We perform further calibration of the ensemble weights with a Gaussian negative log-likelihood loss as done in Kellner and Ceriotti,[38] resulting in a UQ profile that is far more informative and accurate (see Fig. 12 of the SI). Furthermore, the UQ profile also accurately reflects the adaptive evaluation window used in the loss function for training. The model uncertainty increases significantly when extrapolating the DOS to high energies, as observed in Fig. 12 of the SI.

Author contributions

WBH worked on the development of the loss function and integration into metatrain, training PET-MAD-DOS, bespoke and LoRA models, developed and trained the CNN models and denoising workflows, performed the accuracy, performance, and speed benchmarks, calculated the ensemble-average DOS and electronic heat capacity on MD trajectories of the three material systems, DFT re-computations of a subset of MAD and DFT computations for the Matbench sample. PF guided the methodology for the determination of the electronic heat capacity, denoising and bandgap from the DOS. SC integrated the UQ strategy described in 4.8 into metatrain, performed the calibration of LLPR-based last-layer ensembles for UQ, and aided the evaluation and analysis of the prediction uncertainties of PET-MAD-DOS. AM worked on the creation of the MAD dataset, sample selection and DFT calculation of the external datasets, implemented the LoRA training procedure on metatrain, integrated PET-MAD-DOS within the PET-MAD repository and performed the MD simulations of surface segregation in CoCrFeMnNi. FB worked on implementation of the metatrain infrastructure for training and evaluating the PET-MAD-DOS model, and supported the development of the

infrastructure for UQ. MK performed on the MD simulations of Ga, GaAs, and As in the solid and liquid phases. SP developed the original version of PET architecture and the shift invariant loss function. MC designed and guided the project, and provided theoretical support. All authors contributed to the writing of the manuscript.

Conflicts of interest

There are no competing interests to declare.

Data availability

The MAD dataset, benchmarks, and simulation input files are available as a record [https://doi.org/10.24435/materialscloud:vd-e8] on the Materials Cloud Archive [https://doi.org/10.1038/s41597-020-00637-5]. A dataset containing the curated DOS data, including also the structures recomputed with a larger number of empty states and training scripts for PET-MAD-DOS, uncertainty quantification, and finetuning are available as a record [https://doi.org/10.24435/materialscloud:gs-z7] on the Materials Cloud Archive [https://doi.org/10.1038/s41597-020-00637-5]. Furthermore, the record also contains the necessary scripts to replicate the results reported in this publication. The upet (universal PET) repository on GitHub [https://doi.org/10.5281/zenodo.18629889] contains the code to reproduce the results in this paper, in the form of a convenient API to the pretrained PET-MAD-DOS model [https://huggingface.co/lab-cosmo/pet-mad-dos]. The repository also contains instructions on how to use the code.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d5dd00557d.

Acknowledgements

References

1 J. Schmidt, M. R. G. Marques, S. Botti and M. A. L. Marques, *npj Comput. Mater.*, 2019, **5**, 83.

2 G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa and A. Fazzio, *J. Phys.: Mater.*, 2019, **2**, 032001.

3 J. Wei, X. Chu, X.-Y. Sun, K. Xu, H.-X. Deng, J. Chen, Z. Wei and M. Lei, *InfoMat*, 2019, **1**, 338–358.

4 A. Chandrasekaran, D. Kamal, R. Batra, C. Kim, L. Chen and R. Ramprasad, *npj Comput. Mater.*, 2019, **5**, 1–7.

5 O. T. Unke, S. Chmiela, H. E. Sauceda, M. Gastegger, I. Poltavsky, K. T. Schütt, A. Tkatchenko and K.-R. Müller, *Chem. Rev.*, 2021, **121**, 10142–10186.

6 L. Gigli, M. Veit, M. Kotiuga, G. Pizzi, N. Marzari and M. Ceriotti, *npj Comput. Mater.*, 2022, **8**, 209.

7 W. B. How, B. Wang, W. Chu, A. Tkatchenko and O. V. Prezhdo, *J. Phys. Chem. Lett.*, 2021, 12026–12032.

8 S. P. G, M. N. Mattur, N. Nagappan, S. Rath and T. Thomas, *J. Materiomics*, 2022, **8**, 937–948.

9 Y. Zhuo, A. Mansouri Tehrani and J. Brgoch, *J. Phys. Chem. Lett.*, 2018, **9**, 1668–1673.

10 W. B. How, B. Wang, W. Chu, S. M. Kovalenko, A. Tkatchenko and O. V. Prezhdo, *J. Chem. Phys.*, 2022, **156**, 054110.

11 A. M. Lewis, A. Grisafi, M. Ceriotti and M. Rossi, *J. Chem. Theory Comput.*, 2021, **17**, 7203–7214.

12 J. Nigam, M. J. Willatt and M. Ceriotti, *J. Chem. Phys.*, 2022, **156**, 014115.

13 H. Li, Z. Wang, N. Zou, M. Ye, R. Xu, X. Gong, W. Duan and Y. Xu, *Nat. Comput. Sci.*, 2022, **2**, 367–377.

14 F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti and L. Emsley, *Nat. Commun.*, 2018, **9**, 4501.

15 W. Gerrard, L. A. Bratholm, M. J. Packer, A. J. Mulholland, D. R. Glowacki and C. P. Butts, *Chem. Sci.*, 2020, **11**, 508–515.

16 C. Ben Mahmoud, A. Anelli, G. Csányi and M. Ceriotti, *Phys. Rev. B*, 2020, **102**, 235130.

17 K. Bang, B. C. Yeo, D. Kim, S. S. Han and H. M. Lee, *Sci. Rep.*, 2021, **11**, 11604.

18 M. Neumann, J. Gin, B. Rhodes, S. Bennett, Z. Li, H. Choubisa, A. Hussey, and J. Godwin, Orb: A Fast, Scalable Neural Network Potential, *arXiv*, 2024, preprint arXiv:2410.22570, DOI: 10.48550/2410.22570.

19 I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin *et al.*, *arXiv*, 2023, preprint arXiv:2401.00096, DOI: 10.48550/2401.00096.

20 H. Yang, C. Hu, Y. Zhou, X. Liu, Y. Shi, J. Li, G. Li, Z. Chen, S. Chen, and C. Zeni *et al.*, *arXiv*, 2024, preprint arXiv:2405.04967, DOI: 10.48550/2405.04967.

21 R. Ruff, P. Reiser, J. Stühmer and P. Friederich, Connectivity Optimized Nested Graph Networks for Crystal Structures, *arXiv*, 2023, preprint arXiv:2302.14102, DOI: 10.48550/2302.14102.

22 S. S. Omee, S.-Y. Louis, N. Fu, L. Wei, S. Dey, R. Dong, Q. Li and J. Hu, Scalable deeper graph neural networks for high-performance materials property prediction, *arXiv*, 2021,

preprint arXiv:2109.12283 [cond-mat], DOI: 10.48550/2109.12283.

23 K. Choudhary and B. DeCost, *npj Comput. Mater.*, 2021, **7**, 185.

24 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chem. Mater.*, 2019, **31**, 3564–3572.

25 Y. Zhong, H. Yu, J. Yang, X. Guo, H. Xiang and X. Gong, Universal Machine Learning Kohn-Sham Hamiltonian for Materials, *arXiv*, 2024, preprint arXiv:2402.09251, DOI: 10.48550/2402.09251.

26 Y. Wang, Y. Li, Z. Tang, H. Li, Z. Yuan, H. Tao, N. Zou, T. Bao, X. Liang, Z. Chen, S. Xu, C. Bian, Z. Xu, C. Wang, C. Si, W. Duan and Y. Xu, *Sci. Bull.*, 2024, **69**, 2514–2521.

27 S. Kong, F. Ricci, D. Guevarra, J. B. Neaton, C. P. Gomes and J. M. Gregoire, *Nat. Commun.*, 2022, **13**, 949.

28 V. Fung, P. Ganesh and B. G. Sumpter, *Chem. Mater.*, 2022, **34**, 4848–4855.

29 N. Lee, H. Noh, S. Kim, D. Hyun, G. S. Na, and C. Park, Predicting Density of States *via* Multi-modal Transformer, *arXiv*, 2023, preprint arXiv:2303.07000 [cond-mat, physics:physics], DOI: 10.48550/2109.12283.

30 A. Mazitov, F. Bigi, M. Kellner, P. Pegolo, D. Tisi, G. Fraux, S. Pozdnyakov, P. Loche and M. Ceriotti, PET-MAD, a universal interatomic potential for advanced materials modeling, *arXiv*, 2025, preprint arXiv:2503.14118 [cond-mat], DOI: 10.48550/2503.14118.

31 S. Pozdnyakov and M. Ceriotti, *Adv. Neural. Inf. Process. Syst.*, 2023, 79469–79501.

32 A. Mazitov, S. Chorna, G. Fraux, M. Bercx, G. Pizzi, S. De and M. Ceriotti, Massive Atomic Diversity: a compact universal dataset for atomistic machine learning, *arXiv*, 2025, preprint arXiv:2506.19674 [cond-mat], DOI: 10.48550/2506.19674.

33 M. Y. Toriyama, A. M. Ganose, M. Dylla, S. Anand, J. Park, M. K. Brod, J. M. Munro, K. A. Persson, A. Jain and G. J. Snyder, *Mater. Today Electron.*, 2022, **1**, 100002.

34 N. W. Ashcroft and N. D. Mermin, *Solid State Physics*, Holt, Rinehart and Winston, 1976.

35 C. Ben Mahmoud, F. Grasselli and M. Ceriotti, *Phys. Rev. B*, 2022, **106**, L121116.

36 N. Lopanitsyna, C. Ben Mahmoud and M. Ceriotti, *Phys. Rev. Mater.*, 2021, **5**, 043802.

37 F. Bigi, S. Chong, M. Ceriotti and F. Grasselli, *Mach. Learn.: Sci. Technol.*, 2024, **5**, 045018.

38 M. Kellner and M. Ceriotti, *Mach. Learn.: Sci. Technol.*, 2024, **5**, 035006.

39 S. Huber, M. Bercx, N. Hörmann, M. Uhrin, G. Pizzi, N. Marzari, *Materials Cloud Archive*, 2022, DOI: 10.24435/materialscloud:rw-t0.

40 D. Campi, N. Mounet, M. Gibertini, G. Pizzi and N. Marzari, *ACS Nano*, 2023, **17**, 11268–11278.

41 M. Cordova, E. A. Engel, A. Stefaniuk, F. Paruzzo, A. Hofstetter, M. Ceriotti and L. Emsley, *J. Phys. Chem. C*, 2022, **126**, 16710–16720.

42 C. R. Groom, I. J. Bruno, M. P. Lightfoot and S. C. Ward, *Acta Crystallogr., Sect. B:Struct. Sci., Cryst. Eng. Mater.*, 2016, **72**, 171–179.

43 B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel and G. Ceder, *Nat. Mach. Intell.*, 2023, **5**, 1031–1041.

44 A. Dunn, Q. Wang, A. Ganose, D. Dopp and A. Jain, *npj Comput. Mater.*, 2020, **6**, 138.

45 J. Schmidt, N. Hoffmann, H.-C. Wang, P. Borlido, P. J. M. A. Carriço, T. F. T. Cerqueira, S. Botti and M. A. L. Marques, *Adv. Mater.*, 2023, **35**, 2210788.

46 P. Eastman, P. K. Behara, D. L. Dotson, R. Galvelis, J. E. Herr, J. T. Horton, Y. Mao, J. D. Chodera, B. P. Pritchard, Y. Wang, G. De Fabritiis and T. E. Markland, *Sci. Data*, 2023, **10**, 11.

47 S. Chmiela, V. Vassilev-Galindo, O. T. Unke, A. Kabylda, H. E. Sauceda, A. Tkatchenko and K.-R. Müller, *Sci. Adv.*, 2023, **9**, eadf0873.

48 L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, *et al.*, *ACS Catal.*, 2021, **11**, 6059–6072.

49 M. F. Langer, S. N. Pozdnyakov and M. Ceriotti, *Mach. Learn.: Sci. Technol.*, 2024, **5**, 04LT01.

50 R. C. Sharma, R. Nandal, N. Tanwar, R. Yadav, J. Bhardwaj and A. Verma, *J. Phys.: Conf. Ser.*, 2023, **2426**, 012008.

51 A. Kwade, W. Haselrieder, R. Leithoff, A. Modlinger, F. Dietrich and K. Droeder, *Nat. Energy*, 2018, **3**, 290–300.

52 F. N. Forrester, J. A. Quirk, T. Famprikis and J. A. Dawson, *Chem. Mater.*, 2022, **34**, 10561–10571.

53 L. Gigli, D. Tisi, F. Grasselli and M. Ceriotti, *Chem. Mater.*, 2024, **36**, 1482–1496.

54 K. Homma, M. Yonemura, T. Kobayashi, M. Nagao, M. Hirayama and R. Kanno, *Solid State Ionics*, 2011, **182**, 53–58.

55 J.-W. Yeh, S.-K. Chen, S.-J. Lin, J.-Y. Gan, T.-S. Chin, T.-T. Shun, C.-H. Tsau and S.-Y. Chang, *Adv. Eng. Mater.*, 2004, **6**, 299–303.

56 B. Cantor, I. Chang, P. Knight and A. Vincent, *J. Mater. Sci. Eng. A*, 2004, **375–377**, 213–218.

57 Y. Sun and S. Dai, *Sci. Adv.*, 2021, 7(20), DOI: 10.1126/sciadv.abg1600.

58 N. Kumar Katiyar, K. Biswas, J.-W. Yeh, S. Sharma and C. Sekhar Tiwary, *Nano Energy*, 2021, **88**, 106261.

59 G. Imbalzano and M. Ceriotti, *Phys. Rev. Mater.*, 2021, **5**, 063804.

60 A. Mazitov, M. A. Springer, N. Lopanitsyna, G. Fraux, S. De and M. Ceriotti, *J. Phys.: Mater.*, 2024, **7**, 025007.

61 V. L. Deringer, N. Bernstein, G. Csányi, C. Ben Mahmoud, M. Ceriotti, M. Wilson, D. A. Drabold and S. R. Elliott, *Nature*, 2021, **589**, 59–64.

62 N. Lopanitsyna, C. Ben Mahmoud and M. Ceriotti, *Phys. Rev. Mater.*, 2021, **5**, 043802.

63 F. Bigi, P. Loche, G. Fraux, A. Mazitov, D. Tisi, S. Pozdnyakov and S. Chong, Training and evaluating machine learning models for atomistic systems, 2025, https://github.com/metatensor/metatrain.

64 P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto,

C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari and R. M. Wentzcovitch, *J. Phys.:Condens. Matter*, 2009, **21**, 395502–395519.

65 G. Prandini, A. Marrazzo, I. E. Castelli, N. Mounet and N. Marzari, *npj Comput. Mater.*, 2018, **4**, 72.

66 N. Marzari, D. Vanderbilt, A. De Vita and M. C. Payne, *Phys. Rev. Lett.*, 1999, **82**, 3296–3299.

67 L. Kleinman, *Phys. Rev. B*, 1981, **24**, 7412–7414.

68 W. B. How, S. Chong, F. Grasselli, K. K. Huguenin-Dumittan and M. Ceriotti, *Phys. Rev. Mater.*, 2025, **9**, 013802.

69 C. Ben Mahmoud, A. Anelli, G. Csányi and M. Ceriotti, *Phys. Rev. B*, 2020, **102**, 235130.

70 D. P. Kingma and J. Ba, *arXiv*, 2014, preprint arXiv:1412.6980, DOI: 10.48550/1412.6980.

71 E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang and W. Chen, *arXiv*, 2021, preprint arXiv:2106.09685, DOI: 10.48550/2106.09685.

72 S. Chong, F. Grasselli, C. Ben Mahmoud, J. D. Morrow, V. L. Deringer and M. Ceriotti, *J. Chem. Theory Comput.*, 2023, **19**, 8020–8031.

73 S. Chong, F. Bigi, F. Grasselli, P. Loche, M. Kellner and M. Ceriotti, *Faraday Discuss.*, 2025, **256**, 322–344.