Phonon-accurate machine-learning potentials from automated workflows

Christina Ertural[1], Aakash A. Naik[1,3], Yuanbin Liu[2], Jonas Grandel[1], Natascia Fragapane[2], Joe D. Morrow[2], Daniel F. Thomas du Toit[2], Yuxing Zhou[2], Philipp Benner[1], Volker L. Deringer[2,*], and Janine George[1,3,*]

[1]Federal Institute for Materials Research and Testing (BAM), Unter den Eichen 87, 12205 Berlin, Germany.
[2]Inorganic Chemistry Laboratory, Department of Chemistry, University of Oxford, Oxford OX1 3QR, UK
[3]Friedrich Schiller University Jena, Institute of Condensed Matter Theory and Solid-State Optics, Max-Wien-Platz 1, 07743 Jena, Germany.
*Corresponding authors: Volker L. Deringer (volker.deringer@chem.ox.ac.uk), Janine George (janine.george@bam.de)

ABSTRACT

The stability, thermal expansion and heat transport of crystalline materials are substantially dependent on their vibrational properties. While stabilities and thermal expansion can be obtained within the quasi-harmonic approximation (QHA) and density functional theory (DFT) frameworks, accurate prediction of thermal conductivity requires inclusion of anharmonic effects such as three-phonon and four-phonon scattering. Even the QHA within DFT is a resource-consuming method, and it does not even include effects of temperature-dependent phonons or high-order force constants. Machine learning-driven interatomic potentials (MLIP) open an alternative and faster route to phonons; however, in most cases, the phonon-accurate MLIPs that can also capture anharmonic effects are specifically tailored and often manually curated for a certain compound or material. While universal MLIPs exist and nowadays show quite remarkable accuracies for ground-state phonons close to the training data, they still lack accuracy for anharmonic properties necessary for thermal conductivity or thermal expansion predictions. In this work, we automate the generation and fine-tuning of MLIPs for (anharmonic) phonon calculations in a Python-based workflow as part of our recently introduced open-source software autoplex, benefiting from automation tools such as atomate2 and data from the Materials Project. The workflow combines automated training data generation for anharmonic phonons at the DFT level with automated MLIP fitting and fine-tuning, enabling easier testing, benchmarking, and validation.

Introduction

The stability, thermal expansion, and thermal conductivity of phonon-related properties are crucial for many applications in materials science, including electronic devices such as batteries and thermoelectric devices[1]. Phonon properties are therefore regularly assessed in materials screening, and several large-scale harmonic phonon databases and some first anharmonic ones exist.[2–4] Stabilities and thermal expansion can be obtained within the quasi-harmonic approximation (QHA), and the thermal conductivity prediction requires including anharmonic effects such as three- and four-phonon scattering.[5–7] However, pure DFT-based approaches are very costly for large-scale screenings.

Machine-learned interatomic potentials (MLIPs) now offer an alternative, as they can reliably predict energies and forces at a fraction of the cost of DFT. Various architectures exist that can be classified into two major categories of how the environments are represented: explicit ones, where an explicit set of features is defined for each atomic environment, and implicit ones, where these representations are learned.[8] In the former category, Gaussian Approximation Potentials would be included, while in the latter, graph-based architectures like[9,10] would be categorized. Graph-based architectures enable the creation of universal machine learning potentials because the learned features scale much more favorably with the number of elements in the potentials.[11–13] Some pre-trained versions have also shown acceptable predictions of harmonic phonon properties near the ground state volume and close to the training data.[14] However, MLIPs are often trained explicitly for anharmonic phonon properties since they require high-quality DFT data[3]. Pre-trained potentials can be improved through distillation methods (which can

also enhance speed) or fine-tuning[15,16]. Both approaches typically need additional DFT datasets optimized explicitly for the target property. Recent research also demonstrates that MLIPs (including automatically trained ones) can be used to predict phase diagrams, where phonon properties play a critical role.[17]

Tight feedback between DFT data point selection, MLIP training/fine-tuning, and benchmarking against target properties is essential for efficient training and fine-tuning of MLIPs for anharmonic phonon properties. Such a workflow, therefore, requires a combination of parallel DFT runs on high-performance CPU clusters and training/fine-tuning of MLIPs on GPUs or CPUs, depending on the architecture. Traditional one-code solutions are not optimal. In addition, new MLIP architectures are still being developed frequently and are difficult to integrate directly into one-code solutions. Therefore, Python-based automation frameworks and workflow tools offer an optimal way to orchestrate such a workflow. We have recently shown that such a framework can also be used to perform random structure search-based training of MLIPs.[18] We envision a future where such automation frameworks largely orchestrate MLIP training based on DFT computations, benchmarking, and application of MLIPs. Such frameworks can also be used in connection with agents. Our work here represents another step towards such a future.

In this work, we will establish a robust implementation of a Python-based, automated workflow to train machine-learned interatomic potentials for entire compounds (in arbitrary configuration space) to compute accurate (anharmonic) phonon data. This is done by fully automated data generation, MLIP fit, and benchmark procedures. This significantly accelerates the training procedure and makes it more accessible and user-friendly for non-experts.

We tested our automatic training/fine-tuning workflow on five different challenges related to phonon properties, with varying levels of difficulty. 1) the harmonic phonon properties of a diverse and large number of silicon allotropes, 2) the harmonic phonon properties of the chemically more complex binary phase-change material Sb_2Se_3 where we expect long-range forces, 3) the quasi-harmonic properties of the Sn and NaTl systems, and 4) the thermal conductivity of β-Ga_2O_3. For the first two examples, we use one-shot training; for the last three, we use iterative training with intermediate benchmarks. The latter allows one to stop once a certain quality of the phonons in the benchmarks is reached.

Results

(i) A workflow for automated generation and evaluation of machine-learned interatomic potentials

Our automated workflow tool that we call "Automated machine-learned Potential Landscape Explorer" (AUTOPLEX)[18] aims to accelerate and enhance the machine-learned interatomic potential (MLIP) fitting process. We are starting with Gaussian approximation potentials (GAP)[19], in combination with the SOAP parameter[20] for atomic environments description, as our choice from the available MLIP architectures, and then expanding this to MACE potentials.[21] The workflow itself relies on the open source Python software packages atomate2,[22] pymatgen,[23] ASE,[24] jobflow,[25] that are closely associated with the Materials Project.[26] jobflow-remote,[27] and FireWorks[28] are used to manage workflow execution.

The workflow is divided into several stages, as shown in Fig. 1: data generation, MLIP fit, benchmark, and application step.

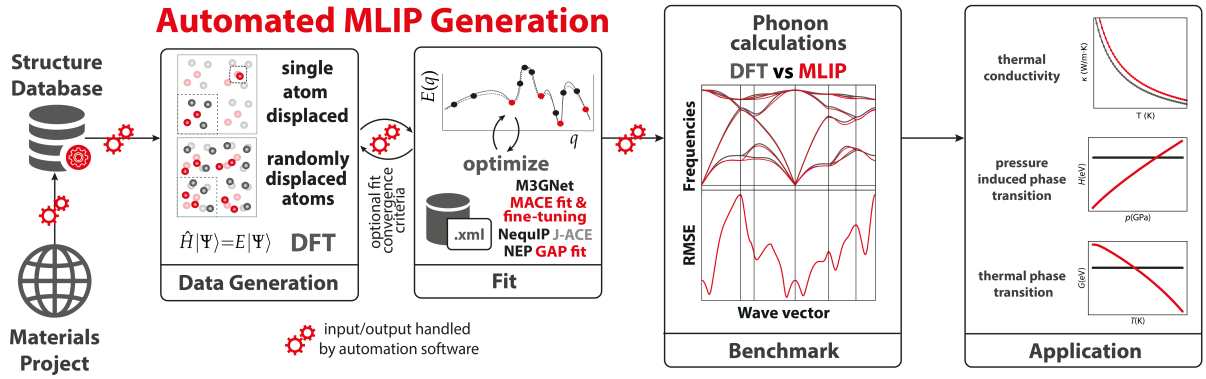

Figure 1. The workflow consists of four stages: (1) Data Generation, where atomic structures from a database (e.g., Materials Project) are perturbed (single-atom and random displacements) and their energies and forces computed via DFT; (2) MLIP Fitting, where (multiple) machine-learning interatomic potential models (GAP, M3GNet, MACE, NequIP, J-ACE, NEP) are trained and optimized; (3) Benchmarking, which compares phonon dispersion and RMSE between DFT and MLIP predictions; and (4) Application, where validated MLIPs are used for property predictions such as thermal conductivity, pressure-induced phase transitions, and thermal phase transitions. Automation software manages input/output and optional iterative fit convergence checks throughout the workflow..

In the data generation stage, we use the Materials Project database to build the dataset at the density functional theory (DFT) level, considering only Materials Project IDs (mpids) of crystalline phases. In principle, this dataset can be extended with additional sources, such as other (online) databases or alternative simulation techniques like Monte Carlo or molecular dynamics. The `autoplex` framework also includes a random structure search (RSS) automation workflow, enabling the construction of MLIPs applicable to a broad range of structural phases—crystalline, liquid, and amorphous.[29] Such random structure searches could also serve as a starting point for building a phonon database.

To create the training and test dataset, `autoplex` automatically generates two types of supercells for each structure. The first type consists of single-atom displaced supercells, which remain close to the pristine lattice and contain only one displaced atom. To enrich the dataset with more diverse chemical environments, rattled supercells—where all atoms are randomly displaced—are also added automatically. The number of rattled supercells can be controlled by the user, following the strategy employed in earlier work by some of the authors.[30] Future extensions may integrate uncertainty quantification and active learning approaches.[31]

In the fitting stage, the workflow interfaces with multiple MLIP frameworks (e.g., GAP, MACE, NequIP, J-ACE, NEP, M3GNet). The fitting procedure may include hyperparameter adjustment. For example, GAP training allows tuning the atom-wise regularization parameter f,[30] while full hyperparameter optimization for all models is planned for future integration via XPOT.[32] For MACE, we employ a fixed hyperparameter configuration that provides robust starting potentials for fine-tuning. Additionally, we utilize the pre-trained MACE-MP-0-3b model[12] and MPA[9,10], fine-tuning it on our dataset. It is worth noting that MACE-MP-0-3b and MPA were originally trained on PBE-level data, while our present DFT calculations predominantly use the PBEsol functional (with exceptions detailed in the Methods section). Despite this difference, fine-tuning significantly reduces the need for additional training data when building accurate MLIPs.

The benchmark stage assesses the accuracy of the MLIP by comparing phonon properties obtained from MLIPs against reference DFT calculations. Key metrics include the root-mean-square error (RMSE) over all phonon frequencies and q-point-resolved RMSE plots to capture detailed discrepancies. Additional benchmarking approaches can be integrated as needed.[33] In iterative mode, the workflow can automatically add new training data until the benchmark achieves a target RMSE for all phonon tasks or a pre-defined maximum number of iterations is reached.

Finally, the application stage deploys the validated MLIPs for advanced property predictions beyond phonons, such as lattice thermal conductivity, pressure-induced phase transitions, and temperature-driven phase transformations. These applications allow high-throughput exploration of materials behavior with near-DFT accuracy at a

fraction of the computational cost.

Overall, the workflow is designed for maximum flexibility: it supports easy extension with additional datasets, alternative MLIP models (e.g., ACE (as implemented in ACEpotentials.jl[34], NequIP[35], NEP[36], M3GNet[37]), as well as customization of DFT calculation parameters, hyperparameters, and automation strategies.[38]

The main results will focus on the following systems: Si, Sn, and Sb_2Se_3. Si and Sn will serve mainly as quality markers for benchmarking against literature concerning the comparison with DFT and existing MLIP (e.g., GAP-18[39] for Si) and to study their atomistic mechanisms further. Sb_2Se_3 is in the focus as an up-and-coming candidate for low thermal conducting thermoelectric materials and shows many similarities to SnS and SnSe.[40] Then, we extend the application of `autoplex` to the quasi-harmonic properties of the Zintl phase NaTl and the metal Sn and the thermal conductivity of β-Ga_2O_3.

(ii) Comparing custom GAP and MACE models with the MP0 foundation model and a fine-tuned MACE model

We are training a GAP and train/finetune several MACE potentials automatically using `autoplex` and automatically generated structure databases. While the graph-based MLIPs might be more powerful when it comes to accuracy, the speed and low-data requirements of GAP-based or other descriptor-based models still can have advantages for applications.

The Si database is constructed by drawing 13 different materials from the Materials Project database to represent the different Si allotropes. Then, the single-atom displaced and rattled supercells are constructed as described above and shown in Fig. 1. In summary, the Si database comprises 119 single-atom displaced supercells and 879 rattled supercells. The GAP fits are performed for several hyperparameter setups, and the MACE fits are performed with a fixed hyperparameter set. More details are given in the methods section. First, we compare the performance and accuracy of our implemented workflow by revisiting the state-of-the-art GAP potentials for Si.[30,39] The results of Ref.[30] for Si phases are reproduced using the automated workflow and shown in Fig. 2a). For ease of comparison, we will focus on diamond-type Si (mp-149), clathrate-I (mp-971662), and oS24 (mp-1095269) with the best-performing (hyper)parameter setup (n_sparse = 8000, f = 0.1, SOAP delta = 1.0) for GAP and provide the results for the other allotropes and parameter settings in the SI. The three different modes of data generation (single-atom displaced and rattled supercells, as well as the combined data) yield different performances. The single-atom-displaced and rattled supercells show an RMSE of 0.73 THz, 0.11 THz, 0.32 THz and 0.13 THz, 0.05 THz, 0.14 THz for mp-149, mp-971662, mp-1095269, respectively (cf. SI). With the combined dataset, the same RMSEs of 0.13 THz, 0.05 THz and 0.14 THz are yielded. Taking dia-Si (mp-149) as the benchmark, the GAP fit of the combined dataset performs slightly worse than demonstrated in the literature[30] (best RMSE around 0.08 THz), which can be attributed to different DFT levels (PBEsol here for Si vs. PW91 in the literature). Also, a larger correlation between rattled supercells and combined dataset RMSE can be seen (cf. SI and Ref.[30]), which is explained by a differing database size (998 supercells here vs. 833 supercells in Ref.[30]), while retaining the same number of single-atom displaced supercells as in the literature reference work.

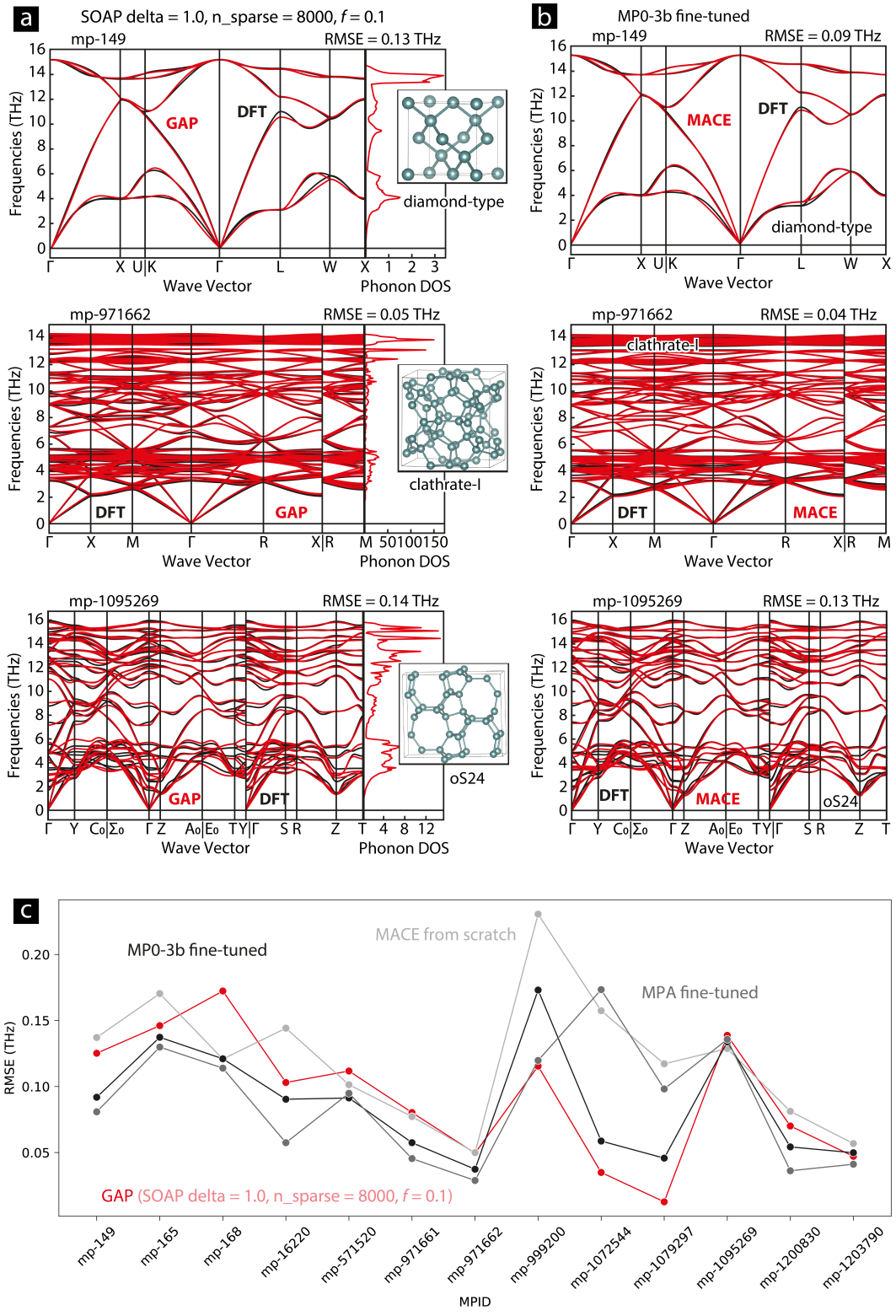

Figure 2. Si phonon bandstructure comparison and DOS for mp-149, mp-971662, and mp-1095269 a) for GAP. b) for fine-tuned MP0-3b. c) RMSE vs. MPID for GAP, MP0-3b fine-tuned, MPA fine-tuned, and MACE from scratch. The insets show the unit cell structures.

Now we take our previous Si database and repeat the MLIP fit using the MACE[21] framework. We will compare

a MACE potential trained from scratch, and fine-tuned MACE-MP-0-3b and MPA potentials[9,10,12]. Fig. 2b) shows the MP0-3b fine-tuned result. The overall MACE results are shown in Fig.S1c) to e). In case of the MACE potential trained from scratch, the energy and force weights were varied in five consecutive runs, until the RMSE was converging to a particular value (more details in the Method section), which is 0.14 THz in the case of mp-149, 0.05 THz for mp-971662 and 0.13 THz for mp-1095269, which are about the same RMSE values as for the GAP trained from scratch. The best result is yielded by the fine-tuned MP0-3b and MPA models as shown in Fig. 2b) and c). Here, the advantages of fine-tuning in contrast to training from scratch become apparent, even in a case where a different level of theory has been used for the fine-tuning than for the foundation model, such as MACE-MP-0, which was trained on PBE data (particularly relevant in the Si comparison which was generated at PBEsol level).

Next, we take a look at Sb_2Se_3. The structure database is constructed in the same way as for Si. It consists of 41 single-atom displaced and 606 rattled supercells, drawn from the three only available structures for Sb_2Se_3 in the Materials Project database. As the other two MPIDs are very close in the structure to mp-2160, and have a similar phonon bandstructure, just that they show some imaginary modes (as shown in the SI), we will focus on mp-2160 for the results. Here again, we will compare GAP vs. MACE trained from scratch vs. MP0-3b vs. fine-tuned MP0-3b. Several hyperparameters and other settings have been tested as well (cf. SI and method section).

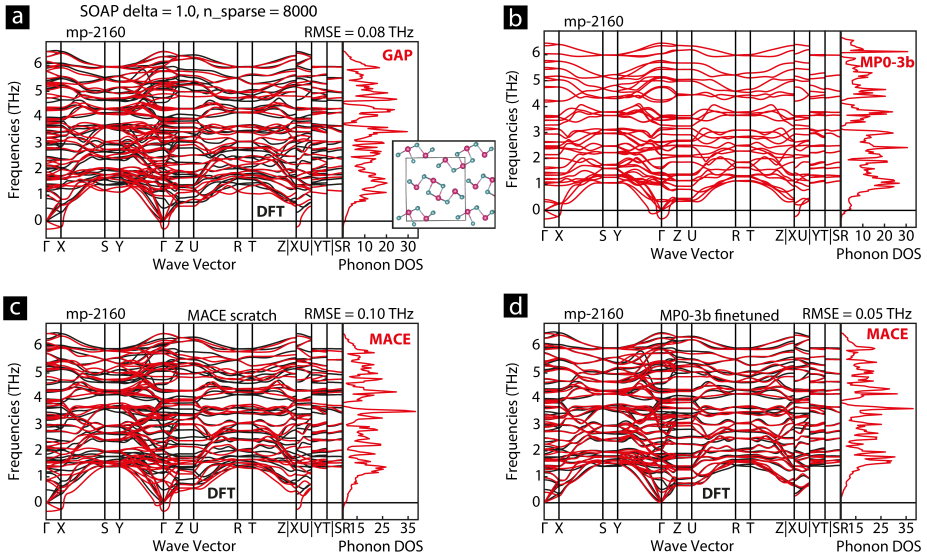

Figure 3. The performance of the different MLIPs for the chemically more complex Sb_2Se_3 system. a) for GAP. b) for MP0. c) for MACE trained from scratch. d) fine-tuned MP0-3b. The inset shows the unit cell structure.

As can be seen in Fig. 3, the GAP and MACE trained from scratch perform similarly again (RMSE = 0.1 THz). The bandstructures derived from those three potentials also show some imaginary modes. Only the fine-tuned MP0-3b model shows the expected bandstructures without any imaginary modes and has an RMSE of 0.05 THz. It is suspected that the GAP and MACE from scratch RMSE and bandstructure would further improve with more diverse data. Also, it demonstrates the clear advantages of fine-tuning an existing foundation model, which enables us (and the users of `autoplex` to save resources while yielding higher accuracies. In the following case study of Sn, we will even demonstrate how little data it takes to achieve a significant improvement by fine-tuning a foundation model. Beyond fine-tuning, the data generation procedure outlined here can also be combined with other existing training databases, e.g., as used for the amorphous properties of Sb_2Se_3, and extended to device-scale dimensions as demonstrated in Ref.41.

Our next case study is Sn. Sn shows structural similarities to Si. α-Sn is diamond-type Sn, and β-Sn is known from the Si configurational space as well. The phase transition of *alpha*-Sn to β-Sn at around 286 K is a well-known example of a structural transition. PBE is known to overestimate the phase transition (e.g., 475 K in the harmonic approximation).[42] Nevertheless, we will use PBE here as it gets the phase transition at least qualitatively right.[42] At

the PBEsol level of theory, α-Sn is less stable than β-Sn.

For the fine-tuned MACE potentials, we take 26 rattled supercells in each iteration and fit one MACE potential using the α-Sn and β-Sn MPIDs only, and a second MACE potential where we will use seven MPIDs (for more details, see the method section). For the MACE potential trained from scratch, we have used the largest training set from the last fine-tuning iteration.

Fig.4 shows the phonon bandstructures for the MACE trained from scratch potential and the two fine-tuned MACE potentials together with a comparison with MP0-3b. Additionally, the aim is to reproduce the typical DFT-based phase diagrams like in Ref.[7] using the MACE MLIPs. A detailed analysis of the Sn data metrics can be found in the SI. We will discuss the results of the potential fits with the best RMSEs.

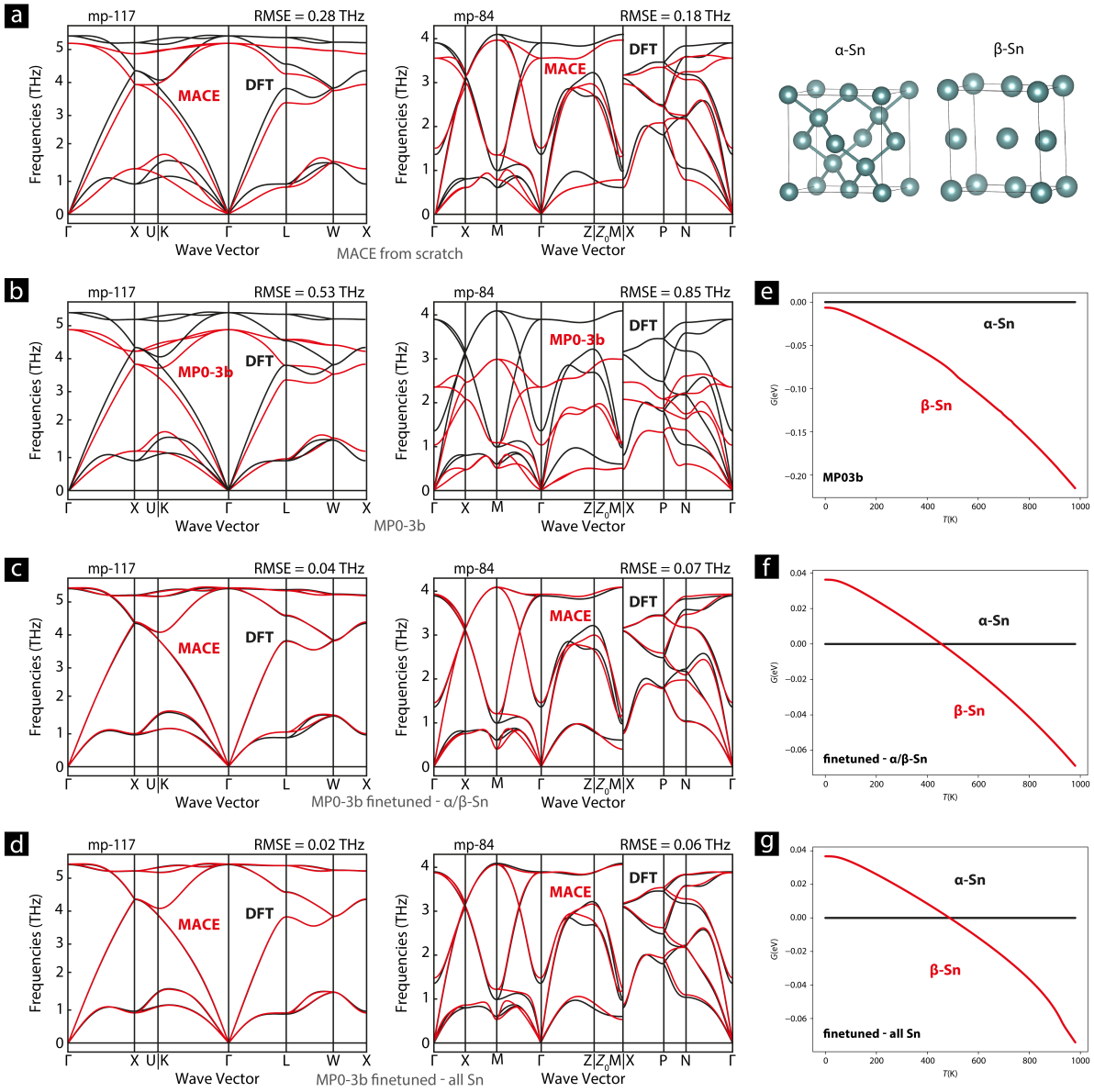

Figure 4. Sn plots a) Phonon bandstructure comparison for a MACE potential trained from scratch. b) Phonon bandstructure comparison MP0-3b. c) Phonon bandstructure comparison fine-tuned MACE α/β-Sn only d) Phonon bandstructure comparison fine-tuned MACE for all seven MPIDs. e+f+g) Temperature phase transition from α to β-Sn for MP0-3b and the fine-tuned MACE potentials, respectively.

The two fientuned MP0-3b potentials (Fig.4c, d) reproduce DFT-accurate results with an RMSE that is distinctly below 0.1 THz, while the potential trained from scratch (a) and the foundation model (b) show higher RMSEs. Fig.4e) also demonstrates that MP0-3b does not predict a temperature-dependent phase transition at all. In case of fine-tuned MP0-3b potentials, a phase transition is predicted at around 450 K and 500 K, respectively. This still overestimates the DFT results of ca. 400 K[7]; however, our aim here was to demonstrate how little data it needs to improve the results of the foundation models by fine-tuning. A next step could be to implement automated workflows for iterating the data and potential generation until a sufficiently accurate transition temperature is reached. The Sn temperature phase transition also shows that the phonon bandstructures can be close to DFT-accuracy and yet yield a difference of 50 K and 100 K, respectively, for the transition temperature.

(iii) Advanced benchmarks

Our first candidate for the advanced benchmark is NaTl. Here, we fine-tuned the foundation model MP0-3b iteratively and aimed at including as little data as possible (169 supercell structures, details in the method section).[43–46] Because of the pressure-induced phase transition from the [NaTl] to [CsCl] structure type,[43,44] we included these two structure types in the training data, as well as the [NaCl] and [ZnS] structure types. Fig. 5 depicts the phonon bandstructures for the fine-tuned MACE (a) and the MP0-3b (b) potentials for the [NaTl] (mp-1564) and [CsCl] structure types, accompanied by the respective pressure-induced phase transition diagram.

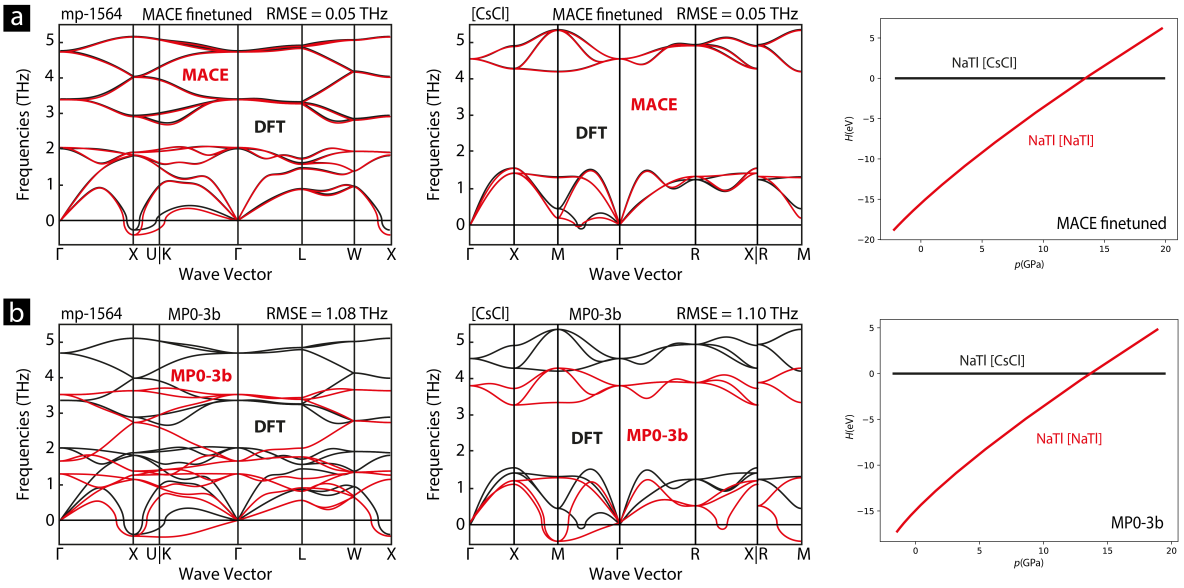

Figure 5. NaTl plots: a) Phonon bandstructure comparison and pressure transition MACE finetuned. b) Phonon bandstructure comparison and pressure transition MP0-3b.

Given the comparatively small training set, NaTl demonstrates the most significant improvement in the fit quality via fine-tuning with an RMSE difference of more than 1 THz. Despite the good phonon bandstructure description by the fine-tuned model, the predicted transition pressure of around 13 GPa is below what can be found in the literature (30 to 40 GPa).[44] As seen in Fig.S13, this value is also already converged given the current data set. It is interesting to note that the predicted transition pressure from the foundation model MP0-3b does not differ much. The MPA fine-tuned model performs the same or slightly worse than the fine-tuned MP0-3b model as can be seen in Fig.S13.

Finally, the last example for the advanced benchmark of `autoplex` generated potentials is β-Ga_2O_3. As a promising superconductor candidate known for the low thermal conductivity of the pristine and defective phase,[47–49] it is a perfect case study for calculating the thermal conductivity from MLIPs. In this case, we only included β-Ga_2O_3 supercell structures in the training data (400 rattled supercell structures, see method section), and again fine-tuned the foundation model MP0-3b. For estimating the thermal conductivity, we have used the linear Boltzmann

transport equation (LBTE) at 300K and the Boltzmann relaxation time approximation (RTA, see Fig.S11, including the non-analytical term correction (NAC).

Fig. 6 shows the comparison between the foundation and fine-tuned model for the phonon bandstructure and the total thermal conductivity (κ_{total}).

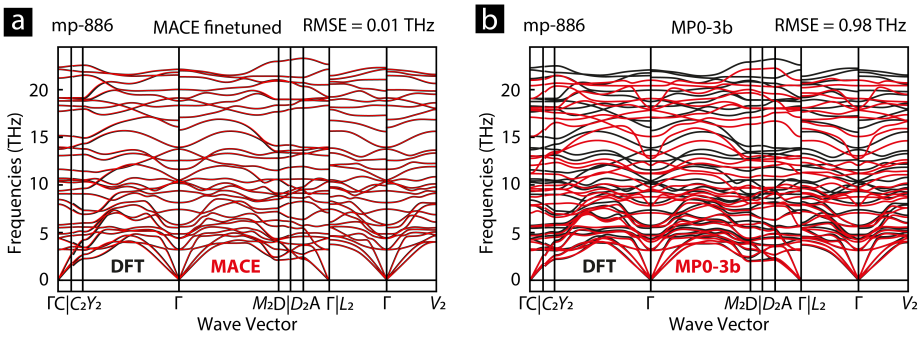

Figure 6. Ga_2O_3 plots: a) Phonon bandstructure comparison MP0-3b finetuned. b) Phonon bandstructure comparison MP0-3b.

Here again, we can demonstrate that the fine-tuning improves the fit quality considerably. As the training data only contained β-Ga_2O_3, it is not too surprising that the fine-tuned RMSE is very low. As for the thermal conductivity, with 9.5 $\mathrm{W/m \cdot K}$, 17.3 $\mathrm{W/m \cdot K}$ and 16.9 $\mathrm{W/m \cdot K}$ for κ_{100}, κ_{010} and κ_{001} respectively (cf. Eq.3, using the fine-tuned MP0-3b model at 300 K, the value is underestimating the experimental[48,50], DFT[48,49] and GAP[47] values by a few (around 3) $\mathrm{W/m \cdot K}$. Therefore, the agreement in predicting the thermal conductivity by the fine-tuned model and the literature is quite good, given the rather restrictive training set.

Discussion

Using the flexibility and modularity of our implemented automated workflow, we compare the replicability, accuracy, duration, and resource consumption of DFT vs. GAP and MACE-based phonon structure calculations. Starting with revisiting Ref.[[30]], we demonstrate that we can successfully reproduce the Si phonon structure therein, using a similarly constructed database and calculation setups. Also, combining single-atom displaced and rattled supercells yielded a comparable increase in the accuracy and the respective reduction of the average RMSE to around 0.1 THz. For Sn and Sb_2Se_3, a similar effect of extending the data set to reduce the RMSE is observed. Especially interesting is that it only takes the addition of a few data points to fine-tune the foundation model MP0-3b for Sn to predict a phase transition temperature that is somewhat close to the DFT literature value.

Regarding the advanced benchmark, it could be well demonstrated to be able to predict further thermal properties of NaTl and Ga_2O_3, while the predicted values for Ga_2O_3 agree better with the literature than for NaTl. This demonstrates a first success for the implementation of `autoplex` and leaves space for improvements and future developments.

With the current rise in using machine-learned interatomic potentials and an automated workflow for high-throughput research,[29] we provide an agile tool to accelerate further and enhance progress in this research area. The current state of the workflow will be extended by means of the data set generation and benchmark metrics. As we make the source code of the automated workflow open-source accessible, it is further customizable and expandable by individual users. The phonon properties are also expandable: Strong quartic anharmonicity, higher order/degree force constants, a systematic analysis of the structural influence on amorphous-like heat conduction, and 3- and 4-phonon interactions will follow.[51] Other extension possibilities are to implement interfaces to DFT software other than VASP and add the possibility of generating student potentials (an MLIP trained from another teacher MLIP) via an ASE interface. In the case of MACE, multi-head fine-tuning can be implemented in addition to the transfer learning.

Methods

General framework

The `autoplex` internal workflows are based on the Materials Project[26] framework and interface with several open-source Python software packages. The focus is on automating workflows and their individual jobs and tasks. For managing DFT settings, as well as the input and output files atomate2,[22] pymatgen,[23] and ASE,[24] are used. jobflow,[25], and jobflow-remote,[27] are used to manage the workflows and their jobs. The current work focuses on applying `autoplex` on crystalline phases for phonon calculations. There already exists an RSS workflow for learning potential-energy surfaces from random structures.[29] An approach to combine both workflows is planned for the future.

DFT calculations

The DFT training data were obtained using projector augmented-wave (PAW) potentials[52,53] as implemented in the Vienna *Ab Initio* Simulation Package (VASP)[53,54] using the generalized gradient approximation (GGA) as parameterized by the Perdew–Burke–Ernzerhof functional revised for solids (PBEsol).[55,56] For Sn and NaTl, the PBE functional was used. The DFT settings in VASP and the `phonopy`[5,6] settings were chosen similar to Ref.[30]. A kinetic energy cutoff of 700 eV and a k-point spacing of 0.2 (as implemented in VASP) were chosen for sampling the Brillouin zone. The Boltzmann transport equation as implemented in `phono3py`[5,6,57] was used for the minimal thermal conductivity calculations for Ga_2O_3 with a q-point density of 100. The supercell size for the `phonopy` calculations was automatically chosen within `autoplex`, and further `phono3py` calculations were conducted with the same settings.

MLIP specifications

The machine learning algorithms are learning the potential-energy surface based on the energy and forces of the reference data (DFT) and the respective representation of the atomic environment of the model. In a kernel-based approach like GAP[19], the atomic energies are interpolated via a kernel function that compares the environment similarity with the reference data. In case of neural network-based models like MACE,[21] atomic energies are learned by message-passing between neighboring atoms to capture local interactions, and this information is propagated through multiple layers of the network to refine the representations and predict the energies iteratively. For GAP and MACE, we are iterating through a set of (hyper)parameters and checking the convergence of the fits as shown in Fig.S1 to S4 for Si, Fig.S5 to S7 for Sn, and Figs.S9, S10 for Sb_2Se_3. As expected, more data and a higher accuracy of the calculation setup lead to lower RMSE values. For Sn, we also trained a GAP potential (see Fig.S5) by adding single-atom displaced and rattled structures with different linear strain ($\pm$20 %) for both α-Sn and β-Sn. In this case, we first added 36 single displaced structures and around 26 rattled structures with a standard deviation of 0.1 in the first iteration. Subsequently, we added 26 additional rattled structures in each iteration.

In the iterative mode of `autoplex` for Sn, NaTl, and Ga_2O_3, the number of data points per iteration is given in Tab.4.

ML phonon settings and benchmark

In case of MLIP-based phonons, we rely on the force field `PhononMaker` in atomate2[22] using the BFGS optimization algorithm.

`autoplex` relies on the RMSE values and the q-dependent RMSE plots for benchmarking. Their definition is given below.

The RMSE is calculated as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{NM}\sum_{i=1}^{N}\sum_{j=1}^{M}\left(\omega_{i,j} - \hat{\omega}_{i,j}\right)^2}. \tag{1}$$

Here, N represents the number of q-points per band, and M denotes the number of bands. $\hat{\omega}_{i,j}$ is the frequency of the i-th k-point and j-th band, while $\omega_{i,j}$ corresponds to the comparable frequencies, such as those obtained through machine learning.

The q-dependent RMSE is defined as:

$$\mathrm{RMSE}_q = \sqrt{\frac{1}{M}\sum_{j=1}^{M}\left(\omega_{q,j} - \hat{\omega}_{q,j}\right)^2}, \tag{2}$$

where we do not sum over N to keep the q-point dependency of the metric.

The LBTE thermoconductivity κ was evaluated as follows:

$$\begin{aligned}
\kappa_{100} &= k_{yy}\cos^2\beta + k_{yz}\sin 2\beta + k_{zz}\sin^2\beta = 9.5075\ \tfrac{\mathrm{W}}{\mathrm{m}\cdot\mathrm{K}} \\
\kappa_{010} &= k_{xx} = 17.3355\ \tfrac{\mathrm{W}}{\mathrm{m}\cdot\mathrm{K}} \\
\kappa_{001} &= k_{zz} = 16.9416\ \tfrac{\mathrm{W}}{\mathrm{m}\cdot\mathrm{K}} \\
T &= 300\ \mathrm{K}
\end{aligned} \tag{3}$$

Data and Code Availability

The open source code of `autoplex` is provided on GitHub.[18]

References

1. Dove, M. T., Putnis, A., Liebermann, R. C. & Hochella, M. F. *Introduction to Lattice Dynamics* (Cambridge University Press, Cambridge, GBR, 1993). OCLC: 958551881.
2. Petretto, G. *et al.* High-throughput density-functional perturbation theory phonons for inorganic materials. *Sci. Data* **5**, 180065, 10.1038/sdata.2018.65 (2018).
3. Lee, H., Hegde, V. I., Wolverton, C. & Xia, Y. Accelerating high-throughput phonon calculations via machine learning universal potentials. *Mater. Today Phys.* **53**, 101688, 10.1016/j.mtphys.2025.101688 (2025).
4. Ohnishi, M. *et al.* Database and deep-learning scalability of anharmonic phonon properties by automated brute-force first-principles calculations, 10.48550/arXiv.2504.21245 (2025). ArXiv:2504.21245 [cond-mat].
5. Togo, A., Chaput, L., Tadano, T. & Tanaka, I. Implementation strategies in phonopy and phono3py. *J. Phys. Condens. Matter* **35**, 353001, 10.1088/1361-648X/acd831 (2023).
6. Togo, A. First-principles phonon calculations with phonopy and phono3py. *J. Phys. Soc. Jpn.* **92**, 012001, 10.7566/JPSJ.92.012001 (2023).
7. Stoffel, R., Wessel, C., Lumey, M.-W. & Dronskowski, R. Ab initio thermochemistry of solid-state materials. *Angewandte Chemie Int. Ed.* **49**, 5242–5266, https://doi.org/10.1002/anie.200906780 (2010). https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.200906780.
8. Jacobs, R. *et al.* A practical guide to machine learning interatomic potentials – Status and future. *Curr. Opin. Solid State Mater. Sci.* **35**, 101214, 10.1016/j.cossms.2025.101214 (2025).
9. Batatia, I., Kovacs, D. P., Simm, G. N. C., Ortner, C. & Csanyi, G. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In Oh, A. H., Agarwal, A., Belgrave, D. & Cho, K. (eds.) *Advances in Neural Information Processing Systems* (2022).
10. Batatia, I. *et al.* The design space of e(3)-equivariant atom-centered interatomic potentials, 10.48550/arXiv.2205.06643 (2022). 2205.06643.
11. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728, 10.1038/s43588-022-00349-3 (2022). Publisher: Springer Science and Business Media LLC.
12. Batatia, I. *et al.* A foundation model for atomistic materials chemistry (2024). 2401.00096.
13. Riebesell, J. *et al.* A framework to evaluate machine learning crystal stability predictions. *Nat. Mach. Intell.* **7**, 836–847, 10.1038/s42256-025-01055-1 (2025). Publisher: Nature Publishing Group.

14. Loew, A., Sun, D., Wang, H.-C., Botti, S. & Marques, M. A. L. Universal machine learning interatomic potentials are ready for phonons. *npj Comput. Mater.* **11**, 178, 10.1038/s41524-025-01650-1 (2025). Publisher: Nature Publishing Group.

15. Gardner, J. L. A. *et al.* Distillation of atomistic foundation models across architectures and chemical domains, 10.48550/arXiv.2506.10956 (2025). ArXiv:2506.10956 [physics].

16. Póta, B., Ahlawat, P., Csányi, G. & Simoncelli, M. Thermal Conductivity Predictions with Foundation Atomistic Models, 10.48550/arXiv.2408.00755 (2025). ArXiv:2408.00755 [cond-mat].

17. Janssen, J. *et al.* pyiron: An integrated development environment for computational materials science. *Comput. Mater. Sci.* **163**, 24–36, 10.1016/j.commatsci.2018.07.043 (2019).

18. Ertural, C. *et al.* Autoplex v0.1.0 (2024).

19. Bartók, A. P., Payne, M. C., Kondor, R. & Csányi, G. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403, 10.1103/PhysRevLett.104.136403 (2010).

20. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115, 10.1103/PhysRevB.87.184115 (2013).

21. Batatia, I., Kovacs, D. P., Simm, G., Ortner, C. & Csanyi, G. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In Koyejo, S. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 35, 11423–11436 (Curran Associates, Inc., 2022).

22. Ganose, A. M. *et al.* Atomate2: modular workflows for materials science, 10.1039/D5DD00019J (2025).

23. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319, https://doi.org/10.1016/j.commatsci.2012.10.028 (2013).

24. Hjorth Larsen, A. *et al.* The Atomic Simulation Environment—a Python library for working with atoms. *J. Phys.: Condens. Matter* **29**, 273002, 10.1088/1361-648x/aa680e (2017).

25. Rosen, A. S. *et al.* Jobflow: Computational Workflows Made Simple. *J. Open Source Softw.* **9**, 5995, 10.21105/joss.05995 (2024).

26. Jain, A. *et al.* Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Mater.* **1**, 011002, 10.1063/1.4812323 (2013).

27. Petretto, G. *et al.* jobflow-remote v0.1.4 (2024).

28. Jain, A. *et al.* FireWorks: a dynamic workflow system designed for high-throughput applications. *Concurr. Comput.* **27**, 5037, 10.1002/cpe.3505 (2015).

29. Liu, Y. *et al.* An automated framework for exploring and learning potential-energy surfaces (2024). 2412.16736.

30. George, J., Hautier, G., Bartók, A. P., Csányi, G. & Deringer, V. L. Combining phonon accuracy with high transferability in gaussian approximation potential models. *J. Chem. Phys.* **153**, 044104, 10.1063/5.0013826 (2020). https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/5.0013826/14719853/044104_1_online.pdf.

31. Erhard, L. C., Rohrer, J., Albe, K. & Deringer, V. L. Modelling atomic and nanoscale structure in the silicon–oxygen system through active machine learning. *Nat. Commun.* **15**, 1927, 10.1038/s41467-024-45840-9 (2024). Publisher: Nature Publishing Group.

32. Thomas du Toit, D. F. & Deringer, V. L. Cross-platform hyperparameter optimization for machine learning interatomic potentials. *The J. Chem. Phys.* **159** (2023).

33. Morrow, J. D., Gardner, J. L. A. & Deringer, V. L. How to validate machine-learned interatomic potentials. *The J. chemical physics* **158** (2023).

34. Witt, W. C. *et al.* ACEpotentials.jl: A Julia implementation of the atomic cluster expansion. *J. Chem. Phys.* **159**, 164101, 10.1063/5.0158783 (2023).

35. Batzner, S. *et al.* E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **13**, 2453, 10.1038/s41467-022-29939-5 (2022).

36. Fan, Z. *et al.* Neuroevolution machine learning potentials: Combining high accuracy and low cost in atomistic simulations and application to heat transport. *Phys. Rev. B* **104**, 104309, 10.1103/PhysRevB.104.104309 (2021).

37. Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.* **2**, 718–728, 10.1038/s43588-022-00349-3 (2022).

38. Developers, A. AutoPLEX documentation. https://autoatml.github.io/autoplex/ (2025). Accessed: 2025-06-20.

39. Bartók, A. P., Kermode, J., Bernstein, N. & Csányi, G. Machine learning a general-purpose interatomic potential for silicon. *Phys. Rev. X* **8**, 041048, 10.1103/PhysRevX.8.041048 (2018).

40. Lanigan-Atkins, T. *et al.* Extended anharmonic collapse of phonon dispersions in sns and snse. *Nat. Commun.* **11**, 4430 (2020).

41. Zhou, Y., Zhang, W., Ma, E. & Deringer, V. L. Device-scale atomistic modelling of phase-change memory materials. *Nat. Electron.* **6**, 746–754 (2023).

42. Legrain, F. & Manzhos, S. Understanding the difference in cohesive energies between alpha and beta tin in dft calculations. *AIP Adv.* **6**, 045116, 10.1063/1.4948434 (2016). https://pubs.aip.org/aip/adv/article-pdf/doi/10.1063/1.4948434/12878635/045116_1_online.pdf.

43. Wang, F. & Miller, G. J. Revisiting the zintl–klemm concept: Alkali metal trielides. *Inorg. Chem.* **50**, 7625–7636, 10.1021/ic200643f (2011). PMID: 21774461, https://doi.org/10.1021/ic200643f.

44. Ozisik, H., Colakoglu, K., Surucu, G. & Ozisik, H. Structural and lattice dynamical properties of zintl nain and natl compounds. *Comput. Mater. Sci.* **50**, 1070–1076, https://doi.org/10.1016/j.commatsci.2010.11.003 (2011).

45. OQMD. The open quantum materials database. https://oqmd.org/materials/entry/1223487 (2025). Accessed: 2025-06-23.

46. Wang, F. & Miller, G. J. Revisiting the zintl–klemm concept: A2aubi (a = li or na). *Eur. J. Inorg. Chem.* **2011**, 3989–3998, https://doi.org/10.1002/ejic.201100312 (2011). https://chemistry-europe.onlinelibrary.wiley.com/doi/pdf/10.1002/ejic.201100312.

47. Liu, Y.-B. *et al.* Machine learning interatomic potential developed for molecular simulations on thermal properties of -ga2o3. *The J. Chem. Phys.* **153**, 144501, 10.1063/5.0027643 (2020). https://pubs.aip.org/aip/jcp/article-pdf/doi/10.1063/5.0027643/13373286/144501_1_online.pdf.

48. Yan, Z. & Kumar, S. Phonon mode contributions to thermal conductivity of pristine and defective -ga2o3. *Phys. Chem. Chem. Phys.* **20**, 29236–29242, 10.1039/C8CP05139A (2018).

49. Safieddine, F., El Haj Hassan, F. & Kazan, M. Theoretical investigation of the thermal conductivity of ga2o3 polymorphs. *Solid State Commun.* **394**, 115715, https://doi.org/10.1016/j.ssc.2024.115715 (2024).

50. Klimm, D. *et al.* The thermal conductivity tensor of -ga2o3 from 300 to 1275 k. *Cryst. Res. Technol.* **58**, 2200204, https://doi.org/10.1002/crat.202200204 (2023). https://onlinelibrary.wiley.com/doi/pdf/10.1002/crat.202200204.

51. Bernstein, N., Csányi, G. & Deringer, V. L. De novo exploration and self-guided learning of potential-energy surfaces. *npj Comput. Mater.* **5**, 99, 10.1038/s41524-019-0236-6 (2019).

52. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979, 10.1103/PhysRevB.50.17953 (1994).

53. Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775, 10.1103/PhysRevB.59.1758 (1999).

54. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186, 10.1103/PhysRevB.54.11169 (1996).

55. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868, 10.1103/PhysRevLett.77.3865 (1996).

56. Perdew, J. P. *et al.* Restoring the Density-Gradient Expansion for Exchange in Solids and Surfaces. *Phys. Rev. Lett.* **100**, 136406 (2008).

57. Togo, A., Chaput, L. & Tanaka, I. Distributions of phonon lifetimes in brillouin zones. *Phys. Rev. B* **91**, 094306, 10.1103/PhysRevB.91.094306 (2015).

Acknowledgements

CE and JG would like to acknowledge the Gauss Centre for Super Computing e.V. (www.gauss-centre.eu) for funding this project by providing generous computing time on the GCS Supercomputer SuperMUC-NG at Leibniz Super Computing Centre (www.lrz.de) (project pn73da). CE thanks Philipp Beckmann, Björn Schrader and Jörg Rädler from the IT department of BAM for the very intense technical support and Lauren Matthews for her feedback on the manuscript in BAM's MatChIngCamp.

DFG and ERC funding need to be added here. This work was supported through a UK Research and Innovation Frontier Research grant [grant number EP/X016188/1].