RESEARCH ARTICLE | AUGUST 22 2024

Derivative learning of tensorial quantities—Predicting finite temperature infrared spectra from first principles

Bernhard Schmiedmayer ; Georg Kresse

J. Chem. Phys. 161, 084703 (2024)
https://doi.org/10.1063/5.0217243

The Journal of Chemical Physics ARTICLE pubs.aip.org/aip/jcp

Derivative learning of tensorial quantities—Predicting finite temperature infrared spectra from first principles

Cite as: J. Chem. Phys. 161, 084703 (2024); doi: 10.1063/5.0217243
Submitted: 3 May 2024 • Accepted: 5 August 2024 •
Published Online: 22 August 2024

Bernhard Schmiedmayer[1,a] and Georg Kresse[1,2]

AFFILIATIONS

[1] Faculty of Physics and Center for Computational Materials Science, University of Vienna, Kolingasse 14-16, A-1090 Vienna, Austria
[2] VASP Software GmbH, Sensengasse 8, A-1090 Vienna, Austria

[a] Author to whom correspondence should be addressed: bernhard.schmiedmayer@univie.ac.at

ABSTRACT

We develop a strategy that integrates machine learning and first-principles calculations to achieve technically accurate predictions of infrared spectra. In particular, the methodology allows one to predict infrared spectra for complex systems at finite temperatures. The method's effectiveness is demonstrated in challenging scenarios, such as the analysis of water and the organic–inorganic halide perovskite $MAPbI_3$, where our results consistently align with experimental data. A distinctive feature of the methodology is the incorporation of derivative learning, which proves indispensable for obtaining accurate polarization data in bulk materials and facilitates the training of a machine learning surrogate model of the polarization adapted to rotational and translational symmetries. We achieve polarization prediction accuracies of about 1% for the water dimer by training only on the predicted Born effective charges.

I. INTRODUCTION

Infrared (IR) spectroscopy is an indispensable methodology for discerning local structures and the intricacies of functional groups within a material. It finds pervasive use across various scientific disciplines, firmly establishing itself as a key analytical instrument.[1] Notably, within the realm of catalysis research, IR spectroscopy is an important technique with the unique capability to elucidate surface structures and mechanisms at the molecular scale *in situ*.[2–4] Beyond catalysis, the application of IR spectroscopy extends its reach into the domains of clinical and biomedical analyses.[5–7] In the field of materials science, IR spectroscopy is also highly valuable, as it provides indirect insight into the bonding properties of the material.[8–10]

Although experimental IR spectroscopy is a mature technique, it is not always a simple matter to relate an experimental IR spectrum to a specific local structural feature. Prior knowledge is almost ubiquitously required to relate experimental data to structures (fingerprints). For instance, the frequencies of certain functional groups often relate to the oxidation state of the group or the electronegativity of the adsorption site. Computer simulations, particularly first-principles (FP) methods, have played a vital role in establishing these relationships. However, such FP calculations are very often limited to zero-temperature simulations,[11–13] whereas *in situ* observations of catalytic reactions almost always involve finite temperature. In this study, we harness recent advancements in machine learning (ML) techniques in combination with FP calculations to develop a method that yields highly accurate computational IR spectra at finite temperatures.

In recent years, ML techniques have emerged as a potent new avenue in computational materials sciences.[14–18] Our study uses an on-the-fly learning method[19–22] to generate transferable machine learned force fields (MLFFs). The first step is thus to harness MLFF to attain the required nano-second-long high-quality molecular dynamics (MD) trajectories. In the present work, we rely on a now fairly standard approach for the MLFF that uses rotational and translationally invariant descriptors and a kernel-based regression, although more refined approaches could be readily adopted.[23–25] Learning the polarization, although, is a more challenging task. The first approach to learn tensorial quantities dates back to

λ-SOAP (Smooth Overlap of Atomic Potentials)[26–28] that we have also decided to adapt in the present work. Equivariant message-passing networks would be equally suitable, but they are matter of fact closely related to the tensorial descriptors in λ-SOAP.

The second key issue that we address in the present work is that the polarization is not uniquely determined in bulk materials.[29–32] This is in stark contrast to molecules, where the polarization can be determined by appropriate integration of the density.[33] Direct learning of the polarization hence requires some curation of the polarization data, e.g., deriving the polarization from Wannier centers imposing some continuity condition[34] or manual "alignment" of the polarization data.

The solution presented in this study to overcome this challenge involves employing derivative learning.[35] In particular, we utilize the derivative of the polarization with respect to the ionic positions, the Born effective charge tensors. The hypothesis is that by computing this derivative information across numerous structures (along any relevant "adiabatic" pathway), it becomes feasible to determine the anti-derivative, i.e., the bulk polarization. This methodology offers a second crucial advantage: akin to the construction of MLFF where learning the forces proved pivotal, the Born effective charges are significantly more expressive yet nearly as computationally efficient to calculate in solid-state codes as a single polarization.

In the Sec. II, we summarize our methodological approach, then demonstrate the feasibility of derivative-based learning for the water dimer, and discuss the results for liquid water, where we find excellent agreement with the experiment only for a functional, including van der Waals corrections. Finally, we demonstrate very good agreement for the IR spectrum of an organic perovskite with experimental data. We finish with discussions and conclusions, as well as identify points worthy of improvement.

II. METHOD

A. General remarks

In general, the energy of a molecule or material in the presence of an electric field is described by

$$E(\mathbf{x},\mathscr{E}) = E_{\mathrm{KS}}(\mathbf{x}) - \mathscr{E}\cdot\mathbf{P}(\mathbf{x},\mathscr{E}), \tag{1}$$

where $\mathscr{E}$ is the electric field, $\mathbf{P}$ is the polarization, and E_{KS} is the Kohn–Sham energy at zero field for the atoms at the position $\mathbf{x}$.[36] (We note that we follow the convention of Ref. 36 and define the polarization to be extensive.) There is an implicit dependence of the Kohn–Sham energy on the electronic field, as the orbitals need to be determined by minimizing the energy in the presence of the field, but thanks to the variational properties and the Hellman–Feynman theorem, variations of the orbitals can be neglected for first derivatives. To calculate second derivatives, only the first derivatives of the orbitals are required. Obviously, the first derivative of the energy with respect to the electric field yields the polarization $\mathbf{P}$. The second derivative of the energy with respect to the field corresponds to the electronic polarizability, and the second mixed derivative with respect to the electric field and the positions yields the Born effective charge tensor,

$$\mathbf{Z}^* = \frac{\partial^2 E(\mathbf{x},\mathscr{E})}{\partial\mathbf{x}\,\partial\mathscr{E}}. \tag{2}$$

This is a second-rank Cartesian tensor. In the present work, we set out to learn polarization as a function of the positions and neglect the dependence of polarization on the electric field (implicitly assuming zero electric field). If we were to evaluate the energy using Eq. (1), this would only be correct to linear order in the field.

B. Green–Kubo relation

Using the Green–Kubo formalism, the ionic contribution to the polarizability, denoted as $\chi(\omega)$, is directly proportional to the Fourier transform of the autocorrelation function of the polarization $\mathbf{P}$ and its time derivative and can be expressed as follows (SI units):[37–39]

$$\chi_{\mu,\nu}(\omega) = \frac{\beta}{V\varepsilon_0}\int_0^T \langle \mathbf{P}_\mu(0)\dot{\mathbf{P}}_\nu(t)\rangle e^{-i(\omega-i\delta)t}\mathrm{d}t. \tag{3}$$

Here, μ and ν represent Cartesian indices, V is the volume, β is the inverse temperature, ε_0 is the vacuum permittivity, ω is the vibrational frequency, and δ denotes a complex shift causing a Lorenzian broadening. It is necessary that the expression $\exp(-\delta T)$ is small to avoid truncation artefacts. The time derivative of the polarization $\dot{\mathbf{P}}$ can be written as

$$\frac{\partial\mathbf{P}}{\partial t} = \frac{\partial\mathbf{P}}{\partial\mathbf{x}}\frac{\partial\mathbf{x}}{\partial t} = \mathbf{Z}^*\dot{\mathbf{x}}. \tag{4}$$

To determine Eq. (3) accurately, three prerequisites exist: first, accurate velocities $\dot{\mathbf{x}}$, i.e., high-quality MD trajectories to describe the time evolution of the system; second, long simulation times T; and third, a reliable method to model the polarization and the Born effective charge tensors.

In infrared reflectivity or transmission experiments, an absorbance α is measured and specified (the absorbance is proportional to ω times the spectral function of the polarizability). We report the product of the absorbance $\alpha(\omega)$ and the refractive index $\eta(\omega)$, isotropically averaged over the three diagonal components of the polarizability tensor,

$$\alpha(\omega)\eta(\omega) = \frac{\beta}{3Vc\varepsilon_0}\Re\mathfrak{e}\int_0^T \langle\dot{\mathbf{P}}(0)\cdot\dot{\mathbf{P}}(t)\rangle e^{-i(\omega-i\delta)t}\mathrm{d}t. \tag{5}$$

In this equation, c is the speed of light. We note that one can also autocorrelate $\mathbf{P}(0)$ with $\mathbf{P}(t)$, or $\mathbf{P}(0)$ with $\dot{\mathbf{P}}(t)$, adding simultaneously factors ω^2 and $-i\omega$, respectively. We have tested all the three approaches and found identical results for the three versions.

C. Molecular dynamics simulations

For disordered materials, the required long simulation times are unattainable using FP simulations, and hence, surrogate models are required. To obtain FP data for constructing the MLFF, we employed the Vienna *Ab initio* Simulation Package (VASP).[40–42] Both the training of the MLFF and its subsequent application were conducted using the ML framework integrated in the VASP code.[19–21] It is essential to emphasize that the quality of an MLFF is not only dependent on the underlying ML algorithm but also the quality and representativeness of the training dataset concerning the problem at hand.[43,44] Ideally, the training dataset should

comprehensively cover all relevant regions of the potential energy surface, and this coverage should be compact to minimize the necessity for computationally expensive FP calculations. To accomplish this, we have employed the on-the-fly learning scheme integrated within VASP.

The training dataset was curated from a collection of multiple MD trajectories. These trajectories were obtained using Langevin thermostats[45–47] with a friction coefficient of 10 ps^{-1}. To comprehensively cover all phases of interest, a series of heating and cooling runs were executed, as discussed in the supplementary material. Notably, the temperature range considered was slightly wider than the region of interest, a choice aimed at enhancing the stability of the MLFF. During training of the MLFF, an active learning scheme relying on Bayesian error estimations was used. Whenever the error estimates were above a threshold, an FP calculation was performed and the MLFF was updated.[19,20,48] This approach yields very robust MLFFs and minimizes the number of FP calculations.

To obtain an IR spectrum, we used multiple MD trajectories within a micro-canonical ensemble to avoid artefacts caused by thermostats. The initial configurations and velocities for these individual MD trajectories were drawn from an isothermal–isobaric ensemble ($MAPbI_3$) or a canonical ensemble (water). The final IR spectrum was then computed as the average of the individual IR spectra obtained from these trajectories, providing a statistically accurate representation of the system's vibrational modes. We note that IR spectra are only sensitive to the long-range vibrational modes, i.e., at zero temperature, the calculations can be performed using the unit cell. However, at finite temperature, supercells are required to account for the finite temperature disorder. We checked carefully that the IR spectra are cell-size converged for all the cases reported here.

D. Polarization model

As highlighted in the introduction, obtaining the polarization for bulk systems presents inherent challenges. This is due to an undetermined modulo resulting from the absence of a unique phase origin or reference point.[29–32] To address this problem, we use derivative learning. In particular, the polarization $\mathbf{P}$ of a system can be understood as the antiderivative of the Born effective charge tensor $\mathbf{Z}^*$.[49] The Born effective charge tensor for ion i is mathematically expressed as

$$\mathbf{Z}^*_{i,\alpha\beta} = Z_{J(i)}\delta_{\alpha,\beta} + \left.\frac{\partial \mathbf{P}^{\mathrm{elec}}_{\alpha}}{\partial \mathbf{x}_{i\beta}}\right|_{\mathscr{E}=0}. \tag{6}$$

Here, $Z_{J(i)}$ represents the bare ionic charge of the ith ion; $J(i)$ signifies the atomic species; and $\mathbf{P}^{\mathrm{elec}}_{\alpha}$ corresponds to the Cartesian component α of the macroscopic electronic polarization. In the equation, $\mathbf{x}_{i\beta}$ denotes the position of the ith ion along the Cartesian component β.

The central idea is that the surrogate ML model describes the polarization $\mathbf{P}^{\mathrm{elec}}_{\alpha}$. However, we aim to avoid training this model on the polarization, but instead train the model on derivative data $\mathbf{Z}^*$. Clearly, the polarization must transform like a vectorial quantity under rotations. To this end, we employ ridge regression, with a linear kernel function K, constructed using the covariance of 3-dimensional descriptors $D^{\mu}_{n}(\mathscr{X})$. Here, $\mathscr{X}$ represents an atomic environment, μ corresponds to a Cartesian component, and n is the feature dimension. The linear kernel function is defined as follows:

$$K_{\mu\nu}(\mathscr{X},\mathscr{X}') = \sum_{n} D^{\mu}_{n}(\mathscr{X}) D^{\nu}_{n}(\mathscr{X}'). \tag{7}$$

This equation captures the similarity between atomic environments $\mathscr{X}$ and $\mathscr{X}'$. To describe the atomic environment, we utilized the λ-SOAP (Smooth Overlap of Atomic Potentials) descriptors developed by Grisafi *et al.*[26] These descriptors conserve the rotational symmetry of tensorial quantities,

$$\hat{S} D^{\mu}_{n}(\mathscr{X}) = D^{\mu}_{n}(\hat{S}\mathscr{X}), \tag{8}$$

where $\hat{S}$ is a generalized symmetry operator of the $SO(3)$ group. The descriptor is tailored to describe the surroundings of an atom. In the present work, we use two- and three-body descriptors and Bessel functions as radial basis sets, as in the original MLFF implementation of VASP.[19,20] To ensure smoothness in derivatives and avoid abrupt discontinuities, we incorporated a Behler and Parrinello cutoff function.[14] The radial cutoffs are typically set to 5.5 Å.

The polarization $\mathbf{P}$ of a configuration, characterized by atomic environments $\mathscr{X}_j$, is determined using descriptors $D^{\nu}_{n}(\mathscr{X}_{I_{\mathrm{ref}}})$ of reference atomic environments $\mathscr{X}_{I_{\mathrm{ref}}}$ following the equation:

$$\mathbf{P}_{\alpha} = \sum_{\nu I_{\mathrm{ref}}} \omega^{\nu}_{I_{\mathrm{ref}}} \sum_{jn} D^{\alpha}_{n}(\mathscr{X}_j) D^{\nu}_{n}(\mathscr{X}_{I_{\mathrm{ref}}}). \tag{9}$$

Here, $\omega^{\nu}_{I_{\mathrm{ref}}}$ represents weights that are paired with each reference descriptor $D^{\nu}_{n}(\mathscr{X}_{I_{\mathrm{ref}}})$. These weights are determined through derivative learning of the Born effective charge tensor, as given by the equation,

$$\mathbf{Z}^*_{\alpha\beta}(i) = \sum_{\nu I_{\mathrm{ref}}} \omega^{\nu}_{I_{\mathrm{ref}}} \sum_{jn} \frac{\partial D^{\alpha}_{n}}{\partial \mathbf{x}_{i\beta}}(\mathscr{X}_j) D^{\nu}_{n}(\mathscr{X}_{I_{\mathrm{ref}}}). \tag{10}$$

The weights are obtained by utilizing a linear regression model using the least squares method, allowing for simultaneous calculation of $\omega^{\nu}_{I_{\mathrm{ref}}}$ across all training configurations and all components of the Born effective charge tensor. We use sparse regression, i.e., reduce the number of kernel-basis functions $\mathscr{X}_{I_{\mathrm{ref}}}$ using farthest point sampling.

Our investigation revealed a substantial improvement in the quality of the fit by specially treating the diagonal elements of the Born effective charge tensor. To achieve this improvement, we applied a preprocessing step that involved subtracting the mean value of the diagonal elements of the Born effective charge tensor for each atomic species $\bar{Z}^*_J$ before the training process. Here, J represents the atomic species. The quantity $\bar{Z}^*_J$ is stored, and in a postprocessing step, we add back to the polarization $\mathbf{P}$ the following term:

$$\tilde{\mathbf{P}}_{\alpha} = \sum_{i=1}^{N_{\mathrm{atom}}} \bar{Z}^*_{J(i)} \mathbf{x}_{i\alpha}. \tag{11}$$

It is known that for finite systems, the polarization is well defined, but for a constant vectorial offset. Consequently, the origin for $\mathbf{x}_{i\alpha}$

is arbitrary, but must be chosen consistently for a trajectory. For example, one could subtract the positional coordinates of each atom for a reference centro-symmetric structure. For simulating IR spectra, any reference point can be chosen. However, the atom coordinates must not change abruptly during the simulations when atoms leave the simulation box on one side and re-enter on the other side.

Although derivative learning has been used to learn and predict the Born effective charges before,[50] it was only done for a single Cartesian component of the polarization, i.e., treating one component as an invariant. This means that when the three Cartesian components are merged into a vector, they are not guaranteed to transform like a vector. In fact, each component will transform like a scalar. In contrast, the present model is fully equivariant and the predicted polarizations transform like vectorial quantities, and consequently, the Born effective charges transform like tensors of rank two.

The computation of Born effective charges was carried out using density functional perturbation theory (DFPT)[51–53] by computing the static ion-clamped dielectric matrix, as discussed by Baroni and Resta[54] and Gajdoš *et al.* for the projector-augmented wave (PAW) method.[52] It is noteworthy that the calculation of the Born effective charge tensor requires three linear response calculations corresponding to the first derivative of the orbitals with respect to the Cartesian components of the external fields $\mathscr{E}$. This enables one to obtain the mixed derivatives, as defined in Eq. (2). Typically, linear response calculations are somewhat slower than ground state calculations, so these calculations take somewhat longer than the three density functional theory (DFT) ground state calculations. Alternatively, one could also determine the first derivative of the orbitals with respect to the positions and then predict the mixed second derivatives; however, this scales linearly with system size and is thus computationally more involved [the results for the Born effective charges are independent of the order of differentiation in Eq. (2)]. To the best of our knowledge, any electronic structure code can determine the Born effective charges via the orbital derivative with respect to the field. The present approach is, therefore, applicable to most FP codes. The inclusion of derivatives provides considerably more information, in particular, an additional $3N_{\text{atom}}$ of data, so that only a small number of FP calculations are required to achieve a high degree of accuracy, as demonstrated in the following.

E. Results and discussion

To demonstrate the versatility of the developed methodology, we have applied it to three distinct systems. First, we consider the water dimer $2(H_2O)$. In the case of molecules, the polarization is a well-defined property. Second, we examine water. Finally, we extend our analysis to a complex solid state system, focusing on an organic perovskite $MAPbI_3$, since it exhibits sizable anharmonicities.

F. Water dimer

We start our analysis with a water dimer $2(H_2O)$. Given that molecules possess a well-defined polarization, we can make a comparative assessment between the derivative learning approach and the conventional method of directly learning the polarization. This provides a valuable benchmark for assessing the reliability of the methodology.

The training and test configurations were extracted from an MD trajectory. We commenced with a water dimer placed in a simulation box with ample vacuum space. One of the oxygen atoms was constrained using selective dynamics to anchor the system. To simulate thermalization, we executed a heating run, spanning temperatures from 10 to 320 K, utilizing a Langevin thermostat. Importantly, the MD run was enhanced with on-the-fly machine learning to speed up the computational process. For the FP calculations, we opted for a revised Perdew–Burke–Ernzerhof (RPBE) functional[55] with van der Waals (vdW) dispersion energy corrections of Grimme *et al.* with zero-damping function[56] (RPBE-D3). This choice ensured that our calculations account for vdW interactions.

During the MD, the mass of the hydrogen atom was increased to 8 a.u., to allow for larger time steps of 1.5 fs. Of a total of 200 000 MD time steps, we selected 1000 uniformly distributed configurations for the computation of the polarization. For calculating the Born effective charge tensor, 150 configurations were chosen. To determine the polarization, we integrated the total (electronic and ionic) charge times the position operator by switching on dipole corrections in VASP (e.g., see Refs. 57 and 58). To improve the accuracy of the Born effective charge tensor, we used a strict convergence criterion of 1×10^{-7} eV for the electronic self-consistency loop and large cells. To confirm the internal consistency between polarization and Born effective charges, we also calculated numerical derivatives of the polarization. These derivatives showed an excellent agreement with the Born charges with a root mean square error (RMSE) less than $5 \times 10^{-6}|e|$, confirming the reliability of the resulting database.

To construct the learning curves, we used a dataset consisting of 1000 polarization calculations. The dataset was split in half by alternately selecting configurations for training and validation. In addition, we selected 50 Born effective charge calculations as validation data, chosen to be uniformly distributed across the dataset of 150 Born effective charge tensor calculations. The remaining 100 Born effective charge calculations served as the training dataset. During the combined training process, the polarization data were weighted ten times higher than the Born effective charge data. To determine the optimal number of fitting parameters, we selected the number of kernel functions that minimized the error in polarization prediction. It is worth noting that, as proposed by Cortes *et al.* for regression,[59] there should be a relationship between the RMSE and the number of training data, characterized by a power-law decay. The learning curves, which provide insights into the model's performance, are shown in Fig. 1.

The learning curves demonstrate the power-law relation between the RMSE and the number of training configurations. However, it is important to note that with a large number of training configurations, the improvement in the mean squared errors seems to plateau somewhat, indicating that our linear regression with two- and three-body descriptors will likely yield some residual (but acceptable) model errors. Overall, the most favorable results are achieved through combined learning, with percentage errors below 0.5% and 3% for the polarization and the Born effective charges, respectively. Crucially, learning only Born charges from 100 training configurations also attains a high relative accuracy of ~1% for the

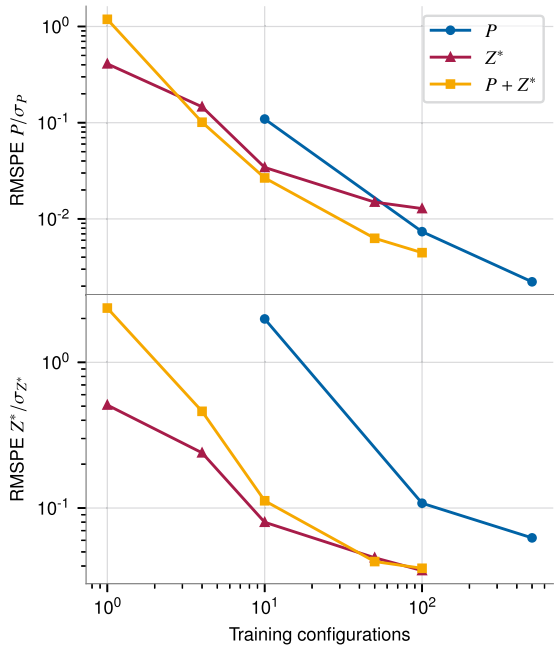

FIG. 1. Learning curves for the polarization (P) and Born effective charges (Z^*) of a water dimer. Errors on the test set are expressed in root mean square percentage errors (RMSPE), where the RMSE is normalized by the standard deviation (σ) of the test set to yield a dimensionless quantity (it should be noted that 10^{-1} and 10^{-2} at the y axis correspond to 10% and 1%, respectively). The blue line (circle) represents training solely on polarization (P), and the red line (triangle) represents training solely on Born effective charges (Z^*). The yellow line (square) demonstrates combined learning of polarization and Born effective charges ($P + Z^*$).

polarization and is as accurate as combined learning for the Born effective charges. We also note that there is no noticeable offset in the ML polarization compared to the FP calculations. Likely this is so since the molecule is free to rotate (and does rotate during the MD) and the surrogate model reliably determines the offset. On the other hand, attempting to train on the polarization data only requires more training data (blue line), but even with 500 training configurations, the Born effective charges still show twice the errors as combined training does using 100 training configurations. It should be noted that a training configuration contains exactly one vectorial data entry for P with three components but tensorial data entries for each atom for Z^*, resulting in a total of six times nine data entries.

G. Liquid water

For our second proof of concept, we selected liquid H_2O at room temperature. While the IR spectrum of water has been extensively studied and is well understood,[60–65] theoretical interpretations of these spectra remain challenging.[66–73] Even with the use of modern DFT methods, accurately reproducing the experimental properties of water is a complex task and often yields results that are far from accurate.[74] As a result, this system serves as an ideal test case for validating the reliability and effectiveness of the methodology and gleaning some insight into the underlying dynamics of water.

We chose the training configurations from an MD trajectory using an on-the-fly learning scheme. The simulation was conducted within a cubic box with periodic boundary conditions, containing a total of 64 H_2O molecules. The lattice constant of the cubic box was adjusted to attain the density of ~997 kg m^{-3}, closely resembling the density of water at room temperature.[75] Data were collected during multiple heating and cooling runs conducted within a canonical MD ensemble. We employed various temperatures, spanning from 270 to 420 K. The MLFF was trained using FP data obtained from the DFT calculations. In particular, we again employed the RPBE-D3 functional but found it necessary to use hard PAW potentials and a cutoff of 800 eV to obtain accurate stretch frequencies.[76] The MLFF trained on the RPBE-D3 data has a training set RMSE of 0.8 meV/atom and 60 meV/Å for energies and forces, respectively. The errors are twice as large as those achieved recently for a more extensive dataset using a similar method.[76] The main reason for the larger errors here is that we did not use single-value decomposition to fit the dataset and that the hyperparameters were not optimized. In addition, we trained an MLFF using strongly constrained and appropriately normed (SCAN)[77] to determine whether SCAN offers a reasonable description of the water dynamics. For this MLFF, the training set RMSE for energies is 0.7 meV/atom and about 65 meV/Å for forces.

The training configurations for the tensorial machine learning framework were once more chosen from a canonical MD ensemble. This ensemble was maintained at a constant temperature of 298.2 K and controlled by a Langevin thermostat. A total of 10 000 MD steps, accelerated by the MLFF, were executed. From the MD trajectory, 100 configurations were uniformly selected as the training dataset for the Born effective charges. A scatter plot of the trained and predicted Born effective charges is shown in Fig. 2. The Born effective charges vary significantly for the two distinct atomic species, particularly for the diagonal elements. This can be visually observed as two distinct clusters for positive and negative values in the scatter plot and relatively less data around $(-0.5, -0.5)e$. For the training set, the RMSE of the Born effective charge tensor is 32 m$|e|$.

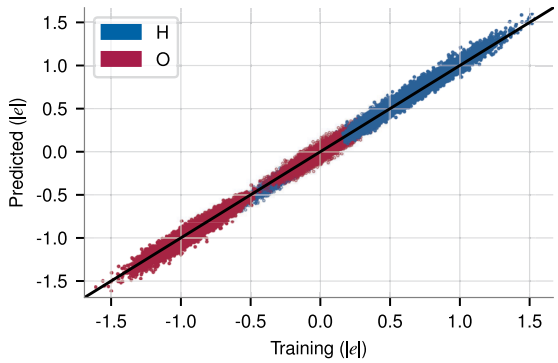

FIG. 2. Scatter plot of the trained and predicted Born effective charges for water. All the nine components of the Born effective charge tensor are shown as individual points.

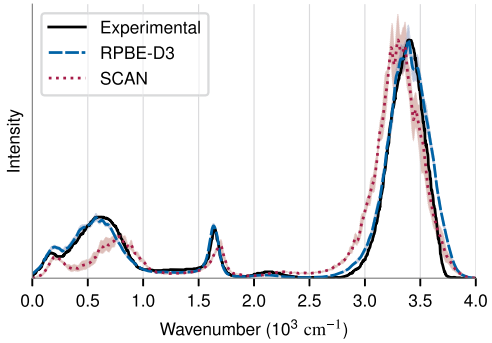

FIG. 3. Experimental and computational IR spectra for liquid water. The experimental reference data are from Ref. 64. The statistical uncertainties are represented by the colored areas centered around the calculated spectra (we show the 95% confidence interval, i.e., $\pm 2\sigma$, where σ is the standard error of the sample mean).

The computational IR spectrum presented alongside experimental data shown in Fig. 3 was computed by averaging the results over 20 individual IR spectra for RPBE-D3 and five individual IR spectra for SCAN. Each of these individual spectra was calculated from a micro-canonical MD trajectory. For each MD run, we initiated the simulation from an uncorrelated starting configuration and starting velocities obtained from a canonical ensemble, driven by Langevin thermostats, ensuring the appropriate average corresponding to room temperature. In each run, we performed a total of 100 000 MD steps, with a time step of 0.25 fs. This strategy of conducting multiple calculations from uncorrelated starting configurations was employed to enhance the statistical accuracy of our results. Instead of Lorenzian broadening, a Gaussian filter was applied before performing the Fourier transform. This improves the feature sharpness in the computed IR spectra.

As shown in Fig. 3, the present methodology allows the computation of the IR spectrum of liquid water with a remarkable level of agreement with the experimental data. There are two important conclusions to draw from this result. The close alignment between the intensity and the experimental data is a consequence of an accurate description of the Born effective charges. Since we train on the Born effective charges, the good agreement is likely not astonishing. Second, RPBE+D3 yields an excellent description of the dynamics of liquid water, both for the high-frequency stretch and for the medium-frequency bending motions. Most important are the lower frequency modes that are related to intra-molecular motions. It is important to note that the results for the SCAN functional are significantly worse for these modes. In particular, the low-frequency mode around 500 cm^{-1} is wrong by almost 40% indicating serious deficiencies in the description of the molecular motion of water molecules encaged by the four surrounding water molecules.

Previous studies, such as those by Sommers *et al.*,[28] Zhang *et al.*,[34] and Gastegger *et al.*,[33] have undertaken similar approaches utilizing MD trajectories in combination with symmetry-preserving machine learning frameworks to obtain Raman spectra for water and IR spectra for molecules. In these prior studies, the focus has been on learning the polarization directly. This required significantly more training data and a careful calculation of the polarization to avoid any discontinuities. In the work of Schienbein *et al.*[78,79] the Born effective charges are learned directly and are used in combination with velocities from MD trajectories to obtain the time derivative of the polarization. As previously discussed by others in the context of learning energies and forces, this misses important properties.[35] In the present work, the Born effective charges are constrained to be the derivatives of a vectorial quantity (the polarization) with respect to the positions. This implies, for instance, that the sum of all effective charges is zero, a property automatically fulfilled by our surrogate model. Furthermore, the existence of an anti-derivative P is not guaranteed if the tensorial quantities are learned directly. Falletta *et al.*[80] recently proposed to learn the polarization, the Born effective charges and the polarizability as the first and second derivatives of the electric enthalpy and applied this approach to the IR spectrum of quartz. This approach is elegant and concise but can be slow because it requires evaluating second derivatives during inference (e.g., using auto-differentiation).

H. Anharmonic solids

Our final test system is the organic–inorganic halide perovskite, methylammonium lead iodide ($MAPbI_3$). This material has been the subject of numerous experimental and theoretical studies, including state-of-the-art vibrational studies.[20,75,81–87] Notably, $MAPbI_3$ undergoes three entropy-driven phase transitions: from an orthorhombic phase to a tetragonal phase at 160 K and from the tetragonal phase to a cubic phase at 330 K. The thermodynamic nature of this material makes it challenging to model the IR spectra of the tetragonal phase using traditional 0 K methods such as DFPT.[51–53] Therefore, $MAPbI_3$ serves as a valuable test case for validating the reliability of the scheme for strongly anharmonic solids.

The training process for the MLFF applied to $MAPbI_3$ closely mirrored the methodology employed for water and previous studies of MAPbI3.[20] Multiple heating runs, encompassing the two phases—orthorhombic and tetragonal—were conducted using Langevin thermostat-driven MD simulations. The temperature range spanned from 80 to 430 K. Initially, fixed cell volumes were used, followed by simulating an isothermal–isobaric ensemble using the Parrinello–Rahman method.[88,89] The training was performed using a strongly constrained and appropriately normed (SCAN) meta-gradient corrected functional.[77] We note that we found in previous studies that SCAN is better suited for the simulation of $MAPbI_3$[90] than say RPBE-D3, as RPBE-D3 does not account for screening of the vdW interactions by the strongly polarizable cage atoms. The RMSE of the MLFF on the training data is 0.4 meV/atom for the energies and 18 meV/Å for the forces, while the tensorial ML framework archives a training dataset RMSE of 40 m$|e|$. Singular value decomposition was used to solve the regression problem, and the cutoff radii were optimized to minimize the error.

In Fig. 4, we present the IR spectra for the orthorhombic phase at 107 K and the tetragonal phase at 228 K, alongside the experimental results for comparison. The spectrum of the well-ordered orthorhombic phase and the tetragonal phase was calculated using a $4 \times 4 \times 4$ supercell for better statistics and to allow for some disorder and rearrangement of the methylammonium molecules. The starting configurations for the individual MD runs were chosen from an isothermal–isobaric ensemble. Aside from using starting configurations with varying cell vectors, the procedure for calculating

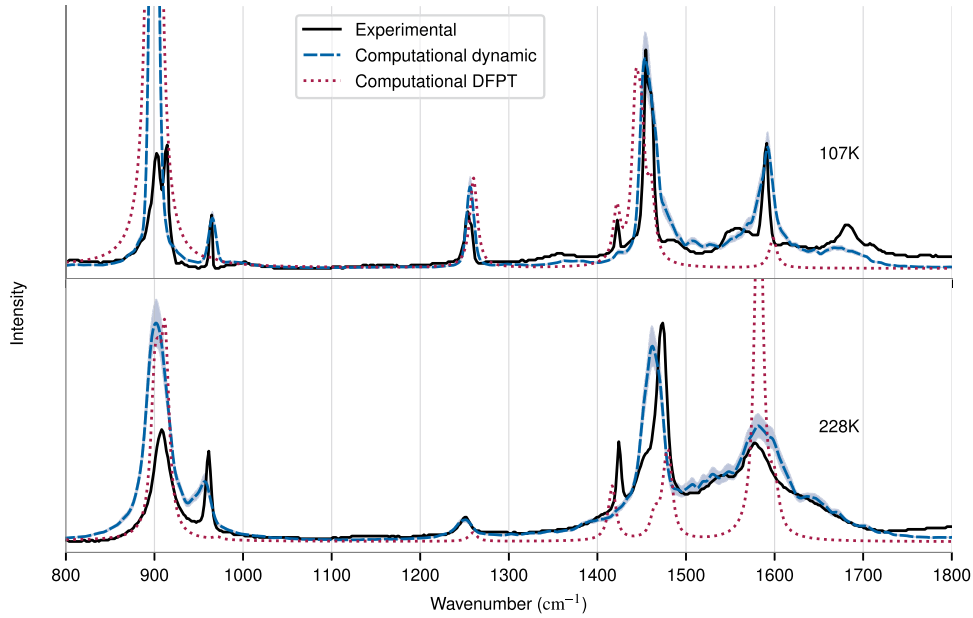

FIG. 4. Experimental and computational IR spectra for the orthorhombic and tetragonal phases of $MAPbI_3$. The vibrational frequencies of the computational IR spectra have been redshifted by 1.5% for alignment with experimental data. The experimental reference data are from Ref. 87. The statistical uncertainties are represented by the colored areas centered around the calculated spectra (we show $\pm 2\sigma$, where σ is the standard error of the sample mean).

the IR spectrum closely mirrors the one described for water. The dynamic IR spectra are also compared to the results calculated using DFPT. To be consistent with our ML results, we used the same DFT parameters as used for training of the surrogate models.

The analysis of the computed IR spectra reveals excellent agreement with the experimental data but also some small discrepancies, particularly in the intensities of the individual peaks. Notably, the peaks around 900 cm^{-1} appear with higher intensity. We note that errors in the Born effective charges could potentially affect the intensities, but also errors in the MLFF could cause slight shifts in the peaks or intensities. These are difficult to disentangle in the present framework. However, it should also be noted that the experimental reference spectra, obtained from a single crystal,[87] may be influenced by surface effects and crystal structure orientation, contributing to the intensity variations between computational and experimental results.

We start with a comparison for the orthorhombic low-temperature phase (107 K). The agreement between the DFPT and the finite temperature simulation is very good, but there are some marked improvements in the finite temperature data. The first peak around 900 cm^{-1} shows a double peak in both the DFPT and finite temperature (FT) simulation, in agreement with the experiment. The peak at 970 cm^{-1} is completely missing in the DFPT data but visible in the finite temperature data. It is a result of anharmonic interactions. The shoulder at 1420 cm^{-1} is pronounced using DFPT but washed out at finite temperature. This peak corresponds to the CH_3-bending motions.[53] In the supplementary material, we show that using the ionic charges only, this peak is visible in the spectrum, but the intensity of the peak is overestimated. Calculating the electronic contribution (el) only shows an almost identical peak; however, the phase of the electronic polarization is opposite to the ionic contribution, so when autocorrelating the sum of the electronic and ionic dipoles, the peak is strongly suppressed. This brings better overall agreement with the experiment, where the peak is also weak, but likely our Born effective charges are not quite sufficiently accurate (linear regression). The physics behind the vanishing of the peak is fairly simple: in this particular mode, the H atoms move orthogonal to their bond direction, a direction in which the total Born effective charges $\mathbf{Z}^*$ are small to start with (electronic contribution cancels ionic one). Furthermore, the movement of three hydrogen atoms is concerted being close to a helicopter motion, which reduces the IR intensity further.

In the experiment, we see quite some intensity between 1450 and 1600 cm^{-1}, which is also nicely reproduced by the FT simulations. We note that this frequency range becomes even more populated in the tetragonal phase in the FT simulations, and this population is a result of the strongly anharmonic rattling motion of the molecules in the cage, in turn, affecting the bending motion of the hydrogens.

The tetragonal phase spectrum (228 K) is also in excellent agreement with the experimental spectrum. We note that the DFPT spectrum now shows many deficiencies, with a complete lack of peaks at 950 cm^{-1} and a tiny peak around 1250 cm^{-1}, as well as sharp features around 1580 cm^{-1}. The FT data resolve these issues, albeit the two main peaks around 900 and 1480 cm^{-1} are somewhat too broad and washed out, and again, the peak around 1480 cm^{-1} is largely suppressed by the electronic contribution and only visible as a weak shoulder.

III. CONCLUSION

The present study showcases the effectiveness of a computational framework combining first-principles simulations with machine-learning methodologies. Machine-learned force fields can be used to access timescales that were previously very difficult to attain. This allows us to obtain vibrational spectra with excellent statistical accuracy, which would be very expensive to calculate without force fields. The main advance of the present work is, however, that we learn the polarization from its derivative, the Born effective charges. As mentioned in the main text, in VASP—but likely so in any plane wave-based code—the calculation of the Born effective charges via the first derivative of the orbitals with respect to external fields is only roughly three times more costly than a ground state Kohn–Sham calculation. Learning of force fields using first-principle nuclear derivatives is now ubiquitous, and learning the polarization from its nuclear derivative is a natural extension to this idea. Crucially, we have shown that the inclusion of polarization data, which is difficult to determine without some arbitrary modulus, is not required. As polarization is the anti-derivative of the Born effective charges, it can be directly calculated from the machine-learning model as long as derivative data are supplied along all the relevant adiabatic pathways. We successfully applied this approach to challenging scenarios, including water and the organic–inorganic halide perovskite $MAPbI_3$. In both cases, our method demonstrates excellent agreement with the experimental results, highlighting its capacity to capture the vibrational properties of diverse materials.

We finish with a few comments on further developments. In the present work, we have only used linear regression with two-body and three-body descriptors. Although the prediction accuracies are good (condensed matter) to excellent (molecules), we feel that the inclusion of higher body-order terms, or non-linear kernel-based regression might further improve the predicted Born effective charges. Furthermore, the subtraction of a diagonal component from the Born effective charges, albeit not particularly cumbersome, seems somewhat unsatisfactory and one would prefer to avoid it.

Overall, the present methodology is already very robust and can be readily applied to many relevant problems. Further research could be directed toward infrared simulations of more complex adsorbates on surfaces or of water interacting with electrodes.

SUPPLEMENTARY MATERIAL

The supplementary material consists of a document with detailed equations for the machine learning part as well as equations for the calculation of IR spectra. In addition, detailed training procedures and a table of all PAW potentials used in the DFT calculations are given. Finally, a comparison between the ionic and ML contributions to the IR spectrum of $MAPbI_3$ is shown.

ACKNOWLEDGMENTS

This research was funded in whole by the Austrian Science Fund (FWF) Grant No. 10.55776/F81. For open access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission. The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC).

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Bernhard Schmiedmayer: Data curation (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal). **Georg Kresse**: Conceptualization (equal); Funding acquisition (equal); Methodology (equal); Project administration (equal); Resources (equal); Supervision (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are openly available in Zenodo at http://doi.org/10.5281/zenodo.12622340.

The code, developed for this work is available at https://doi.org/10.5281/zenodo.12626935 and on GitHub at https://github.com/schmiedmayer/polipy4vasp_public.

REFERENCES

[1] B. H. Stuart, *Infrared Spectroscopy: Fundamentals and Applications* (John Wiley & Sons, 2004).
[2] A. Vimont, F. Thibault-Starzyk, and M. Daturi, "Analysing and understanding the active site by IR spectroscopy," Chem. Soc. Rev. **39**, 4928–4950 (2010).
[3] Y. J. Chabal, "Surface infrared spectroscopy," Surf. Sci. Rep. **8**, 211–357 (1988).
[4] J. Ryczkowski, "IR spectroscopy in catalysis," Catal. Today **68**, 263–381 (2001).
[5] A. Barth, "Infrared spectroscopy of proteins," Biochim. Biophys. Acta, Bioenerg. **1767**, 1073–1101 (2007).
[6] L. M. Ng and R. Simmons, "Infrared spectroscopy," Anal. Chem. **71**, 343–350 (1999).
[7] J. Luypaert, D. Massart, and Y. Vander Heyden, "Near-infrared spectroscopy applications in pharmaceutical analysis," Talanta **72**, 865–883 (2007).
[8] T. Theophile, *Infrared Spectroscopy: Materials Science, Engineering and Technology* (Books on Demand, 2012).
[9] L. Fernández-Carrasco, D. Torrens-Martín, L. Morales, and S. Martínez-Ramírez, "Infrared spectroscopy in the analysis of building and construction materials," in *Infrared Spectroscopy: Materials Science, Engineering and Technology* (IntechOpen, 2012), Vol. 510.
[10] S. M. Silva, C. R. Braga, M. V. Fook, C. M. Raposo, L. H. Carvalho, and E. L. Canedo, "Application of infrared spectroscopy to analysis of chitosan/clay nanocomposites," in *Infrared Spectroscopy: Materials Science, Engineering and Technology* (IntechOpen, 2012), pp. 43–62.
[11] M. Biczysko, J. Bloino, and C. Puzzarini, "Computational challenges in astrochemistry," Wiley Interdiscip. Rev.: Comput. Mol. Sci. **8**, e1349 (2018).
[12] T. L. Jansen, "Computational spectroscopy of complex systems," J. Chem. Phys. **155**, 170901 (2021).
[13] K. B. Beć, J. Grabska, and C. W. Huck, "Current and future research directions in computer-aided near-infrared spectroscopy: A perspective," Spectrochim. Acta, Part A **254**, 119625 (2021).

[14]J. Behler and M. Parrinello, "Generalized neural-network representation of high-dimensional potential-energy surfaces," Phys. Rev. Lett. **98**, 146401 (2007).
[15]A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, "Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons," Phys. Rev. Lett. **104**, 136403 (2010).
[16]A. P. Bartók, J. Kermode, N. Bernstein, and G. Csányi, "Machine learning a general-purpose interatomic potential for silicon," Phys. Rev. X **8**, 041048 (2018).
[17]L. Bonati and M. Parrinello, "Silicon liquid structure and crystal nucleation from *ab initio* deep metadynamics," Phys. Rev. Lett. **121**, 265701 (2018).
[18]T. Morawietz, A. Singraber, C. Dellago, and J. Behler, "How van der Waals interactions determine the unique properties of water," Proc. Natl. Acad. Sci. U. S. A. **113**, 8368–8373 (2016).
[19]R. Jinnouchi, F. Karsai, and G. Kresse, "On-the-fly machine learning force field generation: Application to melting points," Phys. Rev. B **100**, 014105 (2019).
[20]R. Jinnouchi, J. Lahnsteiner, F. Karsai, G. Kresse, and M. Bokdam, "Phase transitions of hybrid perovskites simulated by machine-learning force fields trained on the fly with Bayesian inference," Phys. Rev. Lett. **122**, 225701 (2019).
[21]R. Jinnouchi, F. Karsai, C. Verdi, R. Asahi, and G. Kresse, "Descriptors representing two- and three-body atomic distributions and their effects on the accuracy of machine-learned inter-atomic potentials," J. Chem. Phys. **152**, 234102 (2020).
[22]R. Jinnouchi, K. Miwa, F. Karsai, G. Kresse, and R. Asahi, "On-the-fly active learning of interatomic potentials for large-scale atomistic simulations," J. Phys. Chem. Lett. **11**, 6946–6955 (2020).
[23]R. Drautz, "Atomic cluster expansion for accurate and transferable interatomic potentials," Phys. Rev. B **99**, 014104 (2019).
[24]S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky, "E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials," Nat. Commun. **13**, 2453 (2022).
[25]I. Batatia, D. P. Kovacs, G. Simm, C. Ortner, and G. Csányi, "MACE: Higher order equivariant message passing neural networks for fast and accurate force fields," in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., 2022), Vol. 35, pp. 11423–11436.
[26]A. Grisafi, D. M. Wilkins, G. Csányi, and M. Ceriotti, "Symmetry-adapted machine learning for tensorial properties of atomistic systems," Phys. Rev. Lett. **120**, 036002 (2018).
[27]D. M. Wilkins, A. Grisafi, Y. Yang, K. U. Lao, R. A. DiStasio, Jr., and M. Ceriotti, "Accurate molecular polarizabilities with coupled cluster theory and machine learning," Proc. Natl. Acad. Sci. U. S. A. **116**, 3401–3406 (2019).
[28]G. M. Sommers, M. F. Calegari Andrade, L. Zhang, H. Wang, and R. Car, "Raman spectrum and polarizability of liquid water from deep neural networks," Phys. Chem. Chem. Phys. **22**, 10592–10602 (2020).
[29]R. King-Smith and D. Vanderbilt, "Theory of polarization of crystalline solids," Phys. Rev. B **47**, 1651 (1993).
[30]D. Vanderbilt and R. King-Smith, "Electric polarization as a bulk quantity and its relation to surface charge," Phys. Rev. B **48**, 4442 (1993).
[31]R. Resta, "Macroscopic polarization in crystalline dielectrics: The geometric phase approach," Rev. Mod. Phys. **66**, 899 (1994).
[32]R. Resta, "Quantum-mechanical position operator in extended systems," Phys. Rev. Lett. **80**, 1800 (1998).
[33]M. Gastegger, J. Behler, and P. Marquetand, "Machine learning molecular dynamics for the simulation of infrared spectra," Chem. Sci. **8**, 6924–6935 (2017).
[34]Y. Zhang, S. Ye, J. Zhang, C. Hu, J. Jiang, and B. Jiang, "Efficient and accurate simulations of vibrational and electronic spectra with symmetry-preserving neural network models for tensorial properties," J. Phys. Chem. B **124**, 7284–7290 (2020).
[35]S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, "Machine learning of accurate energy-conserving molecular force fields," Sci. Adv. **3**, e1603015 (2017).
[36]P. Umari and A. Pasquarello, "*Ab initio* molecular dynamics in a finite homogeneous electric field," Phys. Rev. Lett. **89**, 157602 (2002).
[37]R. Kubo, "Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems," J. Phys. Soc. Jpn. **12**, 570–586 (1957).
[38]R. Zwanzig, "Time-correlation functions and transport coefficients in statistical mechanics," Annu. Rev. Phys. Chem. **16**, 67–102 (1965).
[39]D. Sangalli, A. Marini, and A. Debernardi, "Pseudopotential-based first-principles approach to the magneto-optical Kerr effect: From metals to the inclusion of local fields and excitonic effects," Phys. Rev. B **86**, 125139 (2012).
[40]G. Kresse and J. Furthmüller, "Efficiency of ab-initio total energy calculations for metals and semiconductors using a plane-wave basis set," Comput. Mater. Sci. **6**, 15–50 (1996).
[41]G. Kresse and J. Furthmüller, "Efficient iterative schemes for *ab initio* total-energy calculations using a plane-wave basis set," Phys. Rev. B **54**, 11169 (1996).
[42]G. Kresse and D. Joubert, "From ultrasoft pseudopotentials to the projector augmented-wave method," Phys. Rev. B **59**, 1758 (1999).
[43]J. Li, B. Jiang, and H. Guo, "Permutation invariant polynomial neural network approach to fitting potential energy surfaces. II. Four-atom systems," J. Chem. Phys. **139**, 204103 (2013).
[44]V. Botu and R. Ramprasad, "Adaptive machine learning framework to accelerate *ab initio* molecular dynamics," Int. J. Quantum Chem. **115**, 1074–1083 (2015).
[45]M. P. Allen and D. J. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, 2017).
[46]W. G. Hoover, A. J. Ladd, and B. Moran, "High-strain-rate plastic flow studied via nonequilibrium molecular dynamics," Phys. Rev. Lett. **48**, 1818 (1982).
[47]D. J. Evans, "Computer 'experiment' for nonlinear thermodynamics of Couette flow," J. Chem. Phys. **78**, 3297–3302 (1983).
[48]C. M. Bishop and N. M. Nasrabadi, *Pattern Recognition and Machine Learning* (Springer, 2006).
[49]C.-Z. Wang, R. Yu, and H. Krakauer, "Polarization dependence of Born effective charge and dielectric constant in $KNbO_3$," Phys. Rev. B **54**, 11161 (1996).
[50]K. Shimizu, R. Otsuka, M. Hara, E. Minamitani, and S. Watanabe, "Prediction of Born effective charges using neural network to study ion migration under electric fields: Applications to crystalline and amorphous Li_3PO_4," Sci. Technol. Adv. Mater.: Methods **3**, 2253135 (2023).
[51]X. Wu, D. Vanderbilt, and D. Hamann, "Systematic treatment of displacements, strains, and electric fields in density-functional perturbation theory," Phys. Rev. B **72**, 035105 (2005).
[52]M. Gajdoš, K. Hummer, G. Kresse, J. Furthmüller, and F. Bechstedt, "Linear optical properties in the projector-augmented wave methodology," Phys. Rev. B **73**, 045112 (2006).
[53]M. A. Pérez-Osorio, R. L. Milot, M. R. Filip, J. B. Patel, L. M. Herz, M. B. Johnston, and F. Giustino, "Vibrational properties of the organic–inorganic halide perovskite $CH_3NH_3PbI_3$ from theory and experiment: Factor group analysis, first-principles calculations, and low-temperature infrared spectra," J. Phys. Chem. C **119**, 25703–25718 (2015).
[54]S. Baroni and R. Resta, "*Ab initio* calculation of the macroscopic dielectric constant in silicon," Phys. Rev. B **33**, 7017 (1986).
[55]B. Hammer, L. B. Hansen, and J. K. Nørskov, "Improved adsorption energetics within density-functional theory using revised Perdew-Burke-Ernzerhof functionals," Phys. Rev. B **59**, 7413 (1999).
[56]S. Grimme, J. Antony, S. Ehrlich, and H. Krieg, "A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu," J. Chem. Phys. **132**, 154104 (2010).
[57]G. Makov and M. C. Payne, "Periodic boundary conditions in *ab initio* calculations," Phys. Rev. B **51**, 4014 (1995).
[58]J. Neugebauer and M. Scheffler, "Adsorbate-substrate and adsorbate-adsorbate interactions of Na and K adlayers on Al(111)," Phys. Rev. B **46**, 16067 (1992).
[59]C. Cortes, L. D. Jackel, S. Solla, V. Vapnik, and J. Denker, "Learning curves: Asymptotic values and rate of convergence," in *Advances in Neural Information Processing Systems* (Morgan-Kaufmann, 1993), Vol. 6.
[60]M. Falk and T. Ford, "Infrared spectrum and structure of liquid water," Can. J. Chem. **44**, 1699–1707 (1966).
[61]D. A. Draegert, N. Stone, B. Curnutte, and D. Williams, "Far-infrared spectrum of liquid water," J. Opt. Soc. Am. **56**, 64–69 (1966).

[62]G. Walrafen, in *Water: A Comprehensive Treatise*, edited by F. Franks (Prenum Press, New York, 1972), Vol. 1, p. 151.
[63]J. Hasted, S. Husain, F. Frescura, and J. Birch, "Far-infrared absorption in liquid water," Chem. Phys. Lett. **118**, 622–625 (1985).
[64]J. E. Bertie and Z. Lan, "Infrared intensities of liquids XX: The intensity of the OH stretching band of liquid water revisited, and the best current values of the optical constants of H_2O(l) at 25°C between 15,000 and 1 cm^{-1}," Appl. Spectrosc. **50**, 1047–1057 (1996).
[65]D. A. Schmidt and K. Miki, "Structural correlations in liquid water: A new interpretation of IR spectroscopy," J. Phys. Chem. A **111**, 10119–10122 (2007).
[66]P. Madden and R. Impey, "On the infrared and Raman spectra of water in the region 5–250 cm^{-1}," Chem. Phys. Lett. **123**, 502–506 (1986).
[67]R. Bansil, T. Berger, K. Toukan, M. Ricci, and S. Chen, "A molecular dynamics study of the OH stretching vibrational spectrum of liquid water," Chem. Phys. Lett. **132**, 165–172 (1986).
[68]B. Guillot, "A molecular dynamics study of the far infrared spectrum of liquid water," J. Chem. Phys. **95**, 1543–1551 (1991).
[69]G. Corongiu, "Molecular dynamics simulation for liquid water using a polarizable and flexible potential," Int. J. Quantum Chem. **42**, 1209–1235 (1992).
[70]P. L. Silvestrelli, M. Bernasconi, and M. Parrinello, "Ab initio infrared spectrum of liquid water," Chem. Phys. Lett. **277**, 478–482 (1997).
[71]B. Auer and J. Skinner, "IR and Raman spectra of liquid water: Theory and interpretation," J. Chem. Phys. **128**, 224511 (2008).
[72]J. Xu, M. Chen, C. Zhang, and X. Wu, "First-principles study of the infrared spectrum in liquid water from a systematically improved description of H-bond network," Phys. Rev. B **99**, 205123 (2019).
[73]G. R. Medders and F. Paesani, "Infrared and Raman spectroscopy of liquid water through 'first-principles' many-body molecular dynamics," J. Chem. Theory Comput. **11**, 1145–1154 (2015).
[74]M. J. Gillan, D. Alfe, and A. Michaelides, "Perspective: How good is DFT for water?," J. Chem. Phys. **144**, 130901 (2016).
[75]M. Tanaka, G. Girard, R. Davis, A. Peuto, and N. Bignell, "Recommended table for the density of water between 0 C and 40 C based on recent experimental reports," Metrologia **38**, 301 (2001).
[76]P. Montero de Hijes, C. Dellago, R. Jinnouchi, B. Schmiedmayer, and G. Kresse, "Comparing machine learning potentials for water: Kernel-based regression and Behler–Parrinello neural networks," J. Chem. Phys. **160**, 114107 (2024).
[77]J. Sun, A. Ruzsinszky, and J. P. Perdew, "Strongly constrained and appropriately normed semilocal density functional," Phys. Rev. Lett. **115**, 036402 (2015).
[78]P. Schienbein, "Spectroscopy from machine learning by accurately representing the atomic polar tensor," J. Chem. Theory Comput. **19**, 705–712 (2023).
[79]K. Joll, P. Schienbein, K. M. Rosso, and J. Blumberger, "Molecular dynamics simulation with finite electric fields using perturbed neural network potentials," arXiv:2403.12319 (2024).
[80]S. Falletta, A. Cepellotti, C. W. Tan, A. Johansson, A. Musaelian, C. J. Owen, and B. Kozinsky, "Unified differentiable learning of the electric enthalpy and dielectric properties with exact physical constraints," arXiv:2403.17207 (2024).
[81]D. Weber, "$CH_3NH_3PbX_3$, ein Pb(II)-system mit kubischer perowskitstruktur/$CH_3NH_3PbX_3$, a Pb(II)-system with cubic perovskite structure," Z. Naturforsch., B **33**, 1443–1445 (1978).
[82]N. Onoda-Yamamuro, T. Matsuo, and H. Suga, "Calorimetric and IR spectroscopic studies of phase transitions in methylammonium trihalogenoplumbates (II)," J. Phys. Chem. Solids **51**, 1383–1395 (1990).
[83]Y. Kawamura, H. Mashiyama, and K. Hasebe, "Structural study on cubic–tetragonal transition of $CH_3NH_3PbI_3$," J. Phys. Soc. Jpn. **71**, 1694–1697 (2002).
[84]C. C. Stoumpos, C. D. Malliakas, and M. G. Kanatzidis, "Semiconducting tin and lead iodide perovskites with organic cations: Phase transitions, high mobilities, and near-infrared photoluminescent properties," Inorg. Chem. **52**, 9019–9038 (2013).
[85]T. Baikie, Y. Fang, J. M. Kadro, M. Schreyer, F. Wei, S. G. Mhaisalkar, M. Graetzel, and T. J. White, "Synthesis and crystal chemistry of the hybrid perovskite $(CH_3NH_3)PbI_3$ for solid-state sensitised solar cell applications," J. Mater. Chem. A **1**, 5628–5641 (2013).
[86]M. Bokdam, T. Sander, A. Stroppa, S. Picozzi, D. Sarma, C. Franchini, and G. Kresse, "Role of polar phonons in the photo excited state of metal halide perovskites," Sci. Rep. **6**, 28618 (2016).
[87]G. Schuck, D. M. Többens, M. Koch-Müller, I. Efthimiopoulos, and S. Schorr, "Infrared spectroscopic study of vibrational modes across the orthorhombic–tetragonal phase transition in methylammonium lead halide single crystals," J. Phys. Chem. C **122**, 5227–5237 (2018).
[88]M. Parrinello and A. Rahman, "Crystal structure and pair potentials: A molecular-dynamics study," Phys. Rev. Lett. **45**, 1196 (1980).
[89]M. Parrinello and A. Rahman, "Polymorphic transitions in single crystals: A new molecular dynamics method," J. Appl. Phys. **52**, 7182–7190 (1981).
[90]M. Bokdam, J. Lahnsteiner, B. Ramberger, T. Schäfer, and G. Kresse, "Assessing density functionals using many body theory for hybrid perovskites," Phys. Rev. Lett. **119**, 145501 (2017).