Letter

Spectroscopy from Machine Learning by Accurately Representing the Atomic Polar Tensor

Philipp Schienbein*

Read Online

ACCESS | 📊 Metrics & More | 📄 Article Recommendations

ABSTRACT: Vibrational spectroscopy is a key technique to elucidate microscopic structure and dynamics. Without the aid of theoretical approaches, it is, however, often difficult to understand such spectra at a microscopic level. Ab initio molecular dynamics has repeatedly proved to be suitable for this purpose; however, the computational cost can be daunting. Here, the E(3)-equivariant neural network `e3nn` is used to fit the atomic polar tensor of liquid water *a posteriori* on top of existing molecular dynamics simulations. Notably, the introduced methodology is general and thus transferable to any other system as well. The target property is most fundamental and gives access to the IR spectrum, and more importantly, it is a highly powerful tool to directly assign IR spectral features to nuclear motion—a connection which has been pursued in the past but only using severe approximations due to the prohibitive computational cost. The herein introduced methodology overcomes this bottleneck. To benchmark the machine learning model, the IR spectrum of liquid water is calculated, indeed showing excellent agreement with the explicit reference calculation. In conclusion, the presented methodology gives a new route to calculate accurate IR spectra from molecular dynamics simulations and will facilitate the understanding of such spectra on a microscopic level.
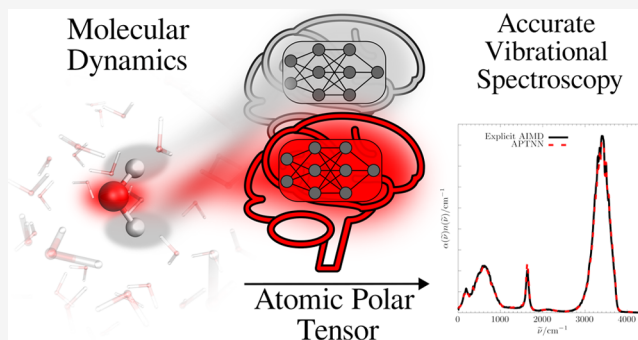
Vibrational spectroscopy, be it IR or Raman spectroscopy, is one of the most important techniques to unveil microscopic properties of matter, such as structure determination[1] or structural dynamics of water.[2,3] It can also be used to time-dependently monitor dynamical processes, e.g. intramolecular couplings in peptides,[4] or proton transfer mechanisms.[5] Attenuated total reflection (ATR) IR or sum-frequency generation (SFG)[6] can selectively probe molecules at interfaces, such as metal/liquid water interfaces[7,8] or water/air interfaces.[9] Using THz spectroscopy, the chemical environment of molecules can directly be probed, and it elucidated solvation dynamics in aqueous solutions from ambient[10,11] to extreme thermodynamic conditions, such as high-pressures,[12] supercritical phases,[13] or confined systems.[14]

Because of the broad applicability and power of vibrational spectroscopy there is a demand to calculate accurate vibrational spectra from ab initio techniques which are predictive and also aid to understand experiments at the molecular level. When it comes to spectroscopy of condensed phase systems at finite temperature, ab initio molecular dynamics (AIMD),[15] where the electronic structure is calculated on-the-fly at every time step, is the prime technique for several reasons. First, the potential energy surface can accurately be represented—clearly depending on the underlying electronic structure theory setup employed to drive the MD. Second, as the electronic structure

is available at every time step, charge transfer and polarization effects are naturally included, and dipole moments can directly be obtained, e.g. from maximally localized Wannier functions[15] or from the electron density.[16] Third, anharmonic effects are also naturally taken into account which can break the standard normal-mode analysis[17] if they are too large. Notably, such large anharmonicities can even be present for single molecules (e.g., peptides) in the gas-phase,[18,19] where AIMD simulations are required because the normal-mode analysis fails to correctly reproduce the measured spectra.
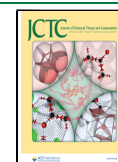
From a MD trajectory, the frequency dependent Beer–Lambert absorption coefficient of IR spectroscopy

$$\alpha(\omega) = \frac{\pi \beta \omega^2}{3 V c \varepsilon_0 n(\omega)} \frac{1}{2\pi} \int_{-\infty}^{\infty} dt \, e^{-i\omega t} \langle \mathbf{M}(0)\mathbf{M}(t) \rangle \quad (1)$$

can be calculated from the time auto correlation function of the total dipole moment vector $\mathbf{M}(t)$ of the simulation box (see e.g.

ref [20]), where $\beta = 1/k_B T$, k_B is the Boltzmann constant, T is the temperature, V is the volume of the periodic simulation box, c is the speed of light in vacuum, ϵ_0 is the dielectric constant, and $n(\omega)$ is the frequency dependent refractive index. Note that the so-called "quantum correction factor"[21] has already been included. The main disadvantage of this technique is that the MD simulations need to be quite long, also to reduce the statistical noise to a minimum. Clearly, this is a problem for AIMD which can be highly demanding computationally, especially if more expensive techniques, such as hybrid DFT, are used to drive the MD.

Over the last decades, machine learning (ML) approaches have been introduced with the aim to accelerate AIMD simulations. Therein, the expensive electronic structure calculations are replaced with a cheaper machine learning model, while retaining the same accuracy.[22−25] Arguably, ML techniques have repeatedly proved to reliably represent the potential energy surface from explicit electronic structure calculations at a fraction of the cost. One apparent problem of these pioneering ML techniques usually is that only the potential energy surface is trained (which is generally sufficient to run MD simulations), but all information on the electronic structure itself is lost. Therefore, total dipole moments at the quality of the underlying electronic structure theory cannot be obtained along the "MLMD" (machine learning molecular dynamics) simulation. One way to circumvent this issue was to extract single snapshots from the MLMD trajectory and explicitly calculate the electronic structure for those snapshots again, e.g. to calculate polarizability tensors for Raman spectra.[26] However, formally, time correlation functions (as in e.g. eq 1 for IR spectra), require the electronic structure at each time step or at least frequently enough such that all vibrations present in the system are correctly sampled. Note that the sampling theorem can be employed to determine how frequently time-dependent data needs to be provided;[27] however, the fastest vibration needs to be known. For example, in the case of liquid water, the fastest vibration is the O−H stretch at roughly 3500 cm^{-1}. According to the sampling theorem, data needs to be provided at least every roughly 4.5 fs to correctly sample this vibration. This introduces a huge bottleneck for MLMD simulations, if a significant amount of configurations needs to be explicitly recalculated anyways to get exact vibrational spectra.

In recent years, training atomic or molecular properties using ML has been an extremely active field, and the calculation of vibrational spectra by ML is no exception. In the following, some key methodological ideas are summarized as to how the computation of vibrational spectra can be accelerated by ML. Since the approach introduced herein aims to calculate vibrational spectra from MD simulations, i.e. via time correlation functions, the following discussion is restricted to accelerating these methods only. Notably, ML approaches have been used previously to accelerate complementary approaches, too, e.g. the normal-mode analysis or vibrational Hamiltonians. ML approaches have also been used for the reverse "Spec-to-Struc" process, where a given spectrum is used to gain information on the underlying structure. The interested reader is referred to ref [28] for a detailed review on ML in the context of these methods and on applications of ML in the context of vibrational spectroscopy in general.

Partial atomic charges have been introduced in third generation NNPs,[25] but with the main purpose to include long-range interactions. Such trained atomic partial charges could potentially also be used as an output parameter to calculate dipole moments along an MLMD trajectory. An apparent problem of partial charges in general is that there are no physical observables. As such, their magnitude depends on the chosen partitioning scheme employed, e.g. Mulliken,[29] Hirshfeld,[30] or Bader[31] (incomplete list). It could be shown that different partitioning schemes can yield very different results.[32−34] Moreover, choosing an unsuitable scheme for a given problem can lead to wrong molecular dipole moments[34] and can yield unphysical atomic charges.[35] Clearly, these caveats potentially also affect the quality of the IR spectrum calculated from such partial atomic charges.

A complementary approach is to train partial charges such that molecular dipole moments are reproduced correctly[33,36] or to train the positions of Wannier centers.[37,38] Molecular dipole moments are generally measurable and can thus be validated against experiments. In the case of gas-phase systems (without periodic boundary conditions), the total dipole moment vector has recently been trained as a whole to predict IR spectra of protonated water clusters[39] as well as of an ethanol and an aspirin molecule.[40] For small single molecules, also the polarizability tensor has been trained[40,41] which can then be used to calculate Raman spectra of these molecules. Finally, even the full electron density of single molecules has been trained[42,43] as well as transition dipole moments to excited states[44] which allow the calculation of UV/vis spectra.

In condensed phase systems with periodic boundary conditions, the total dipole moment is, however, multivalued[45] which can possibly lead to ambiguities in the training and prediction process. This problem can be circumvented by considering molecular dipole moments instead. The total dipole moment vector required to calculate the IR spectrum according to eq 1 is then the sum of all molecular dipole moment vectors in the system. Molecular dipole moment vectors have been trained directly by symmetry-adapted Gaussian process regression to accurately calculate the IR spectrum of liquid ambient water.[46] Using the same approach, it was shown that the molecular polarizability tensor can also directly be trained which enables one to calculate machine learned Raman[46−48] and SFG[48] spectra from molecular dynamics simulations.

In this work, I introduce a machine learning model to train the atomic polar tensor (APT)[49] which is then utilized to accurately calculate an IR spectrum. The APT is a proper physical observable which does not rely on any charge partitioning scheme or any definition of molecules or molecular reference frames. Its definition is therefore also perfectly valid when covalent bonds are broken during an MD simulation and the molecular composition changes, e.g. during proton transfer in water. Conceptually, the herein introduced APT neural network (APTNN) is therefore transferable to any system with or without periodic boundary conditions. It is noted in passing that nuclear quantum effects are essential to describe e.g. proton transfer in water correctly which are not considered in this work. However, the introduced APTNN is readily applicable to path integral trajectories since the APT centroid can straightforwardly be computed.

Importantly, the APT itself is a highly relevant property for spectroscopy, because it can be utilized to assign specific atomic motion to spectral features. As it will be laid out in the following, the APT represents nothing else than the definition of the IR selection rule. As such, any velocity spectrum (e.g., the vibrational density of states) can be promoted to a proper IR spectrum by weighting with the respective APTs. This has been done several times in the past, e.g. to decompose the IR

spectrum of liquid water into translational, rotational, and vibrational contributions.[50] IR spectra of peptides have also been dissected in terms of atomic velocities in the past,[51] even using sophisticated decompositions based on graph theory to understand the origin of low-frequency backbone vibrations.[18] Moreover, it has also been used to calculate SFG spectra.[52] None of these works have yet utilized the full power of the APT simply due to the enormous computational cost: A single APT requires six additional single point calculations to obtain the necessary finite differences (if the APT is calculated numerically), see below. As a result, severe approximations have been used so far, such as parametrizing the APT,[52] using the instantaneous normal mode (INM) approximation,[50] or calculating the APT not at every time step, under the approximation that it does not change much as a function of time.[53] These approximations clearly counteract its potential power: The INM approximation typically introduces imaginary frequencies (similar to the standard normal-mode analysis) which need to be dealt with in some ad hoc way. Similarly, calculating the APT not frequently enough can induce spurious signals in the IR spectrum.[53] Having a machine learning approach available to specifically predict the APT at each time step is therefore highly beneficial for all above-mentioned problems. To the best of my knowledge, such an ML model does not exist yet.

The derivation of the APT has already been presented in the literature, see e.g. refs 52 and 53. To set the stage, its derivation is, however, summarized here. First, a Fourier transform identity is used on eq 1, such that the equation can be rewritten

$$\alpha(\omega) = \frac{\pi\beta}{3Vc\varepsilon_0 n(\omega)} \frac{1}{2\pi} \int_{-\infty}^{\infty} dt \ e^{-i\omega t} \langle \dot{\mathbf{M}}(0)\dot{\mathbf{M}}(t) \rangle \quad (2)$$

where $\dot{\mathbf{M}}(t)$ is the time derivative of the total dipole moment. In an effort to express the dipolar velocity as a function of atomic velocities, the chain rule can be applied to express the ξ-th component of $\dot{\mathbf{M}}(t)$

$$\dot{M}_\xi(t) = \sum_i^{N_{nuc}} \sum_\zeta \frac{\partial M_\xi(t)}{\partial r_{i_\zeta}(t)} \cdot \frac{\partial r_{i_\zeta}(t)}{\partial t} \quad (3)$$

where ζ and ξ represent the three Cartesian coordinates. In matrix notation, this equation can be rewritten in a more compact way

$$\dot{\mathbf{M}}(t) = \sum_i \mathbf{P}_i(t)\cdot\mathbf{v}_i(t) \quad (4)$$

where $\mathbf{v}_i(t)$ is the velocity of the i-th nuclei, and

$$\mathbf{P}_i(t) = (\nabla_i \otimes \mathbf{M}(t))^{\mathrm{T}} \quad (5)$$

is the APT of atom i. The atom velocities $\mathbf{v}_i(t)$ are readily available along any MD trajectory. As shown in eq 3, the APT is the spatial derivative of the total dipole moment vector, which is indeed nothing else than the IR selection rule. Any IR spectrum can therefore be expressed based on atomic velocities weighted by the corresponding APTs. The atomic velocities in turn can intuitively be dissected a la carte using classical mechanics. The herein introduced APTNN is meant to be trained on existing MD trajectories (AIMD, MLMD). After the training, the APTs of each sampled configuration can be predicted, and the time correlation function can be sampled. Note that IR spectra have usually been computed using the autocorrelation function depicted in eq 1 from explicit AIMD simulations where electric dipole moments are straightforwardly available. Training the

APT therefore only becomes relevant either if the APT is used for the spectral analysis itself, see e.g. refs 18, 50, 52, and 53, or if sampling is performed by MLMD simulations where the total dipole moment is usually not naturally available. The latter is especially important when longer time scales are required which exceed the ones accessible by explicit AIMD or when more expensive electronic structure calculations, e.g. hybrid DFT or beyond, are applied.

The APT of atom i is the spatial derivative of the total dipole moment with respect to a displacement of that atom (eq 5). It is thus a 3 × 3 tensor which is invariant with respect to translations but equivariant with respect to rotations. This means that if the atomic coordinates are translated in space, the corresponding APTs do not change. However, if the set of atomic coordinates rotates in space, the APTs also rotate accordingly. Here, the derivative is evaluated numerically from central (or "two-sided") finite differences as it has frequently been done before, see e.g. refs 18, 50, and 52, using a displacement of 0.01 Å. Due to the finite differences, 6 additional single point calculations are necessary to calculate the APT of a single atom. Previously, it could already be shown that the chosen displacement is small enough for liquid water.[50] Indeed, I also computed the APT explicitly with a displacement of 0.04 Å and did not find any significant difference. It might, however, not be sufficiently small enough for other systems. Notably, it is also possible to calculate an APT analytically using Density Functional Perturbation Theory (DFPT) as implemented in CP2k.[54] This way, only one electronic structure calculation is required per atom per Cartesian coordinate to calculate the APT. The total number of required electronic structure calculations would thus be reduced compared to using numerical derivatives. More importantly, the displacement for the numerical derivative is a convergence parameter which could be omitted completely when using analytical derivatives. However, the herein employed numerical derivatives render the presented methodology general, such that it can directly be applied even if analytical APTs are not available, e.g. when other codes or electronic structure methods beyond DFT are used.

An apparent problem when training an APT (see eq 5) is that it is an equivariant property as already mentioned, while most machine learning models can only infer invariant properties. Recently, the e3nn framework has been introduced[55] for PyTorch which can be used to train E(3)-equivariant graph Neural Networks and thus enables one to infer also equivariant properties by a machine learning model. Its introduction also caused a huge boom in the field, and numerous works have been published, where equivariant properties have been modeled. For example, the *NequIP* package was recently developed[56] to model potential energy surfaces using equivariant message passing. Moreover, it was also used recently to train the electron density as a whole for gas-phase molecules.[43] Besides such equivariant message passing Neural Networks, also Gaussian process regression can be used to create equivariant machine learning models.[57] This approach was utilized recently to train molecular dipole moment vectors and polarizability tensors.[46−48]

Here, I now use the so-called *SimpleNetwork* (v2106) from the e3nn toolkit to create an APTNN for liquid ambient water which is capable to predict the APTs of all atoms at a given MD snapshot. The predicted APTs are then used to calculate the IR spectrum via eq 2 and eq 4. Although the IR spectrum of liquid ambient water has already been machine learned by training molecular dipole moments,[46,48] even explicitly considering nuclear quantum effects, it is merely done here to demonstrate
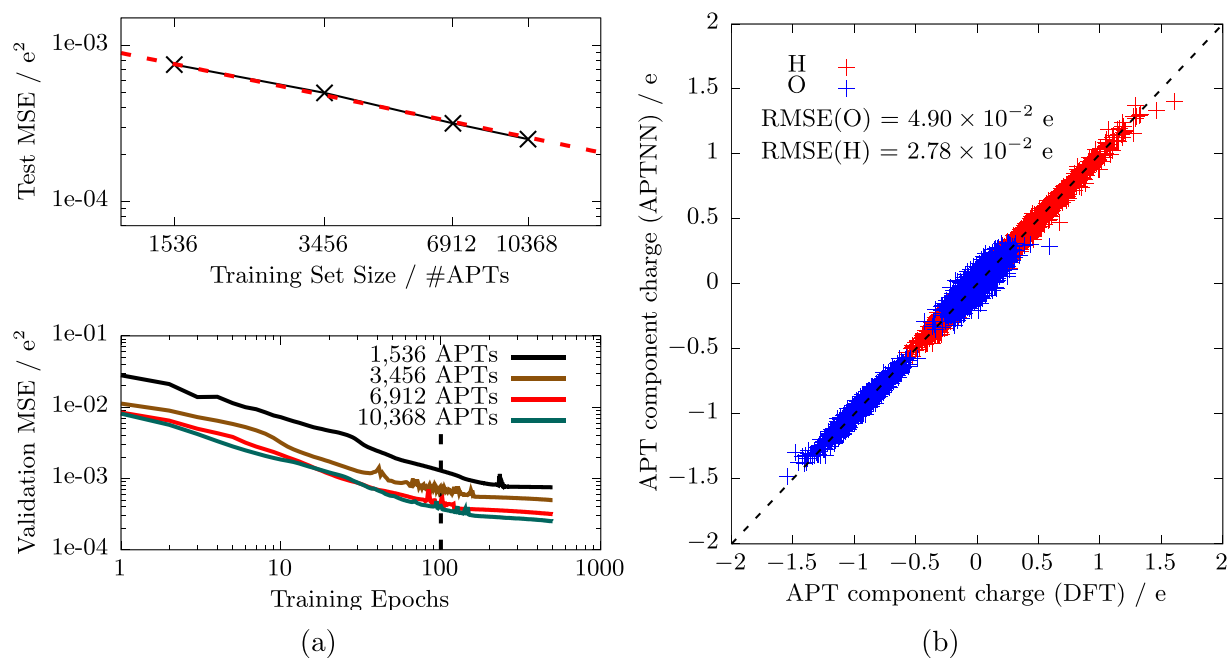
Figure 1. (a) Test error evaluated on an unknown test set containing 3840 APTs stemming from 10 randomly sampled MD snapshots as a function of training data set size (learning curve, top panel). Mind the log−log scale of the plot and that the expected linear relation (indicated by the red dashed line) is recovered. In the bottom panel, the MSE on the respective validation set is shown as a function of training epoch and training data set size, i.e. four (black), 9 (brown), 18 (red), and 27 (green) MD snapshots, corresponding to 1536, 3456, 6912, and 10368 APTs, respectively. (b) Performance of the APTNN trained on 3840 APTs (9 MD snapshots) on an unknown test set containing 3840 APTs stemming from 10 randomly sampled MD snapshots. The figure compares all components of the APT matrix individually for O (blue) and H (red) atoms. Note that the test set used to benchmark the APTNN in the top panel of (a) and in (b) is the same.

the applicability of the herein introduced technique. The introduced APTNN methodology and the training protocol is generally transferable to any other system, with and without periodic boundary conditions. As the underlying electronic structure theory, I opt to use the RPBE functional[58] supplemented by D3 dispersion corrections.[59] It could be shown repeatedly in the past that this electronic structure model reproduces fluid water excellently, even far away from ambient conditions.[13,60−62] Previously, liquid ambient water has been simulated,[60] and these data (16 trajectories, 20 ps each using a time step of 1 fs) are used in this work. Note that the APTNN was meant to be trained on top of existing MD trajectories. This means that these underlying MD trajectories are entirely responsible for sufficiently sampling the configuration space. As a starting point, I randomly selected 10 statistically independent configurations from the available AIMD trajectories and calculated the APT. The calculation setup is exactly the same as before,[60] and I refer to that reference for an elaborate description. All electronic structure calculations have been performed using version 8.0 of the `CP2k` program package[63] and the `Quickstep module`.[64] Each configuration contains 384 atoms, and therefore 2304 single point calculations are required. This is undoubtedly a substantial computational commitment; however, this way 384 APTs are obtained from a single MD snapshot, which is quite a lot of training data which will become apparent in the following.

Having 10 MD snapshots with explicitly calculated APTs for all atoms available, an APTNN is trained by randomly selecting 9 (90%) configurations for its training set. The remaining configuration is used to validate how well the model generalizes during the training. Recall that the APT is calculated for each atom in each of the 10 snapshots and a single configuration

contains 128 water molecules, i.e. 384 atoms. The training and test sets therefore consist of 3456 and 384 APTs, respectively. I use an `e3nn`*SimpleNetwork* (v2106) model consisting of two message passing layers and a technical feature configuration "`20x0o+20x0e+20x1o+20x1e+20x2o+20x2e`", describing the feature set of each atom. The latter string encodes that each atom is represented by a feature vector containing 20 scalars (tensor rank 0), 20 vectors (tensor rank 1), and 20 tensors of rank 2 with even ("e") and odd ("o") parity each. In a nutshell, the input geometry is translated into a graph representation, where each atom is represented by a node and each interatomic connection is represented by an edge. Through the message passing layers, the feature vector of each node (atom) is iteratively refined, taking the graph edges (interatomic distance vectors) and the feature vectors of all neighboring nodes (atoms) into account. Thereby the feature vectors are optimized, such that they contain a unique representation of the environment of each atom. After the message passing phase, the feature vectors are then used to predict an APT in a given configuration. The interested reader is referred to ref 56 for an elaborate discussion on how the feature vectors are iteratively refined in the `e3nn` framework. Here, I use a radial cutoff of 6 Å to create the graph from the input geometry. This means that only edges between two nodes are added to the graph, if the interatomic distance is smaller than 6 Å. The radial cutoff is mainly implemented for computational efficiency. It ensures that each atom has approximately the same number of neighbors in the graph representation, irrespective of the total system size. This effectively reduces the computational cost, since the number of graph edges (connections between atoms) is limited to the local environment of each atom only. Moreover, the computational cost then scales only linearly with the total

number of atoms in the system.[56] The radial cutoff clearly is a convergence parameter which needs to be tested to be large enough. Here, the actual cutoff value of 6 Å was chosen because it was proved previously that it is large enough to correctly reproduce the potential energy surface of liquid ambient water using High-Dimensional NNPs[65] as well as graph Neural Networks.[56] Note that I will also show in the following that the cutoff is large enough to train the APT and to accurately reproduce the IR spectrum of liquid ambient water.

The model has been trained using the Adam optimizer[66] implemented in pyTorch.[67] The Adam optimizer is one of the most frequently used optimizers in pytorch and has been successfully applied when training e3nn based models in the past, see e.g. ref 56. An initial learning rate of 0.01 is used which is automatically reduced by a factor of 0.1, when the loss of the validation data set does not decrease further over the last 10 training epochs. The hyperparameters have been chosen manually based on the validation set performance during the training and have then been fixed. The overall performance of the trained model was finally evaluated on an unrelated test set and on the predicted IR spectrum compared with available reference and experimental data, see below.

The learning curve of a APTNN model as a function of the data set size (4, 9, 18, and 27 training configurations, corresponding to 1536, 3456, 6912, and 10368 APTs, respectively) is shown in the top panel of Figure 1a. All APTNNs have been trained according to the above-described training procedure. The presented test mean squared error (MSE) is computed on an unknown test set containing 3840 APTs in total stemming from 10 randomly sampled MD snapshots. The test error systematically decreases as a function of provided training data set size, following a linear behavior in the log−log plot as expected.

All data presented in the following are calculated from the APTNN trained using 9 configurations only (solid brown learning curve in the bottom panel of Figure 1a) after 100 epochs (marked by a vertical dashed line). To benchmark this model, I randomly select 10 configurations from the available AIMD simulations of liquid water which have not been included in the training set ("test set"). For those configurations, the APT is explicitly calculated using the electronic structure method as before and compared with the prediction of the APTNN. The direct component-wise comparison is shown in Figure 1b, and the overall component-wise RMSE is 3.49×10^{-2} e, where e is the electron charge.

As a second quality benchmark, I now use the trained APTNN to predict the IR spectrum of liquid ambient water. In practical terms, I take all the available AIMD trajectories of RPBE-D3 liquid ambient water (16 trajectories, 20 ps each using a time step of 1 fs, see above) and predict the APT for each atom at each time step. Recall that this is a daunting task for explicit electronic structure calculations. Having all these APTs available, I then calculate the total dipole moment derivative at each time step according to eq 4. Finally, the machine learned IR spectrum is obtained by eq 2 and presented in Figure 2. The figure also shows the reference IR spectrum which has been obtained directly from the explicit AIMD simulations. Note that this reference IR spectrum of RPBE-D3 water has already been published before, see e.g. ref 13. The comparison shows that the machine learned spectrum reproduces the explicitly calculated reference spectrum exactly in shape and intensity.

Remarkably, this accuracy is reached by only considering 9 randomly selected configurations for training, i.e. without even
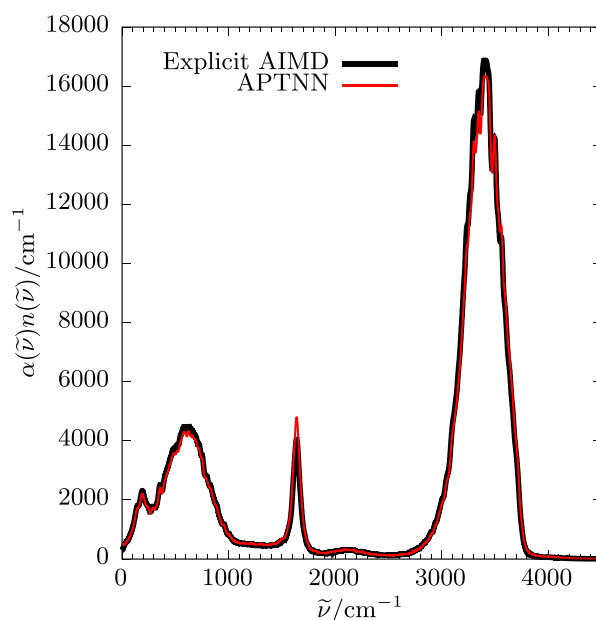

Figure 2. Machine learned IR spectrum of liquid ambient water (red) is compared with the explicit AIMD reference spectrum (black), see text.

using more sophisticated active learning techniques. Indeed, the NequIP package which also utilizes the e3nn library to train the potential energy surface also showed a high data efficiency.[56] This means that a very good agreement with the reference data could be achieved with providing a comparably small training set. Interestingly, the data efficiency could directly be traced back to the equivariance of the model, since the performance of the model significantly declined as the equivariance was disabled.[56] This data efficiency foreshadows that training an APTNN with even more computationally expensive electronic structure theory will be possible; indeed, potential energy surfaces[39,56,68] and polarizability tensors[41] of coupled cluster accuracy have been successfully trained already. This could for example be achieved using a committee of APTNNs which straightforwardly enables active learning. Committees have recently shown to be an efficient tool to obtain a highly accurate model while the training set size is minimized.[69−71]

The acceleration gained by the APTNN is significant. To have a fair comparison, a benchmark was conducted on the very same machine (16 Intel(R) Xeon(R) E5-2640 v3 CPUs, 128 cores in total). A single MD step including wave function extrapolation took about 8 s using the GGA functional RPBE. Calculating an APT for a single atom requires six single points and therefore amounts to 48 s. Predicting an APT for a single atom using the APTNN however only takes about 8 ms, giving an acceleration of more than 3 orders of magnitude. Note that the acceleration becomes even larger when more expensive electronic structure methods are used, e.g. hybrid DFT or beyond. In that case, the computational time to calculate a single point increases dramatically, while the time needed to predict a single APT from the APTNN remains constant. Moreover, pyTorch and e3nn are designed to run more efficiently on GPUs than CPUs. A significant additional acceleration is therefore expected when switching to GPUs; however, this has not been tested yet.

In conclusion, an equivariant neural network has been used to model the atomic polar tensors in liquid ambient water. From this machine learned model, the IR spectrum was calculated which showed excellent agreement with the explicit reference

calculation. Thereby it was demonstrated that the training of the atomic polar tensors is indeed possible. The presented methodology is transferable to any other system class as well. This transferability simply arises by the atomic polar tensor itself: It reduces an IR spectrum to a very fundamental basis, namely atomic motion (given by the atom velocity) and the dipolar changes caused by this motion (given by the atomic polar tensor), corresponding to the IR selection rule. It is therefore a fundamental physical property which is rigorously defined for any atom. Moreover, it does not rely on any definition of molecules or charge partitioning schemes. This virtue is preserved by the herein introduced model, and the latter thus formally *must* be able to describe any atomistic system. Indeed, explicit ab initio molecular dynamics simulations showed that vibrational spectra of e.g. molecules or clusters,[18,53,54] solids,[54] liquids,[50] or solid/liquid interfaces[52] can faithfully be described by the atomic polar tensor.[19]

Atomic polar tensors have been used in the past to understand IR spectra at the microscopic level since they allow one to dissect the spectrum in terms of atomic velocities. This feature was utilized several times in the past already,[18,19,50,53] however, only using severe approximations due to the prohibitive computational cost. This computational bottleneck can be overcome with the herein presented model. Importantly, it was demonstrated that the latter does not compromise on accuracy compared to the presented exhaustive explicit ab initio molecular dynamics benchmark. Moreover, to achieve that accuracy, a surprisingly small training data set was required which foreshadows that it might even be possible to train an APTNN on more expensive electronic structure calculations, such as hybrid DFT or even beyond. Machine learned MD simulations have already been performed using hybrid DFT[71] or even CCSD(T)[72] and allow one to sample nanoseconds of MD trajectories at that level of theory. Such high level MD trajectories can easily be postprocessed by the APTNN model to obtain well converged vibrational spectra, which are clearly computationally prohibitive otherwise.

Given the generality of the atomic polar tensor and the rather small training data set required to accurately train the herein presented APTNN model, the latter has the potential to significantly contribute toward novel physical findings in various systems, especially where large-scale MD simulations or expensive electronic structure calculations are required. Notably, the herein employed modology can also be generalized to other 3×3 tensors, such as the polarizability tensor which is required for Raman and SFG spectra. A general toolkit to automatically and efficiently train an APTNN on top of existing (machine learned) molecular dynamics trajectories is currently being developed.

AUTHOR INFORMATION

Corresponding Author

 Philipp Schienbein − *Department of Physics and Astronomy and Thomas Young Centre, University College London, London WC1E 6BT, United Kingdom;* ⓞ orcid.org/0000-0003-2417-1472; Email: p.schienbein@ucl.ac.uk

Complete contact information is available at:
https://pubs.acs.org/10.1021/acs.jctc.2c00788

Notes

The author declares no competing financial interest.

ACKNOWLEDGMENTS

REFERENCES

(1) Huth, F.; Schnell, M.; Wittborn, J.; Ocelic, N.; Hillenbrand, R. Infrared-spectroscopic nanoimaging with a thermal source. *Nat. Mater.* **2011**, *10*, 352−356.

(2) Bakker, H. J.; Skinner, J. L. Vibrational Spectroscopy as a Probe of Structure and Dynamics in Liquid Water. *Chem. Rev.* **2010**, *110*, 1498−1517.

(3) Perakis, F.; De Marco, L.; Shalit, A.; Tang, F.; Kann, Z. R.; Kühne, T. D.; Torre, R.; Bonn, M.; Nagata, Y. Vibrational Spectroscopy and Dynamics of Water. *Chem. Rev.* **2016**, *116*, 7590−7607.

(4) Kolano, C.; Helbing, J.; Kozinski, M.; Sander, W.; Hamm, P. Watching hydrogen-bond dynamics in a *β*-turn by transient two-dimensional infrared spectroscopy. *Nature* **2006**, *444*, 469−472.

(5) Wolke, C. T.; Fournier, J. A.; Dzugan, L. C.; Fagiani, M. R.; Odbadrakh, T. T.; Knorke, H.; Jordan, K. D.; McCoy, A. B.; Asmis, K. R.; Johnson, M. A. Spectroscopic snapshots of the proton-transfer mechanism in water. *Science* **2016**, *354*, 1131−1135.

(6) Shen, Y. R.; Ostroverkhov, V. Sum-Frequency Vibrational Spectroscopy on Water Interfaces: Polar Orientation of Water Molecules at Interfaces. *Chem. Rev.* **2006**, *106*, 1140−1154.

(7) Nayak, S.; Erbe, A. Mechanism of the potential-triggered surface transformation of germanium in acidic medium studied by ATR-IR spectroscopy. *Phys. Chem. Chem. Phys.* **2016**, *18*, 25100−25109.

(8) Gardner, A. M.; Saeed, K. H.; Cowan, A. J. Vibrational sum-frequency generation spectroscopy of electrode surfaces: studying the mechanisms of sustainable fuel generation and utilisation. *Phys. Chem. Chem. Phys.* **2019**, *21*, 12067−12086.

(9) Kusaka, R.; Nihonyanagi, S.; Tahara, T. The photochemical reaction of phenol becomes ultrafast at the air-water interface. *Nat. Chem.* **2021**, *13*, 306−311.

(10) Heyden, M.; Sun, J.; Funkner, S.; Mathias, G.; Forbert, H.; Havenith, M.; Marx, D. Dissecting the THz spectrum of liquid water from first principles via correlations in time and space. *Proc. Natl. Acad. Sci. U.S.A.* **2010**, *107*, 12068−12073.

(11) Schienbein, P.; Schwaab, G.; Forbert, H.; Havenith, M.; Marx, D. Correlations in the Solute−Solvent Dynamics Reach Beyond the First Hydration Shell of Ions. *J. Phys. Chem. Lett.* **2017**, *8*, 2373−2380.

(12) Imoto, S.; Kibies, P.; Rosin, C.; Winter, R.; Kast, S. M.; Marx, D. Toward Extreme Biophysics: Deciphering the Infrared Response of Biomolecular Solutions at High Pressures. *Angew. Chem., Int. Ed.* **2016**, *55*, 9534−9538.

(13) Schienbein, P.; Marx, D. Supercritical Water is not Hydrogen Bonded. *Angew. Chem., Int. Ed.* **2020**, *59*, 18578−18585.

(14) Ruiz-Barragan, S.; Sebastiani, F.; Schienbein, P.; Abraham, J.; Schwaab, G.; Nair, R. R.; Havenith, M.; Marx, D. Nanoconfinement effects on water in narrow graphene-based slit pores as revealed by THz spectroscopy. *Phys. Chem. Chem. Phys.* **2022**, *24*, 24734.

(15) Marx, D.; Hutter, J. *Ab Initio Molecular Dynamics: Basic Theory and Advanced Methods*; Cambridge University Press: Cambridge, 2009; DOI: 10.1017/CBO9780511609633.

(16) Thomas, M.; Brehm, M.; Kirchner, B. Voronoi dipole moments for the simulation of bulk phase vibrational spectra. *Phys. Chem. Chem. Phys.* **2015**, *17*, 3207−3213.

(17) Wilson, E. B., Jr.; Decius, J. C.; Cross, P. C. *Molecular Vibrations: The Theory of Infrared and Raman Vibrational Spectra*; McGraw-Hill Publishing Company Ltd.: New York, London, Toronto, 1955.

(18) Galimberti, D. R.; Bougueroua, S.; Mahé, J.; Tommasini, M.; Rijs, A. M.; Gaigeot, M.-P. Conformational assignment of gas phase peptides and their H-bonded complexes using far-IR/THz: IR-UV ion dip

experiment, DFT-MD spectroscopy, and graph theory for mode assignment. *Faraday Discuss.* **2019**, *217*, 67−97.

(19) Gaigeot, M.-P. Some opinions on MD-based vibrational spectroscopy of gas phase molecules and their assembly: An overview of what has been achieved and where to go. *Spectrochim. Acta, Part A* **2021**, *260*, 119864.

(20) McQuarrie, D.*Statistical Mechanics*; University Science Books: Sausalito, 2000.

(21) Ramírez, R.; López-Ciudad, T.; Kumar, P. P.; Marx, D. Quantum corrections to classical time-correlation functions: Hydrogen bonding and anharmonic floppy modes. *J. Chem. Phys.* **2004**, *121*, 3973−3983.

(22) Behler, J.; Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys. Rev. Lett.* **2007**, *98*, 146401.

(23) Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Phys. Rev. Lett.* **2010**, *104*, 136403.

(24) Schütt, K. T.; Arbabzadah, F.; Chmiela, S.; Müller, K. R.; Tkatchenko, A. Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **2017**, *8*, 13890.

(25) Behler, J. Four Generations of High-Dimensional Neural Network Potentials. *Chem. Rev.* **2021**, *121*, 10037−10072.

(26) Morawietz, T.; Marsalek, O.; Pattenaude, S. R.; Streacker, L. M.; Ben-Amotz, D.; Markland, T. E. The Interplay of Structure and Dynamics in the Raman Spectrum of Liquid Water over the Full Frequency and Temperature Range. *J. Phys. Chem. Lett.* **2018**, *9*, 851−857.

(27) Shannon, C. Communication in the Presence of Noise. *Proc. IRE* **1949**, *37*, 10−21.

(28) Han, R.; Ketkaew, R.; Luber, S. A Concise Review on Recent Developments of Machine Learning for the Prediction of Vibrational Spectra. *J. Phys. Chem. A* **2022**, *126*, 801−812.

(29) Mulliken, R. S. Electronic Population Analysis on LCAO−MO Molecular Wave Functions. I. *J. Chem. Phys.* **1955**, *23*, 1833−1840.

(30) Hirshfeld, F. Bonded-atom fragments for describing molecular charge densities. *Theoretica chimica acta* **1977**, *44*, 129−138.

(31) Bader, R. F. W. Atoms in molecules. *Acc. Chem. Res.* **1985**, *18*, 9−15.

(32) Wiberg, K. B.; Rablen, P. R. Comparison of atomic charges derived via different procedures. *J. Comput. Chem.* **1993**, *14*, 1504−1518.

(33) Sifain, A. E.; Lubbers, N.; Nebgen, B. T.; Smith, J. S.; Lokhov, A. Y.; Isayev, O.; Roitberg, A. E.; Barros, K.; Tretiak, S. Discovering a Transferable Charge Assignment Model Using Machine Learning. *J. Phys. Chem. Lett.* **2018**, *9*, 4495−4501.

(34) Han, B.; Isborn, C. M.; Shi, L. Determining Partial Atomic Charges for Liquid Water: Assessing Electronic Structure and Charge Models. *J. Chem. Theory Comput.* **2021**, *17*, 889−901.

(35) Ahart, C. S.; Rosso, K. M.; Blumberger, J. Implementation and Validation of Constrained Density Functional Theory Forces in the CP2K Package. *J. Chem. Theory Comput.* **2022**, *18*, 4438.

(36) Gastegger, M.; Behler, J.; Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **2017**, *8*, 6924−6935.

(37) Gao, A.; Remsing, R. C. Self-consistent determination of long-range electrostatics in neural network potentials. *Nat. Commun.* **2022**, *13*, 1572.

(38) Cools-Ceuppens, M.; Dambre, J.; Verstraelen, T. Modeling Electronic Response Properties with an Explicit-Electron Machine Learning Potential. *J. Chem. Theory Comput.* **2022**, *18*, 1672−1691.

(39) Beckmann, R.; Brieuc, F.; Schran, C.; Marx, D. Infrared Spectra at Coupled Cluster Accuracy from Neural Network Representations. *J. Chem. Theory Comput.* **2022**, *18*, 5492−5501.

(40) Schütt, K.; Unke, O.; Gastegger, M.Equivariant message passing for the prediction of tensorial properties and molecular spectra. *Proceedings of the 38th International Conference on Machine Learning*; 2021; pp 9377−9388.

(41) Wilkins, D. M.; Grisafi, A.; Yang, Y.; Lao, K. U.; DiStasio, R. A.; Ceriotti, M. Accurate molecular polarizabilities with coupled cluster theory and machine learning. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 3401−3406.

(42) Grisafi, A.; Fabrizio, A.; Meyer, B.; Wilkins, D. M.; Corminboeuf, C.; Ceriotti, M. Transferable Machine-Learning Model of the Electron Density. *ACS Cent. Sci.* **2019**, *5*, 57−64.

(43) Unke, O.; Bogojeski, M.; Gastegger, M.; Geiger, M.; Smidt, T.; Müller, K.-R. SE(3)-equivariant prediction of molecular wavefunctions and electronic densities. *Advances in Neural Information Processing Systems* **2021**, 14434−14447.

(44) Westermayr, J.; Marquetand, P. Deep learning for UV absorption spectra with SchNarc: First steps toward transferability in chemical compound space. *J. Chem. Phys.* **2020**, *153*, 154112.

(45) Spaldin, N. A. A beginner's guide to the modern theory of polarization. *J. Solid State. Chem.* **2012**, *195*, 2−10.

(46) Kapil, V.; Wilkins, D. M.; Lan, J.; Ceriotti, M. Inexpensive modeling of quantum dynamics using path integral generalized Langevin equation thermostats. *J. Chem. Phys.* **2020**, *152*, 124104.

(47) Raimbault, N.; Grisafi, A.; Ceriotti, M.; Rossi, M. Using Gaussian process regression to simulate the vibrational Raman spectra of molecular crystals. *New J. Phys.* **2019**, *21*, 105001.

(48) Shepherd, S.; Lan, J.; Wilkins, D. M.; Kapil, V. Efficient Quantum Vibrational Spectroscopy of Water with High-Order Path Integrals: From Bulk to Interfaces. *J. Phys. Chem. Lett.* **2021**, *12*, 9108−9114.

(49) Person, W. B.; Newton, J. H. Dipole moment derivatives and infrared intensities. I. Polar tensors. *J. Chem. Phys.* **1974**, *61*, 1040−1049.

(50) Imoto, S.; Marx, D. Pressure Response of the THz Spectrum of Bulk Liquid Water Revealed by Intermolecular Instantaneous Normal Mode Analysis. *J. Chem. Phys.* **2019**, *150*, 084502.

(51) Jaeqx, S.; Oomens, J.; Cimas, A.; Gaigeot, M.-P.; Rijs, A. M. Gas-Phase Peptide Structures Unraveled by Far-IR Spectroscopy: Combining IR-UV Ion-Dip Experiments with Born−Oppenheimer Molecular Dynamics Simulations. *Angew. Chem., Int. Ed.* **2014**, *53*, 3663−3666.

(52) Khatib, R.; Sulpizi, M. Sum Frequency Generation Spectra from Velocity-Velocity Correlation Functions. *J. Phys. Chem. Lett.* **2017**, *8*, 1310−1314.

(53) Galimberti, D. R.; Milani, A.; Tommasini, M.; Castiglioni, C.; Gaigeot, M.-P. Combining Static and Dynamical Approaches for Infrared Spectra Calculations of Gas Phase Molecules and Clusters. *J. Chem. Theory Comput.* **2017**, *13*, 3802−3813.

(54) Ditler, E.; Kumar, C.; Luber, S. Analytic calculation and analysis of atomic polar tensors for molecules and materials using the Gaussian and plane waves approach. *J. Chem. Phys.* **2021**, *154*, 104121.

(55) Geiger, M.; Smidt, T.; M, A; Miller, B. K.; Boomsma, W.; Dice, B.; Lapchevskyi, K.; Weiler, M.; Tyszkiewicz, M.; Batzner, S.; Madisetti, D.; Uhrin, M.; Frellsen, J.; Jung, N.; Sanborn, S.; Wen, M.; Rackers, J.; Rød, M.; Bailey, M.*Euclidean neural networks: e3nn*; 2022; DOI: 10.5281/zenodo.6459381.

(56) Batzner, S.; Musaelian, A.; Sun, L.; Geiger, M.; Mailoa, J. P.; Kornbluth, M.; Molinari, N.; Smidt, T. E.; Kozinsky, B. E. (3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.* **2022**, *13*, 2453.

(57) Grisafi, A.; Wilkins, D. M.; Csányi, G.; Ceriotti, M. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys. Rev. Lett.* **2018**, *120*, 036002.

(58) Hammer, B.; Hansen, L. B.; Nørskov, J. K. Improved Adsorption Energetics within Density-Functional Theory using revised Perdew-Burke-Ernzerhof Functionals. *Phys. Rev. B* **1999**, *59*, 7413−7421.

(59) Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A Consistent and Accurate Ab Initio Parametrization of Density Functional Dispersion Correction (DFT-D) for the 94 Elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104.

(60) Imoto, S.; Forbert, H.; Marx, D. Water structure and solvation of osmolytes at high hydrostatic pressure: pure water and TMAO solutions at 10 kbar versus 1 bar. *Phys. Chem. Chem. Phys.* **2015**, *17*, 24224−24237.

(61) Schienbein, P.; Marx, D. Assessing the properties of supercritical water in terms of structural dynamics and electronic polarization effects. *Phys. Chem. Chem. Phys.* **2020**, *22*, 10462−10479.

(62) Groß, A.; Sakong, S. Ab Initio Simulations of Water/Metal Interfaces. *Chem. Rev.* **2022**, *122*, 10746.

(63) Kühne, T. D.; Iannuzzi, M.; Ben, M. D.; Rybkin, V. V.; Seewald, P.; Stein, F.; Laino, T.; Khaliullin, R. Z.; Schütt, O.; Schiffmann, F.; Golze, D.; Wilhelm, J.; Chulkov, S.; Bani-Hashemian, M. H.; Weber, V.; Borštnik, U.; Taillefumier, M.; Jakobovits, A. S.; Lazzaro, A.; Pabst, H.; Müller, T.; Schade, R.; Guidon, M.; Andermatt, S.; Holmberg, N.; Schenter, G. K.; Hehn, A.; Bussy, A.; Belleflamme, F.; Tabacchi, G.; Glöß, A.; Lass, M.; Bethune, I.; Mundy, C. J.; Plessl, C.; Watkins, M.; VandeVondele, J.; Krack, M.; Hutter, J. CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *J. Chem. Phys.* **2020**, *152*, 194103.

(64) VandeVondele, J.; Krack, M.; Mohamed, F.; Parrinello, M.; Chassaing, T.; Hutter, J. Quickstep: Fast and Accurate Density Functional Calculations Using a mixed Gaussian and Plane Waves Approach. *Comput. Phys. Commun.* **2005**, *167*, 103−128.

(65) Morawietz, T.; Singraber, A.; Dellago, C.; Behler, J. How van der Waals interactions determine the unique properties of water. *Proc. Natl. Acad. Sci. U.S.A.* **2016**, *113*, 8368−8373.

(66) Kingma, D. P.; Ba, J.Adam: A Method for Stochastic Optimization. 2014, arXiv:1412.6980. *arXiv Preprint*. https://arxiv.org/abs/1412.6980 (accessed 2023-01-23).

(67) Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; Chintala, S. In *Advances in Neural Information Processing Systems 32*; Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché Buc, F., Fox, E., Garnett, R., Eds.; Curran Associates, Inc.: 2019; pp 8024−8035.

(68) Schran, C.; Behler, J.; Marx, D. Automated Fitting of Neural Network Potentials at Coupled Cluster Accuracy: Protonated Water Clusters as Testing Ground. *J. Chem. Theory Comput* **2020**, *16*, 88−99.

(69) Schran, C.; Brezina, K.; Marsalek, O. Committee neural network potentials control generalization errors and enable active learning. *J. Chem. Phys.* **2020**, *153*, 104105.

(70) Schran, C.; Thiemann, F. L.; Rowe, P.; Müller, E. A.; Marsalek, O.; Michaelides, A. Machine learning potentials for complex aqueous systems made simple. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, No. e2110077118.

(71) Schienbein, P.; Blumberger, J. Nanosecond solvation dynamics of the hematite/liquid water interface at hybrid DFT accuracy using committee neural network potentials. *Phys. Chem. Chem. Phys.* **2022**, *24*, 15365−15375.

(72) Daru, J.; Forbert, H.; Behler, J.; Marx, D. Coupled cluster molecular dynamics of condensed phase systems enabled by machine learning potentials: Liquid water benchmark. *Phys. Rev. Lett.* **2022**, *129*, 226001.