Machine learning potentials for complex aqueous systems made simple

Christoph Schran[a,b,c,d,1], Fabian L. Thiemann[a,b,c,d,e,2], Patrick Rowe[a,b,c,d,2], Erich A. Müller[e], Ondrej Marsalek[f], and Angelos Michaelides[a,b,c,d,1]

[a]Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom; [b]Thomas Young Centre, University College London, London WC1E 6BT, United Kingdom; [c]London Centre for Nanotechnology, University College London, London WC1E 6BT, United Kingdom; [d]Department of Physics and Astronomy, University College London, London WC1E 6BT, United Kingdom; [e]Department of Chemical Engineering, Sargent Centre for Process Systems Engineering, Imperial College London, London SW7 2AZ, United Kingdom; and [f]Charles University, Faculty of Mathematics and Physics, 121 16 Prague 2, Czech Republic

Edited by Angel Rubio, Max-Planck-Institut fur Struktur und Dynamik der Materie, Hamburg, Germany, and approved July 27, 2021 (received for review June 1, 2021)

Simulation techniques based on accurate and efficient representations of potential energy surfaces are urgently needed for the understanding of complex systems such as solid–liquid interfaces. Here we present a machine learning framework that enables the efficient development and validation of models for complex aqueous systems. Instead of trying to deliver a globally optimal machine learning potential, we propose to develop models applicable to specific thermodynamic state points in a simple and user-friendly process. After an initial ab initio simulation, a machine learning potential is constructed with minimum human effort through a data-driven active learning protocol. Such models can afterward be applied in exhaustive simulations to provide reliable answers for the scientific question at hand or to systematically explore the thermal performance of ab initio methods. We showcase this methodology on a diverse set of aqueous systems comprising bulk water with different ions in solution, water on a titanium dioxide surface, and water confined in nanotubes and between molybdenum disulfide sheets. Highlighting the accuracy of our approach with respect to the underlying ab initio reference, the resulting models are evaluated in detail with an automated validation protocol that includes structural and dynamical properties and the precision of the force prediction of the models. Finally, we demonstrate the capabilities of our approach for the description of water on the rutile titanium dioxide (110) surface to analyze the structure and mobility of water on this surface. Such machine learning models provide a straightforward and uncomplicated but accurate extension of simulation time and length scales for complex systems.

machine learning potentials | solid–liquid systems | aqueous phase

There is a great need for a better understanding of complex aqueous systems, in particular those involving solid–liquid interfaces, to promote progress in fields as diverse as heterogeneous catalysis, material design, biotechnology, and energy conversion or storage (1, 2). For this purpose, atomistic insight provided by computational approaches is urgently required, but off-the-shelf simulation techniques come with important limitations. Ab initio–based methods, such as ab initio molecular dynamics (AIMD), struggle in terms of the accessible time and length scales, while traditional force field approaches are complicated to develop and often not accurate enough to provide reliable answers for complex interface problems. In recent years, machine learning potentials (MLPs) have become a promising alternative, bypassing expensive ab initio calculations and extending length and time scales in molecular simulations (3–7). This is exemplified in studies on the understanding of the unique properties of water (8–10), structural and electronic transitions in disordered silicon (11), and phase transitions of hybrid perovskites (12) to name but a few. The success of MLPs is grounded in a number of distinct approaches that have been introduced over the years, notably, using artificial neural networks (13–17) or kernel-based methods (18–23).

Despite compelling advances toward data-driven and automated techniques mostly in the context of active learning (24–31), the construction of a successful model, in particular for complex systems, remains a difficult task. This becomes most apparent when trying to achieve a high degree of transferability or generality, as, for example, recently shown in the development of general purpose MLPs for silicon (32), carbon (33), or phosphorous (34). The examples cited and many similar ones reported in the literature can take years to develop as broad regions of phase space have to be sampled by an appropriate balance of training points to provide reliable predictions across the board. As a consequence, relatively few studies exist in which complex solid–liquid systems have been described with MLPs (35–39). These limitations have hampered progress in understanding solid–liquid interfaces, where accurate MLPs are urgently needed and offer many opportunities for deepening our understanding of processes like wetting, ice formation, or liquid flow and friction under confinement.

Significance

Understanding complex materials, in particular those with solid–liquid interfaces, such as water on surfaces or under confinement, is a key challenge for technological and scientific progress. Although established simulation approaches have been able to provide important atomistic insight, ab initio techniques struggle with the required time and length scales, while force field methods can often be limited in terms of their accuracy. Here we show how these limitations can be overcome in a simple and automated machine learning procedure to provide accurate models of interactions at the ab initio level, as illustrated for a variety of complex aqueous systems. These developments open up the prospect of the straightforward exploration of many technologically relevant systems by molecular simulations.

Author contributions: C.S., P.R., and A.M. designed research; C.S., F.L.T., P.R., and O.M. performed research; C.S. and O.M. contributed new reagents/analytic tools; C.S., F.L.T., P.R., E.A.M., O.M., and A.M. analyzed data; and C.S., O.M., and A.M. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: cs2121@cam.ac.uk or am452@cam.ac.uk.

[2]F.L.T. and P.R. contributed equally to this work.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2110077118/-/DCSupplemental.

Published September 13, 2021.

However, the high degree of generality that most MLPs try to achieve is in practice not always needed to answer the scientific question at hand. Often, it is sufficient to sample just a small region of configuration space under specific thermodynamic and boundary conditions, while reaching the time and length scales at appropriate ab initio accuracy represents the true challenge. This is the main motivation behind the very promising on-the-fly learning techniques (22, 40, 41) [or surrogate models for global structure optimizations (42)] that do not aim for a high degree of generality. Yet, even these approaches have so far not had wide uptake for complex interfaces, and they have been developed and validated on a system-specific case by case basis.

Here we present another approach to generate MLPs in a simple and fast fashion that is particularly suited for complex systems. To achieve this task, we make use of a (1) widely applicable, robust, and scalable machine learning model for the representation of the structure–energy relation; (2) a general strategy for the generation and selection of representative training data; and finally, (3) a comprehensive and automated validation procedure. By design we concentrate the development on a specific thermodynamic condition. This inherent loss of generality is counterbalanced by the speed of the development as well as the local (as opposed to global) accuracy of the MLP. The workflow to develop and apply such models follows broadly the following simple and computationally inexpensive steps. The relevant thermodynamic condition is initially sampled with a small-scale reference ab initio simulation. This trajectory is screened by a data-driven and automated active learning procedure to construct the machine learning model. The resulting model is then validated through an automated validation protocol and afterward applied in large-scale simulations to answer the relevant scientific question. We show how these models can be developed with minimum human effort, while retaining reliable predictions over long time scales for complex aqueous systems at orders of magnitude lower computational cost than the original ab initio method. This methodology is applied to six exemplary aqueous systems, comprising the fluoride and sulfate ions in aqueous solution, water in carbon and hexagonal boron nitride nanotubes, water on a titanium dioxide surface, and water under molybdenum disulfide confinement.

Due to the efficient nature of this approach, from both a computational and a user perspective, such readily developed models can afterward be applied in extensive molecular simulations to evaluate properties of interest. We demonstrate this in the present case for the description of water on the rutile titanium dioxide surface, for which we investigate structural and dynamical properties with extended molecular dynamics simulations. We believe that the change of paradigm of generating a machine learning model in a cheap and simple process as described here will lead to an increase in the adoption of MLPs to simulate complex systems. These concepts are also expected to be transferable to other machine learning approaches that can be easily coupled to the open-source active learning package (43), which we make freely accessible. Having shown that this approach is able to correctly capture the properties of various aqueous systems under confinement and at interfaces, we suggest that it outlines a straightforward strategy for the uncomplicated but accurate investigation of many technologically relevant systems.

Rapid Development of Committee Neural Network Potentials

The implementation of our MLP framework relies on committee neural network potentials (C-NNPs) (44), in our case built from Behler–Parrinello NNPs (13, 45). While these models only consider the local environment of each atom explicitly within a finite range, long-range contributions can in principle be incorporated in MLPs as demonstrated in the literature (46–49). The main idea behind the current approach is the combination of multiple NNPs in a committee model, where the committee members are separately trained from independent random initial conditions to a subset of the total training set. The resulting committee model has multiple benefits over its individual members. While the committee average, which is used as the prediction of the whole model, has been shown to provide better performance than the individual NNPs, the committee disagreement, which is defined as the SD between the committee member predictions, grants access to an estimate of the error of the model. If scaled by a constant obtained by comparison to the true validation error (50), the committee disagreement provides an objective measure of the accuracy of the underlying model. To construct a training set of such a model in an automated and data-driven way, new configurations with the highest disagreement can be added to the training set. This is an active learning strategy called query by committee and can be used to systematically improve a machine learning model, while making efficient use of the limited data available, which provides an important advantage for the current application. Further details on the C-NNP methodology can be found in ref. 44.

These concepts can be used for the rapid development of MLPs as described in the following and schematically depicted in Fig. 1, *Top*. Initially, a small-scale reference AIMD simulation is performed to sample the system of interest under the selected thermodynamic condition. Small-scale refers here mostly to the envisaged time and length scales of the final application, while they might still be considered large from the ab initio perspective. In practice, AIMD trajectories with a length of 30 ps have been sufficient for this purpose with system sizes that are large enough to cover all required local chemical environments. Given this trajectory, which can also come from existing previous work, we then construct the training set of the C-NNP model by selecting the most representative configurations for the system and condition of interest in an iterative active learning procedure. In the beginning, a small set of 20 structures is randomly selected from the reference trajectory in order to train an initial C-NNP model. Next, query by committee is used to actively select 20 configurations based on the highest committee disagreement in the atomic forces from the set of candidate structures provided by the reference trajectory. These points are added to the training set, and an improved C-NNP model is trained to the extended training set. No additional ab initio reference calculations are required in this process, as the potential energy and atomic forces are already available from the reference AIMD simulation. Such iterations are repeated until convergence of the committee disagreement is observed, indicated by marginal improvements of the disagreement in subsequent iterations and no substantial difference in disagreement between the selected points and those already present in the training set. This implies that a sufficient variety of structures has been added to the training set to yield an accurate and robust C-NNP model.

The uncomplicated construction of MLPs for new systems requires a general set of atomic descriptors. In the case of the atom-centered symmetry functions (51), employed here, we make that possible by generating a systematic set of radial and angular functions. This set consists of 10 equidistantly shifted radial functions with a fixed width and 4 angular functions applied to each pair and triple of atoms, respectively. In addition, we apply the same hyperparameters, such as number of committee members, hidden layers and nodes, and neural network optimization parameters, to every system to remove as much user input from our procedure as possible. Thanks to the active learning procedure that adapts the training set to the flexibility of a particular model, we have in practice observed no limitations of such an application of a general set of parameters and settings. The remaining task of the user is to provide the reference

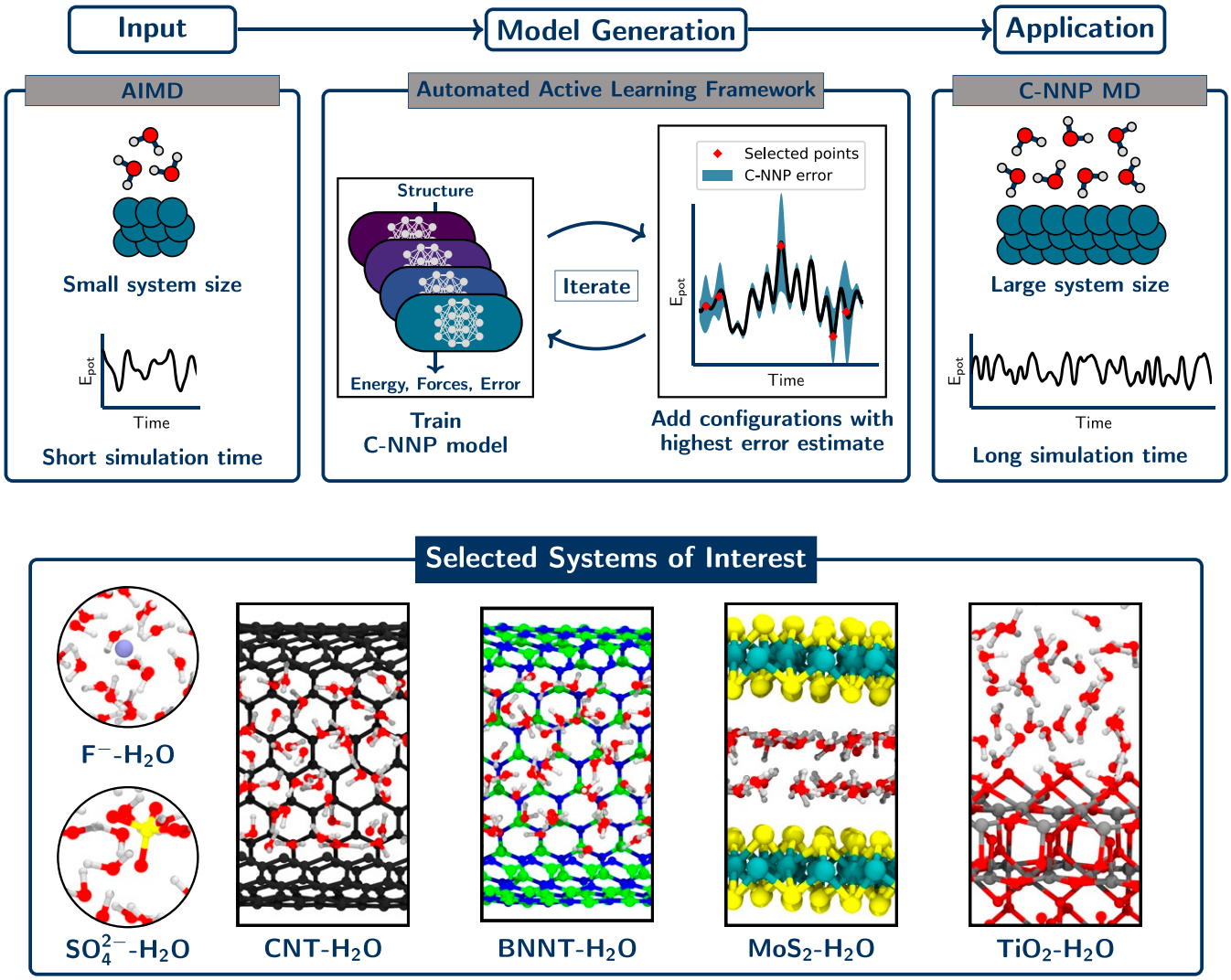

Fig. 1. Schematic depiction of the rapid development process of machine learning models with C-NNPs. (*Top*) The workflow used to generate a C-NNP model starting from a single reference trajectory. Using a small-scale AIMD simulation as input, the C-NNP model is constructed in an active learning cycle that selects the most important configurations for an improvement of the model. This is achieved in an automated iterative process of first training the model and then screening of a large set of candidate configurations for structures with largest error estimate, which are added to the training set. Subsequently, the C-NNP model can be applied to large-scale simulations in order to provide insight into the system of interest. The systems and potential energy curves schematically shown in *Top* are chosen for illustration purposes and do not reflect actual simulation data. (*Bottom*) Representative sections of the simulation cells used for the six aqueous systems chosen in this study for which we successfully applied our machine learning protocol. They are the fluoride ion in solution ($F^- - H_2O$), the sulfate ion in solution ($SO_4^{2-} - H_2O$), water in carbon ($CNT - H_2O$) and hexagonal boron nitride nanotubes ($BNNT - H_2O$), water under molybdenum disulfide confinement ($MoS_2 - H_2O$), and water on a titatium dioxide interface ($TiO_2 - H_2O$).

trajectory for the system of interest under the chosen simulation conditions. Given that input, a C-NNP model can be obtained in practice without further adjustments and in a short amount of time.

Once a new model is trained, it can be applied to the system of interest close to the same state point sampled by the original reference trajectory, reducing the computational cost compared to the ab initio reference by orders of magnitude. During such simulations, the committee disagreement of the C-NNP model is a valuable tool to gauge the validity of the model. It provides an intrinsic error estimate and thus can be monitored over the course of the simulation and compared to the training process. These concepts enable the straightforward extension of both time and length scales in molecular simulations. Per design, the construction in the above fashion will be state dependent. The heart of the matter is that this lack of generality is compensated by the speed and simplicity in its fitting.

Besides this simple and robust framework for the generation of MLPs for complex systems, special emphasis lies on the selection of a suitable electronic structure reference method. MLPs will always only be as good as their underlying reference method, and a user has to make a careful choice for each system of interest. In this context, density functional theory (DFT) has become indispensable for the investigation of aqueous systems, while careful benchmark studies (52–59) can guide the selection of suitable functionals. In addition, there have been promising steps to better understand remaining limitations of existing functionals and provide potential solutions in recent studies (60–62). Combined with the inexpensive yet reliable representation of interactions, as shown in this work, this opens up the possibility for the uncomplicated but accurate investigation of many technologically relevant systems.

Following our active learning workflow, we have obtained C-NNP models for six different aqueous phase systems, which are

shown in Fig. 1, *Bottom*. These diverse systems involve various types of interactions; feature bulk, interface, and confinement regions; and go up to a chemical composition of four elements. In all cases, the training set of the model is exclusively based on a set of structures generated by AIMD simulations, as described in detail in *Materials and Methods*. Given that input, all six models have been generated without further adjustments. For all these systems, convergence was achieved with a compact training set of roughly 300 structures, highlighting the advantages of the active learning procedure.

Automated Quality Assessment of Committee Neural Network Potentials

An automated training procedure calls for an efficient and robust validation protocol. Through extensive comparisons of our models to the underlying ab initio reference trajectories we have identified a general set of properties that serve to provide a stern test of our models. These can be evaluated for any system of interest and provide a broad overview of the performance of a MLP, while further tests are included in *SI Appendix*. The selected properties exemplify the performance of the models for structural and dynamical properties as well as the precision of the force prediction. Specifically, the performance for structural properties is assessed by the match of the radial distribution functions (RDFs) for all involved species comparing the ab initio reference to the model prediction based on molecular dynamics simulations. All RDFs of a given system provide a comprehensive summary of the structural arrangement of the system of interest and are thus ideal to evaluate the performance of the machine learning model for thermodynamic properties. Dynamical properties are validated by comparing the species-resolved vibrational density of states (VDOS) obtained with the model and the reference, which gives a comprehensive overview of intermolecular and intramolecular motions. Finally, the force prediction of the model is validated by the force root mean square error (RMSE) of a randomly selected subset of structures from the ab initio reference simulation. This quantity is chosen since the forces ultimately drive the molecular dynamics simulations when using the model. In order to make these properties comparable for all systems, they are reduced into a score by suitable difference measurements and subsequent averaging over the involved species as described in detail in *SI Appendix*. The entire testing protocol functions in an automated manner and efficiently provides a condensed summary of the accuracy of each model.

The resulting summary of the quality assessment for all six studied systems is shown in Fig. 2. From this analysis it is clear that all models reproduce the three selected properties with rather high precision, where the RDF score ranges between 100 and 98%, the VDOS score between 98 and 96%, and the force score between 95 and 86%. To illustrate the meaning of these values, we depict the individual functions (RDF and VDOS) and the force correlation for the solvated fluoride ion C-NNP model along with the total scores in Fig. 2, while all other properties for the remaining models are compiled in *SI Appendix*. This comparison shows that the selected properties are reproduced with good agreement to the ab initio reference method by our six C-NNP models. In addition, all C-NNP results included in this performance summary are based on substantially extended simulation times compared to the AIMD references as described in detail

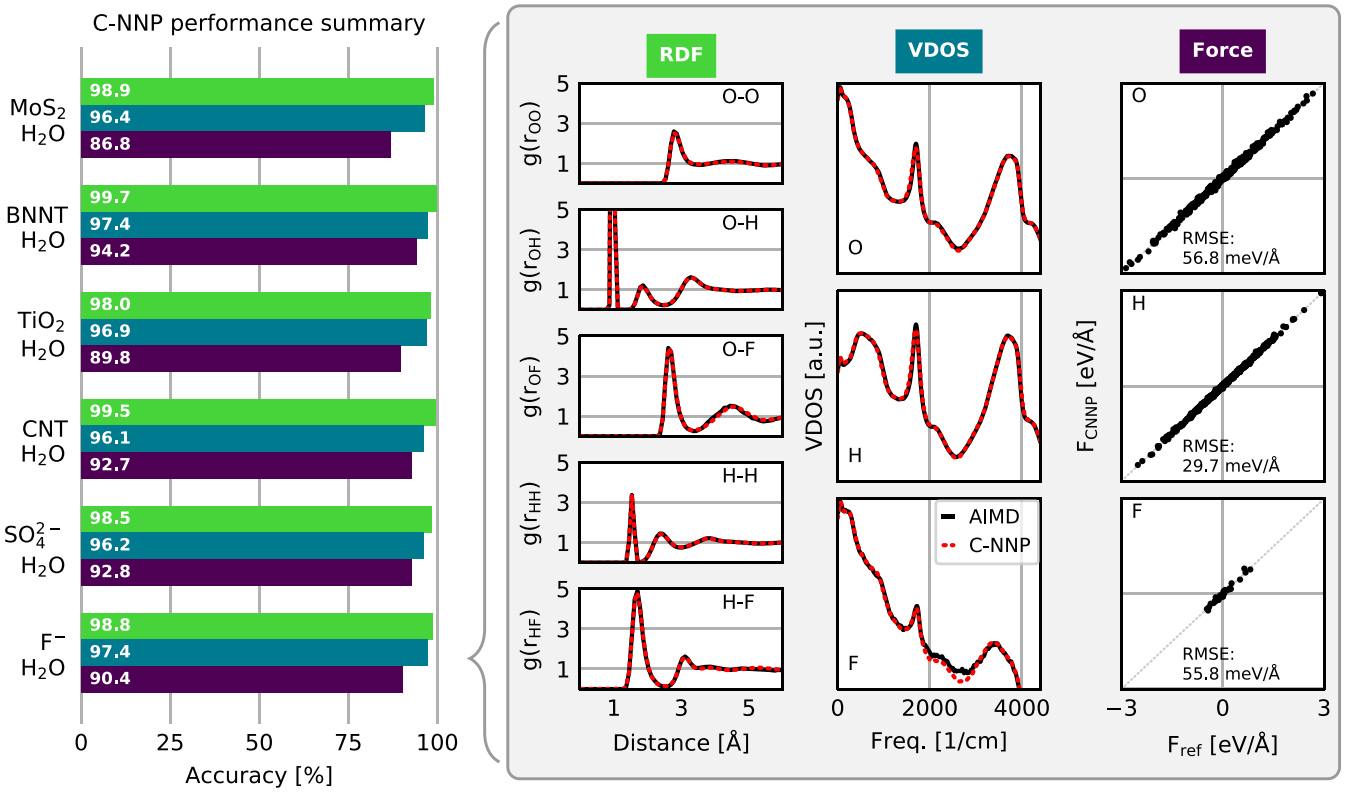

Fig. 2. Performance assessment of the C-NNP for six different aqueous systems. (*Left*) Bar plot featuring the summary of the accuracy for the RDFs, the VDOS, and the force predictions (Force) in percent for each system. (*Right*) The species resolved functions (RDF and VDOS [in logarithmic scale]) and the force correlation of the C-NNP model with respect to the reference method for the solvated fluoride ion ($F^- - H_2O$) C-NNP model, which are condensed into the three scores for the fluoride/water C-NNP model, shown in *Left*. Details of the suitable difference measure and reduction for the three properties can be found in *SI Appendix*.

in *SI Appendix*. This highlights the robust nature of our models, enabling reliable predictions over long time scales. We note that statistical fluctuations due to the more converged nature of our C-NNP simulations account only for very minor changes of the final property scores on the order of 0.5%.

Besides the three sets of properties that we have quantitatively validated here, we have performed additional performance tests, in particular for the more complex systems of water confined in nanotubes and between MoS_2 sheets as well as water on TiO_2. These tests include the detailed analysis of the global structure of the solid and liquid subsystems, as encoded by the density profiles, the hydrogen bonding of water in the various systems, and the orientation of water with respect to the involved interfaces. For all these tests, which are presented in detail in *SI Appendix*, we observe good agreement between our C-NNP results and the AIMD reference simulations within the statistics of those shorter AIMD runs. We are therefore confident that our performance overview, presented in Fig. 2, underlines the high quality of our C-NNP models.

The quality assessment for the six different systems included in this work clearly highlights that our simple and straightforward process to develop MLPs is able to provide robust and accurate models for the selected thermodynamic condition. Compared to the typical DFT setups employed here, the evaluation of the potential energy and atomic forces is usually four to five orders of magnitude faster with the C-NNP model. As a consequence, all chosen systems could now be studied in detail using exhaustive simulations that are accessible with the developed models. Given the focus on general properties in the testing protocol, we expect that it could prove useful for the development of potentials for various other solid–liquid systems of technological and/or scientific interest.

Reaching Longer Length and Time Scales

Let us finally showcase the potential of the presented methodology to extend the length and time scales of molecular simulations and thus further the understanding of a system of interest. For that purpose we investigate structural and dynamical properties of water in contact with rutile TiO_2(110). This system is of scientific and technological importance due to the application of TiO_2, for example, in photocatalysis or self-cleaning coatings and sensors. In addition, it is an established prototypical oxide system in surface science (63) and a rather controversial benchmark system both for theory and experiment (64). For example, the extent of the mobility of water in the contact layers, relevant, e.g., for a detailed understanding of catalytic processes, has been the focus of substantial research (65–69).

In order to shed light on these questions, we have used the developed C-NNP model to simulate rutile TiO_2(110) in contact with water. The model was constructed from a 30-ps AIMD simulation at 300 K with the optB88-vdW functional (70), involving a four O-Ti-O trilayer slab in contact with 80 water molecules, forming a 1.5-nm water film on the surface. After the development and benchmarking of the C-NNP model as shown in the previous section, we made a 2×2 model of the interface (resulting in a TiO_2/H_2O setup with 1,728 atoms) and ran 5 ns of MD. Reaching such length and time scales with AIMD simulations would represent an enormous computational burden, while they can be routinely performed with the C-NNP models. Further details of these simulations can be found in *Materials and Methods*.

First, we analyze the water structuring by looking at the density profile of water on the TiO_2 surface shown in Fig. 3*B*. The first important observation is the close match between the crude density profile obtained for our AIMD simulation and the statistically converged profile obtained with the significantly extended C-NNP simulation. Overall, we observe a highly

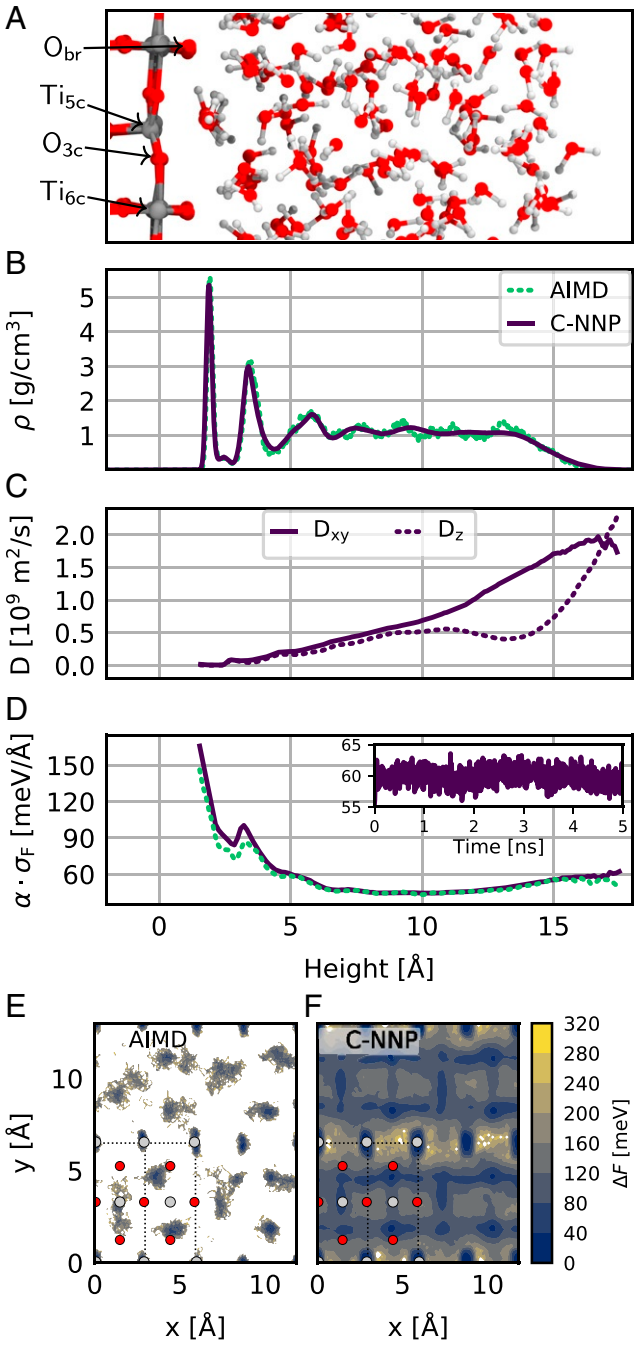

Fig. 3. Properties of water on the rutile TiO_2 (110) surface. (*A*) A representative section of the simulation cell including the four distinct adsorption sites at the interface (Ti_{5c}, fivefold coordinated titanium; Ti_{5c}, sixfold coordinated titanium; O_{3c}, threefold coordinated oxygen; and O_{br}, oxygen bridge site). (*B*) The mass density profile based on all water atoms, (*C*) the water diffusion constant separated into parallel (*xy*) and perpendicular (*z*) components, and (*D*) as a function of the distance from the surface, the C-NNP atomic force error estimate for structures from the C-NNP simulation and for all structures from the original AIMD simulation. This error estimate is obtained as a direct product of the committee disagreement σ_F and a scaling factor α to match the force RMSE of a validation set as proposed in ref. 50. The *Inset* in *D* depicts the average atomic force error estimate of the water atoms as a function of the simulation time. (*E* and *F*) The free energy profile of the water adsorbed in the two contact layers from AIMD and C-NNP simulations, respectively. Titanium atoms are shown in gray, oxygen atoms are shown in red, and hydrogen atoms are shown in white.

structured arrangement of water in the first two layers, which correspond to the water adsorbed on the 5-coordinated titanium site for the first peak and the water above the bridging oxygen site for the second peak, as shown in the snapshot in Fig. 3*A*. This density profile is substantially more structured than for AIMD simulations of water on rutile (66) with the PBE functional, which highlights the complex dependence of interfacial properties on the chosen functional. Given the improved understanding of the dependence of the properties of water on the DFT functional, in particular regarding the inclusion of dispersion interactions (55), we conclude that modern DFT approaches predict a highly structured arrangement of water on the rutile surface reaching up to about 1 nm into the liquid. This is also in good agreement with the density profiles obtained from MLP simulations of water on the anatase (101) surface (37) that used the SCAN functional as the reference method.

In a next step, we evaluate the water diffusion coefficient resolved by the distance from the TiO_2 surface. Specifically, we make use of the mean square displacement, which we spatially decompose based on the position of each water molecule at zero delay (71)—an approach made feasible by the extensive statistics provided by our C-NNP model. We then obtain a local estimate of the diffusion coefficient by the well-known Einstein relation, which can be evaluated separately for the parallel (*xy*) and perpendicular (*z*) directions with respect to the interface. Fig. 3*C* depicts the resulting water diffusion constant $\mathrm{D_{xy}}$ and $\mathrm{D_z}$ as a function of the distance from the TiO_2 surface. As anticipated from the very structured density profile, the mobility close to the surface is reduced substantially, where essentially no diffusion is observed in the strongly adsorbed contact layer. Beyond the first and second layer, the diffusivity in the *xy* direction increases steadily up to the water–vacuum interface. At the same time, $\mathrm{D_z}$ features a plateau around 1 nm from the surface and a substantial increase in diffusion toward the vapor interface. Overall, this analysis highlights the strong influence of the TiO_2 interface on the water diffusion stretching more than 1 nm into the liquid.

Next, we address the accuracy of our extended C-NNP simulations by analyzing the intrinsic error estimate of our model, given by the atomic force committee disagreement σ_F. Fig. 3*D* resolves this local error estimate of the atomic force components for all water atoms as a function of the distance from the surface. This error estimate is a product of the committee disagreement σ_F, as directly provided by our C-NNP simulations, and a scaling factor α determined to match the force RMSE of a validation set as proposed in ref. 50. In addition, we have evaluated this error estimate with our C-NNP model for all configurations of the original AIMD simulation to assess if the increased system size or C-NNP generated structures lead to higher errors. Overall, we observe error estimates between 40 and 80 meV/Å over the entire water region with only slightly higher values close to the TiO_2 surface, indicating the increased complexity of the involved interactions in this inhomogeneous region. Furthermore, the error (averaged over all water atoms) does not deteriorate over the course of the 5-ns-long trajectory, fluctuating around an average of 60 meV/Å, as shown in Fig. 3*D*, *Inset*). Such atomic force errors are similar or even smaller than those reported for other developed MLPs, e.g., for pure water (8–10, 44). At the same time, the error estimate obtained for the AIMD configurations features essentially the same distance resolved profile, which reveals that our C-NNP simulations are able to conserve their predictive power, while substantially extending both time and length scales of the simulations.

Finally, we analyze the free energy profile of water adsorbed in the first two contact layers on the TiO_2 surface, as depicted in Fig. 3 *E* and *F* for the AIMD and C-NNP simulation, respectively. From the direct comparison between the AIMD and C-NNP results it is clear that the limited statistics of the shorter AIMD simulation is insufficient to provide reliable insight into this property. Only with the extensive sampling enabled by the C-NNP model, the free energy profile can be fully converged. The C-NNP free energy profile clearly underlines the strong preference for water adsorption above the fivefold coordinated titanium sites in the first contact layer and the slightly weaker adsorption of water around the sixfold coordinated titanium and threefold coordinated oxygen sites in the second contact layer. In between these adsorption sites, substantial free energy barriers are observed, highlighting the immobile nature of the two contact layers as also revealed by the analysis of the water diffusion.

In summary, the extensive simulations with our C-NNP model, obtained in a straightforward and efficient workflow, provide detailed insight into the properties of water on the rutile surface. We observe a pronounced water layering effect with strong density fluctuations and clear evidence of a highly structured arrangement of water in the first adsorption layers. In addition, our analysis of the water dynamics reveals almost no water diffusion close to the interface and a strong influence on the diffusion stretching more than 1 nm into the liquid. The treatment of a complex interfacial system such as this one requires an accurate description of the binding at the various adsorption sites as well as long-time sampling of the dynamics. The C-NNP model developed here delivers on both fronts, which highlights the potential of our approach to deepen understanding of technologically relevant solid–liquid systems.

Conclusion

In this work, we have presented a machine learning framework that makes the generation of MLPs simple. We have also showcased its versatility for the description of a range of complex aqueous systems. Making use of committee neural network potentials, we have shown how MLPs can be obtained in a straightforward and robust process from a single reference simulation. By essentially removing the need to adjust any hyperparameters, a new system of interest can be tackled in a direct, data-driven way. We have demonstrated the potential of this approach employing it for six complex liquid and solid–liquid systems and have evaluated the quality of the resulting models in detail for various properties underlining the high accuracy of our models. This important final step is realized with an automated validation protocol that is fully integrated into our framework. These developments are directly accessible to the community as they build exclusively on open-source solutions, and we make our underlying software package and all templates available (43).

In its spirit similar to on-the-fly learning techniques, we depart from the goal of a high degree of transferability or generality to concentrate exclusively on the thermodynamic condition relevant for the chosen scientific question. Under these constraints, we have shown how a robust and accurate MLP can be obtained with limited user input in an uncomplicated process. We note that we have also explored the application of our models for elevated and lowered temperatures, which was possible without problems in a $\pm$30 K regime. During such simulations, the intrinsic error estimate of our approach is a very useful tool to gauge the validity of the results. Due to these promising signs, we plan to systematically validate the robustness of our models to venture beyond the chosen thermodynamic state point, for example, to describe variations in pressure.

We note that we built our development on established components, such as an efficient ML structure–energy representation and active learning concepts. At the same time, we expect that the concepts laid out in this work are transferable to other MLPs

and active learning approaches, making it possible to achieve similar results. Our work presents a change in perspective, where relatively little effort is put into creating an intermolecular potential which, despite concentrating on a subsection of phase space, is still robust and accurate enough to be used to describe the nontrivial behavior of complex molecular systems. The importance clearly is not in the individual pieces but rather in the end-to-end framework and the broad range of applications made possible. This work therefore enables simulations that were not possible a short time ago, pushing forward the straightforward and reliable application of MLPs.

Looking to the future, we have limited ourselves to aqueous systems with four species in this work. However, we are confident that systems with more element types can also be tackled with our approach. In addition, we have not explored reactive processes in this study. Since MLPs are able to describe bond breaking and bond making events by design, we again anticipate the straightforward application of our methodology to such situations. This will be especially important to investigate interesting surface reaction phenomena, such as water dissociation on reactive surfaces or the recently reported reversible hydrolysis of zeolites in contact with water (72). Key to a successful application in such situations will be the sampling of the relevant reactive process by the initial AIMD simulation used as input to our active learning protocol. Finally, we are currently exploring a training protocol in which structures are generated by classical molecular dynamics as input for our active learning protocol to minimize the need for expensive ab initio calculations. In this approach the expensive quantum computing engine is only used to obtain the ab initio potential energies and atomic forces for those configurations identified to be most important for the generation of the model. Such an approach has potential for additional cost savings over the one presented here and does not require expertise in AIMD simulations, thus more readily opening up the approach to researchers from the classical force field community.

In our six showcase applications of complex aqueous systems we have developed models with different DFT functionals, all of which represent reasonable choices for the aqueous systems studied. However, this illustrates a broader issue which is that there is currently no perfect DFT functional for water and complex aqueous interfaces. We believe that the approach developed here could become a valuable tool in this long-standing quest to find suitable DFT functionals. Since our procedures provide the ability to reveal the true converged thermal performance of any given functional for a realistic system, at a modest cost, the systematic exploration of the performance of DFT methods for complex disordered systems becomes feasible. This makes it possible to go beyond the usual energetic benchmarks of relatively small systems in the absence of temperature and thus facilitates direct comparison with experiment. Due to the moderate size of our training sets, our framework is also expected to be easily extendable to more expensive ab initio methods, e.g., at the hybrid DFT level or considering explicit electron correlation, thus making these methods available for the realistic simulation of complex interfacial systems.

Overall, the developments reported herein will enable the investigation of complex aqueous processes such as water structuring in contact with interfaces and wetting or ice formation on surfaces in a straightforward manner. Although here we applied it to aqueous systems, we believe that the methodology will also prove useful for other materials and liquids in contact with solids, as well as general solvation phenomena, enabling the fast screening of different materials at ab initio accuracy. It will also be particularly useful for situations where long sampling is required as for the exploration of free energy surfaces, or calculations of dynamical properties, such as the friction or viscosity of liquids in contact with interfaces. In summary, this work outlines a straightforward strategy for the uncomplicated yet accurate investigation of many technologically and scientifically relevant systems by molecular simulations.

Materials and Methods

The introduced machine learning framework has been implemented in the AML Python package, which interleaves the required simulation packages and data manipulation steps in a user-friendly environment. The AML package is available free of charge at https://github.com/MarsalekGroup/aml and enables the straightforward generation of a C-NNP model given a reference trajectory as input. With this code all six C-NNP models were developed for the various aqueous phase systems studied here. NNP optimizations were performed with the open-source n2p2 code (73) using the optimization parameters and symmetry functions as provided in the template file in the associated data repository for this paper. All additional information on the C-NNP fitting procedure can be found in *SI Appendix*, while all training input files, training sets, and parameters of the final models are publicly available at https://doi.org/10.5281/zenodo.5235246 (74). The reference AIMD simulations used as the starting point for our C-NNP models employed quite different DFT settings, while all having been performed with the CP2K simulation package (75). We provide full detail about these reference simulations in *SI Appendix*, but the typical simulation setups consist of 64 to 110 water molecules reaching simulation times between 30 and 130 ps. MD simulations using the C-NNP models were also performed with CP2K, which features an open-source implementation of the C-NNP methodology since release 8.1. All C-NNP simulations for our validation protocol were propagated for at least 0.5 ns to allow for the converged computation of the RDF and VDOS. Further details of the validation protocol and the associated simulations can be found in *SI Appendix*, while the validation can be performed with the AML Python package. The C-NNP simulations of the TiO_2 water system were propagated for 5 ns for a $2 \times 2 \times 1$ supercell of the original AIMD setup, resulting in total in 1,728 atoms making up 320 water molecules on four O-Ti-O trilayers in a 23.672, 25.988, 42.0 Å periodic box. This simulation employed a molecular dynamics time step of 1 fs, while using deuterium masses for the hydrogen atoms. The temperature of 300 K was maintained with a canonical sampling through velocity rescaling thermostat (76).

Data Availability. Algorithms and computer codes have been deposited on GitHub (https://github.com/MarsalekGroup/aml), while datasets are available in Zenodo (https://doi.org/10.5281/zenodo.5235246).

ACKNOWLEDGMENTS. We thank Christopher Penschke for providing the water TiO_2 AIMD trajectory. C.S. acknowledges partial financial support from the Alexander von Humboldt-Stiftung. This work was partially supported by the Operational Programme Research, Development and Education project (CZ.02.2.69/0.0/0.0/18_070/0010462), International Mobility of Researchers at Charles University (Marie Skłodowska-Curie Individual Fellowships II). We are grateful to the UK Materials and Molecular Modelling Hub for computational resources, which is partially funded by Engineering and Physical Sciences Research Council (EPSRC) (grants EP/P020194/1 and EP/T022213/1). We are also grateful for computational support from the UK national high-performance computing service, Advanced Research Computing High End Resource (ARCHER), for which access was obtained via the UK Car-Parrinello consortium, funded by EPSRC grant reference EP/P022561/1.

1. F. Zaera, Probing liquid/solid interfaces at the molecular level. *Chem. Rev.* **112**, 2920–2986 (2012).
2. O. Björneholm *et al*., Water at interfaces. *Chem. Rev.* **116**, 7698–7726 (2016).
3. J. Behler, Perspective: Machine learning potentials for atomistic simulations. *J. Chem. Phys.* **145**, 170901 (2016).
4. K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, A. Walsh, Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
5. V. L. Deringer, M. A. Caro, G. Csányi, Machine learning interatomic potentials as emerging tools for materials science. *Adv. Mater.* **31**, e1902765 (2019).
6. P. L. Kang, C. Shang, Z. P. Liu, Large-scale atomic simulation via machine learning potentials constructed by global potential energy surface exploration. *Acc. Chem. Res.* **53**, 2119–2129 (2020).
7. J. Behler, Four generations of high-dimensional neural network potentials. *Chem. Rev.*, 10.1021/acs.chemrev.0c00868 (2021).

8. T. Morawietz, A. Singraber, C. Dellago, J. Behler, How van der Waals interactions determine the unique properties of water. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 8368–8373 (2016).
9. B. Cheng, E. A. Engel, J. Behler, C. Dellago, M. Ceriotti, Ab initio thermodynamics of liquid and solid water. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 1110–1115 (2019).
10. T. E. Gartner 3rd *et al*., Signatures of a liquid-liquid transition in an ab initio deep neural network model for water. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 26040–26046 (2020).
11. V. L. Deringer *et al*., Origins of structural and electronic transitions in disordered silicon. *Nature* **589**, 59–64 (2021).
12. R. Jinnouchi, J. Lahnsteiner, F. Karsai, G. Kresse, 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).
13. J. Behler, M. Parrinello, Generalized neural-network representation of high-dimensional potential-energy surfaces. *Phys. Rev. Lett.* **98**, 146401 (2007).
14. S. A. Ghasemi, A. Hofstetter, S. Saha, S. Goedecker, Interatomic potentials for ionic systems with density functional accuracy based on charge densities obtained by a neural network. *Phys. Rev. B Condens. Matter Mater. Phys.* **92**, 045131 (2015).
15. K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, A. Tkatchenko, Quantum-chemical insights from deep tensor neural networks. *Nat. Commun.* **8**, 13890 (2017).
16. L. Zhang, J. Han, H. Wang, R. Car, W. E, Deep potential molecular dynamics: A scalable model with the accuracy of quantum mechanics. *Phys. Rev. Lett.* **120**, 143001 (2018).
17. O. T. Unke, M. Meuwly, PhysNet: A neural network for predicting energies, forces, dipole moments, and partial charges. *J. Chem. Theory Comput.* **15**, 3678–3693 (2019).
18. A. P. Bartók, M. C. Payne, R. Kondor, G. Csányi, Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Phys. Rev. Lett.* **104**, 136403 (2010).
19. M. Rupp, A. Tkatchenko, K. R. Müller, O. A. von Lilienfeld, Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
20. A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, G. J. Tucker, Spectral neighbor analysis method for automated generation of quantum-accurate interatomic potentials. *J. Comput. Phys.* **285**, 316–330 (2015).
21. A. V. Shapeev, Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Model. Simul.* **14**, 1153–1173 (2015).
22. Z. Li, J. R. Kermode, A. De Vita, Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Phys. Rev. Lett.* **114**, 096405 (2015).
23. S. Chmiela *et al*., Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
24. M. Gastegger, J. Behler, P. Marquetand, Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci. (Camb.)* **8**, 6924–6935 (2017).
25. E. V. Podryabinkin, A. V. Shapeev, Active learning of linearly parametrized interatomic potentials. *Comput. Mater. Sci.* **140**, 171–180 (2017).
26. L. Zhang, D.-Y. Lin, H. Wang, R. Car, W. E, Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Mater.* **3**, 023804 (2019).
27. J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, A. E. Roitberg, Less is more: Sampling chemical space with active learning. *J. Chem. Phys.* **148**, 241733 (2018).
28. V. L. Deringer, C. J. Pickard, G. Csányi, Data-driven learning of total and local energies in elemental boron. *Phys. Rev. Lett.* **120**, 156001 (2018).
29. K. T. Schütt, H. E. Sauceda, P. J. Kindermans, A. Tkatchenko, K. R. Müller, SchNet - A deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
30. F. Musil *et al*., Machine learning for the structure-energy-property landscapes of molecular crystals. *Chem. Sci. (Camb.)* **9**, 1289–1300 (2017).
31. C. Schran, J. Behler, D. Marx, Automated fitting of neural network potentials at coupled cluster accuracy: Protonated water clusters as testing ground. *J. Chem. Theory Comput.* **16**, 88–99 (2020).
32. A. P. Bartók, J. Kermode, N. Bernstein, G. Csányi, Machine learning a general-purpose interatomic potential for silicon. *Phys. Rev. X* **8**, 041048 (2018).
33. P. Rowe, V. L. Deringer, P. Gasparotto, G. Csányi, A. Michaelides, An accurate and transferable machine learning potential for carbon. *J. Chem. Phys.* **153**, 034702 (2020).
34. V. L. Deringer, M. A. Caro, G. Csányi, A general-purpose machine-learning force field for bulk and nanostructured phosphorus. *Nat. Commun.* **11**, 5461 (2020).
35. S. K. Natarajan, J. Behler, Neural network molecular dynamics simulations of solid-liquid interfaces: Water at low-index copper surfaces. *Phys. Chem. Chem. Phys.* **18**, 28704–28725 (2016).
36. M. Hellström, V. Quaranta, J. Behler, One-dimensional *vs.* two-dimensional proton transport processes at solid-liquid zinc-oxide-water interfaces. *Chem. Sci. (Camb.)* **10**, 1232–1243 (2018).
37. F. Marcos *et al*, Free energy of proton transfer at the water-TiO_2 interface from: Ab initio deep potential molecular dynamics. *Chem. Sci. (Camb.)* **11**, 2335–2341 (2020).
38. H. Ghorbanfekr, J. Behler, F. M. Peeters, Insights into water permeation through hBN nanocapillaries by ab initio machine learning molecular dynamics simulations. *J. Phys. Chem. Lett.* **11**, 7363–7370 (2020).
39. N. Artrith, Machine learning for the modeling of interfaces in energy storage and conversion materials. *J. Phys. Energy* **1**, 032002 (2019).
40. R. Jinnouchi, K. Miwa, F. Karsai, G. Kresse, R. Asahi, On-the-fly active learning of interatomic potentials for large-scale atomistic simulations. *J. Phys. Chem. Lett.* **11**, 6946–6955 (2020).
41. J. Vandermause *et al*., On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Comput. Mater.* **6**, 1–11 (2020).
42. M. K. Bisbo, B. Hammer, Efficient global structure optimization with a machine-learned surrogate model. *Phys. Rev. Lett.* **124**, 086102 (2020).
43. C. Schran, O. Marsalek, MarsalekGroup/aml: AML Python package. GitHub. https://github.com/MarsalekGroup/aml. Deposited 31 May 2021.
44. C. Schran, K. Brezina, O. Marsalek, Committee neural network potentials control generalization errors and enable active learning. *J. Chem. Phys.* **153**, 104105 (2020).
45. J. Behler, First principles neural network potentials for reactive simulations of large molecular and condensed systems. *Angew. Chem. Int. Ed. Engl.* **56**, 12828–12840 (2017).
46. N. Artrith, T. Morawietz, J. Behler, High-dimensional neural-network potentials for multicomponent systems: Applications to zinc oxide. *Phys. Rev. B Condens. Matter Mater. Phys.* **83**, 153101 (2011).
47. A. Grisafi, M. Ceriotti, Incorporating long-range physics in atomic-scale machine learning. *J. Chem. Phys.* **151**, 204105 (2019).
48. T. W. Ko, J. A. Finkler, S. Goedecker, J. Behler, A fourth-generation high-dimensional neural network potential with accurate electrostatics including non-local charge transfer. *Nat. Commun.* **12**, 398 (2021).
49. S. Yue *et al*., When do short-range atomistic machine-learning models fall short? *J. Chem. Phys.* **154**, 034111 (2021).
50. G. Imbalzano *et al*., Uncertainty estimation for molecular dynamics and sampling. *J. Chem. Phys.* **154**, 074102 (2021).
51. J. Behler, Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *J. Chem. Phys.* **134**, 074106 (2011).
52. J. C. Grossman, E. Schwegler, E. W. Draeger, F. Gygi, G. Galli, Towards an assessment of the accuracy of density functional theory for first principles simulations of water. *J. Chem. Phys.* **120**, 300–311 (2004).
53. R. A. DiStasio, Jr, B. Santra, Z. Li, X. Wu, R. Car, The individual and collective effects of exact exchange and dispersion interactions on the ab initio structure of liquid water. *J. Chem. Phys.* **141**, 084502 (2014).
54. K. Forster-Tonigold, A. Groß, Dispersion corrected RPBE studies of liquid water. *J. Chem. Phys.* **141**, 064501 (2014).
55. M. J. Gillan, D. Alfè, A. Michaelides, Perspective: How good is DFT for water? *J. Chem. Phys.* **144**, 130901 (2016).
56. M. Chen *et al*., Ab initio theory and modeling of water. *Proc. Natl. Acad. Sci. U.S.A.* **114**, 10846–10851 (2017).
57. O. Marsalek, T. E. Markland, Quantum dynamics and spectroscopy of ab initio liquid water: The interplay of nuclear and electronic quantum effects. *J. Phys. Chem. Lett.* **8**, 1545–1551 (2017).
58. J. G. Brandenburg, A. Zen, D. Alfè, A. Michaelides, Interaction between water and carbon nanostructures: How good are current density functional approximations? *J. Chem. Phys.* **151**, 164702 (2019).
59. P. Schienbein, D. Marx, Supercritical water is not hydrogen bonded. *Angew. Chem. Int. Ed. Engl.* **59**, 18578–18585 (2020).
60. K. Sharkas *et al*., Self-interaction error overbinds water clusters but cancels in structural energy differences. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 11283–11288 (2020).
61. K. Wagle *et al*., Self-interaction correction in water-ion clusters. *J. Chem. Phys.* **154**, 094302 (2021).
62. T. T. Duignan, S. M. Kathmann, G. K. Schenter, C. J. Mundy, Toward a first-principles framework for predicting collective properties of electrolytes. *Acc. Chem. Res.* **54**, 2833–2843 (2021).
63. C. L. Pang, R. Lindsay, G. Thornton, Structure of clean and adsorbate-covered single-crystal rutile TiO_2 surfaces. *Chem. Rev.* **113**, 3887–3948 (2013).
64. U. Diebold, Perspective: A controversial benchmark system for water-oxide interfaces: H_2O/TiO_2(110). *J. Chem. Phys.* **147**, 040901 (2017).
65. M. Předota, P. T. Cummings, D. J. Wesolowski, Electric double layer at the rutile (110) surface. 3. Inhomogeneous viscosity and diffusivity measurement by computer simulations. *J. Phys. Chem. C* **111**, 3071–3079 (2007).
66. M. L. Li, C. Zhang, G. Thornton, A. Michaelides, Structure and dynamics of liquid water on rutile TiO_2(110). *Phys. Rev. B Condens. Matter Mater. Phys.* **82**, 161415 (2010).
67. E. C. Spencer *et al*., Inelastic neutron scattering study of confined surface water on rutile nanoparticles. *J. Phys. Chem. A* **113**, 2796–2800 (2009).
68. N. J. English, R. S. Kavathekar, J. M. D. MacElroy, Hydrogen bond dynamical properties of adsorbed liquid water monolayers with various TiO_2 interfaces. *Mol. Phys.* **110**, 2919–2925 (2012).
69. L. Agosta, E. G. Brandt, A. P. Lyubartsev, Diffusion and reaction pathways of water near fully hydrated TiO_2 surfaces from ab initio molecular dynamics. *J. Chem. Phys.* **147**, 024704 (2017).
70. J. Klimeš, D. R. Bowler, A. Michaelides, Chemical accuracy for the van der Waals density functional. *J. Phys. Condens. Matter* **22**, 022201 (2010).
71. E. Pluhařová, P. Jungwirth, N. Matubayasi, O. Marsalek, Structure and dynamics of the hydration shell: Spatially decomposed time correlation approach. *J. Chem. Theory Comput.* **15**, 803–812 (2019).
72. C. J. Heard *et al*., Fast room temperature lability of aluminosilicate zeolites. *Nat. Commun.* **10**, 4690 (2019).
73. A. Singraber, T. Morawietz, J. Behler, C. Dellago, Parallel multistream training of high-dimensional neural network potentials. *J. Chem. Theory Comput.* **15**, 3075–3092 (2019).
74. C. Schran, water-ice-group/simple-MLP: Supporting data for published paper. Zenodo. https://doi.org/10.5281/zenodo.5235246. Deposited 23 August 2021.
75. T. D. Kühne *et al*., CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *J. Chem. Phys.* **152**, 194103 (2020).
76. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *J. Chem. Phys.* **126**, 014101 (2007).