






Cite this: *Phys. Chem. Chem. Phys.*,  
2025, 27, 9169

# From flat to stepped: active learning frameworks for investigating local structure at copper–water interfaces

Johannes Schörghuber,  Nina Bučková, Esther Heid  and  
Georg K. H. Madsen  \*

Understanding processes at solid–liquid interfaces at the atomic level is important for applications such as electrocatalysis. Here we explore the effects of different step densities on the structure of interfacial water at the copper–water interface. Utilizing spatially resolved uncertainties, we develop an active learning framework and train a machine-learning force field (MLFF) based on dispersion-corrected density functional theory data. Using molecular dynamics simulations, we investigate structural properties of water molecules in the contact layer, including density profiles, angular distributions, and 2D pair correlation functions. In accordance with previous studies, we observe the formation of two sublayers within the contact layer at the Cu(111) surface, whereas the structure on surfaces with a high step density is dominated by the undercoordinated ridge atoms. By systematically decreasing the step density, we identify the cross-over to when the behavior observed at the flat surface can be locally recovered. Using dimensionality reduction, we identify four distinct types of Cu environments at the interfaces, providing insights into analyzing less idealized surfaces with MLFFs.

Received 29th January 2025,  
Accepted 3rd April 2025

DOI: 10.1039/d5cp00396b

rsc.li/pccp

## 1 Introduction

Metal–water interfaces are of large importance in technological applications, for example in devices used in electrochemical energy conversion and storage.<sup>1,2</sup> In heterogeneous catalysis not only the composition of the material, but also the surface morphology has a profound effect on performance and efficiency. In the case of electrochemical CO<sub>2</sub> reduction on Cu, just changing the orientation of a single crystal already has a large impact on the selectivity toward C<sub>2</sub>- or C<sub>1</sub>-products.<sup>3–5</sup> Further, the structure of both the surface and in conjunction also interfacial water has been shown to be an integral descriptor for other catalytic processes on Cu surfaces, such as the hydrogen evolution reaction and CO reduction, using *in situ* measurements at the solid–liquid interface.<sup>6,7</sup>

The applicability of classical force fields for explicit simulations of metal–water interfaces in molecular dynamics (MD) simulations is limited by the fundamentally different properties of the solid and liquid phases. On the other hand, the computational cost of *ab initio* calculations meant that early work was focused on clusters and ultra-thin layers of frozen water.<sup>1,8,9</sup> As *ab initio* molecular dynamics (AIMD) and machine-learned force fields (MLFFs) have become available they have enabled progressively more realistic

simulations of the bulk water–metal interface. These studies have established the presence of a double-peak structure in the interfacial water density for the Cu(111) surface,<sup>10,11</sup> similar to observations for Pt(111) surface.<sup>11–16</sup> Additionally, the distinct behavior of interfacial water on stepped surfaces has also been demonstrated for Cu–H<sub>2</sub>O and Pt–H<sub>2</sub>O interfaces.<sup>11,14,17</sup> However, while global descriptors such as Miller indices and edge densities can provide insights into the bonding properties of corrugated surfaces, they can obscure the local atomistic details. Bridging the gap between morphological understanding and detailed atomic-level analysis remains a key challenge.<sup>18</sup> MD simulations have historically been limited when it comes to achieving the time scales necessary to generate density profiles and analysis with local resolution. The recent rapid development of MLFFs has made it possible to obtain reliable statics for local resolution through higher-dimensional pair-correlation functions and free energy calculations.<sup>16</sup> This can open up possibilities for bridging the gap between morphological insights and local atomistic understanding.

The quality of MLFF predictions is highly dependent on the data it was trained on, as the sampled structures determine the region of configuration space for which accurate predictions can be obtained. As even small representative systems for solid–liquid interfaces contain hundreds of atoms and the *ab initio* reference calculations are consequently costly, it is important to sample new configurations efficiently and only perform as many calculations as necessary. This makes it a prime task for

Institute of Materials Chemistry, TU Wien, A-1060 Vienna, Austria.  
E-mail: georg.madsen@tuwien.ac.at



active learning (AL).<sup>19</sup> In AL procedures, new configurations are added to the training database based on the inaccuracy of the model predictions. The model is then retrained in order to improve model predictions in the region of the configuration space in which the newly selected structures lie. This procedure is iterated until a desired convergence is reached. When studying interfaces it is clearly desirable that the AL procedure can be built on preexisting data for the individual constituents,<sup>20</sup> and then use an AL strategy to sample from model-based interface simulations.

In the present study, we develop a MLFF to systematically investigate the flat Cu(111) and stepped Cu( $n + 1$ ,  $n$ ,  $n$ )-H<sub>2</sub>O interfaces ( $1 \leq n \leq 3$ ) and elucidate the effect of different step densities on the structure of interfacial water. Beginning with reference datasets for bulk water and Cu, we employ an AL procedure to efficiently construct a dataset and a transferable MLFF for the Cu-H<sub>2</sub>O interfaces. We show how the use of spatially resolved uncertainties<sup>21</sup> allows to finely resolve the quality of model predictions in the different regions of interface structures. MD simulations are then conducted to obtain atomically resolved structural properties of the H<sub>2</sub>O network in the contact layer at the various Cu-H<sub>2</sub>O interfaces. A data-driven classification of the local geometries reveals four distinct types of Cu atom environments at the interface. Notably, the Cu(433)-H<sub>2</sub>O interface can be identified as the cross-over where the local structure characteristics of the flat Cu(111) surface can be recovered on a stepped surface.

## 2 Computational details

DFT calculations for the initial Cu database were carried out using VASP version 6.2.0,<sup>22</sup> with the RPBE functional being used to model the XC contributions to the total energy.<sup>23</sup> The default PAW setups provided with VASP with a core radius of 2.3 Å for Cu were used. The plane wave energy cutoff was set to 400 eV and the second order Methfessel-Paxton smearing scheme was employed with a smearing width of 0.05 eV. The Brillouin zone was sampled with  $k$ -point densities corresponding to a  $11 \times 11 \times 11$   $\Gamma$ -centered mesh for the one-atom primitive unit cell of fcc-Cu. For surface calculations, only a single  $k$ -point was considered in the surface normal direction.

Calculations for water were run using VASP version 6.4.2. Due to the short bond lengths in water the hard PAW setups provided with VASP were used, the core radii being 0.8 Å for hydrogen and 1.1 Å for oxygen. The energy cutoff was set to 850 eV, the width for Gaussian smearing to 0.05 Å and only the  $\Gamma$ -point of the Brillouin zone was sampled. To account for van-der-Waals interactions, D3 corrections<sup>24</sup> were computed using the zero damping scheme following previously reported results.<sup>25,26</sup>

Energies and forces for Cu-H<sub>2</sub>O interface structures were calculated using VASP version 6.4.2 with the calculation parameters for bulk water as described above, VASP default Cu PAW setups and the  $k$ -point grid being taken corresponding to the  $k$ -point grid for Cu. Only a single  $k$ -point was considered in the surface normal direction. To approximate the screening of D3 interactions by the metal, only the water molecules and the top

Cu layers were included in the evaluation of the D3 contributions to the total energy.<sup>27–29</sup>

MD simulations were carried out using MACE version 0.3.5<sup>30</sup> models to calculate energies and forces and LAMMPS version 2023.3.28 as the simulation engine.<sup>31</sup> MACE models were trained using the package as provided with hyperparameters set as described in the following. Models were constructed using a cutoff radius of 5 Å, two layers with rank zero even parity and rank one odd parity hidden features of size 64 each, and a maximum radial order of  $l_{\max} = 2$ . Radial features were constructed using eight Bessel functions and a polynomial cutoff of order  $p = 5$ . Messages were generated using a MLP with three layers of 64 nodes each and SiLU as the non-linear transfer function. The readouts were performed using a single-layer MLP with 16 nodes. Trainings were run using the AMSGrad optimizer<sup>32</sup> with hyperparameters and learning rates as given by the defaults provided with the MACE package. First, model parameters were optimized with energy and force weights of 1.0 and 100.0 respectively for a maximum of 1200 epochs with an early stopping patience of 50 epochs. Subsequently, energy and force weights were set to 1000.0 and 100.0 respectively and another maximum of 400 epochs were performed.

The timestep for MD simulations was set to 0.5 fs for all runs, the temperature set to 300 K using a Nosé-Hoover thermostat for simulations in the NVT and NPT ensembles and the pressure set to 1 bar using a Nosé-Hoover barostat for simulations in the NPT ensemble. The characteristic time scales were set to 50 fs for the thermostat and 500 fs for the barostat. The barostat was only coupled to the surface normal direction to not artificially strain the slabs in the directions parallel to the surface.

MD simulations for structural investigations were run using a MACE model trained on the dataset obtained after the AL cycles for Cu-H<sub>2</sub>O interfaces. Interfaces were set up with an initial target water film diameter of 40 Å at a density of 1.0 g cm<sup>-3</sup>. After performing an energy minimization of the initial system and a 10 ps equilibration run in the NVT ensemble, a 200 ps simulation in the NPT ensemble was run. The equilibrium density was then determined from the last 100 ps of the NPT trajectory the same way as was done during the AL iterations. After setting up a new initial system at the equilibrium density and a subsequent energy minimization, 4 ns were simulated in the NVT ensemble. The Cu(111)-H<sub>2</sub>O interface was modelled using a  $12 \times 12$  Cu(111) slab and 1104 water molecules. For the Cu(211)-H<sub>2</sub>O interface, a  $12 \times 4$  Cu(211) slab and 1088 water molecules were used. The Cu(322)-H<sub>2</sub>O interface was represented using a  $12 \times 3$  Cu(322) slab and 1251 water molecules. Finally, a  $12 \times 2$  Cu(433) slab and 1185 water molecules were used to model the Cu(433)-H<sub>2</sub>O interface. The step densities are 1.57 nm<sup>-1</sup>, 0.93 nm<sup>-1</sup>, and 0.66 nm<sup>-1</sup> for the Cu(211), Cu(322) and Cu(433) slabs respectively.

## 3 Reference databases

Active learning runs for Cu-H<sub>2</sub>O interfaces were initiated based on the combination of three reference databases: one for bulk Cu and Cu slabs in vacuum, one for bulk water and one



containing a small number of naively set up Cu(111)–H<sub>2</sub>O interfaces. The generation process for each of these databases is described in this section.

To start the generation of bulk Cu data, the lattice parameter was determined by relaxing the primitive unit cell of fcc Cu. This yielded a value of 3.67 Å, which lies above the experimental lattice parameter of 3.61 Å.<sup>33</sup> An overestimation compared to the experimental lattice parameter of Cu using the RPBE functional has been reported before.<sup>34</sup> Subsequently, rattled structures of a  $2 \times 2 \times 2$  bulk supercell were generated following the procedure reported in ref. 35 and assuming a Debye temperature of 343 K for Cu. In total, 400 structures were generated, 200 for each temperature of 500 K and 1000 K. To provide information about structures with a non-optimized lattice parameter, bulk structures based on  $4 \times 4 \times 4$  supercells with scaled lattice parameters were added to the database: Five structures rattled at a temperature of 500 K were added for each scaling in the range of  $\{-5.0, -2.5, 2.5, 5.0\}$  % of the optimal lattice parameter.

The addition of Cu surface slabs to the database was performed in steps based on the maximum Miller index (MMI) determining the surface orientations. In the first step, only (111), (110) and (100) slabs were included. Bulk-terminated slabs for all symmetrically distinct combinations of orientations were generated based on the relaxed lattice parameter of bulk Cu using the slab generation algorithm implemented in pymatgen.<sup>36,37</sup> The number of layers were chosen such that a minimum slab thickness of 10 Å was achieved and a vacuum of 10 Å was added in surface normal direction. Slabs were relaxed by performing a geometry optimization starting from the bulk-terminated positions while keeping the unit cell fixed. Both the bulk-terminated slabs and slabs with relaxed atomic positions were added to the database. Additionally, as a starting point for active learning, slabs with MMI one perturbed by random displacements drawn from a normal distributions with standard deviations 0.03 Å, 0.05 Å and 0.10 Å respectively were added. Two sets of ten structures per surface orientation and standard deviation were generated. In the first set, only the outer layers were perturbed. In the second set, displacements were added to all atoms. In total this results in 180 structures obtained by adding random displacements. Further structures were added by adopting the active learning approach based on adversarial loss maximizations.<sup>35,38</sup> Ensembles of ten NeuralIL<sup>39</sup> models with a three layer ResNet<sup>40</sup> core structure with widths [128, 64, 32], a cutoff radius of 4.0 Å,  $n_{\text{max}} = 5$  and a two dimensional embedding for the atom types were used at each iteration. In a first batch, adversarial loss maximizations with initial displacements drawn from a Gaussian with zero mean and standard distribution of 0.1 Å were performed for 50 replicas each of  $1 \times 1$  and  $2 \times 2$  slabs with the orientations mentioned above. In a second iteration, adversarial loss maximizations were performed with initial displacements drawn with a standard distribution of 0.2 Å. DFT calculations were run for all structures obtained from the optimization procedures and the configurations added to the database. The same active learning procedure described for MMI one surfaces was then repeated for surfaces with orientations with a MMI of two, these

being (210), (211) and (221). For surfaces with a MMI of three, adversarial loss maximizations were only performed with initial displacements drawn with a standard distribution of 0.1 Å.

After discarding non-converged calculations, a database consisting of a total of 2276 structures covering bulk Cu and surfaces with symmetrically distinct orientations up to a MMI of three was obtained. As will be discussed below, AL procedures for interfaces were run using MACE, which often requires significantly less training data than NeuralIL. In order to efficiently train neural-network models it is desirable to only use as many data points as necessary to accurately model the region of interest of the potential energy hypersurface. To reduce the total amount of data points in the database, structures were randomly sampled from each of the individual batches to obtain a smaller dataset: 25 structures each from the randomly displaced  $2 \times 2 \times 2$  bulk supercells (total 50), rattled  $4 \times 4 \times 4$  bulk supercells at non-equilibrium volumes (total 20), bulk-terminated and relaxed slab structures for all symmetrically distinct surface orientations (total 26), MMI one slabs with random perturbations with standard deviation 0.05 Å (total 30), 15 structures for each orientations from adversarial loss optimizations for surfaces with MMI one and MMI two and 10 structures from adversarial loss optimizations for surfaces with MMI three (total 160). After subsampling, a database containing a total of 286 data points was obtained. Note that this database was generated using NeuralIL models and further used to train MACE models, which has been shown to be justified due to the good correlation of uncertainties based on NeuralIL and MACE ensembles.<sup>41</sup>

To make use of the work that has already gone into the construction of databases for water, we chose the 1593 structures of 64 water molecules each published by Cheng *et al.*, originally computed at the revPBE0-D3 level of theory.<sup>42</sup> The set contains five duplicate structures in the sense of having identical atomic positions, which were removed. Energies and forces were then recomputed for the remaining 1588 structures at the RPBE-D3 level of theory.

In addition to the Cu and water databases, a database of ten Cu(111)–H<sub>2</sub>O interface structures was created by packing water molecules above a  $4 \times 4$  Cu(111) slab at a density of  $1.0 \text{ g cm}^{-3}$  using GROMACS version 2024.1.<sup>43</sup> The number of inserted water molecules was chosen such that the density matched the prescribed value and the height of the water film is as close as possible to 20 Å. Such generated configurations each contain 263 atoms in total. A gap of 1.4 Å was assumed for the distance between the outer Cu layer and the region for which the water density, and thus the exact cell height and number of water molecules to insert, was calculated.

## 4 Active learning for copper–water interfaces

A reference dataset with a total of 1884 structures was assembled by combining the Cu, bulk water and Cu(111)–H<sub>2</sub>O databases. Energies and forces for the Cu database were recomputed with the VASP setup used for Cu–H<sub>2</sub>O interfaces to



avoid systematic errors arising from different DFT calculation parameters. The reference database served as a starting point for a MD-based AL cycle. At each iteration the dataset was split 90 : 10 into training and validation sets, and an ensemble of five MACE models was trained.

The AL cycle started from a  $4 \times 4$  bulk-terminated Cu(111) slab consisting of five layers. Water molecules were added above the slab such that the bulk water density equaled  $1.0 \text{ g cm}^{-3}$  and the initial water film diameter was  $20 \text{ \AA}$ . This resulted in the addition of 64 water molecules, yielding structures with 272 atoms in the unit cell. To find the equilibrium density at a given iteration, a 200 ps NPT run was performed after an energy minimization of the initial configuration and a 10 ps equilibration run in the NVT ensemble. The equilibrium volume was then calculated as the mean volume of every 50th frame in the last 100 ps of the NPT simulation run. Subsequently, five new initial configurations with the box volume set to the equilibrium volume were set up. After an initial energy minimization, NVT simulations were run for 200 ps for each of the replicas.

To avoid sampling correlated frames each NPT and NVT trajectory was divided into two 100 ps segments resulting in 12 segments total. One configuration was selected from each segment by determining the frame featuring the highest locally aggregated force uncertainty.<sup>21</sup> While the use of structure-wide aggregation is common practice,<sup>20,35,44–46</sup> it can fail to identify sub-regions featuring high-error.<sup>41</sup> This has recently been addressed by aggregating only within a defined cutoff radius around each atom. Thereby local atomic uncertainties that still correlate with the actual error are obtained.<sup>21</sup> We thus calculate local uncertainties by aggregating atomic uncertainties in a neighborhood  $N_i = \{j \in N \mid \|\mathbf{r}_i - \mathbf{r}_j\|_2 < r_{\text{agg}}\}$  of each atom  $i$  with aggregation cutoff radius  $r_{\text{agg}}$  ( $N$  denotes the set of all atoms).

$$s_{f,i}^{\text{local}} = \frac{1}{|N_i|} \sum_{j \in N_i} \frac{1}{3} \sum_{k \in \{x,y,z\}} s_{f,j}^k, \quad (1)$$

where  $s_{f,j}^k$  denotes the uncertainty of a single force component as obtained from the committee and  $k$  refers to the Cartesian axis. For the present study the aggregation cutoff radius was set to  $4 \text{ \AA}$ . Since both bulk copper and bulk water are already well represented in the initial dataset, uncertainties, and by proxy errors, are expected to be highest at the Cu–H<sub>2</sub>O interface. As visualized in Fig. 1, this is resolved well by the local uncertainties.

An AL procedure for a given interface was considered converged if the maximum local force uncertainty observed for all atoms and all timesteps in a 200 ps NPT trajectory was smaller than  $0.02 \text{ eV \AA}^{-1}$ . According to this criterion, four AL iterations were performed at a temperature of 300 K, Fig. 2, resulting in 48 Cu(111)–H<sub>2</sub>O interface structures being added to the database. After the AL iterations for the Cu(111)–H<sub>2</sub>O interface were completed, the same procedure as detailed above was repeated for the Cu(211)–H<sub>2</sub>O interface. This interface was modelled using a  $4 \times 2$  Cu(211) slab and 86 water molecules. Again, four AL cycles were needed to reach the convergence criterion, resulting in 48 Cu(211)–H<sub>2</sub>O structures that were added to the database. The Cu(322)–H<sub>2</sub>O interface was represented using a

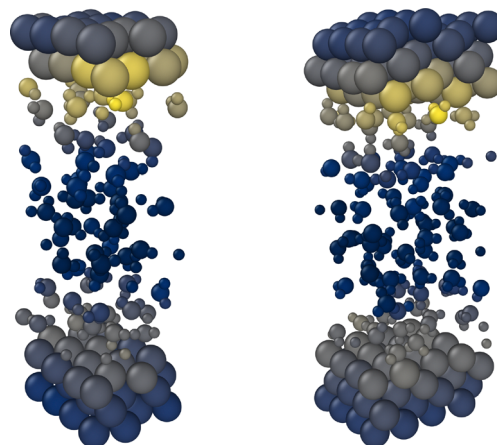


Fig. 1 Spatially resolved uncertainties for a Cu(111)–H<sub>2</sub>O (left) and a Cu(211)–H<sub>2</sub>O interface structure (right) selected from NVT trajectories in iterations 3 and 6. A color gradient from dark blue to yellow indicates the lowest and highest local uncertainties respectively. Separate color scales are used for the individual structures.

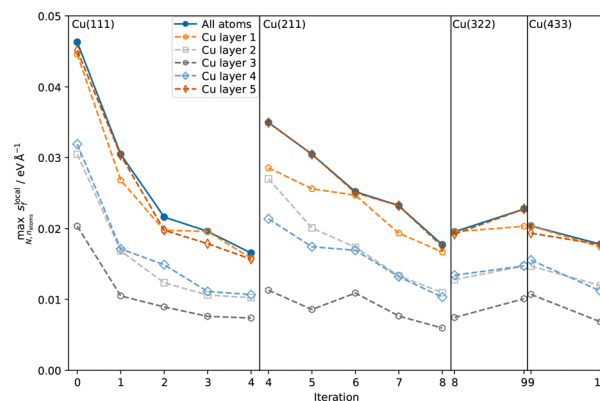


Fig. 2 Maximum local force uncertainties observed for the NPT trajectories at each active learning iteration. The solid curves represent the maximum over all atoms, including the water molecules. The dashed curves show the maximum local uncertainty of the Cu atoms resolved layer-wise (inner layer index 3, sublayers indices 2 and 4, interfacial layers indices 1 and 5).

$4 \times 1$  Cu(322) slab and 72 water molecules. A  $4 \times 1$  Cu(433) slab and 92 water molecules were used to model the Cu(433)–H<sub>2</sub>O interface for AL. As will be discussed below, the observed uncertainties for the Cu(322)–H<sub>2</sub>O and Cu(433)–H<sub>2</sub>O systems were already low in the first AL iteration, suggesting that adding additional reference data for these structures can be omitted. However, in the present study we added twelve data points from the first AL cycle for the Cu(322)–H<sub>2</sub>O interface after completing the Cu(211)–H<sub>2</sub>O cycles. Similarly, twelve Cu(433)–H<sub>2</sub>O structures obtained from one AL cycle were added after finishing the procedure for the Cu(322)–H<sub>2</sub>O interface. In total this yielded 120 interface structures that were added to the database over ten subsequent AL cycles. Training a MACE model on this database yields an RMSE of  $0.76 \text{ meV atom}^{-1}$  for the energies and  $20.42 \text{ meV \AA}^{-1}$  for the forces when evaluating errors on the whole training set.





In all iterations the highest uncertainties are observed directly at the interface, while both in the bulk Cu and bulk H<sub>2</sub>O regions, uncertainties are low, see Fig. 1. The difference in uncertainties for the interface and bulk regions is further apparent in Fig. 2. At all AL iterations the atomic uncertainties are highest for an outer layer Cu atom or an atom in a water molecule in the contact layer. Since the initial database contains no structures sampled from MD trajectories, high maximum local uncertainties are observed when starting the AL process. After converging for the Cu(111)–H<sub>2</sub>O interface, a jump is observed when moving on to the Cu(211) surface, as no data for stepped interfaces is yet present in the database. However, even for the first Cu(211) iteration, uncertainties for Cu atoms in the bulk layer are already low and no systematic reduction is observed in further AL iterations. The databases used to train the models used in the AL iterations for Cu(322)–H<sub>2</sub>O and Cu(433)–H<sub>2</sub>O interfaces did not contain any training structures of the respective interfaces they were applied on. Still, uncertainties already satisfy the specified cutoff criterion in the first iteration on the respective surfaces. A small increase local force uncertainty is observed for the Cu(322)–H<sub>2</sub>O interface after adding data, but was not investigated further and no additional AL cycles were run.

## 5 Structure at the interface

Water density profiles were obtained from the 4 ns MD simulations of Cu–H<sub>2</sub>O interfaces using the MACE model trained on the final dataset. They are shown in Fig. 3, along with snapshots of representative structures of the water contact layer. The Cu(111)–H<sub>2</sub>O interface exhibits a distinct double-peak structure with maxima of 3.62 g cm<sup>−3</sup> and 1.75 g cm<sup>−3</sup> around 2.99 Å and 2.41 Å above the outermost Cu layer, in agreement with previous studies.<sup>10,11</sup> A similar ordering is also found for *e.g.* Pt(111)<sup>11–16</sup> but not for the more noble Au(111).<sup>11,15,16</sup> The tendency towards ordering of interfacial water molecules within the water contact layer indicates chemisorbed water. This is illustrated in the snapshot in Fig. 3 where a water molecule at approximately 2.4 Å with the oxygen atom oriented towards the Cu slab.

Separating the density profiles for the individual atom types, as shown in Fig. 4, also points towards the first density maximum representing chemisorbed water, as only a single peak is observed for the hydrogen atoms. This peak represents both chemisorbed water molecules, with hydrogen atoms pointing away from the surface, and water molecules oriented with the hydrogen atoms towards the surface, which are mapped to the global maximum of the density curve for water, Fig. 3. This is additionally evident from the density profiles weighted with  $\cos \phi$ , the cosine of the angle between the dipole vector and the surface normal, as visualized in Fig. 5. A positive first peak shows the water molecules closest to the slab to be oriented primarily with the oxygen towards the surface, while the negative peak at 2.99 Å indicates the opposite orientation for the corresponding water molecules. This is also similar to

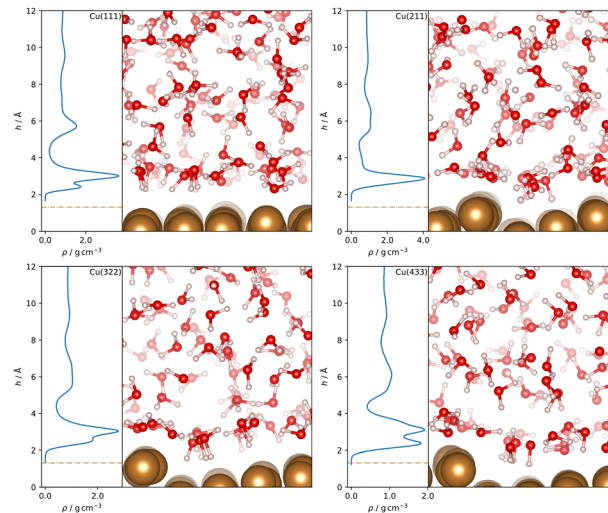


Fig. 3 Water density profiles as a function of the distance  $h$  from the topmost surface layer for Cu–H<sub>2</sub>O interfaces with different Cu surface orientations. The density profiles were obtained by averaging the distances of the centers of mass of the water molecules from the surface over the trajectory after discarding the first 10 ps. The position  $h = 0$  Å of the topmost surface layer in surface normal direction was computed as the mean  $z$ -position of all Cu atoms in the top layer at each timestep. Visualized frames were obtained from the trajectories by selecting the snapshot featuring the lowest instantaneous hydrogen distance from the surface.

the Pt(111)–H<sub>2</sub>O interface, for which a compensation of the dipoles of chemisorbed water by the outer layer is reported.<sup>14</sup>

A double peak structure is not observed for the Cu(211)–H<sub>2</sub>O interface, where only a strong single peak is present at a distance of 2.86 Å from the surface. Qualitatively, this is understood as a consequence of the step edge combined with a short plateau, since water molecules primarily adsorb on the ridge atoms due to their lower coordination. As illustrated by the Cu(211)–H<sub>2</sub>O interface snapshot in Fig. 3, water primarily adsorbing at the ridge sites induces a strict H-down orientation of water molecules at the adjacent crevice to facilitate hydrogen bonding to the

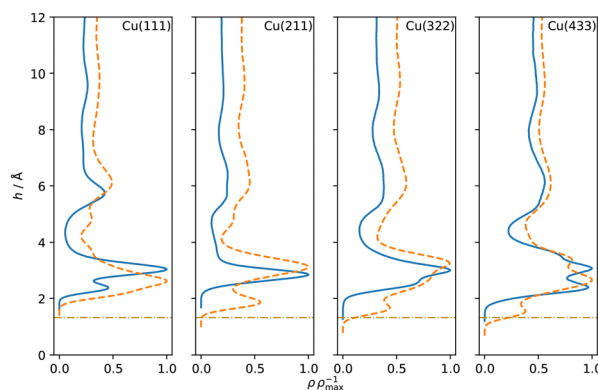


Fig. 4 Density profiles of oxygen (solid blue) and hydrogen (dashed orange) as a function of the distance  $h$  from the topmost surface layer for Cu–H<sub>2</sub>O interfaces with different Cu surface orientations. The curves were normalized with respect to their global maximum. Dash-dotted lines indicate the covalent radius of Cu.



molecules adsorbed at the ridge sites. This is reflected in the density profile for hydrogen (Fig. 4), which now features two peaks in contrast to the Cu(111)–H<sub>2</sub>O interface. The first of these maxima corresponds to water molecules at the step crevices. For all stepped interfaces, hydrogen densities are non-zero in the region bounded by the covalent radius of Cu, which is a consequence of both the definition of the reference height as the instantaneous mean of the heights of the Cu atoms in the top layer in surface normal direction and the aforementioned arrangement around the step. The local structure at the steps furthermore explains the negative first peak in the angular weighted density profiles for all stepped interfaces (Fig. 5). In contrast to results reported for the Pt(211)–H<sub>2</sub>O interface,<sup>14</sup> the Cu(211)–H<sub>2</sub>O interface exhibits a small positive peak, attributed to the water molecules at the ridge sites.

Decreasing the step density leads to the formation of a second peak in the contact layer, Fig. 3. Specifically, a weak shoulder at the Cu(322)–H<sub>2</sub>O interface and a distinct second peak at the Cu(433)–H<sub>2</sub>O interface are observed. However, the origin of the two peaks is different than for the flat Cu(111) surface, since the structure is still strongly influenced by the

undercoordinated ridge sites. The depth of the first minimum in the cosine-weighted density curves, Fig. 5, is lowered with increasing step density. As can be seen in the snapshot of the Cu(433)–H<sub>2</sub>O interface in Fig. 3 and will be discussed in more detail below, a sufficiently long plateau allows for water molecules to orient with the oxygen towards the surface, similar as at the Cu(111)–H<sub>2</sub>O interface.

Density curves such as those shown in Fig. 3–5 can also be obtained by AIMD.<sup>11,15</sup> The comparatively longer time-scales that become accessible through MLFF-backed MD make it possible to obtain reliable statics for local resolution. Here we investigate interfacial water structure in the surface parallel directions by calculating the oxygen–oxygen 2D pair correlation functions (2D PCF) as given by<sup>16,47</sup>

$$g_{\text{oo}}(\Delta\mathbf{r}) = \frac{1}{n_t} \sum_t \frac{1}{\rho_{\text{ads}}(t)(|N_{\text{ads}}(t)| - 1)} \times \sum_{i \neq j}^{N_{\text{ads}}(t)} \delta(\Delta\mathbf{r} - \Delta\mathbf{r}_{ij}(t)), \quad (2)$$

where  $\Delta\mathbf{r}_{ij}$  is a two-dimensional vector that denotes the pairwise distance in the directions parallel to the surface,  $\rho_{\text{ads}}(t)$  the surface number density of oxygen atoms in the contact layer at time  $t$ ,  $N_{\text{ads}}(t)$  the set of oxygen atoms in the contact layer at time  $t$  and  $n_t$  the total number of timesteps. The 2D PCFs for both bulk water and the different interfaces are visualized in Fig. 6. On the flat Cu(111) surface, a similar structure as for the Pt(111) surface is found.<sup>16</sup> A ring of large  $g_{\text{oo}}(\Delta\mathbf{r})$  values indicating the first solvation shell is observed, followed by weaker peaks arising from the second and third solvation shells. The 2D PCF does however not converge to a constant value of 1.0 with increasing distance, as is the asymptotic limit for of bulk water, but exhibits peaks induced by the underlying Cu atoms. Notably, the third solvation shell matches the distance of 6.86 Å of the fourth nearest in-plane neighbor of a Cu atom in the top layer and therefore shows pronounced peaks at these locations. As evident from Panel Q of Fig. 1 in ref. 16, this is in contrast to Pt, for which such a lattice parameter match is not observed. The introduction of a step

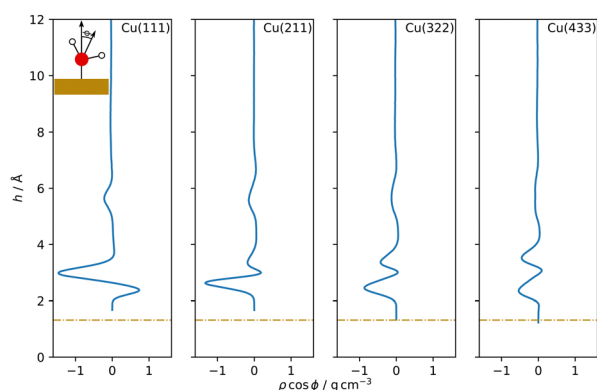


Fig. 5 Dipole orientation distribution profiles as a function of the distance  $h$  from the topmost surface layer for Cu–H<sub>2</sub>O interfaces with different Cu surface orientations.  $\cos \phi$  is the cosine of the angle between the dipole vector, calculated as the bisector of the water molecule orientated from the oxygen to the mid point of the two hydrogen atoms, and the surface normal (see insert). Dash-dotted lines indicate the covalent radius of Cu.

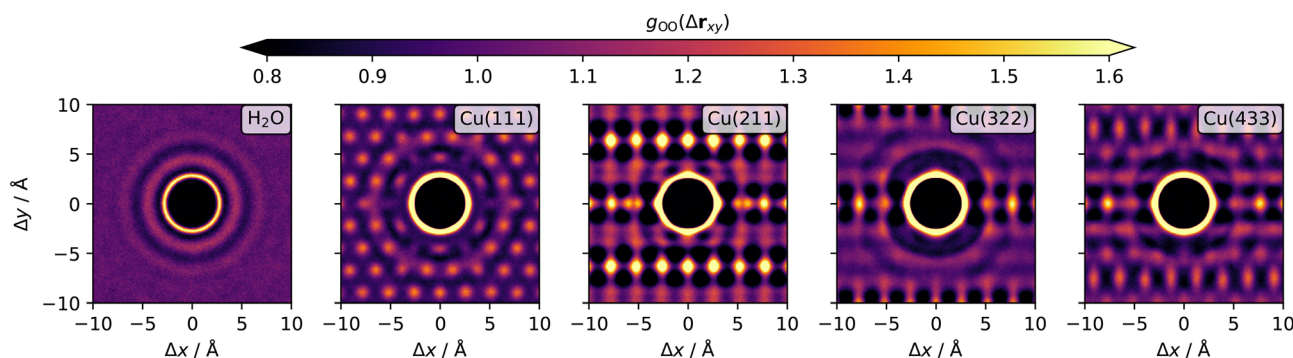


Fig. 6 2D oxygen–oxygen pair correlation functions for bulk water or water in the contact layers of the interfaces respectively. The maximum distance  $h$  defining the contact layer was chosen as the local minimum of the density profiles (Fig. 3) following the the global maximum. The 2D-PCF for bulk water is computed based on a slab with thickness 5 Å in  $z$ -direction.



on the surface leads to pronounced changes in the shapes of the 2D PCFs. On the Cu(211) surface, which features the highest step density of all investigated surfaces, a highly anisotropic profile is observed. The 2D PCF is mainly dominated by the steps, even in the short-range regime, with also the first solvation shell being distorted. Strong peaks are observed at the ridge Cu atoms even at longer distances indicating the steps as the preferred adsorption sites. Anisotropic profiles are also observed for the Cu(322)-H<sub>2</sub>O and Cu(433)-H<sub>2</sub>O interfaces, but the effects are less pronounced due to the decreasing step densities. The 2D oxygen-oxygen PCF for the water molecules in the contact layer on the Cu(433) surface already shows that the pattern observed on the flat Cu(111) surface is partially recovered due the longer (111)-like plateau of the Cu(433) surface.

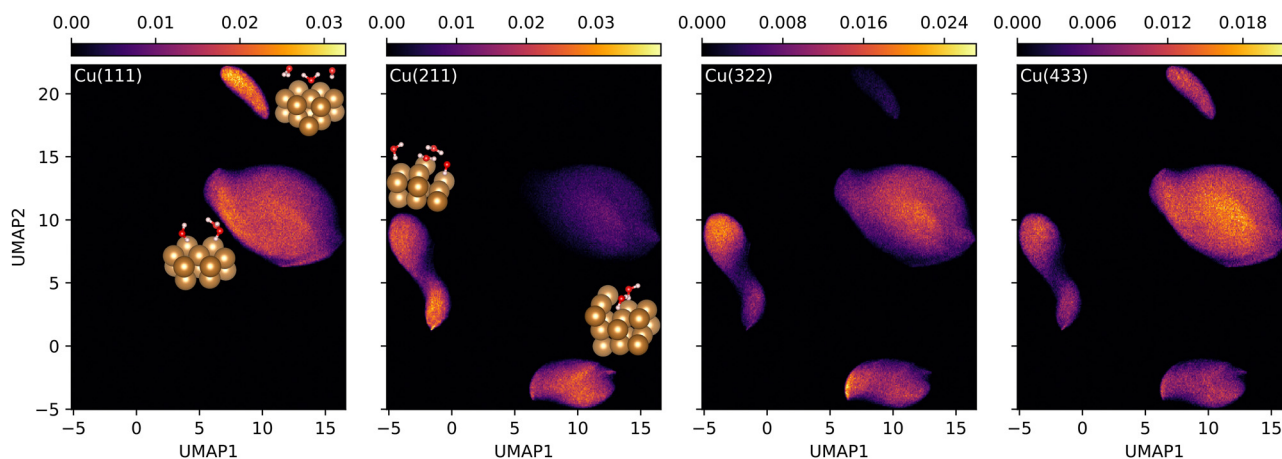
To illustrate how decreasing the step density recovers structural features observed for the flat Cu(111)-H<sub>2</sub>O interface we investigate the local environments of the atoms in the contact layer sampled during the MD simulations. Using spherical Bessel descriptors<sup>39,48</sup> to represent the local environments of Cu atoms in the top layer, we encode the local environment in a rotationally invariant manner. Snapshots were taken at 1 ps intervals from the simulations for each interface. The combined set of descriptors is visualised using the dimensionality reduction technique UMAP<sup>49</sup> in Fig. 7.

At the Cu(111)-H<sub>2</sub>O interface, two distinct types of local Cu environments are observed, corresponding to the two types of water orientations producing the opposite peaks observed in Fig. 5. The smaller patch in Fig. 7 corresponds to chemisorbed water with hydrogen atoms oriented away from the surface (positive  $\cos \phi$ ) and shorter Cu-O distances. The larger patch corresponds to environments where the hydrogen atoms are oriented towards the surface. Environments in the latter region are also observed at the Cu(211)-H<sub>2</sub>O interface. On Cu(211), chemisorption primarily takes place at the undercoordinated ridge sites, while water molecules in the crevice of the step are almost exclusively oriented with the hydrogen atoms oriented

towards the surface (Fig. 3), forming distinct UMAP patches, Fig. 7. The high step density of the Cu(211) surface geometrically restricts interfacial water molecules from orienting in a (111)-like chemisorption geometry. This results in the absence of the corresponding UMAP patch, in agreement with the behavior observed in Fig. 3–5. Reducing the step density exposes longer plateaus, thereby alleviating these restrictions and making Cu(111)-like chemisorption geometries possible, as seen for the Cu(322) and Cu(433) panels of Fig. 7. The Cu(322) surface, featuring a shorter plateau than Cu(433), exhibits a sparser population of the Cu(111)-like chemisorption patch. The Cu(433) surface clearly shows this, consistent with the oxygen atoms oriented towards the surface in the corresponding snapshot Fig. 3.

Interestingly, no new regions of configuration space are visited during the simulations of the Cu(322)-H<sub>2</sub>O and Cu(433)-H<sub>2</sub>O interfaces compared to Cu(111)-H<sub>2</sub>O and Cu(211)-H<sub>2</sub>O. This is in accordance with the already low uncertainties in the first AL cycles on the higher-index surfaces, Fig. 2. We therefore also expect interfaces to less idealized surfaces, which can be conceptualized as being composed of individual elements present in simpler model systems, to be predicted well by an MLFF trained only on these model interfaces.

A more fine-grained analysis of the environments is achieved by separating the data according to rows parallel to the step edge, as shown for the Cu(433)-H<sub>2</sub>O interface in Fig. 8. The environments of the crevice and ridge Cu atoms are visualized in panels a and g respectively and form two separate clusters of environment types as discussed above. Due to the geometric restriction of the water molecules in the direct vicinity of the step, where the ridge atoms are the preferred sites for chemisorption, no environments indicating (111)-like chemisorption geometries are found for Cu atoms in the rows adjacent to the crevice and ridge rows (panels b and f respectively). As discussed in the context of Fig. 6, the interfacial structure of water approaches that of the Cu(111)-H<sub>2</sub>O interface



**Fig. 7** Two-dimensional density histograms of the UMAP representations of the spherical Bessel descriptors of the top layer Cu environments. A single UMAP has been computed using 40 neighbors, a minimum distance of 0.2 and the Euclidean distance metric for the whole set of data and was subsequently split into subsets representing the individual interfaces, for which individual histograms were calculated. Representative environments taken from snapshots illustrate the types of substructures constituting the individual clusters.





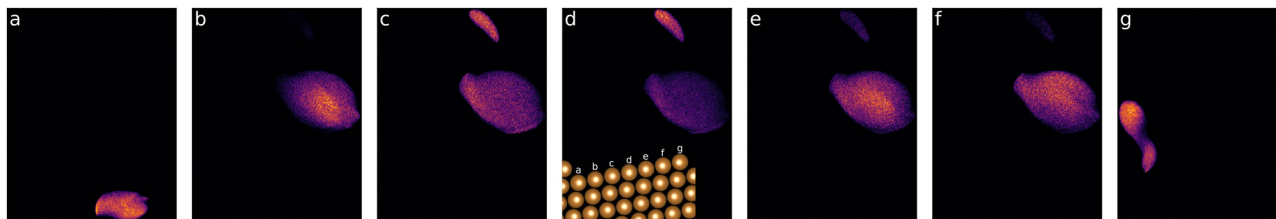


Fig. 8 Two-dimensional density histogram of the UMAP representations of the spherical Bessel descriptors of the top layer Cu atom environments of the Cu(433)–H<sub>2</sub>O interface. Panels (a–g) visualize the environments observed at the individual rows, which are enumerated as shown in the inset of panel d. The individual histograms sum to the values visualized in Fig. 7, individual color scales were used for visibility.

for high-index Cu( $n + 1, n$ ) surfaces. This is further supported by the environments of the top layer Cu atoms located in rows further from the step being clustered similarly to those of the Cu(111)–H<sub>2</sub>O interface, as evident in panels c–e in Fig. 8.

Characterizing the local structure of water is a long-standing challenge, with different local order parameters often giving ambiguous results.<sup>50,51</sup> The UMAP analysis highlights how a data-driven approach based on local invariant descriptors, enables a clear classification of the 32 million geometries sampled during the MD simulations without relying on, potentially biased, a-priori intuition of which geometric features to probe.

## Conclusions

In the present study we have demonstrated that high step densities lead to a fundamentally different structuring of interfacial water compared to the ideal, flat surface and that lowering the step density gradually recovers behavior similar to the interface to the flat surface. We recover previously reported results for the Cu(111)–H<sub>2</sub>O interface, such as the double peak structure in the density and the Cu ontop sites being preferred. On all stepped surfaces, we observe chemisorption taking place mainly on the step sites, as opposed to the terrace sites. Further, we connect observed results and a data-driven approach to identify common types of atomic environments across the differently oriented surfaces. In order to obtain these results we both make use of existing databases and also developed a locally sensitive active learning workflow for Cu–H<sub>2</sub>O interfaces that makes use of the high data efficiency of equivariant MLFFs. This minimizes the amount of computationally demanding reference calculations, in this case even removing the need for AIMD trajectories entirely. The workflow allows for systematic extension of the database to investigate a wider range of systems. By identifying the relevant local environments and training on according model systems, we conceive this approach to be suitable for modelling interfaces to less idealized surfaces. In order to achieve more realistic models, some aspects remain to be addressed. For the present system in particular, including different oxidization states of copper is essential to move towards a more realistic system.<sup>52,53</sup> Furthermore, no water dissociation is observed in the present study and no dissociated species are considered. These aspects can already be tackled with currently available MLFFs,<sup>54</sup> but may require the use of enhanced sampling techniques to be simulated efficiently.<sup>55</sup> Additionally, the treatment of long-range interactions

and the inclusion of electric fields in MLFFs is still an open research question and obviously of vital importance for modelling electrochemical interfaces.

## Author contributions

Johannes Schörghuber: conceptualization, methodology, software, validation, investigation, visualization, writing – original draft Nina Bučková: Investigation Esther Heid: conceptualization, methodology, supervision, writing – review & editing Georg K. H. Madsen: conceptualization, methodology, supervision, writing – review & editing, project administration.

## Data availability

The generated dataset, models, input files, initial structures for MD simulation runs, and a Jupyter notebook for the presented trajectory analysis are available on Zenodo at <https://doi.org/10.5281/zenodo.14755563>.

## Conflicts of interest

There are no conflicts to declare.

## Acknowledgements

The authors acknowledge Ralf Wanzenböck for his helpful comments and feedback regarding the visualization of the results. This research was funded in part by the Austrian Science Fund (FWF) 10.55776/COE5 and 10.55776/F81. For open access purposes, the authors have applied a CC BY public copyright license to any author accepted manuscript version arising from this submission. The computational results presented have been achieved in part using the Vienna Scientific Cluster (VSC).

## Notes and references

- O. Björneholm, M. H. Hansen, A. Hodgson, L.-M. Liu, D. T. Limmer, A. Michaelides, P. Pedevilla, J. Rossmeisl, H. Shen, G. Tocci, E. Tyrode, M.-M. Walz, J. Werner and H. Bluhm, *Chem. Rev.*, 2016, **116**, 7698–7726.
- Z. W. Seh, J. Kibsgaard, C. F. Dickens, I. Chorkendorff, J. K. Nørskov and T. F. Jaramillo, *Science*, 2017, **355**, eaad4998.





- 3 Y. Hori, I. Takahashi, O. Koga and N. Hoshi, *J. Phys. Chem. B*, 2002, **106**, 15–17.
- 4 Y. Hori, I. Takahashi, O. Koga and N. Hoshi, *J. Mol. Catal. A: Chem.*, 2003, **199**, 39–47.
- 5 S. Nitopi, E. Bertheussen, S. B. Scott, X. Liu, A. K. Engstfeld, S. Horch, B. Seger, I. E. L. Stephens, K. Chan, C. Hahn, J. K. Nørskov, T. F. Jaramillo and I. Chorkendorff, *Chem. Rev.*, 2019, **119**, 7610–7672.
- 6 A. Auer, F. J. Sarabia, D. Winkler, C. Griesser, V. Climent, J. M. Feliu and J. Kunze-Liebhäuser, *ACS Catal.*, 2021, **11**, 10324–10332.
- 7 D. Winkler, M. Leitner, A. Auer and J. Kunze-Liebhäuser, *ACS Catal.*, 2024, **14**, 1098–1106.
- 8 J. Carrasco, A. Hodgson and A. Michaelides, *Nat. Mater.*, 2012, **11**, 667–674.
- 9 A. Groß and S. Sakong, *Chem. Rev.*, 2022, **122**, 10746–10776.
- 10 S. K. Natarajan and J. Behler, *Phys. Chem. Chem. Phys.*, 2016, **18**, 28704–28725.
- 11 H. H. Heenen, J. A. Gauthier, H. H. Kristoffersen, T. Ludwig and K. Chan, *J. Chem. Phys.*, 2020, **152**, 144703.
- 12 J. Lan, V. V. Rybkin and M. Iannuzzi, *J. Phys. Chem. Lett.*, 2020, **11**, 3724–3730.
- 13 A. E. G. Mikkelsen, J. Schiøtz, T. Vegge and K. W. Jacobsen, *J. Chem. Phys.*, 2021, **155**, 224701.
- 14 A. Chen, J.-B. Le, Y. Kuang and J. Cheng, *J. Chem. Phys.*, 2022, **157**, 094702.
- 15 S. Liu, S. Vijay, M. Xu, A. Cao, H. Prats, G. Kastlunger, H. H. Heenen and N. Govindarajan, *J. Chem. Phys.*, 2023, **159**, 084702.
- 16 J. Gäding, V. Della Balda, J. Lan, J. Konrad, M. Iannuzzi, R. H. Meißner and G. Tocci, *Proc. Natl. Acad. Sci. U. S. A.*, 2024, **121**, e2407877121.
- 17 S. K. Natarajan and J. Behler, *J. Phys. Chem. C*, 2017, **121**, 4368–4383.
- 18 R. Arrigo, *Curr. Opin. Green Sustainable Chem.*, 2022, **34**, 100601.
- 19 J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev and A. E. Roitberg, *J. Chem. Phys.*, 2018, **148**, 241733.
- 20 P. M. Piaggi, A. Selloni, A. Z. Panagiotopoulos, R. Car and P. G. Debenedetti, *Faraday Discuss.*, 2024, **249**, 98–113.
- 21 E. Heid, J. Schörghuber, R. Wanzénböck and G. K. H. Madsen, *J. Chem. Inf. Model.*, 2024, **64**, 6377–6387.
- 22 G. Kresse and J. Furthmüller, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1996, **54**, 11169–11186.
- 23 B. Hammer, L. B. Hansen and J. K. Nørskov, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 1999, **59**, 7413–7421.
- 24 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.
- 25 T. Morawietz, A. Singraber, C. Dellago and J. Behler, *Proc. Natl. Acad. Sci. U. S. A.*, 2016, **113**, 8368–8373.
- 26 P. Montero de Híjes, C. Dellago, R. Jinnouchi, B. Schmiedmayer and G. Kresse, *J. Chem. Phys.*, 2024, **160**, 114107.
- 27 E. R. McNellis, J. Meyer and K. Reuter, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2009, **80**, 205414.
- 28 G. Mercurio, *Phys. Rev. Lett.*, 2010, **104**, 036102.
- 29 K. Tonigold and A. Groß, *J. Comput. Chem.*, 2012, **33**, 695–701.
- 30 I. Batatia, D. P. Kovacs, G. Simm, C. Ortner and G. Csányi, *Adv. Neural. Inf. Process. Syst.*, 2022, **35**, 11423–11436.
- 31 A. P. Thompson, H. M. Aktulga, R. Berger, D. S. Bolintineanu, W. M. Brown, P. S. Crozier, P. J. in't Veld, A. Kohlmeyer, S. G. Moore, T. D. Nguyen, R. Shan, M. J. Stevens, J. Tranchida, C. Trott and S. J. Plimpton, *Comput. Phys. Commun.*, 2022, **271**, 108171.
- 32 S. J. Reddi, S. Kale and S. Kumar, *arXiv*, 2019, preprint, arXiv:1904.09237, DOI: [10.48550/arXiv.1904.09237](https://doi.org/10.48550/arXiv.1904.09237).
- 33 M. E. Straumanis and L. S. Yu, *Acta Crystallogr., Sect. A*, 1969, **25**, 676–682.
- 34 S. Luo, Y. Zhao and D. G. Truhlar, *J. Phys. Chem. Lett.*, 2012, **3**, 2975–2979.
- 35 J. Carrete, H. Montes-Campos, R. Wanzénböck, E. Heid and G. K. H. Madsen, *J. Chem. Phys.*, 2023, **158**, 204801.
- 36 R. Tran, Z. Xu, B. Radhakrishnan, D. Winston, W. Sun, K. A. Persson and S. P. Ong, *Sci. Data*, 2016, **3**, 160080.
- 37 S. P. Ong, W. D. Richards, A. Jain, G. Hautier, M. Kocher, S. Cholia, D. Gunter, V. L. Chevrier, K. A. Persson and G. Ceder, *Comput. Mater. Sci.*, 2013, **68**, 314–319.
- 38 D. Schwalbe-Koda, A. R. Tan and R. Gómez-Bombarelli, *Nat. Commun.*, 2021, **12**, 5104.
- 39 H. Montes-Campos, J. Carrete, S. Bichelmaier, L. M. Varela and G. K. H. Madsen, *J. Chem. Inf. Model.*, 2022, **62**, 88–101.
- 40 K. He, X. Zhang, S. Ren and J. Sun, *arXiv*, 2015, preprint, arXiv:1512.03385, DOI: [10.48550/arXiv.1512.03385](https://doi.org/10.48550/arXiv.1512.03385).
- 41 R. Wanzénböck, E. Heid, M. Riva, G. Franceschi, A. M. Imre, J. Carrete, U. Diebold and G. K. H. Madsen, *Digit. Discovery*, 2024, **3**, 2137–2145.
- 42 B. Cheng, E. A. Engel, J. Behler, C. Dellago and M. Ceriotti, *Proc. Natl. Acad. Sci. U. S. A.*, 2019, **116**, 1110–1115.
- 43 M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1–2**, 19–25.
- 44 L. Kahle and F. Zipoli, *Phys. Rev. E*, 2022, **105**, 015311.
- 45 P. B. Jørgensen, J. Busk, O. Winther and M. N. Schmidt, *arXiv*, 2023, preprint, arXiv:2312.04174, DOI: [10.48550/arXiv.2312.04174](https://doi.org/10.48550/arXiv.2312.04174).
- 46 M. Kellner and M. Ceriotti, *Mach. Learn.: Sci. Technol.*, 2024, **5**, 035006.
- 47 E. P. Bernard and W. Krauth, *Phys. Rev. Lett.*, 2011, **107**, 155704.
- 48 E. Kocer, J. K. Mason and H. Erturk, *AIP Adv.*, 2020, **10**, 015021.
- 49 L. McInnes, J. Healy, N. Saul and L. Großberger, *J. Open Source Softw.*, 2018, **3**, 861.
- 50 E. Duboué-Dijon and D. Laage, *J. Phys. Chem. B*, 2015, **119**, 8406–8418.
- 51 H. Doi, K. Z. Takahashi and T. Aoyagi, *J. Chem. Phys.*, 2021, **154**, 164505.
- 52 M. Favaro, H. Xiao, T. Cheng, W. A. Goddard, J. Yano and E. J. Crumlin, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 6706–6711.
- 53 S. H. Lee, J. C. Lin, M. Farmand, A. T. Landers, J. T. Feaster, J. E. Avilés Acosta, J. W. Beeman, Y. Ye, J. Yano, A. Mehta, R. C. Davis, T. F. Jaramillo, C. Hahn and W. S. Drisdell, *J. Am. Chem. Soc.*, 2021, **143**, 588–592.
- 54 V. Quaranta, J. Behler and M. Hellström, *J. Phys. Chem. C*, 2019, **123**, 1293–1304.
- 55 Z. Zeng, F. Wodaczek, K. Liu, F. Stein, J. Hutter, J. Chen and B. Cheng, *Nat. Commun.*, 2023, **14**, 6131.

