

# Learning a reactive potential for silica-water through uncertainty attribution

Received: 31 July 2023

Accepted: 3 July 2024

Published online: 17 July 2024

Swagata Roy<sup>1</sup>, Johannes P. Dürholt<sup>2</sup>, Thomas S. Asche<sup>2</sup>, Federico Zipoli<sup>3</sup> & Rafael Gómez-Bombarelli<sup>1</sup>✉

The reactivity of silicates in aqueous solution is relevant to various chemistries ranging from silicate minerals in geology, to the C-S-H phase in cement, nanoporous zeolite catalysts, or highly porous precipitated silica. While simulations of chemical reactions can provide insight at the molecular level, balancing accuracy and scale in reactive simulations in the condensed phase is a challenge. Here, we demonstrate how a machine-learning reactive interatomic potential trained on PaiNN architecture can accurately capture silicate-water reactivity. The model was trained on a dataset comprising 400,000 energies and forces of molecular clusters at the  $\omega$ B97X-D3/def2-TZVP level. To ensure the robustness of the model, we introduce a general active learning strategy based on the attribution of the model uncertainty, that automatically isolates uncertain regions of bulk simulations to be calculated as small-sized clusters. The potential reproduces static and dynamic properties of liquid water and solid crystalline silicates, despite having been trained exclusively on cluster data. Furthermore, we utilize enhanced sampling simulations to recover the self-ionization reactivity of water accurately, and the acidity of silicate oligomers, and lastly study the silicate dimerization reaction in a water solution at neutral conditions and find that the reaction occurs through a flanking mechanism.

One of the most abundant materials on the planet, silica<sup>1</sup> and its many polymorphs are used in applications like catalysts<sup>2</sup>, pharmaceuticals<sup>3</sup>, nanotechnology<sup>4</sup>, and additives<sup>5</sup>. Amorphous silica, in particular, has dozens of commercial applications and the global precipitated silica market is expected to reach US\$3.49 billion by 2023<sup>6</sup>.

Precipitated silica or silica gel is produced by the polymerization of water-soluble silicate salts in the presence of mineral acids, resulting in a precipitate or gel<sup>7</sup>. Similarly, zeolites are crystallized from a precursor gel in hydrothermal conditions<sup>8</sup>. While various experimental techniques can characterize nanosized crystallites and precipitates, they cannot resolve the individual silicate aggregates or the network of silicon connectivity<sup>9,10</sup>.

Molecular simulations of reactivity in the aqueous condensed phase can bridge this knowledge gap on silica condensation. The key

challenge is describing covalent bond breaking and formation accurately while being able to simulate large length- and time-scales. Numerous classical interatomic potentials for anhydrous silica have been generated to simulate silica glasses or other polymorphs<sup>11,12</sup>, including the well-known BKS<sup>11</sup> pair potential for  $\alpha$ -quartz and silica glasses, and more recent reparametrizations for nanoclusters and amorphous silica<sup>12,13</sup>. These models, along with ReaxFF models parameterized for water-silica<sup>14,15</sup> cannot accurately replicate the condensation reactions in aqueous solution. The recent advances in machine learning (ML) techniques and the expansion of neural networks (NN) and automatic differentiation routines<sup>16,17</sup> have provided a branch between physical sciences and statistical learning. NN-based interatomic potentials (NNIP) built on a variety of molecular representations retain the accuracy of the ab initio training data and can be

<sup>1</sup>Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA. <sup>2</sup>Evonik Operations GmbH, Essen, North Rhine-Westphalia, Germany. <sup>3</sup>IBM Research Europe, Säumerstrasse 4, 8803 Rüschlikon, Switzerland. ✉e-mail: [rafagb@mit.edu](mailto:rafagb@mit.edu)

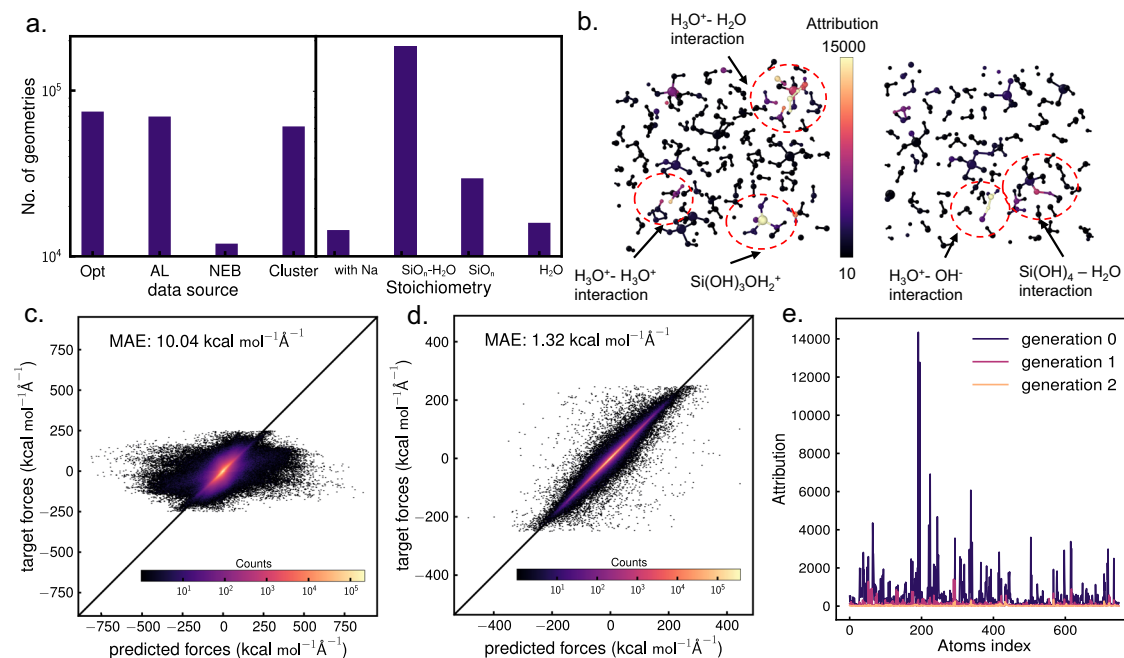


**Fig. 1 | Training dataset and active learning method.** **a** Distribution of data based on data source - Cluster: silicate and water clusters based on random position guesses. Opt: molecule clusters optimized at a lower level of Density Functional Theory (DFT) or Semiempirical Extended Tight-Binding (xtb). AL: data produced through active learning and NEB: transition states using Nudged elastic band theory (NEB); and based on stoichiometry -  $\text{SiO}_n$ : silicate clusters in a vacuum.  $\text{SiO}_n\text{-H}_2\text{O}$ : silicate clusters surrounded by water molecules,  $\text{H}_2\text{O}$ : water molecules only and with Na: molecular clusters of silicate or water with Na in it. **b** Two examples of attribution. The color bar is based on attributes. The red circles show the atomic environments added for active learning with atoms with high attributes at the

center. The  $\text{H}_3\text{O}^+ \text{--} \text{H}_2\text{O}$ ,  $\text{H}_3\text{O}^+ \text{--} \text{H}_3\text{O}^+$ ,  $\text{H}_3\text{O}^+ \text{--} \text{OH}^-$ ,  $\text{Si(OH)}_4 \text{--} \text{H}_2\text{O}$  interactions and  $\text{Si(OH)}_3\text{OH}_2^+$  complex lead to high attributes as they were not present in the training data. Parity plot of forces of a set of molecular clusters obtained through attribution-based active learning by **c** first generation NN-based inter-atomic potential (NNIP) and **d** final generation NNIP. The first-generation NNIP performs poorly on them while the final generation NNIP performs better. MAE: Mean absolute error. **e** Consecutive generation-wise improvement of attributions on a held-out condensed phase system. Attribution-based active learning thus improves our model. Source data are provided as a Source Data file.

executed at a computational cost much lower than first principle calculations. Several NNIP have been constructed for pure water<sup>18–20</sup>, amorphous and liquid silica<sup>21,22</sup> and water-filled aluminosilicate zeolites<sup>23</sup>. However, no NNIP exists that models reactive silica-water interactions in acids or bases, which are required to understand the fundamental steps of silica precipitation, crystallization, and gel formation.

The initial step of silicate polymerization or oligomerization is dimer formation, whose mechanism depends on pH and temperature<sup>24,25</sup>. In neutral conditions or in a slightly acidic medium, the dimerization can occur through neutral monomers<sup>24</sup>. Aqueous reactions such as these are hard to understand at an atomic level, given the need to simulate the role of explicit solvent in the condensed phase, various bond breaking and formation, and proton-transfers in the solvent and reactants. Quantum chemical simulations can provide accurate energies for systems up to hundreds of atoms but are too costly to address free energy questions, that arise in the presence of full explicit solvation. A couple of possible molecular mechanisms have been reported for the neutral dimerization reaction<sup>26</sup>. One involves a  $S_N2$ -backside reaction<sup>27</sup> whereas the other involves a lateral flank-side attack<sup>28</sup>.

In this study, we trained a reactive water-silica NNIP through an active learning loop. Our ground truth data consists of orbital-based hybrid DFT calculations on molecular clusters of various sizes. The trained NNIP generalizes to condensed phase simulation of thousands of atoms, both aqueous and solid silica. Since production simulations contain many more atoms than can be simulated with DFT we introduce an uncertainty attribution technique that builds on adversarial attacks on uncertainty<sup>29</sup> to locate clusters of uncertainty within large simulation boxes for annotating with DFT. The NNIP was validated

successfully against experimental properties of liquid water, and silica, both physical (diffusivity, structure, vibrational) and chemical (dissociation constants). We then used the potential to evaluate the dimerization of orthosilicic acid with over 2 ns of enhanced-sampling molecular dynamics (MD) simulations and resolve the previously unsettled mechanism in an aqueous solution.

## Results

### Reference Data Set

A preliminary training data set was constructed containing molecular silica clusters with explicit aqueous solvation. Molecular graphs for linear and cyclic  $\text{Si}_x\text{O}_y\text{H}_z$  were generated with  $x < 10$  programmatically, and embedded into 3D conformers with RDkit<sup>30</sup>. Solvating water molecules were added at random with geometrical constraints to avoid clashes and to maximize hydrogen bonding. A subset of structures were deprotonated in one or more O-H bonds and calculated as anion, or balanced with  $\text{Na}^+$  cations. We also added secondary and common building units of zeolites from the international zeolite database<sup>31</sup>. All structures were pre-relaxed with GFN2-xtb as implemented in xtb-6.4.1 and then refined with  $\omega\text{B97X-D3/def2-TZVP}$  level of theory in Orca<sup>32</sup>, which served as our ground truth forces and energies. Reactive geometries for proton transfer and covalent Si-O-Si reactivity were obtained using a nudged elastic band at the GFN2-xtb level of theory with implicit solvent effect, followed by gradient DFT calculations at  $\omega\text{B97X-D3/def2-TZVP}$ . Figure 1a shows the distribution of the data set comprising 210K geometries based on sources as well as based on the stoichiometry of the molecule clusters.

Hybrid DFT is typically more accurate for bond-breaking than the GGA functionals used to train interatomic potentials on solids, but orbital-based simulations are restricted to a few hundred atoms in

molecule clusters. However, considering the local, atom-centered architecture of most NNIPs, they can learn from cluster data and run predictions on periodic systems. Most of the interactions in our concerning phase space are short-range coupled with electrostatic long-range interactions which are relatively shielded in a high dielectric medium like water<sup>20,33</sup>.

### Active learning strategies

Our active learning methods rely on prediction uncertainty. An ensemble of three NNIP models was trained with varying parameters initialized with different random seeds and the variance of force predictions was used as the uncertainty metric. An uncertainty-based adversarial attack was applied by performing gradient-based optimization of the differentiable uncertainty metric<sup>29</sup>. New molecular conformations from the data set were sampled by back-propagating atomic displacements to find local optima that maximize the uncertainty of the NNIP committee while balancing thermodynamic likelihood. These new configurations were then evaluated using quantum mechanical calculations and used to retrain the NNIPs in an active learning loop.

### Differentiable uncertainty attribution

After early loops of adversarial attacks starting on the gen-0 dataset, the NNIP became stable enough to run MD simulations of condensed phases. An adversarial attack on condensed phases will only generate condensed phase configurations, while our data set comprises molecular clusters. Hence, we adapted the idea of attribution. Attribution is a technique employed to test how a change in a certain input neuron can impact an output neuron, thus enabling some interpretability<sup>34,35</sup>. Similarly, we calculated per-atom attribution of the force uncertainty, which can be qualitatively interpreted as which atoms contribute most to the uncertainty of the NNIP committee. Specifically, it is calculated as the derivative of the variance of forces or energy to atomic positions.

An active learning strategy around attribution allowed the selection of computationally tractable, high-uncertainty atomic clusters from MD frames of condensed phases without manual inspection. Atoms with attributes higher than two times the standard deviation from the mean are chosen as atomic centers and all neighboring molecules with at least one atom within a sphere of radius equal to the NNIP cutoff, the distance used to generate neighbors (6 Å), were chosen as the molecule cluster to add to training data. Hence, the approach collects the full first-solvation shell of any uncertain center, without disconnecting bonded atoms. The active learning approach is exemplified in Fig. 1b. The NNIP which is used to create the first generation of molecular clusters through attribution is tested on the same set of geometries after their DFT energies and forces were calculated. It performs very poorly while our current NNIP performs better as shown in Fig. 1c and Fig. 1d, thus proving the efficiency of attribution to choose atomic environments with high errors. Figure 1e shows a consecutive decrease in attributes over three generations showing that the attribution technique produces generalizable improvements to the potential, beyond the specific environments added in each generation. We also calculated attribution on bonds by taking a derivative to inter-atomic distances or bond lengths and showed it in Section 2 of the Supplementary document. Further, Supplementary Fig. 3 shows that without attribution-based active learning the base generation NNIP cannot simulate accurate structural properties of water.

### Properties of water

Benchmarking force error is not sufficient to quantify the quality of ML potentials<sup>36</sup> and simulation-based statistics should be used to evaluate model performance in production. We begin by testing the properties of liquid water.

### Radial distribution functions

We compared the oxygen-oxygen ( $g_{OO}$ ) and oxygen-hydrogen ( $g_{OH}$ ) radial distribution functions (RDFs) to experimental RDFs obtained from X-ray diffraction measurements<sup>37</sup> and those obtained by reported Gaussian moment neural network (GM-NN) potentials<sup>20</sup> trained on periodic systems of water as well as those trained on water clusters. Figure 2a shows good agreement of  $g_{OO}$  with the experimental results with the first peak slightly overestimated similar to the ones obtained by BLYP and B3LYP potentials trained on clusters. Both the peak positions in  $g_{OH}$  match experimental results as seen in Fig. 2b but are overestimated for all the NN potentials, which can be attributed to the lack of Quantum nuclear effects (QNE)<sup>38</sup> in the benchmark MD simulations. The O-O bond structure of water at ambient conditions is not as affected by QNEs as the bonds involving hydrogen<sup>18,38</sup> and we notice a similar behavior for our NNIP.

### Diffusion coefficient

The mean square displacements (MSDs) at different temperatures vs time are shown in Fig. 2c. At 300 K, diffusion coefficients ( $D$ ) were calculated to be  $0.21 \pm 0.005 \text{ Å}^2/\text{ps}$ , close to an experimental value of  $0.24 \pm 0.015 \text{ Å}^2/\text{ps}$ <sup>39</sup>. Similar to RDFs, we compare our results to the GM-NN potentials whose predicted values are  $0.182 \pm 0.006 \text{ Å}^2/\text{ps}$ ,  $0.215 \pm 0.007 \text{ Å}^2/\text{ps}$ ,  $0.143 \pm 0.004 \text{ Å}^2/\text{ps}$ ,  $0.139 \pm 0.006 \text{ Å}^2/\text{ps}$  respectively. We plotted our diffusion coefficients at different temperatures in Fig. 2d and compared our results with those obtained by another reported NN potential with MP2 accuracy (DP-MP2)<sup>19</sup> and experimental results<sup>40</sup>. Our results are higher than those predicted by the DP-MP2 model for most of the temperatures and are closer to experimental results. These underestimated  $D$  values for the DP-MP2 model are attributed to the overestimation of inter-molecular energy by MP2 with the aug-cc-pVDZ basis set. Thus our potential can perform better at predicting the diffusion properties of water at different temperatures and provide results comparable to experimental ones.

### Vibrational density of states

The vibrational density of states of water (VDOS) is shown in Fig. 2e, where we could observe that the peaks corresponding to O-H bond stretch, bending, and libration of water at 300 K obtained using our NNIP are quite close to the experimental observations<sup>18</sup>. We further employed Ring Polymer molecular dynamics (RPMD) coupled with path integral Langevin equation thermostat (PILE)<sup>41</sup> to introduce QNE. The RPMD simulations shifted the O-H stretch and bending peaks towards the left. The O-H stretch peak is shifted from 3670 to 3600  $\text{cm}^{-1}$  (3380  $\text{cm}^{-1}$  from the experiment) and the bending motion peak is shifted from 1664 to 1628  $\text{cm}^{-1}$  (1637  $\text{cm}^{-1}$  from the experiment). The libration has very minor changes due to QNE, also observed for VDOS obtained by DP-MP2 model<sup>19</sup> and in previous studies<sup>18</sup>.

### Equilibrium density

Our NNIP predicted an equilibrium density of 1.08  $\text{g cc}^{-1}$ . The GM-NN potentials produced liquid water with densities 0.86  $\text{g cc}^{-1}$ , 1.02  $\text{g cc}^{-1}$ , 1.1  $\text{g cc}^{-1}$ , and 1.12  $\text{g cc}^{-1}$  respectively<sup>20</sup>. Our NNIP performed equally well as the GM-NN potentials trained on clusters. Considering our results we can claim that our NNIP though trained on molecular clusters can predict properties of bulk periodic water systems.

### Crystalline silica

We next tested our ability to predict the relative formation energies of crystalline silica like pure siliceous zeolites to take it a notch further and establish the success of our NNIP on periodic systems. We compared the relative formation energies of 236 zeolites to  $\alpha$ -Quartz predicted by our NNIP to the ones calculated using a periodic PBE-D3 level of DFT theory. We showed that our NNIP reproduces structural properties of  $\alpha$ -Quartz accurately in section 1 of the Supplementary document. Figure 3a depicts our NNIP predicting the relative energies

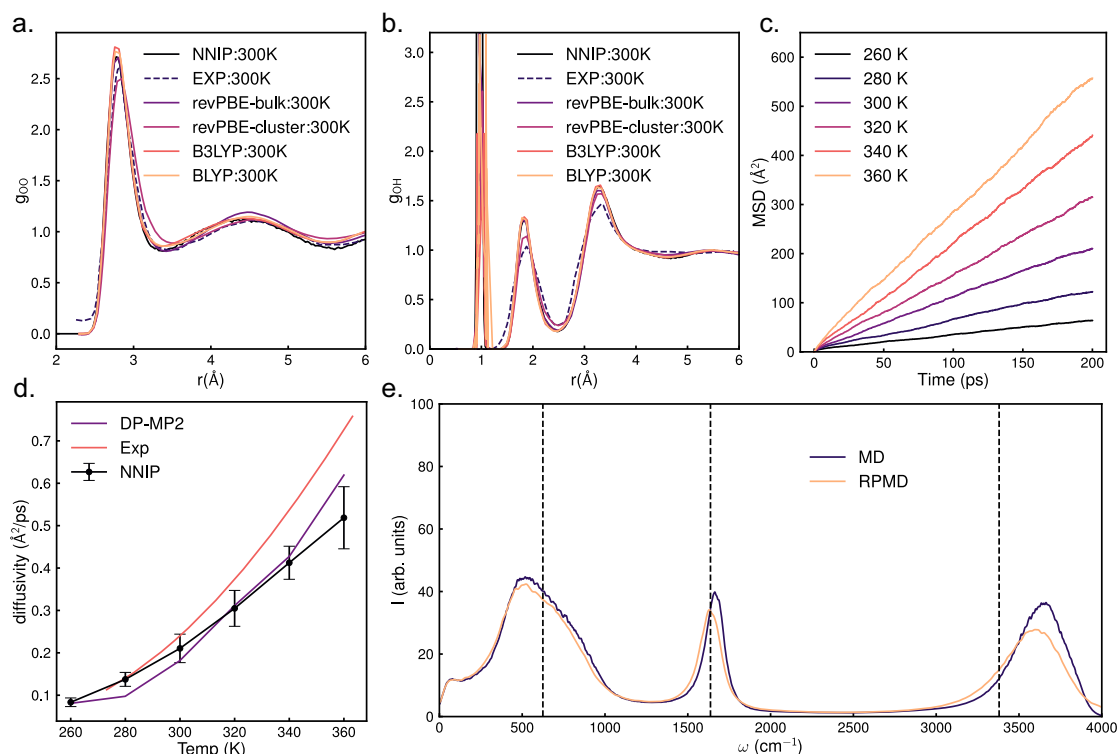


**Fig. 2 | Properties of water.** **a** O-O and **b** O-H radial distribution functions at 300 K were calculated by constant-temperature, constant-volume ensemble (NVT) simulations with our NNIP and compared to experimental and GM-NN potentials' results. NNIP: our NN-based inter-atomic potential; EXP: experimental data; revPBE-bulk: GM-NN potential trained on periodic water box calculated at the revPBE-D3 level of theory; revPBE-cluster: GM-NN potential trained on water clusters calculated at the revPBE-D3 level of theory; B3LYP: GM-NN potential trained on water clusters calculated at the B3LYP-D3 level of theory; BLYP: GM-NN potential trained

on water clusters calculated at the BLYP-D3 level of theory. **c** Mean squared displacements vs time at different temperatures from 260 K to 360 K. **d** diffusivity vs temperature obtained using our NNIP compared with results from a DP-MP2 potential and experiment. The error bar shows the standard deviation of diffusivity prediction by our NNIP. **e** Vibrational density of states of liquid water obtained from classical molecular dynamics (MD) and Ring polymer molecular dynamics (RPMD) with our NNIP. The dashed vertical lines represent the experimental infrared spectrum. Source data are provided as a Source Data file.

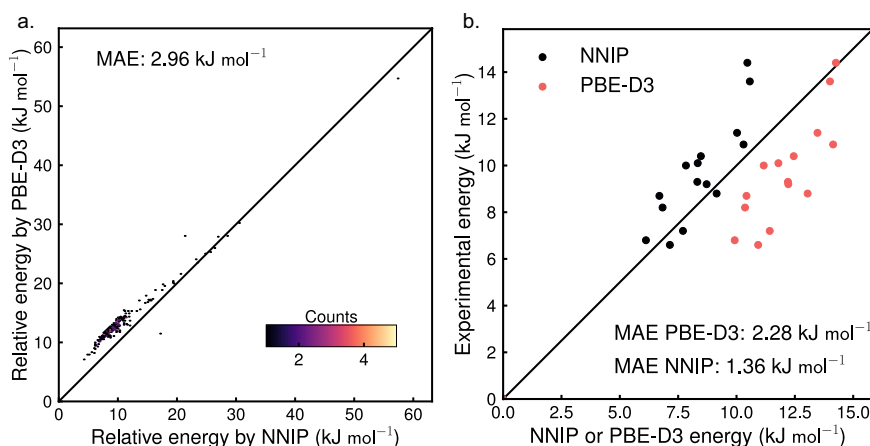


**Fig. 3 | Properties of crystalline silica.** **a** Relative formation energies of pure siliceous zeolites with respect to  $\alpha$ -Quartz calculated using our NN-based interatomic potential (NNIP) compared to those obtained by PBE-D3 level of Density

Functional Theory (DFT). **b** Relative formation energies of fifteen zeolites obtained by NNIP and PBE-D3 compared to their experimental transition enthalpies. MAE Mean absolute error. Source data are provided as a Source Data file.

quite closely to periodic DFT results. We also collected experimental relative transition enthalpies for fifteen siliceous zeolites<sup>42</sup> and compared them to the relative formation energies calculated using periodic PBE-D3 DFT theory and our NNIP in Fig. 3b. The NNIP performed better than the periodic DFT. This better prediction can be because our NNIP is trained on data obtained using a higher level of hybrid functional DFT theory than PBE-D3. Thus, training on cluster data has a lesser impact than the theory of DFT employed to obtain the training

data and our NNIP serves better in predicting the relative formation energies of crystalline zeolites.

### Elastic constants

We also calculated elastic constants of  $\alpha$ -Quartz and siliceous FAU zeolite framework to further solidify our claim on NNIP's capability to predict properties of periodic systems involving silicates. The calculated elastic constants compared to results found in the literature are

**Table 1 | Elastic constants of  $\alpha$ -Quartz and FAU siliceous zeolite framework calculated by our NNIP ensemble and compared to results published in the literature<sup>66</sup>**

| Crystalline silica |               | Bulk Modulus (GPa) | Young Modulus (GPa) | Shear Modulus (GPa) | Poisson's ratio |
|--------------------|---------------|--------------------|---------------------|---------------------|-----------------|
| $\alpha$ -Quartz   | NNIP ensemble | $51.92 \pm 10$     | $112.16 \pm 8$      | $49.20 \pm 3$       | $0.14 \pm 0.06$ |
|                    | literature    | 38.23              | 101.41              | 47.93               | 0.06            |
| FAU                | NNIP ensemble | $44.34 \pm 5$      | $42.11 \pm 4$       | $15.69 \pm 5$       | $0.34 \pm 0.02$ |
|                    | literature    | 61.37              | 50.24               | 18.42               | 0.36            |

The standard deviations obtained from the ensemble are also shown.

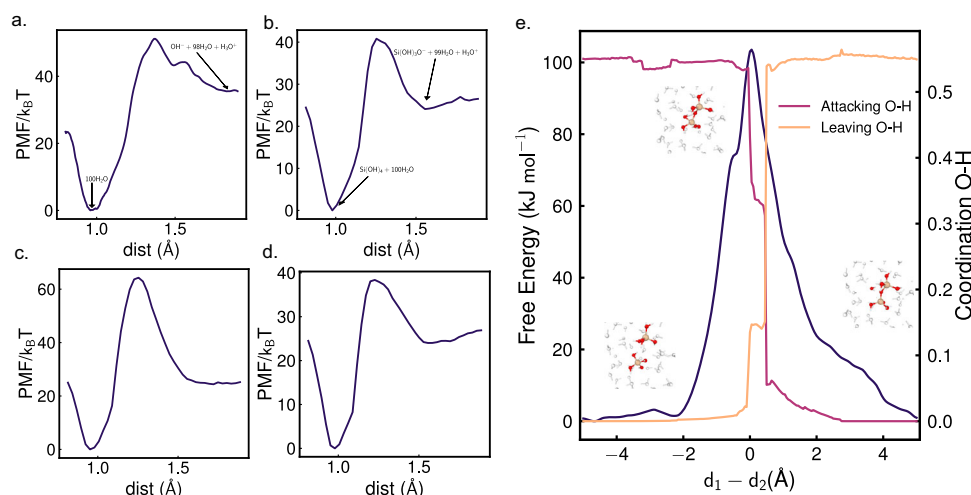


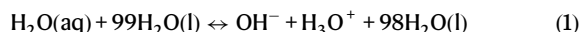
**Fig. 4 | Simulations of chemical reactions.** **a** Potential of mean force (PMF) of water deprotonation. The insets show reactant state ( $0.95 \text{ \AA} < r_{\text{OH}} < 1 \text{ \AA}$ ) and product state ( $r_{\text{OH}} = 1.85 \text{ \AA}$ ). The free energy of the reactant state is set to 0. **b** PMF of orthosilicic acid deprotonation. The insets show reactant state ( $0.95 \text{ \AA} < r_{\text{OH}} < 1 \text{ \AA}$ ) and product state ( $r_{\text{OH}} = 1.55 \text{ \AA}$ ). **c** PMF of dimer deprotonation. **d** PMF of trimer deprotonation. T is temp = 300 K. **e** Free energy profile of silica dimerization with

lateral mechanism, coordination between attacking O and transferring hydrogen and that between leaving oxygen and the transferring hydrogen along the reaction coordinate  $d_1 - d_2$  ( $\text{\AA}$ ).  $d_2$  represents the bond distance between attacking oxygen and the attacked silicon, whereas  $d_1$  represents the distance between the leaving oxygen and the attacked silicon. The insets show the reactant, product, and transition states. Source data are provided as a Source Data file.

shown in Table 1. It can be seen the NNIP ensemble is quite close to obtaining the elastic constants for both  $\alpha$ -Quartz and FAU zeolite framework. Thus, our NNIP can capture the symmetry in periodic silicates adequately and can be used for predicting their properties.

### $\text{pK}_a$ of water

To obtain the  $\text{pK}_a$  of water or  $\text{pK}_w$ , we chose a single  $\text{H}_2\text{O}$  molecule in a cubic box of 100 water molecules and varied one of the O-H bond lengths as our reaction coordinates from 0.8 to 1.9  $\text{\AA}$  with an increment of 0.05  $\text{\AA}$ . The deprotonation reaction is shown in equation (1)

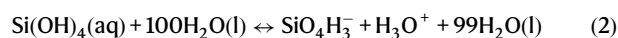


Where  $\text{H}_2\text{O}(\text{aq})$  is the water molecule undergoing dissociation. The potential of mean force (PMF) vs O-H bond length ( $r_{\text{OH}}$ ) is shown in Fig. 4a. As seen from the figure, the PMF falls into a minimum between  $r_{\text{OH}} = 0.95$  and  $1 \text{ \AA}$ . This is the equilibrium O-H bond length in the neutral water molecule, thus signifying the reactant state. We find another equilibrium point at  $r_{\text{OH}} = 1.85 \text{ \AA}$ . Here, the  $\text{H}^+$  has left the dissociated water molecule and has been successfully transferred to  $\text{H}_3\text{O}^+$ . The only interaction between  $\text{OH}^-$  and the  $\text{H}_3\text{O}^+$  is long-range Coulombic now and hence this state signifies the product state where the dissociation is complete. From the difference in PMF, we obtain our  $\text{pK}_w$  as 15.4. The reference  $\text{pK}_w$  value is 14<sup>43</sup> but considering the QNE the expected  $\text{pK}_w$  becomes 17<sup>44</sup>. Our result is also comparable to a reported value of 16 obtained by first principle calculations using

SCAN functional<sup>45</sup>. They used the same procedure as us to find the  $\text{pK}_a$  of water.

### $\text{pK}_a$ of silicate oligomers

We also calculated the  $\text{pK}_a$  of orthosilicic acid, silicate dimer, and trimer. Orthosilicic acid and dimer have all identical O-H bonds. Hence dissociating any one of them is fine. However, for trimer two O-H bonds attached to the middle Si are different than the end ones. We chose one of the end O-H bond dissociations for trimer to get its  $\text{pK}_a$ . Similar to water, we took a single orthosilicic acid, dimer, and trimer in a cubic box of 100 water molecules and varied their chosen O-H bond lengths. The individual PMFs vs  $r_{\text{OH}}$  are shown in Fig. 4b–d. The deprotonation reaction of orthosilicic acid looks as shown in equation (2).



We found the equilibrium reactant state at  $r_{\text{OH}}$  between 0.95 and  $1 \text{ \AA}$  and product state at  $r_{\text{OH}} = 1.55 \text{ \AA}$  for orthosilicic acid and between 1.5 and  $1.55 \text{ \AA}$  for the dimer and the trimer. From the PMF, we obtain the  $\text{pK}_a$  of orthosilicic acid as 10.45. Orthosilicic acid is a weak acid with a known  $\text{pK}_a$  of 9.8<sup>46</sup>. Further, for dimer and trimer, we obtained a  $\text{pK}_a$  of 10.43 and 10.42 respectively, whose  $\text{pK}_a$  values have been reported to vary between 9.5 and 10.7<sup>47</sup>. Our reactive NNIP is thus shown to be capable of replicating complex reactions involving pure water as well as amorphous silicates in water.

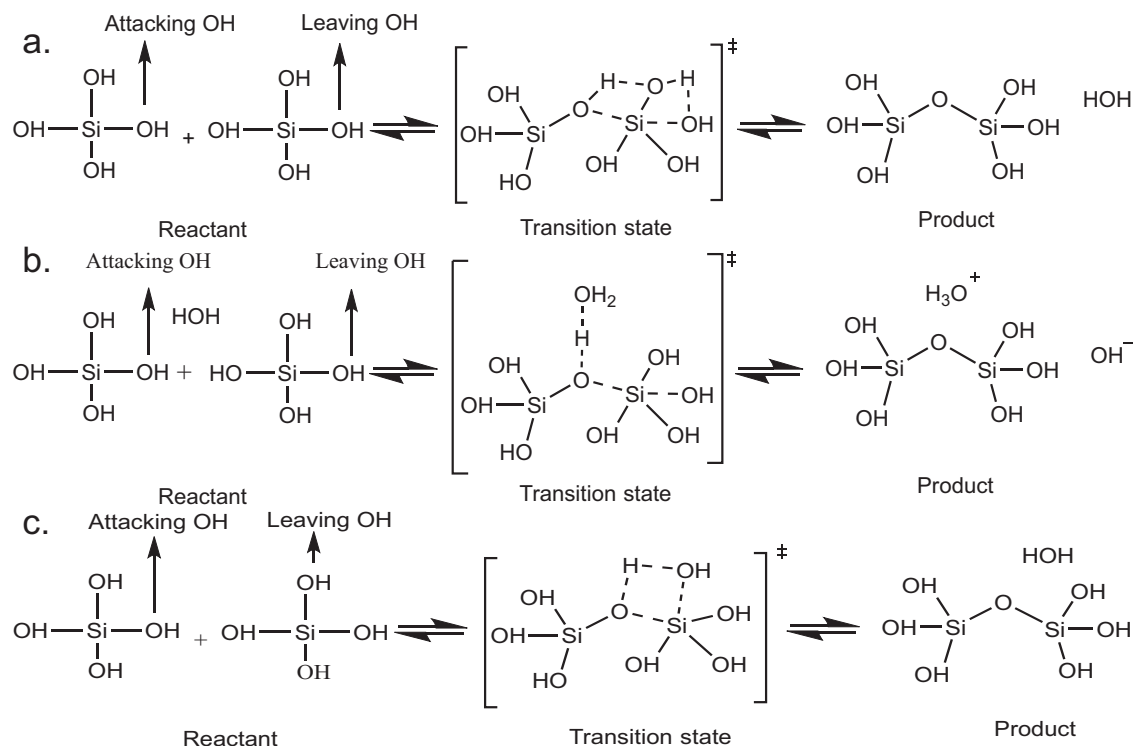


**Fig. 5 | Schemes of silicate dimerization reactions.** **a** Expected scheme of  $S_N2$  back side attack mechanism. The hydrogen from the attacking oxygen goes to the lateral oxygen, which then loses hydrogen to the leaving oxygen. **b** Observed scheme of back-side attack mechanism. When we chose the back side oxygen as leaving oxygen to replicate  $S_N2$  mechanism, a surrounding water molecule

interfered and took up hydrogen from the attacking oxygen forming a  $H_3O^+$  and  $OH^-$ . **c** Observed lateral flank-side attack mechanism. This mechanism was not hampered by surrounding water molecules and occurred as reported in the literature. The transition state has a pentavalent Silicon atom.

## Silicate dimerization

We next simulated a dimerization reaction of two silicate monomers in a neutral solution with 100 water molecules at 300 K. We ran umbrella sampling on a reaction coordinate,  $d_1 - d_2$  varying it from  $-5 \text{ \AA}$  to  $5 \text{ \AA}$  with an increment of  $0.05 \text{ \AA}$ .  $d_2$  represents the bond distance between attacking oxygen and the attacked silicon, whereas  $d_1$  represents the distance between the leaving oxygen and the attacked silicon. The attacking oxygen, attacked silicon, and the leaving oxygen atoms for the  $S_N2$  and the lateral attack mechanisms are shown in Fig. 5. The  $S_N2$  mechanism we observed differed from the predicted one in explicit solvent. We observed the formation of  $H_3O^+$  and  $OH^-$  which results in products with energy  $79 \text{ kJ mol}^{-1}$  higher than the reactants and the activation energy is also higher ( $180 \text{ kJ mol}^{-1}$ ) than the lateral attack mechanism. The free energy profile of the  $S_N2$  mechanism is shown in Supplementary Fig. 2. Hence, instead of the ions coming together to form the desired product, they tend to go back to the reaction state as shown in Fig. 5a–b. However, in the case of the lateral attack mechanism, the reaction in condensed solvent follows the proposed mechanism in continuum solvent<sup>24–26</sup>, where the leaving oxygen leaves with the hydrogen from the attacking oxygen (Fig. 5c). This helps us to conclude that silicate dimerization in an aqueous solution at neutral conditions occurs through a lateral flank-side attack mechanism with pentavalent silicon as a transition state. Now from the free energy profile of the lateral mechanism in Fig. 4e, we see that at the coordinate close to  $5 \text{ \AA}$ , the product dimer is formed. The reactant state with two monomers is at the coordinates between  $-3$  and  $-5 \text{ \AA}$  where the reaction energy profile is almost flat. The product has  $0.2 \text{ kJ mol}^{-1}$  more energy than the reactant phase. In a DFT study with an implicit solvent effect, the authors found the product has an energy  $9 \text{ kJ mol}^{-1}$  higher than the reactant state with an activation energy of  $129 \text{ kJ mol}^{-1}$ <sup>24</sup>. We obtained the activation energy as  $103 \text{ kJ mol}^{-1}$  which can be trusted

more as we used water solvent explicitly. From the plot of the coordination between the attacking oxygen with the transferring hydrogen and that of the leaving oxygen with the transferring hydrogen in Fig. 4e, we further observed that at the transition state, the hydrogen leaves the attacking oxygen with a drop in their coordination value and moves to the leaving oxygen with an increase in their coordination value almost immediately as expected in the predicted lateral flank-side mechanism.

## Discussion

Neural Network potentials have proven to work as a bridge between first principle calculations and classical potentials several times. In this work, we have proved the same by developing a reactive equivariant NNIP trained on a complex domain of silicate-water interactions. Our NNIP though trained on molecule clusters can predict properties of crystalline solids. The NNIP reproduces the physical and dynamic properties of water with high accuracy. We further proposed an active learning strategy based on attribution of differentiable uncertainty which further displays highly uncertain inter-atomic interactions in amorphous systems. This method not only serves as a qualitative measure for unsure atomic environments but also serves to minimize computational costs by extracting only the uncertain atomic environments. The reactive potential is further adept at predicting the reaction path of deprotonation of water, small silicate oligomers, and silica dimerization reaction in a water solution. This NNIP can further be used to probe into other silicate polymerization reactions leading to precipitated silica. We can also study the impact of pH and temperature on these reactions in the future as we have sodium in our data set to maintain charge neutrality or we can add aluminum in our data set to train a potential on preliminary stages of the synthesis of zeolites.

## Methods

### Model architecture

Our NN potential is based on PaiNN architecture<sup>48</sup>, which is an equivariant neural network with message passing as its backbone. Equivariant models<sup>48–50</sup> can act on non-invariant inputs like displacement vectors in a symmetry-respecting way and thus offer good accuracy with a low amount of training data for properties that are equivariant in Euclidean space like forces, thus being more data sufficient than invariant potentials like Schnet<sup>51</sup>. In this method, an invariant feature vector for each atom is generated with their atomic numbers which are then updated through convolutions with “messages” from neighboring atoms which consist of distance, orientation, and features of nearest neighbor atoms within a cutoff distance. Each atomic feature is updated with information from neighbors and through convolutions obtains information from atoms far away as well, thereby generating a representation vector as a function of atomic positions and orientations, which is then mapped to atomic energies through feed-forward NN and summed to obtain the energy of the system. Through automatic differentiation, the forces on each atom can also be obtained. We further modified the model to predict stresses of a system based on the virial theorem<sup>52</sup> as given in equation (3) for pair potentials.

$$\sigma_{ij}^V = \frac{1}{V} \sum_{\alpha} \left[ \sum_{\substack{\beta=1 \\ \beta \neq \alpha}}^N (r_i^{\beta} - r_i^{\alpha}) f_j^{\alpha\beta} - m^{\alpha} v_i^{\alpha} v_j^{\alpha} \right], \quad f_j^{\alpha\beta} = \frac{\partial E}{\partial (r_j^{\beta} - r_j^{\alpha})} \quad (3)$$

where (i,j) signifies x,y, and z directions,  $\beta$  varies from 1 to N neighbors of atom  $\alpha$ ,  $r_i^{\beta}$  and  $r_i^{\alpha}$  are positions of atom  $\beta$  and  $\alpha$  along direction  $i$  respectively,  $f_j^{\alpha\beta}$  is the force on  $\alpha$  by  $\beta$  along direction  $j$  and  $E$  is total energy,  $V$  is total volume,  $m^{\alpha}$  is the mass of atom  $\alpha$  and  $v^{\alpha}$  is the thermal velocity of atom  $\alpha$ .

### Training details

Molecular geometries of the same stoichiometry are split at a ratio of 3:1:1 into training, validation, and test sets. We employed a mini-batch gradient descent optimization with Adam optimizer. The learning drops from  $10^{-4}$  to  $10^{-6}$  at a rate of 0.5 when the validation loss hits a plateau for 20 epochs. The loss function is the weighted sum of square losses of forces and energies as implemented in PaiNN<sup>48</sup>, with a loss coefficient of 0.95 for forces and 0.05 for energies. We fixed the number of convolutional layers at 4 and the cutoff for nearest neighbors at 6 Å. We optimized the other hyperparameters with SigOpt<sup>53</sup> on validation loss. The optimized dimension of the feature vector for each atom is 310; the number of radial basis functions used is 20 and the batch size is chosen as 9.

### Active learning

We ran adversarial attacks on molecular geometries randomly chosen from the available data set for three generations. Our model was now capable of running a few ps of simulations without exploding the system and we used it to run MD simulations on 10–100 orthosilicate molecules with ratios 1:2, 1:3, 1:5, and 1:10 of water molecules at 300, 500, 750, and 1000 K and ran three generations of attribution-based active learning on them. At high temperatures, we could sample silicate polymerization along with their intermediate and transition states. The fourth generation of attributions was run on simulations replicating the deprotonation of water, orthosilicic acid, and other silicates. We then ran one generation of attributions on zeolite crystals and the sixth generation on the neutral dimerization reaction to extract uncertain molecule clusters from them. Our active learning thus consists of a total of nine rounds, involving three from adversarial attacks and then six generations of attribution-based active learning on the various stoichiometries.

### Simulation details

We used the atomic simulation environment (ASE) package coupled with our NNIP to run simulations. All periodic structures with silicates and water molecules are generated using Packmol<sup>54</sup> and rendered using OVITO<sup>55</sup>. The simulation boxes were generated with the equilibrium density of water predicted by our NNIP of 1.08 g cc<sup>-1</sup>. All MD simulations were run with canonical (NVT) ensemble with Nosé-Hoover thermostat and timestep of 0.5 fs.

### Radial distribution functions of water

We obtain the oxygen-oxygen ( $g_{OO}$ ) and oxygen-hydrogen ( $g_{OH}$ ) RDFs from 100 ps long MD simulations of a periodic box containing 400 H<sub>2</sub>O molecules in canonical (NVT) ensemble at 300 K after 20 ps equilibration period. We divide the whole trajectory into ten parts and take the average RDF curve as the estimated one.

### Diffusion coefficients of water

The diffusion coefficients were calculated from the MSDs by Einstein's diffusion equation as given in equation (4)

$$D = \frac{MSD}{6\Delta t} = \frac{\langle |x(t + \Delta t) - x(t)|^2 \rangle}{6\Delta t} \quad (4)$$

Where  $D$  is the diffusion coefficient,  $\Delta t$  is the time interval used for measuring MSDs. We used 10 ps as the time interval. We further obtained the slope of  $\ln(MSD)$  vs  $\ln(t)$  and choose the region for linear fitting where the slope is within  $1 \pm 10^{-4}$ . The MSDs were calculated based on the relative squared displacement of oxygen atoms and averaged over trajectories for all water molecules. We further needed to extrapolate the diffusion coefficient to infinite-system size owing to the finite-size effect<sup>56</sup>. The extrapolated diffusion coefficient ( $D(\infty)$ ) is obtained from finite-system size diffusion coefficient ( $D(L)$ ) with an extra correction term as shown in equation (5)

$$D(\infty) = D(L) + \frac{\zeta}{6\pi\beta\eta L} \quad (5)$$

Where  $L$  is the simulation box size,  $\zeta$  depends on the geometry of the simulation box (2.837297 for a cubic box),  $\beta$  is  $\frac{1}{k_B T}$ , and  $\eta$  is experimental shear viscosity =  $0.8925 \times 10^{-3}$  Pa s. We calculated the diffusion coefficients by first running an NVT simulation for 50 ps to equilibrate the system and then ran five independent NVE simulations for 150 ps at different temperatures of 260, 280, 300, 320, 340, and 360 K and chose the average MSDs.

### Vibrational density of states of water

The VDOS of water can be procured from the Fourier transform of the velocity-velocity autocorrelation function as given in equation (6)

$$C_{vv}(\omega) = \int \langle v(\tau)v(\tau + t) \rangle e^{-i\omega t} dt \quad (6)$$

Where  $v(\tau)$  is the centroid velocity of the atom and  $\omega$  is the vibrational frequency. The RPMD simulations were implemented using SchnetPack<sup>57</sup> with 5 beads and 0.25fs timestep.

### Equilibrium density of water

We ran MD simulations in NPT (constant pressure) ensemble using the isobaric-isothermal form of the Nosé-Hoover dynamics at a temperature of 300 K and a pressure of 1 atm. We chose a cubic box of 400 water molecules and the box was allowed to relax in a hydrostatic manner. A 1 ns long trajectory after 100 ps of equilibration period was divided into ten parts and the time-averaged volume of the simulation box was taken as equilibrium volume. The equilibrium density was then calculated from it.

## Elastic constants of silicates

We provided normal strains of  $-1$  to  $1\%$  in steps of  $0.5\%$  and shear strains of  $-4$  to  $4\%$  in steps of  $2\%$  to the equilibrium structures of  $\alpha$ -Quartz and FAU zeolite and obtained their deformed structures using Pymatgen<sup>58</sup>. We then optimized each deformed structure by the BFGS algorithm implemented in the ASE package. We then obtained the stresses at different strains and their corresponding slopes to measure the elastic tensors for each system. We then used ELATE software<sup>59</sup> to obtain elastic constants from the tensors. We used our NNIP ensemble to procure the average elastic constants and their standard deviation.

## Chemical reactions

We performed constrained MD simulations with umbrella sampling varying our reaction coordinates chosen for different reactions. We then obtained a smooth reaction path from reactant to product state by calculating the PMF as a function of our chosen reaction coordinates. The PMF was calculated using the Multistate Bennett's Acceptance Ratio (MBAR)<sup>60</sup> on the sampled trajectories. The MBAR analysis was carried out using codes implemented in adaptive sampling package<sup>61,62</sup>. We can use the PMF of the deprotonation reactions of water and silicates to obtain their  $pK_a$  values. The difference in the PMF of reactant and product phase ( $\Delta F$ ) is related to the acid dissociation constant ( $K_c$ ) as shown in equation (7) and the relation between  $pK_a$  and  $K_c$  is shown in equation (8), Combining both we can get  $pK_a$  from  $\Delta F$  as shown in equation (9).

$$\Delta F = -k_B T \ln K_c \quad (7)$$

$$pK_a = -\log K_c \quad (8)$$

$$pK_a = \frac{\Delta F}{k_B T \times 2.3026} \quad (9)$$

Where  $k_B$  is the Boltzmann constant and  $T$  is temp =  $300\text{ K}$ .

## Data availability

The silica-water data set and the NNIP models generated during this study are deposited at Materials Data Facility<sup>63</sup> at <https://doi.org/10.18126/PZJR-X7PV>. Source files for all graphs in our figures are provided in the Source Data folder deposited at figshare<sup>64</sup>.

## Code availability

The codes used for this study can be downloaded from the following github repository: <https://github.com/learningmatter-mit/NeuralForceField> under the MIT License (see ref. 65 for permanent link).

## References

- Heaney, P. J., Prewitt, C. T. & Gibbs, G. V. *Silica: Physical behavior, geochemistry, and materials applications*, vol. 29 (Walter de Gruyter GmbH & Co KG, 2018).
- Weitkamp, J. Zeolites and catalysis. *Solid state Ion.* **131**, 175–188 (2000).
- Gao, Y. et al. Multifunctional Role of Silica in Pharmaceutical Formulations. *AAPS PharmSciTech* **23**, 4 23, 1–18 (2022).
- Halas, N. J. Nanoscience under glass: The versatile chemistry of silica nanostructures. *ACS Nano* **2**, 179–183 (2008).
- Bergna, H. & Roberts, W. *Colloidal silica: fundamentals and applications* [https://books.google.com/books?hl=en&lr=&id=d0HuBwAAQBAJ&oi=fnd&pg=PP1&ots=uYjxTawhd\\_&sig=Pf92yXgRxSVaJ1onOL2VdXsvb\\_c](https://books.google.com/books?hl=en&lr=&id=d0HuBwAAQBAJ&oi=fnd&pg=PP1&ots=uYjxTawhd_&sig=Pf92yXgRxSVaJ1onOL2VdXsvb_c) (2005).
- Raza, N. et al. Synthesis and characterization of amorphous precipitated silica from alkaline dissolution of olivine. *RSC Adv.* **8**, 32651–32658 (2018).
- Dewati, R. et al. Precipitated Silica from Pumice and Carbon Dioxide Gas (Co2) in Bubble Column Reactor. *J. Phys.: Conf. Ser.* **953**, 012226 (2018).
- Cundy, C. S. & Cox, P. A. The hydrothermal synthesis of zeolites: History and development from the earliest days to the present time. *Chem. Rev.* **103**, 663–701 (2003).
- Rai, D. K., Beaucage, G., Vogtt, K., Ilavsky, J. & Kammler, H. K. In situ study of aggregate topology during growth of pyrolytic silica. *J. Aerosol Sci.* **118**, 34–44 (2018).
- Meier, M., Sonnick, S., Asylbekov, E., Rädle, M. & Nirschl, H. Multi-scale characterization of precipitated silica. *Powder Technol.* **354**, 45–51 (2019).
- Van Beest, B. W., Kramer, G. J. & Van Santen, R. A. Force fields for silicas and aluminophosphates based on ab initio calculations. *Phys. Rev. Lett.* **64**, 1955 (1990).
- Carré, A., Horbach, J., Ispas, S. & Kob, W. New fitting scheme to obtain effective potential from Car-Parrinello molecular-dynamics simulations: Application to silica. *Europhys. Lett.* **82**, 17001 (2008).
- Flikkema, E. & Bromley, S. T. A new interatomic potential for nanoscale silica. *Chem. Phys. Lett.* **378**, 622–629 (2003).
- Fogarty, J. C., Aktulga, H. M., Grama, A. Y., Van Duin, A. C. & Pandit, S. A. A reactive molecular dynamics simulation of the silica-water interface. *J. Chem. Phys.* **132**, 174704 (2010).
- Rimsza, J. M., Yeon, J., Van Duin, A. C. & Du, J. Water Interactions with Nanoporous Silica: Comparison of ReaxFF and ab Initio based Molecular Dynamics Simulations. *J. Phys. Chem. C* **120**, 24803–24816 (2016).
- Musil, F. et al. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **121**, 1–60 (2021).
- Axelrod, S. et al. Learning Matter: Materials Design with Machine Learning and Atomistic Simulations. *Acc. Mater. Res.* **3**, 343–357 (2022).
- Yao, Y. & Kanai, Y. Nuclear Quantum Effect and Its Temperature Dependence in Liquid Water from Random Phase Approximation via Artificial Neural Network. *J. Phys. Chem. Lett.* **12**, 6354–6362 (2021).
- Liu, J., Lan, J. & He, X. Toward High-level Machine Learning Potential for Water Based on Quantum Fragmentation and Neural Networks. *J. Phys. Chem. A* **126**, 3926–3936 (2022).
- Zaverkin, V., Holzmüller, D., Schuldt, R. & Kästner, J. Predicting properties of periodic systems from cluster data: A case study of liquid water. *J. Chem. Phys.* **156**, 114103 (2022).
- Erhard, L. C., Rohrer, J., Albe, K. & Deringer, V. L. A machine-learned interatomic potential for silica and its relation to empirical models. *npj Comput. Mater.* **2022** 8:1 8, 1–12 (2022).
- Balyakin, I. A., Rempel, S. V., Ryltsev, R. E. & Rempel, A. A. Deep machine learning interatomic potential for liquid silica. *Phys. Rev. E* **102**, 52125 (2020).
- Erlebach, A. et al. A reactive neural network framework for water-loaded acidic zeolites. *Nat Commun* **15**, 4215 <https://doi.org/10.1038/s41467-024-48609-2> (2024).
- Trinh, T. T., Jansen, A. P. & Van Santen, R. A. Mechanism of oligomerization reactions of silica. *J. Phys. Chem. B* **110**, 23099–23106 (2006).
- Zhang, X. Q., Trinh, T. T., Van Santen, R. A. & Jansen, A. P. Mechanism of the initial stage of silicate oligomerization. *J. Am. Chem. Soc.* **133**, 6613–6625 (2011).
- Schaffer, C. L. & Thomson, K. T. Density functional theory investigation into structure and reactivity of prenucleation silica species. *J. Phys. Chem. C* **112**, 12653–12662 (2008).
- Pereira, J. C., Catlow, C. R. & Price, G. D. Silica condensation reaction: an ab initio study. *Chemical Communications* 1387–1388 <https://pubs.rsc.org/en/content/articlelanding/1998/cc/a801816b> (1998).

28. Elanany, M. et al. A quantum molecular dynamics simulation study of the initial hydrolysis step in sol-gel process. *J. Phys. Chem. B* **107**, 1518–1524 (2003).
29. Schwalbe-Koda, D., Tan, A. R. & Gómez-Bombarelli, R. Differentiable sampling of molecular geometries with uncertainty-based adversarial attacks. *Nat. Commun.* **12**, 1–12 (2021).
30. Landrum, G. RDKit: Open-source cheminformatics [www.rdkit.org](http://www.rdkit.org) (2006).
31. Baerlocher, Ch. and McCusker, L.B. Database of Zeolite Structures <http://www.iza-structure.org/databases/> (2021).
32. Neese, F., Wennmohs, F., Becker, U. & Riplinger, C. The ORCA quantum chemistry program package. *J. Chem. Phys.* **152**, 224108 (2020).
33. Morawietz, T., Singraber, A., Dellago, C. & Behler, J. How van der Waals interactions determine the unique properties of water. *Proc. Natl Acad. Sci. USA* **113**, 8368–8373 (2016).
34. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. *34th Int. Conf. Mach. Learn., ICML* **7**, 4844–4866 (2017).
35. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. *34th Int. Conf. Mach. Learn., ICML* **7**, 5109–5118 (2017).
36. Fu, X. et al. Forces are not Enough: Benchmark and Critical Evaluation for Machine Learning Force Fields with Molecular Simulations <https://arxiv.org/abs/2210.07237v1> (2022).
37. Skinner, L. B. et al. Benchmark oxygen-oxygen pair-distribution function of ambient water from x-ray diffraction measurements with a wide Q-range. **138**, 074506 <http://aip.scitation.org/doi/10.1063/1.4790861> (2013).
38. Marsalek, O. & Markland, T. E. 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).
39. Holz, M., Heil, S. R. & Sacco, A. Temperature-dependent self-diffusion coefficients of water and six selected molecular liquids for calibration in accurate 1H NMR PFG measurements. *Phys. Chem. Chem. Phys.* **2**, 4740–4742 (2000).
40. Easteal, A. J., Price, W. E. & Woolf, L. A. Diaphragm cell for high-temperature diffusion measurements. Tracer Diffusion coefficients for water to 363 K. *J. Chem. Soc., Faraday Trans. 1: Phys. Chem. Condens. Phases* **85**, 1091–1097 (1989).
41. Ceriotti, M., Parrinello, M., Markland, T. E. & Manolopoulos, D. E. Efficient stochastic thermostating of path integral molecular dynamics. *J. Chem. Phys.* **133**, 124104 (2010).
42. Erlebach, A., Nachtigall, P. & Grajciar, L. Accurate large-scale simulations of siliceous zeolites by neural network potentials. *npj Comput Mater* **8**, 174 <https://doi.org/10.1038/s41524-022-00865-w> (2022).
43. Silverstein, T. P. & Heller, S. T. PKa Values in the Undergraduate Curriculum: What Is the Real pKa of Water? *J. Chem. Educ.* **94**, 690–695 (2017).
44. Ceriotti, M. et al. Nuclear Quantum Effects in Water and Aqueous Systems: Experiment, Theory, and Current Challenges. *Chem. Rev.* **116**, 7529–7550 (2016).
45. Wang, R., Carnevale, V., Klein, M. L. & Borguet, E. First-Principles Calculation of Water p Ka Using the Newly Developed SCAN Functional. *J. Phys. Chem. Lett.* **11**, 54–59 (2020).
46. Perry, C. C. Biogenic Silica: A Model of Amorphous Structure Control. *Growth, Dissolution and Pattern Formation in Geosystems* 237–251 [https://link.springer.com/chapter/10.1007/978-94-015-9179-9\\_11](https://link.springer.com/chapter/10.1007/978-94-015-9179-9_11) (1999).
47. Belton, D. J., Deschaume, O. & Perry, C. C. An overview of the fundamentals of the chemistry of silica with relevance to biosilicification and technological advances. *FEBS J.* **279**, 1710–1720 (2012).
48. Schütt, K., Unke, O. & Gastegger, M. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *Proc. 38th International Conference on Machine Learning, Proc. Machine Learning Research* Vol. 139 (eds Meila, M. & Zhang, T.) 9377–9388 (PMLR, 2021).
49. Batzner, S. et al. E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nat. Commun.* **2022** 13:113, 1–11 (2021).
50. Musaelian, A. et al. Learning Local Equivariant Representations for Large-Scale Atomistic Dynamics <http://arxiv.org/abs/2204.05249> (2022).
51. Schütt, K. T. et al. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. *Adv. Neural Inf. Process. Syst.* **30**, 992–1002 (2017).
52. Subramaniyan, A. K. & Sun, C. T. Continuum interpretation of virial stress in molecular simulations. *Int. J. Solids Struct.* **45**, 4340–4346 (2008).
53. SigOpt. <https://sigopt.com/>.
54. Martinez, L., Andrade, R., Birgin, E. G. & Martinez, J. M. PACKMOL: a package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem.* **30**, 2157–2164 (2009).
55. Stukowski, A. Visualization and analysis of atomistic simulation data with OVITO—the Open Visualization Tool. *Model. Simul. Mater. Sci. Eng.* **18**, 015012 (2009).
56. Yeh, I. C. & Hummer, G. System-size dependence of diffusion coefficients and viscosities from molecular dynamics simulations with periodic boundary conditions. *J. Phys. Chem. B* **108**, 15873–15879 (2004).
57. Schütt, K. T. et al. SchNetPack: A Deep Learning Toolbox for Atomistic Systems. *J. Chem. Theory Comput.* **15**, 448–455 (2019).
58. Ong, S. P. et al. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
59. Gaillac, R., Pullumbi, P. & Coudert, F. X. ELATE: an open-source online application for analysis and visualization of elastic tensors. *J. Phys.: Condens. Matter* **28**, 275201 (2016).
60. Shirts, M. R. & Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2671659/pdf/JCPA6-000129-124105\\_1.pdf](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2671659/pdf/JCPA6-000129-124105_1.pdf) (2008).
61. Dietschreit, J. C., Diestler, D. J. & Ochsenfeld, C. How to obtain reaction free energies from free-energy profiles. *J. Chem. Phys.* **156**, 114105 (2022).
62. Hulm, A., Dietschreit, J. C. & Ochsenfeld, C. Statistically optimal analysis of the extended-system adaptive biasing force (eABF) method. *J. Chem. Phys.* **157**, 24110 (2022).
63. silica-water dataset with potentials. *Mater Data Facility* <https://doi.org/10.18126/pzjr-x7pv>.
64. Roy, S. Figure\_Source\_Data [https://figshare.com/articles/dataset/Figure\\_Source\\_Data/25928506](https://figshare.com/articles/dataset/Figure_Source_Data/25928506) (2024).
65. simonaxelrod et al. learningmatter-mit/NeuralForceField: NeuralForceField with Uncertainty attribution <https://doi.org/10.5281/zenodo.11391758> (2024).
66. Coudert, F. X. Systematic investigation of the mechanical properties of pure silica zeolites: stiffness, anisotropy, and negative linear compressibility. *Phys. Chem. Chem. Phys.* **15**, 16012–16018 (2013).

## Acknowledgements

R.G-B. and S.R. acknowledge funding support from Evonik AG and the MIT-IBM Watson AI Lab. We acknowledge the MIT Engaging cluster at the Massachusetts Green High-Performance Computing Center (MGHPCC) for providing high-performance computing resources.

## Author contributions

R.G-B. conceived the project. S.R. designed methodologies, curated data, and implemented computer codes and algorithms. J.P.D. contributed to programming. S.R. analyzed the results. R.G-B., J.P.D., and T.S.A. contributed to the analysis and discussion. F.Z. provided some initial datasets.

R.G-B. supervised the research. S.R. wrote the first draft of the manuscript and J.P.D., T.S.A., and R.G-B. contributed to the final version.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-024-50407-9>.

**Correspondence** and requests for materials should be addressed to Rafael Gómez-Bombarelli.

**Peer review information** *Nature Communications* thanks Alvaro Vazquez-Mayagoitia, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024