Supplementary Information

Simulating solvation and acidity in complex mixtures with first-principles accuracy: the case of CH_3SO_3H and H_2O_2 in phenol

Kevin Rossi[a], Veronika Jurásková[b], Raphael Wischert[c], Laurent Garel[d], Clémence Corminbœuf[b], Michele Ceriotti[a]

[a]Laboratory of Computational Science and Modeling (COSMO), Institute of materials, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[b] Laboratory for Computational Molecular Design (LCMD), Institute of Chemical Sciences and Engineering, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

[c]Eco-Efficient Products and Processes Laboratory, Solvay, RIC Shanghai, China

[d]Aroma Performance Laboratory, Solvay, RIC Lyon, France

Contents

1 Neural network training and simulation set ups

The database for training the neural network (*i.e.* and corresponding DFTB-D3H5 and PBE0-D3BJ energies and forces) as well as the CP2K, DFTB, I-PI inputs are stored in the Materials Cloud repository, DOI: https://doi.org/10.24435/materialscloud:z9-zr

2 Technical details

2.1 Generation of the initial structures

The initial simulation boxes are generated in Pacmol starting from boxes (a = 15 Å) containing 20 phenol molecules, one protonated hydrogen peroxide, 1 to 3 hydrogen peroxide molecules, zero to 4 water molecules and the acid counter anion. All the boxes are minimized in Amber 16 using the General Amber Force Field (GAFF) for 100 steepest descent steps followed by 1400 steps with the conjugated gradient algorithm. The minimized structures are then equilibrated in the NPT ensemble for 1 ns at 363.15 K and 1.0 bar using the Langevin thermostat (collision frequency $\gamma = 1\mathrm{ps}^{-1}$) and Berendsen barostat with an isotropic scaling. Equilibrated boxes are used as initial structures for the DFTB simulations used for the NN training.

2.2 Initial structures for the REMD simulation with NN potentials

The structures equilibrated at the force-field level (see 2.1) are reequilibrated using the baselined NN potential in the NVT ensemble for 30 ps with simulation time step 0.5 fs. The temperature is kept by the Langevin thermostat set to 333 K with a box size equal to 14.6 Å (ρ = 1.07 g/cm^3).

2.3 Collective variables for Metadynamics and sketch-map

The coordination number, CN_i is estimated using a smooth, differentiable function. Let p_0 be the maximum reference distance for considering two atoms as bonded, q_0 the width of the descending branch of a sigmoid function f, and n and m the exponents that tune the smoothness and decay speed of f:

$$CN_i = \sum_{i \neq j} f(r_{ij}),$$

$$f(r_{ij}) = \begin{cases} 1 & \text{if } r_{ij} \leq p_0, \\ \dfrac{1 - \left(\dfrac{r_{ij} - p_0}{q_0}\right)^n}{1 - \left(\dfrac{r_{ij} - p_0}{q_0}\right)^m} & \text{if } r_{ij} > p_0. \end{cases} \quad (1)$$

The use of an slowly-decaying rational function for the calculation of the CN has the drawback that atoms may contribute to the coordination of several neighbors, leading to an overall contribution that exceeds one. To reduce the effect of this artifact we use a normalized coordination, $\hat{CN}_i$, so that the contributions of each atom to the coordination number of its neighbors sum to one:

$$\hat{CN}_i = \sum_{i \neq j} \frac{f(r_{ij})}{CN_j} \quad (2)$$

To probe whether the acid is more likely to be observed in its neutral or deprotonated form we monitor the sum of the normalized coordination numbers of the oxygen atoms in the acid, CN^O, accounting only for contributions from the hydrogen atoms:

$$CN^O = \sum_j \hat{CN}_j^O = \sum_{j\,=\,O} \sum_{i'\,=\,H} \frac{f(r_{i'j})}{CN_{i'}} \quad (3)$$

Two minima are expected in the free energy landscape: $CN^O \sim 1$ (neutral acid) or $CN^O \sim 0$ (deprotonated acid).

We also use a second collective variable, the minimum distance between hydrogen atoms bonded to oxygens and the oxygens in the acid, D_{min}, to assess the formation of hydrogen bonding following proton transfer. Let δ be a free parameter, and d_i an array containing all the distances between the hydrogen atoms attached to any oxygen atom in the system and the oxygen atoms in the acid molecule, D_{min} is calculated as:

$$D_{min} = \frac{\delta}{\log \sum_i \exp(\frac{\delta}{d_i}),} \quad (4)$$

The reactive mixture contains different protonated species, which need to be unambiguously identified. In the system under consideration, the oxygen atoms are the most likely to be protonated. The distinction between the different oxygen atoms and their respective protonation state is made using the non-linear dimensionality reduction algorithm sketch-map. [**?**, **?**, **?**] Each oxygen atom in the mixture is labeled with a set of chemically informed parameters P_i defined as a sum of reciprocal distances of neighboring atoms of one type. The distances are estimated for increasing radii (from 1.0 Å to 3.5 Å) around the central oxygen to gather information on bonding and non-bonding environments. The distances higher than the cut-off are multiplied by the corresponding coordination number towards oxygen to ensure a smooth transition to zero.

Mathematically, the parameters P_i are defined as

$$P_i(r_c) = \sum_{j \neq i, j=1}^{N_k} g(r_c, r_{ij}) \tag{5}$$

where r_c is the chosen cut-off radius, N_k is the number of atoms of given element k and

$$g(r_c, r_{ij}) = \begin{cases} \frac{1}{r_{ij}} & \text{if } r_{ij} \leq r_c, \\ \frac{1}{r_{ij}} \dfrac{1 - \left(\dfrac{r_{ij} - r_c}{q_0}\right)^6}{1 - \left(\dfrac{r_{ij} - r_c}{q_0}\right)^{12}} & \text{if } r_{ij} > r_c. \end{cases} \tag{6}$$

3 Error distribution of the baselined-learning

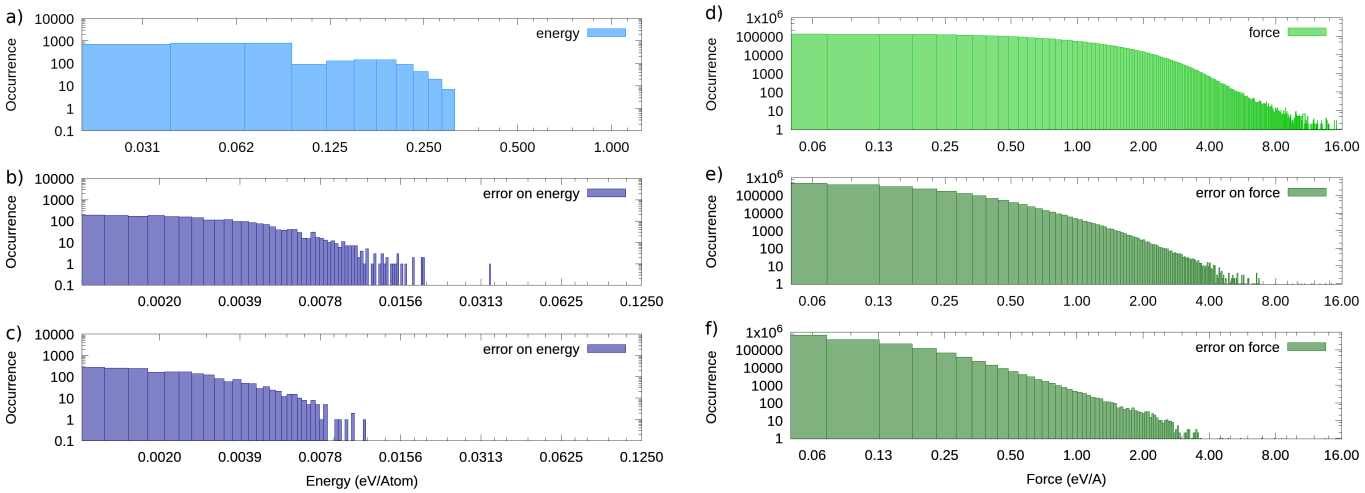

Figure S1: Magnitudes and errors distributions (occurrence of error of a given magnitude vs. error magnitude) baselined NN and direct NN energy predictions are shown in panel (a), (b), and (c). The same quantities are reported for the force predictions in panels (d), (e), and (f). Results are reported by taking the average energy and force predictions of the 5 baselined and direct models trained with 80% of the database structures.

4 Multiple Time Stepping benchmarks

4.1 Multiple Time Stepping stability - conserved quantity and temperature

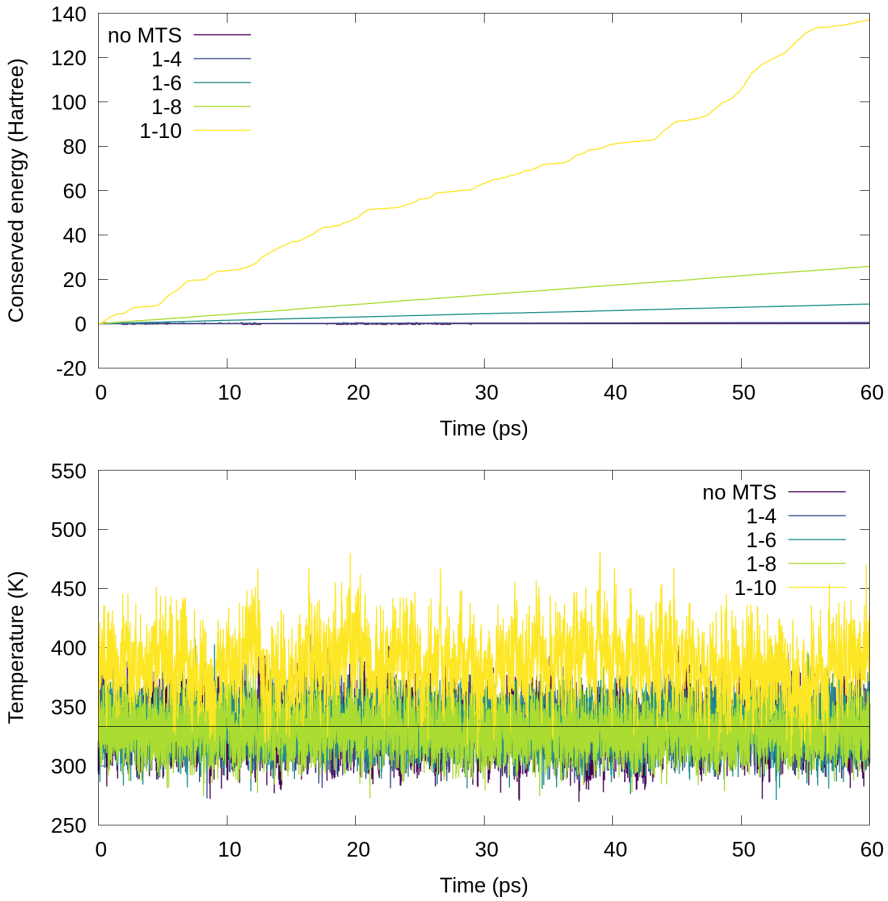

Figure S2: Evolution of the conserved quantity (top panel) and temperature (low panel) in the system containing 20 phenol molecules for different MTS integration schemes with an increasing number of inner steps with respect to the outer correction. The same GLE thermostat reported in the manuscript acts on the system for all simulations set ups. All simulations are stable up to 4 fs outer time step.

4.2 Multiple Time Stepping stability - Radial distribution functions

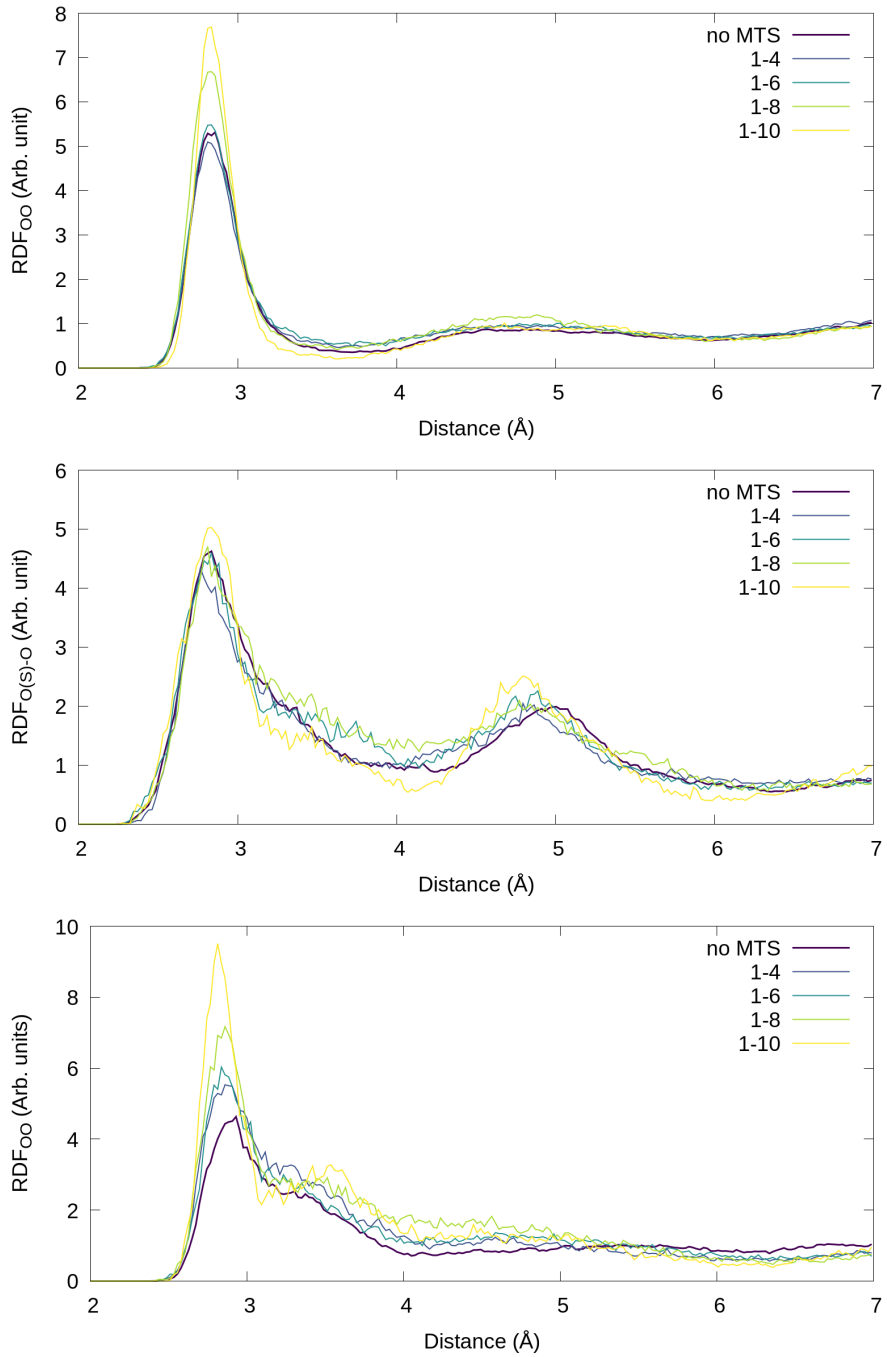

Figure S3: Radial distribution functions for 60 ps trajectories of various mixtures: top panel - oxygen oxygen RDF for 20 phenol molecules, middle panel - phenol oxygen and acid oxygen RDF for 20 phenol molecules and 1 CH_3SO_3H and lower panel - phenol oxygen and hydrogen peroxide oxygen atoms RDF for 20 phenol molecules and 1 H_2O_2. MTS integration schemes with an increasingly large number of inner steps (1-10) are employed.

5 Density distribution computations

We carry out the following transformation before the computations of the density distributions of different atomic species.

- We center the reference of frame on the protonated oxygen.
- We fold each atom in the periodic box so that each of its coordinates lie in between -L/2 and +L/2 distant from the acid protonated O, where L is the length of the box.
- We align each configuration with respect to the axis defined by the H-O-S solid angle.

The corresponding spatial distribution of the atoms in the acid, for the example case of the classical trajectory also including 1 H_2O_2 result as in Figure S5.

The distributions found for the S atom and the hydroxyl group result localized, by construction. The distribution found for the other atoms instead occupy anisotropic but well-defined regions.

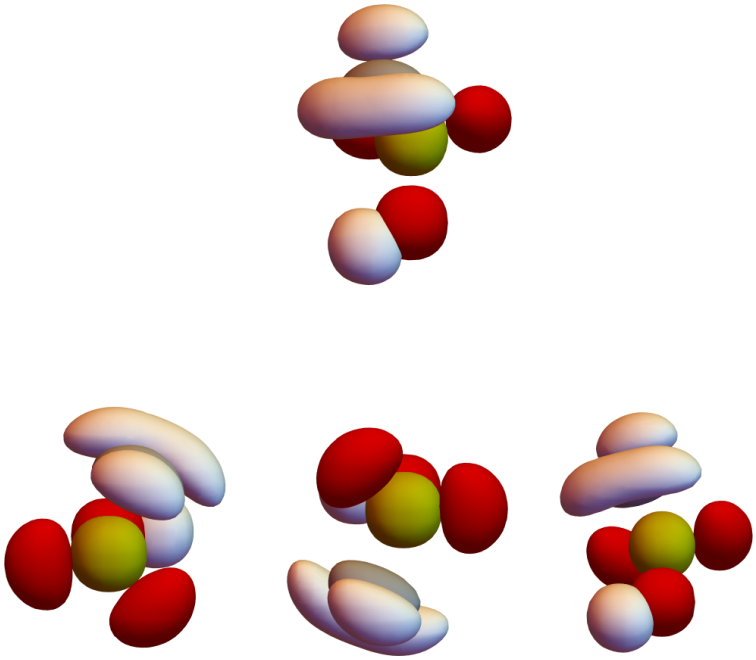

Figure S4: Density distribution of atoms in the methanesulfonic acid sampling carried out with classical REMD in a mixture comprising 1 H_2O_2 and 20 phenols. The S - and hydroxyl group determine the frame of reference. Isovalue 0.0012 is used. The upper panel shows the acid in the same orientation as appearing in the main text. Below, other orientation are shown for completeness.

For reference we also report the atomic distributions found when choosing the reference frame defined by the C-S-O atom closest to the H in the hydroxyl group.

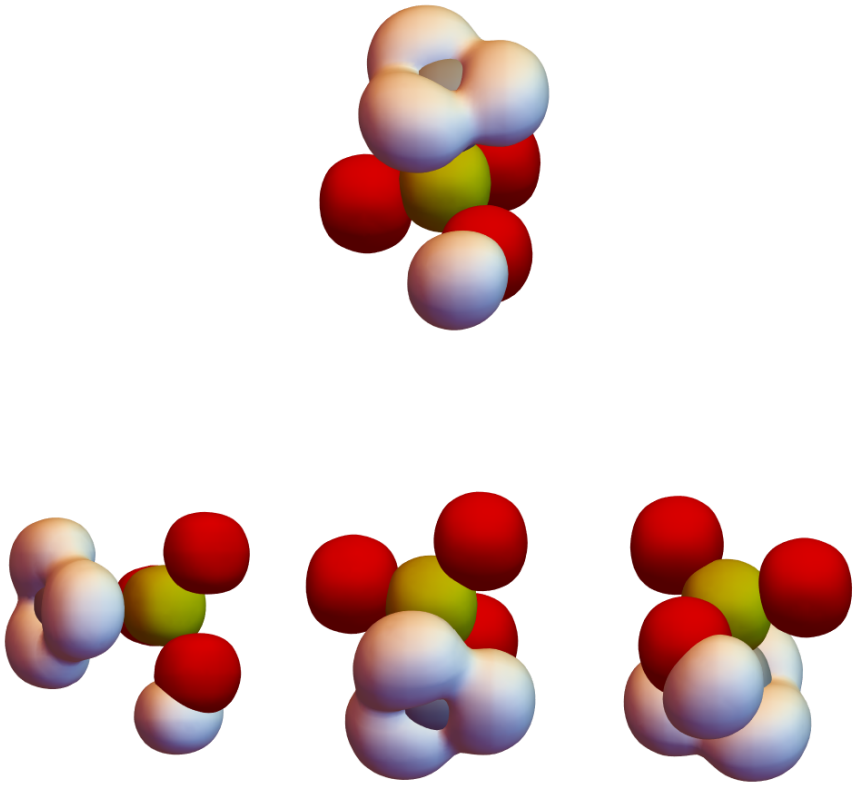

Figure S5: Density distribution of atoms in the methanesulfonic acid sampling carried out with classical REMD in a mixture comprising 1 H_2O_2 and 20 phenol molecules. The C - S - and sulfonyl oxygen closest to the H in the acid hydroxyl group determine the frame of reference. Isovalue 0.0012 is used.

These lead to qualitatively similar but quantitatively different distributions of phenol and peroxide hydroxyl groups around the acid as shown in Figure S6.

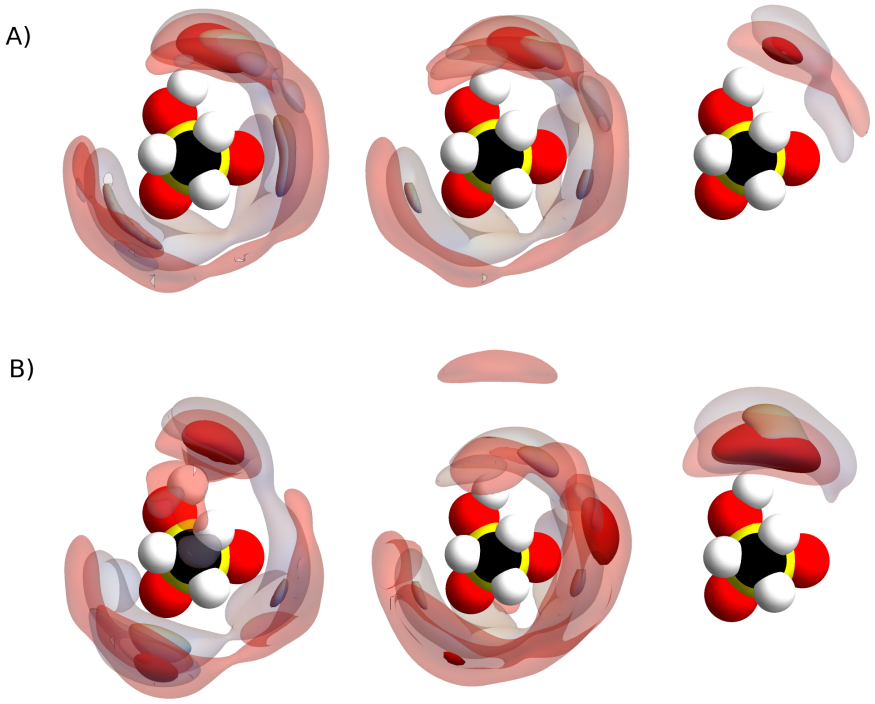

Figure S6: Density distribution of oxygen (red contour) and hydrogen (grey contour) atoms around the acid: from phenol in phenol (left) and mixture of phenol and H_2O_2(middle); from H2O2in mixture of phenol and H2O2(right) for sampling carried out with classicalREMD simulation (panel A) and with path integral MD (panel B). Used isovalues are 0.004(transparent) and 0.012 (opaque). The C - S - and sulfonyl Oxygen closest to the H in the acid hydroxyl group determine the frame of reference.

6 Temperature dependence of Oxygen-Oxygen pair correlation functions from REMD simulations

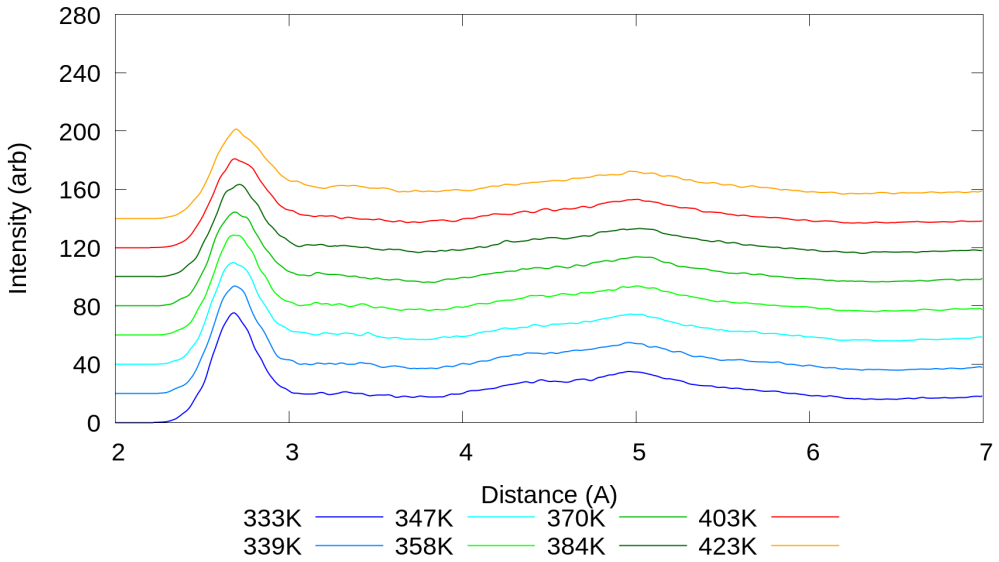

Figure S7: Temperature dependent pair correlation functions between the protonated O atom in the acid and the O in phenol . The $g(r)$ are normalized as in the main text

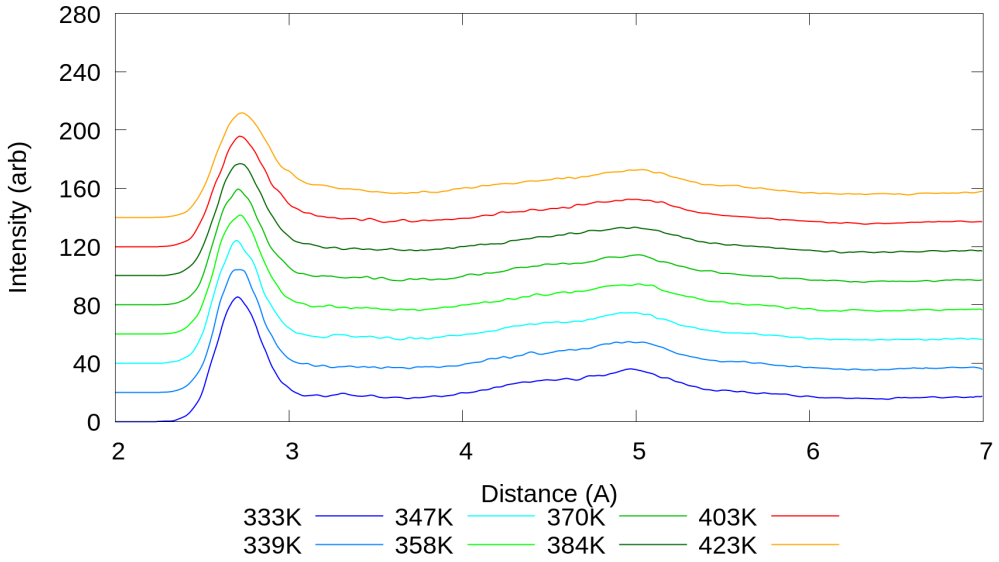

Figure S8: Temperature dependent pair correlation functions between the the protonated O atom in the acid and the O in phenol . The $g(r)$ are normalized as in the main text

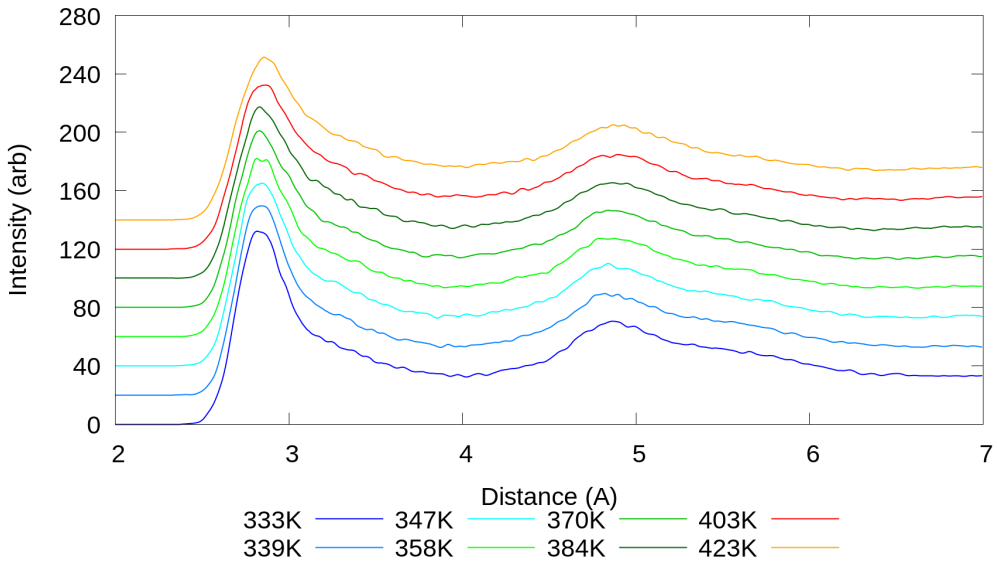

Figure S9: Temperature dependent pair correlation functions the sulfonyl O atoms in the acid and the O in phenol . The $g(r)$ are normalized as in the main text

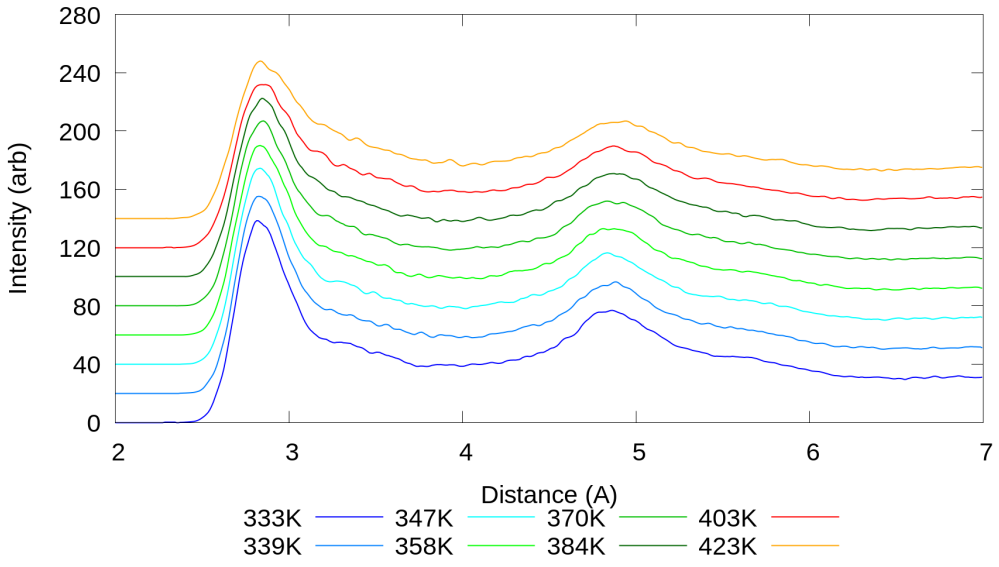

Figure S10: Temperature dependent pair correlation functions between the sulfonyl O atoms in the acid and the O in phenol. The $g(r)$ are normalized as in the main text

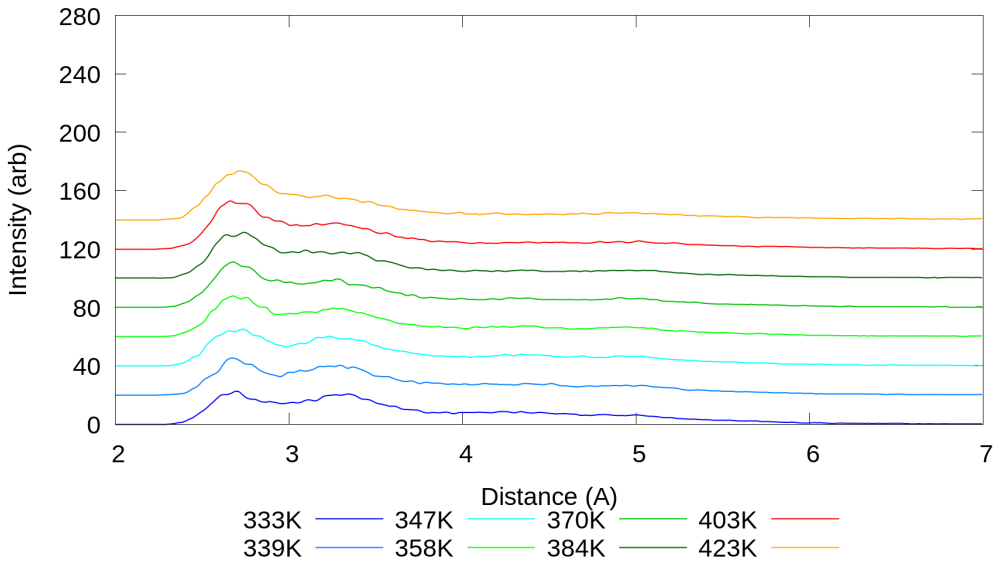

Figure S11: Temperature dependent pair correlation functions between the protonated O in the acid and the O in peroxide. The $g(r)$ are normalized as in the main text

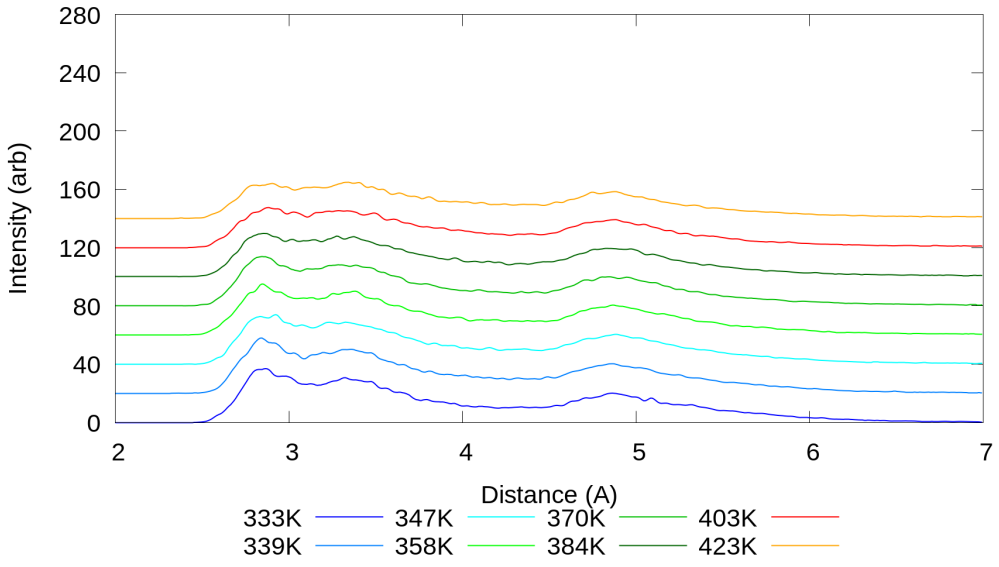

Figure S12: Temperature dependent pair correlation functions between the sulfonyl O in the acid and the O in peroxide . The $g(r)$ are normalized as in the main text

7 Hydrogen bonding network of the acid with peroxide

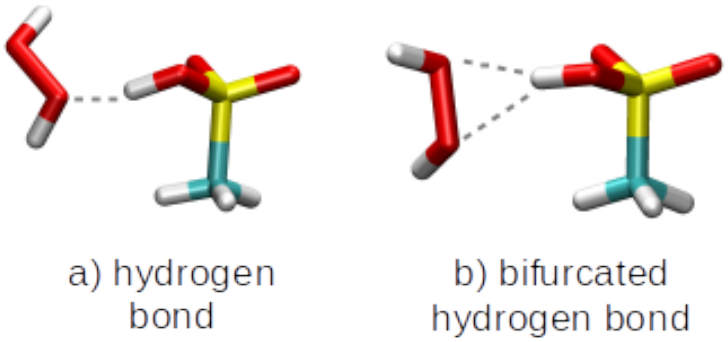

Figure S13: Illustrative snapshots of methanesulfonic acid donating one (left) and two hydrogen bonds (right) towards hydrogen peroxide.

8 NQEs in apolar interaction

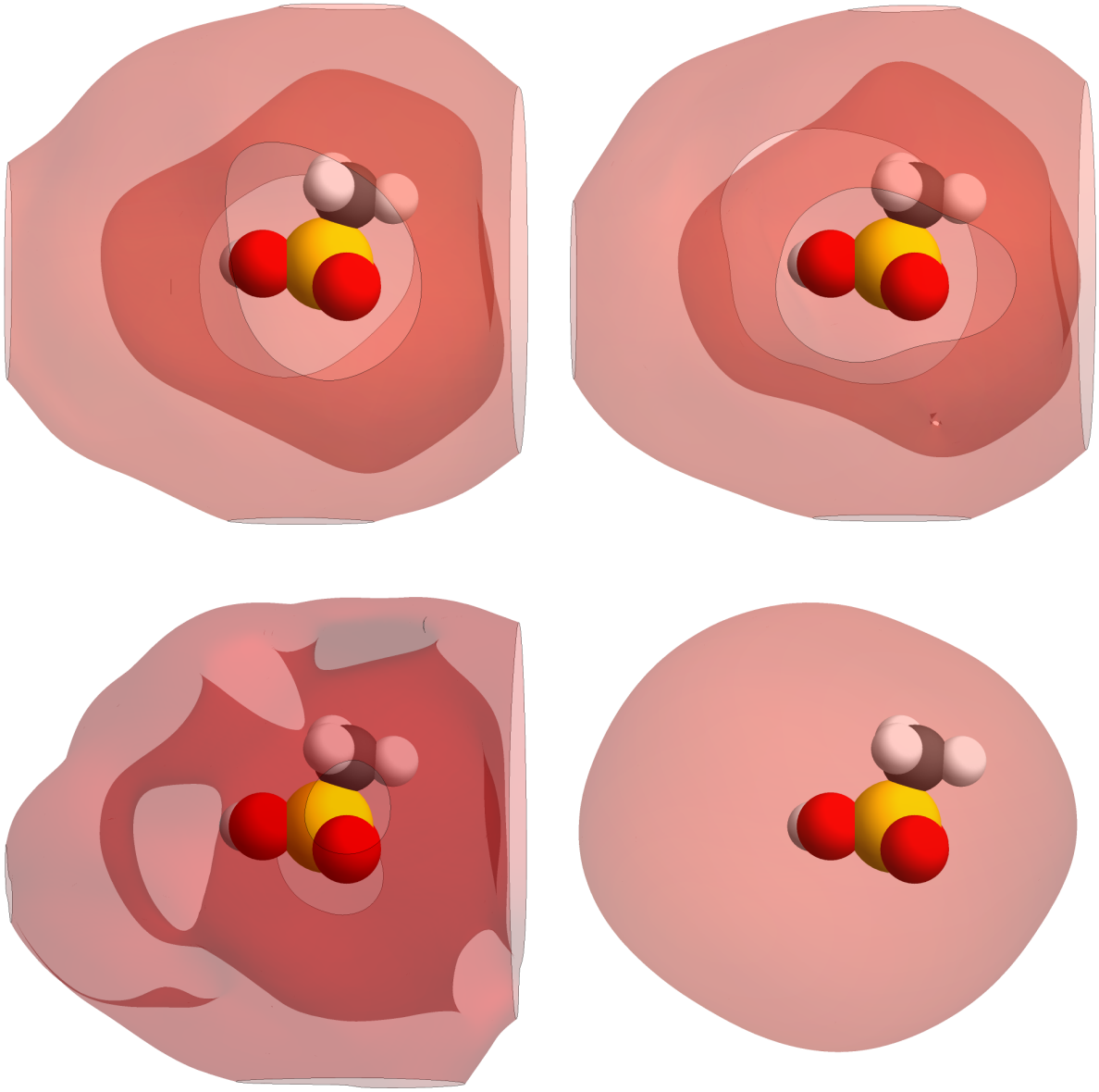

Figure S14: Density distribution of phenol ring centres around the acid in phenol (left) and in a mixture of phenol and H_2O_2 (right) for sampling carried out with classical REMD (upper row) and PIMD (lower row) simulations. Isovalue 0.0012 is used. Phenol ring center of mass spatial distribution density is normalized so that its integral corresponds to the number of phenols in the system, 20.