High Dimensional Fluctuations in Liquid Water: Combining Chemical Intuition with Unsupervised Learning

Supplementary Information

Adu Offei-Danso,[†,‡] Ali Hassanali,[*,†] and Alex Rodriguez[†]

†*The Abdus Salam International Center for Theoretical Physics, Strada Costiera 11, 34151 Trieste, Italy.*

‡*SISSA – International School for Advanced Studies, via Bonomea 265, 34136 Trieste, Italy.*

E-mail: ahassana@ictp.it

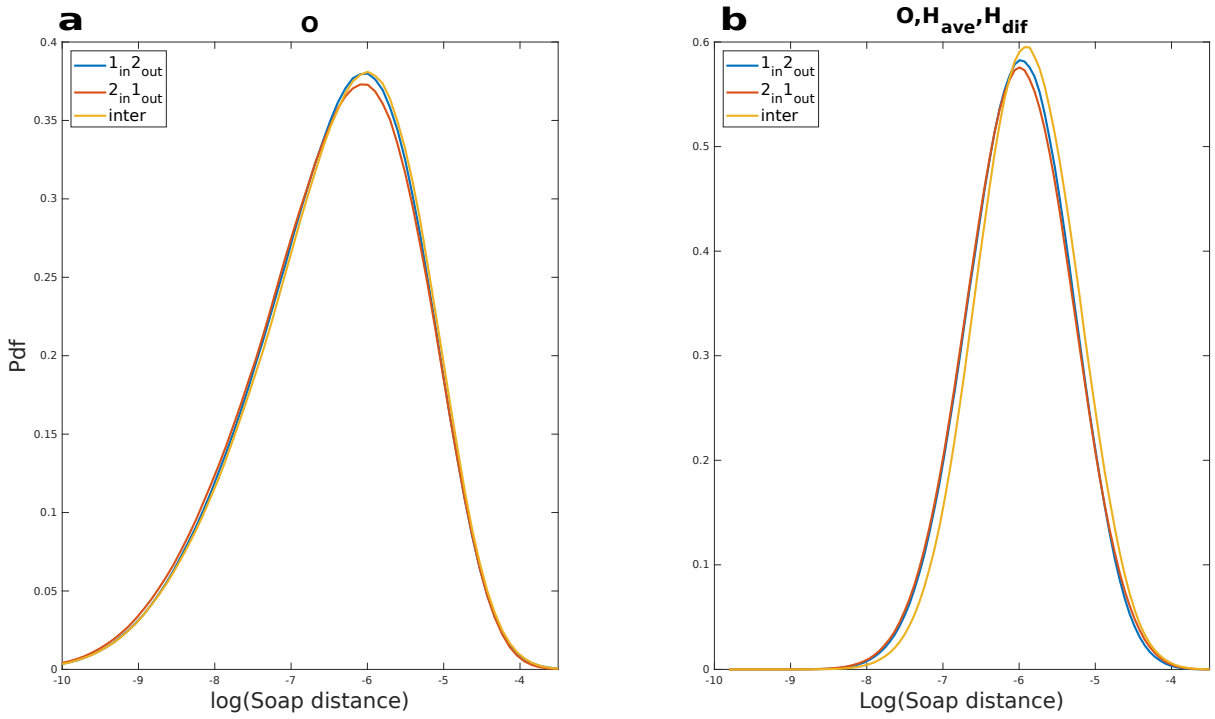

Figure S1: This figure shows the probability density estimates of the logarithm of SOAP distances between $1_{in}2_{out}$ environments(blue) , $2_{in}1_{out}$ environments(red) as well as distances between $1_{in}2_{out}$ and $2_{in}1_{out}$ environments(yellow). Panel a shows that there is a complete overlap between all estimates when only oxygen atoms are used in computing the SOAP distances. In panel b however, we observe that the three distributions exhibit bigger differences when hydrogen atoms are included in the analysis.

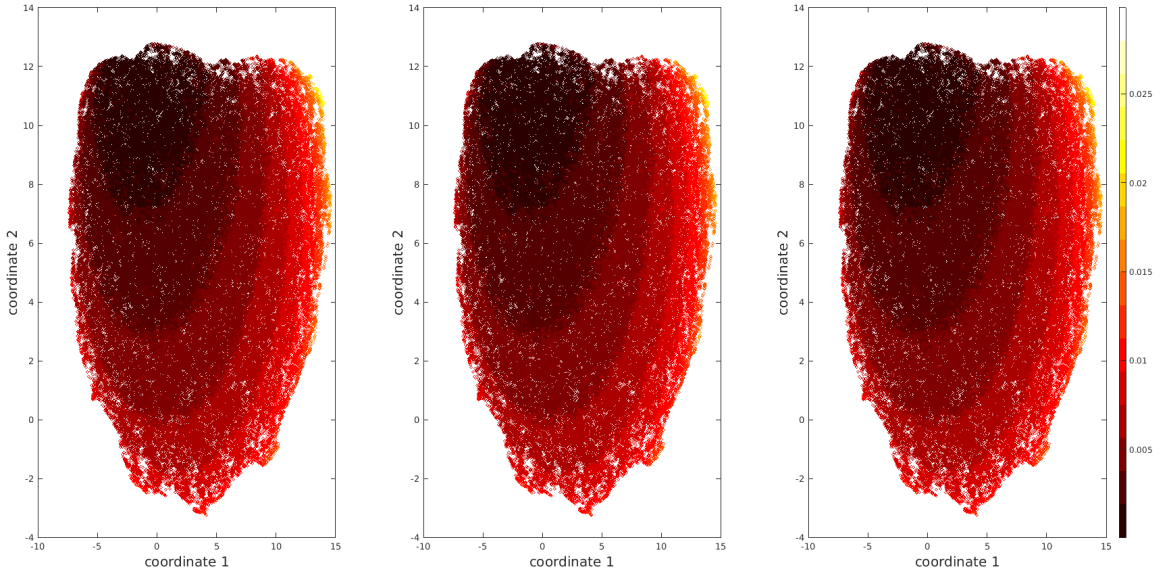

Figure S2: 2D UMAP projection of the environments for three datasets colored by d_{ice}.

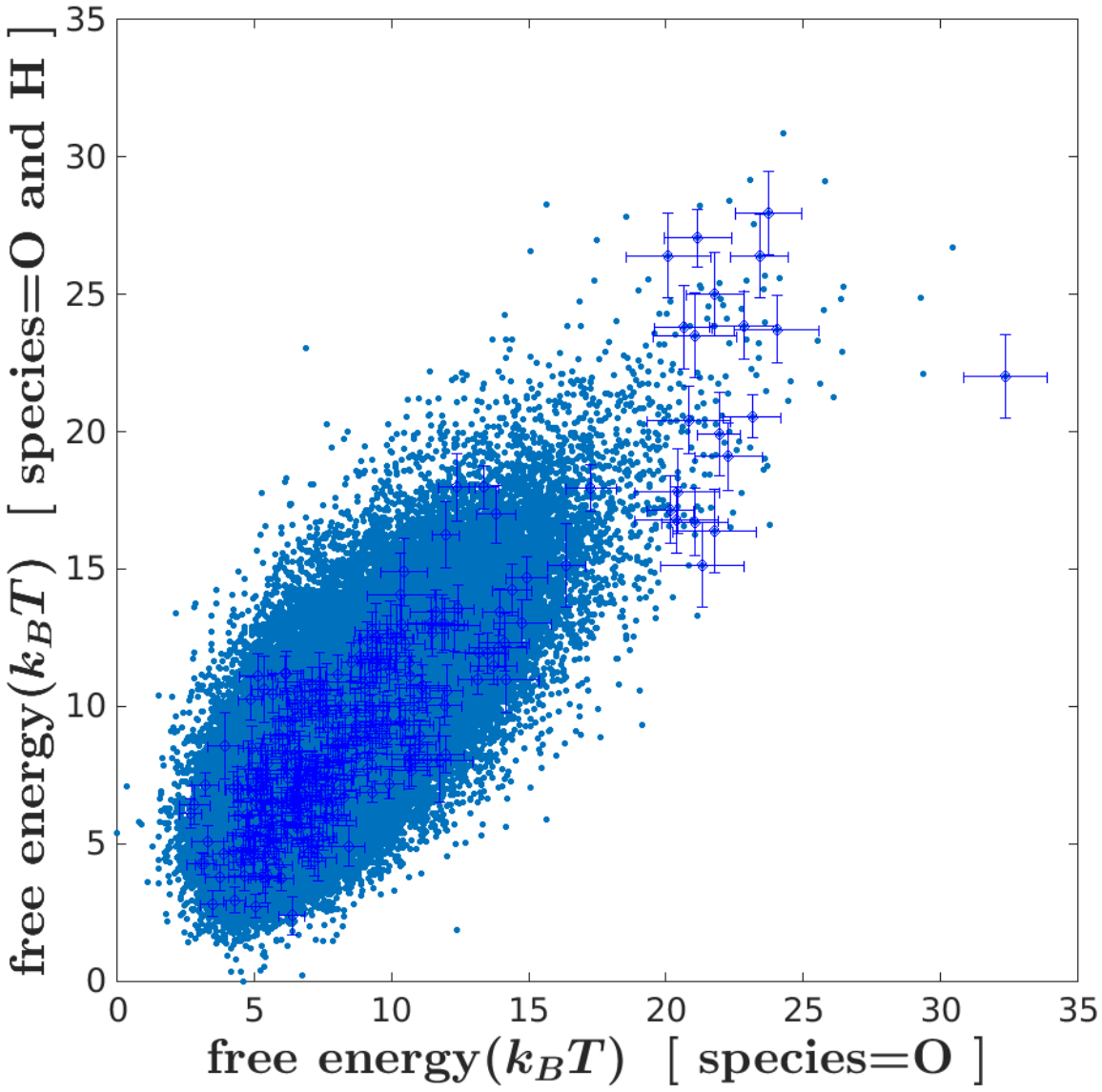

Figure S3: This figure which shows the scatter plot of the free energy values of $\vec{\mathbf{O}}$ environments versus ($\vec{\mathbf{O}}$, $\vec{\mathbf{H}}_{ave}$,$\vec{\mathbf{H}}_{dif}$). There is close to a linear relationship with a correlation coefficient of 0.7 and an RMSE between the two free energies of $\sim 2k_BT$.

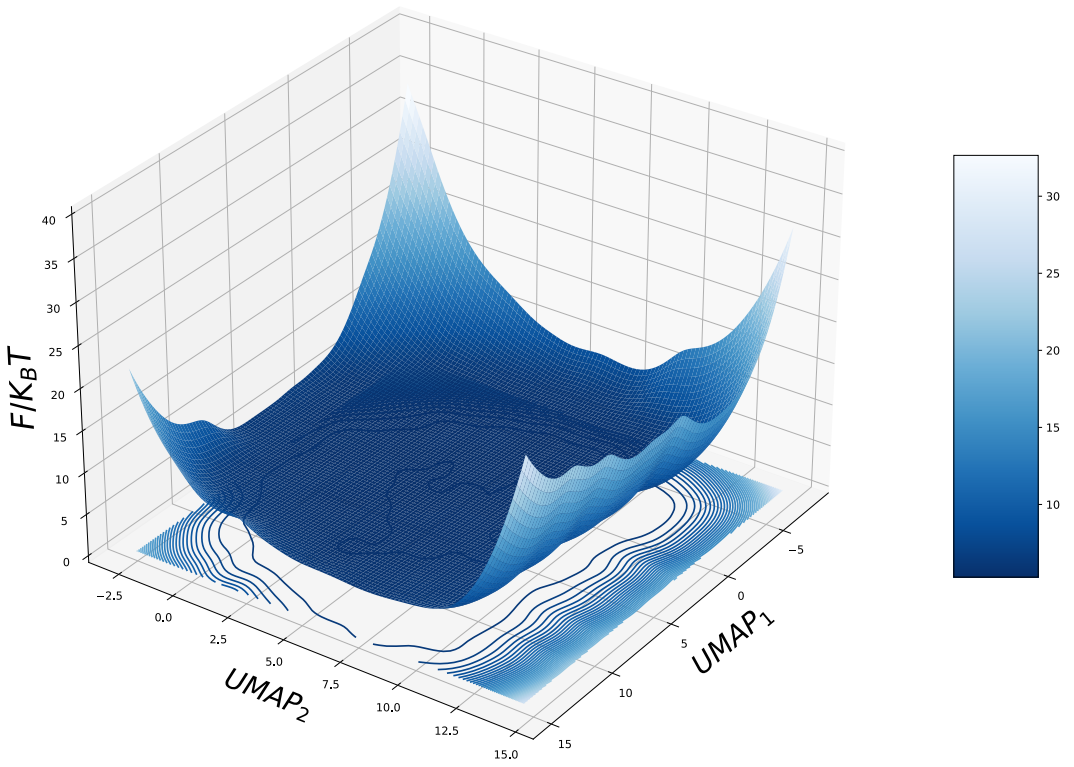

Figure S4: Free energy surface of MB-pol constructed in 2D UMAP manifold reveals a single basin without an appreciable barrier consistent with the results from TIP4P/2005 shown in the manuscript.

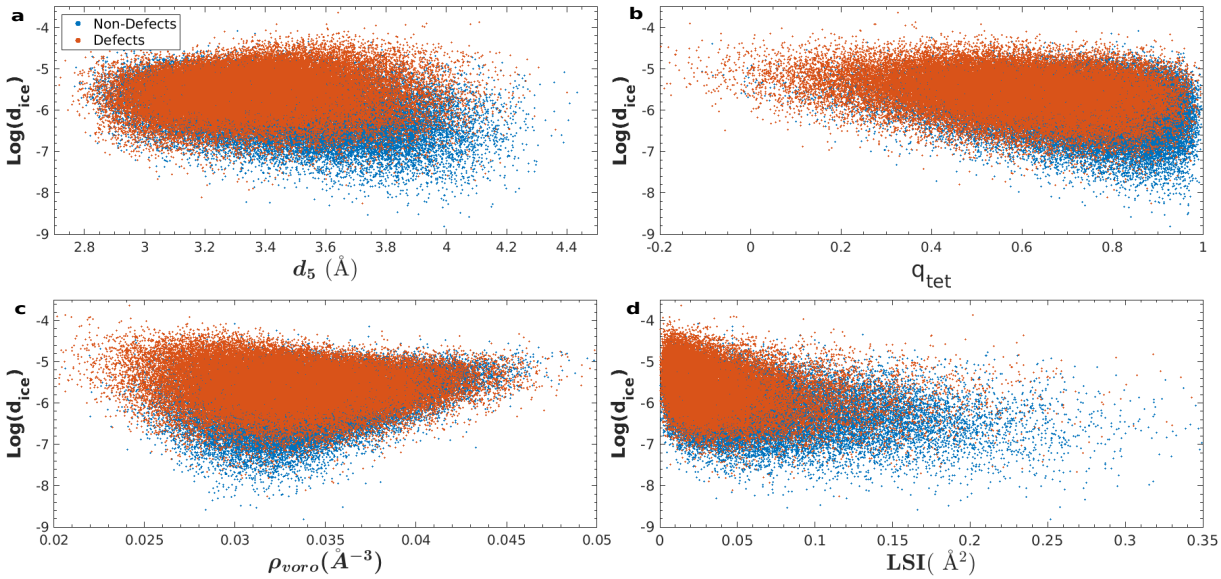

Figure S5: The panels a)-d) show the scatter plots of Log(d_{ice}) at 3.7 Å (including the hydrogen atom SOAP descriptors) versus the chemical-based collective variables for $q_{tet}, d_5, \rho_{voro}$ and LSI.

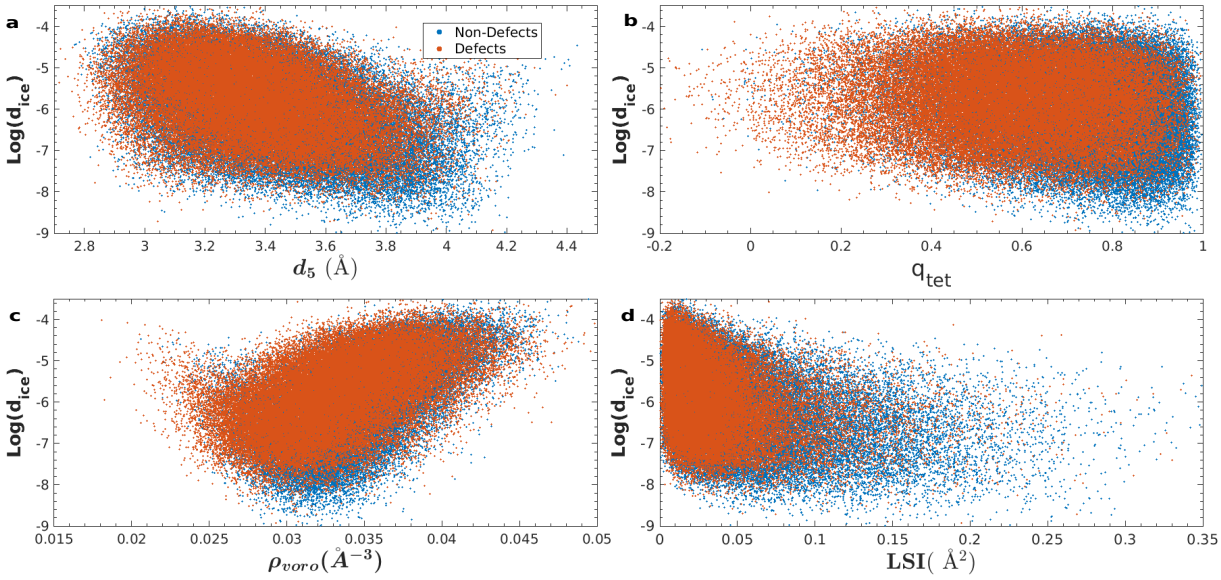

Figure S6: The panels a)-d) show the scatter plots of Log(d_{ice}) at 6.0 Å (including only the oxygen atoms for the SOAP descriptor) versus collective variables for $q_{tet}, d_5, \rho_{voro}$ and LSI.

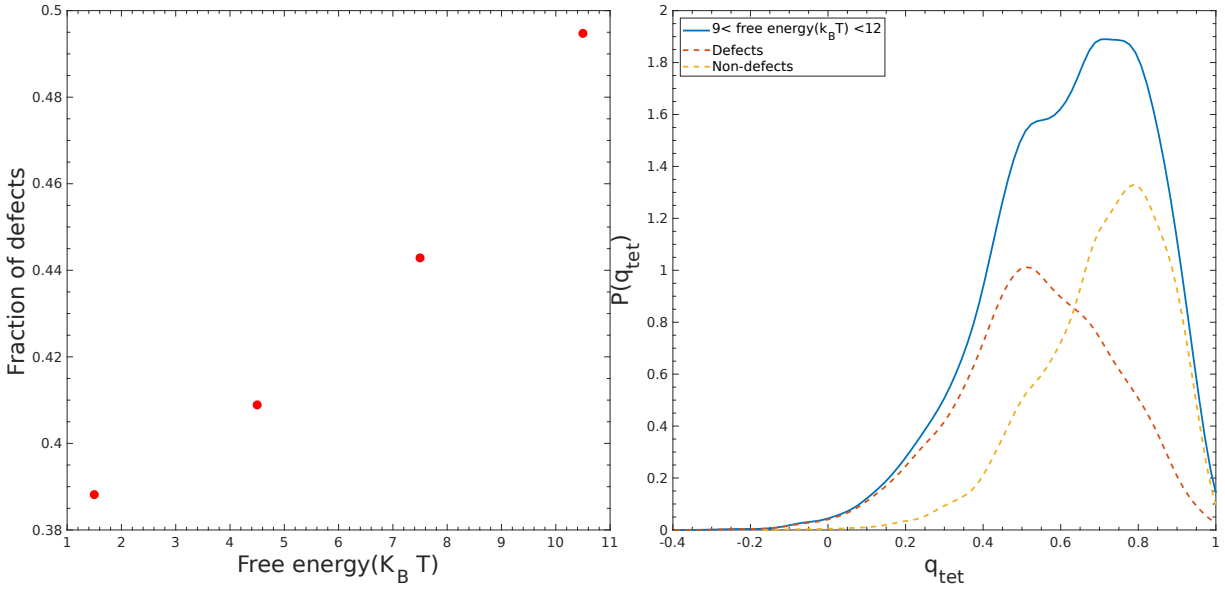

Figure S7: Left panel is the fraction of defects for $3k_BT$ cuts of the free energy. Right panel shows the q_{tet} distribution of points high in free energy. Also shown are the weighted q_{tet} distributions of points high in free energy restricted to defective(red) environments and non-defects(orange).

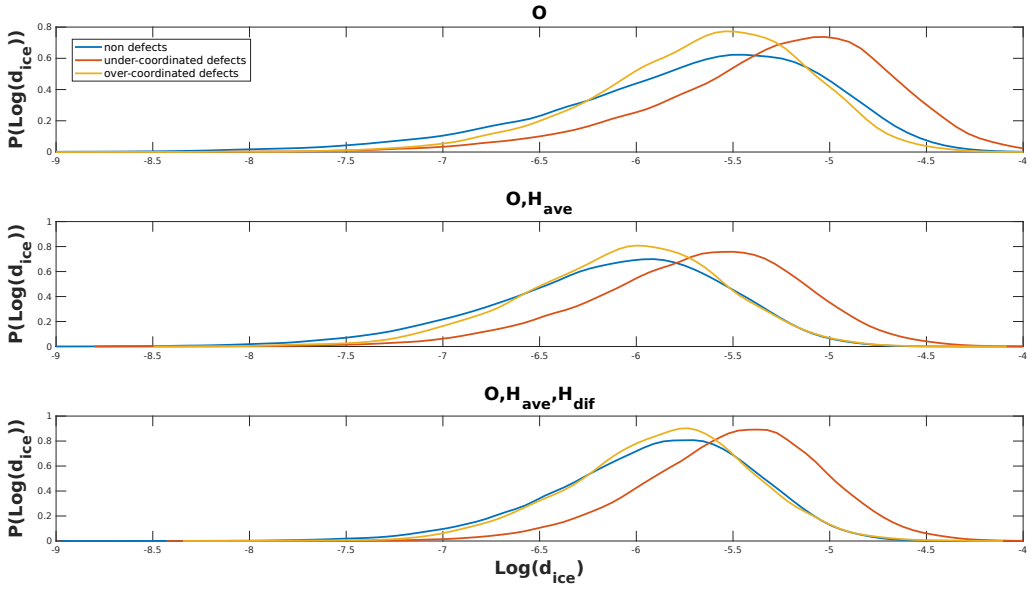

Figure S8: Probability density estimate of $\mathrm{Log}(d_{ice})$ for non-defects, under-coordinated defects and over-coordinated defects.

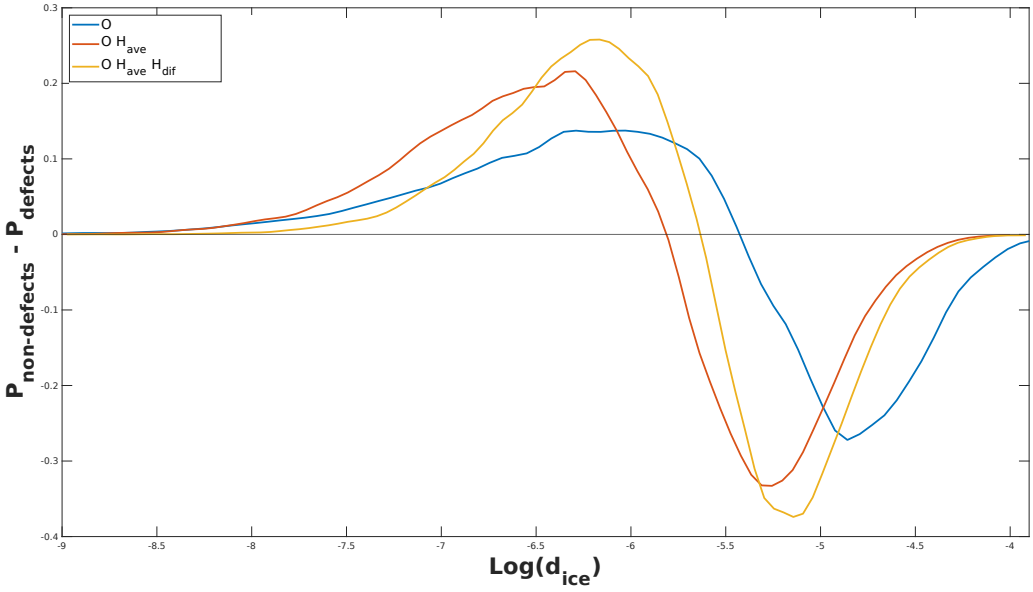

Figure S9: Figure shows the difference in distribution of Log(d_{ice}) of non-defects and defects for the three variations of the SOAP descriptors: ($\mathbf{O}$), ($\mathbf{OH}_{ave}$), ($\mathbf{O}, \mathbf{H}_{ave}, \mathbf{H}_{dif}$). The descriptor including both $\mathbf{H}_{ave}$ and $\mathbf{H}_{dif}$ is found to have the greatest difference between defects and non-defects.

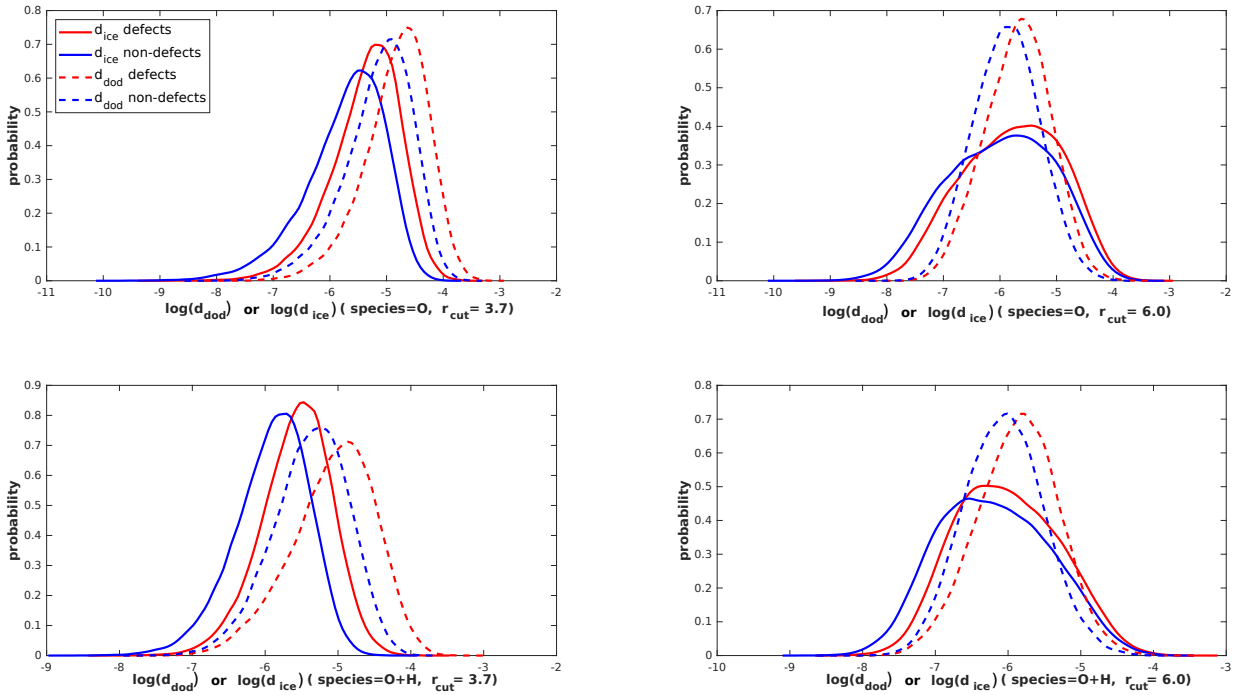

Figure S10: Probability density estimates of $\mathrm{Log}(d_{ice})$ and $\mathrm{Log}(d_{dod})$ restricted to defects and non-defects for radial cutoffs of 3.7 Å and 6.0 Å. The top panels are constructed using only oxygen atoms ($\mathbf{O}$) while bottom panels include the hydrogen atoms in computing the distance ($\mathbf{O}, \mathbf{H}_{ave}, \mathbf{H}_{dif}$).

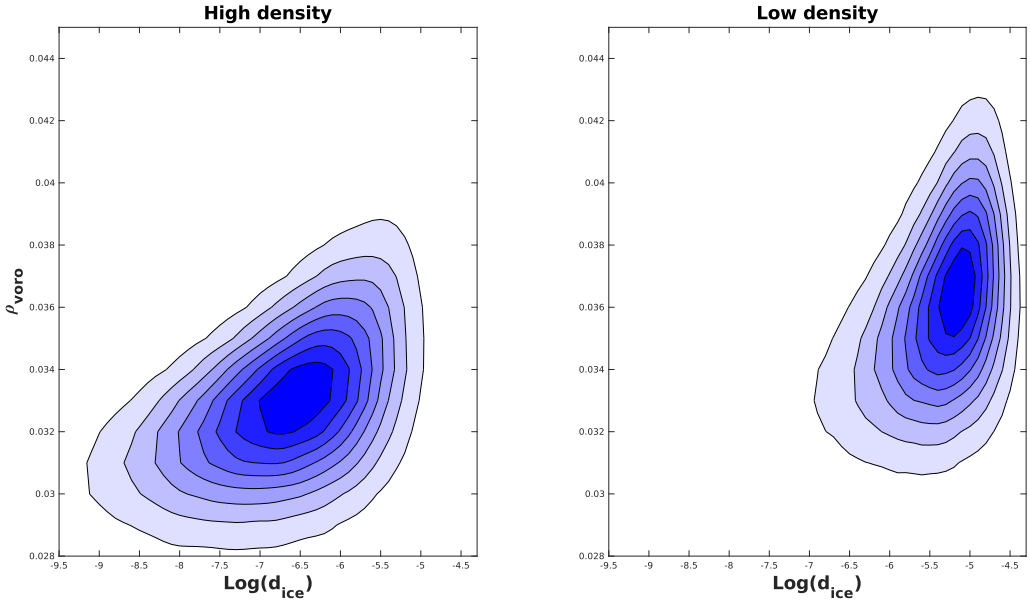

Figure S11: Density plot of $\log(d_{ice})$ versus ρ_{voro} for HD and LD environments of supercooled water.