SISSA
Scuola
Internazionale
Superiore di
Studi Avanzati

Physics Area - PhD course in
Theory and Numerical Simulation of Condensed Matter

Unsupervised Learning of the Structure and Dynamics of Liquid Water

Candidate:
Adu Offei-Danso

Advisors:
Ali Hassanali
Alex Rodriguez

Academic Year 2021-2022

Abstract

The microscopic description of the local structure of water remains an open challenge. Here, we adopt an agnostic approach to understanding water's hydrogen bond network using data harvested from molecular dynamics simulations of an empirical water model. A battery of state-of-the-art unsupervised data-science techniques is used to characterize the free energy landscape of water starting from encoding the water environment using local-atomic descriptors, through dimensionality reduction, and finally the use of advanced clustering techniques. Analysis of the free energy at ambient conditions was found to be consistent with a rough single basin and independent of the choice of the water model. We find that the fluctuations of the water network occur in a high-dimensional space which we characterize using a combination of both atomic descriptors and chemical-intuition-based coordinates. We demonstrate that a combination of both types of variables is needed in order to adequately capture the complexity of the fluctuations in the hydrogen bond network at different length scales both at room temperature and also close to the critical point of water. Our results provide a general framework for examining fluctuations in water under different conditions.

We also explore the collective nature of orientational fluctuations on the free energy landscape. Specifically, we develop an unsupervised protocol for identifying reorientational dynamics in liquid water. We show that large swings are more likely to occur higher up in the free energy landscape than smaller amplitude swings. We show that these orientational fluctuations are collective and occur in waves on the order of tens of picoseconds. These waves of large swings are found to correlate well with the fraction of defects as well as the fluctuations in local density.

Dedication

I am indebted to my supervisors Ali Hassanali , Alex Rodriguez for their patience, expert guidance, insight and support as well as embuing me with unsupervised/agnostic approach to physics.

It would not have been possible without the support of the wonderful team of collaborators Asja Jelic, Uriel Morzan, and Eddy Donkor from whom I learned important skills such as programming and soft skills like scientific communication.

I would also like to acknowledge ICTP and SISSA for the financial and computational support. For this I am particulary indebted to the head of the ICTP CMSP Prof. Rosario Fazio and the head of the PhD. programme at SISSA Prof. Giuseppe E. Santoro.

I am indebted to members ICTP condensed matter group with whom I engaged in very useful scientific discussion such as Prof. Nicola Seriani, Dr. Emiliano Poli, K. Jong, Narjes Ansari, Elham Goliaei, Fernando Iemini, F.M. Surace to name a few.

I would like to express my gratitude to Dr. Eric Abavare, Elliot Menkah and Prof. E Adei for teaching the fundamental principles and skills of condensed matter physics. In similar regard, I would would like to thank the staff of the ICTP Diploma programme.

I would like to thank my friends, Kelsey, Merab, Tram, Vicharit , Daniel, Miha, Martina, Maja, Natasa, Richmond and Claudia(to name a few) for emotional support.

I thank my family for finally supporting my career choice, and for providing me with moral support. I dedicate this thesis to the memory of my beloved Uncle, Kwaku Dua Oyinka, whose contagiously joyful and optimistic outlook on life, spurred me on into academia.

Contents

Chapter 1

Introduction

Water is perhaps the most common solvent for processes in physics, chemistry and biology[1, 2, 3]. Unlike simple liquids, water exhibits rather anomalous behavior along different regions of its phase diagram. For example, upon cooling, liquid water is characterized by a maximum in its density as well as a minimum in the compressibility[4, 5]. Despite long study, the structural and dynamical fingerprints underlying these anomalies remains an open area of study and has been the subject of lively debate and discussion from both experimental and theoretical fronts [6, 7, 8, 9, 10, 11, 12].

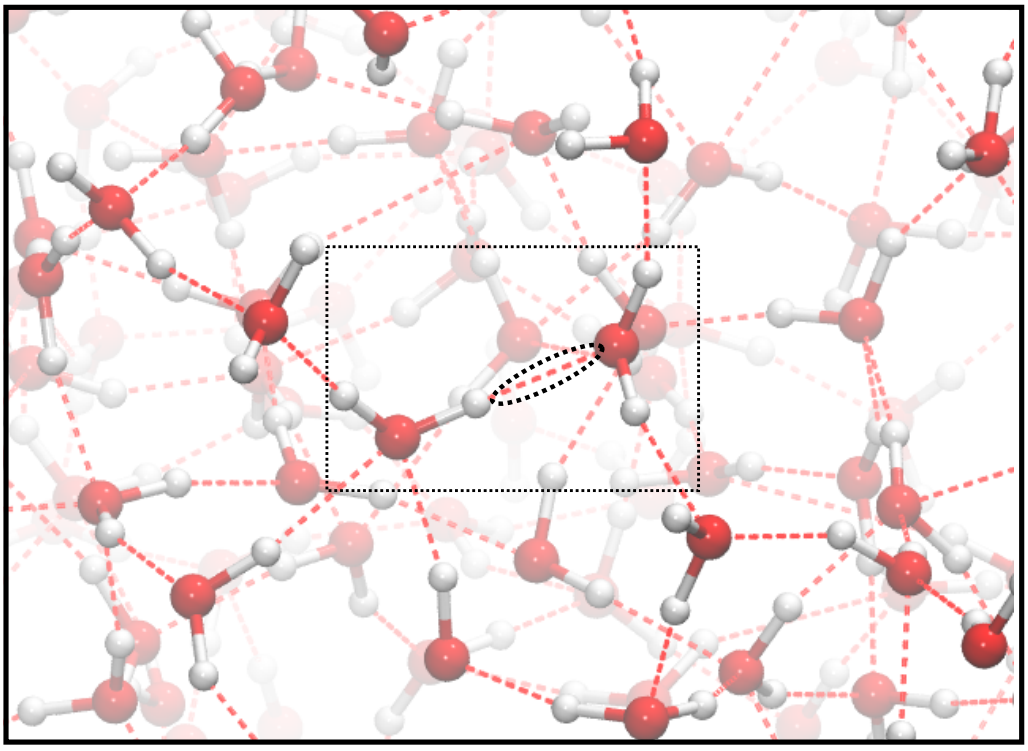


Figure 1.1: This figure illustrates the extended three-dimensional hydrogen bond network of water. The dashed red lines(red) correspond to hydrogen bonds between water molecules. The hydrogen bonds are seen to form an extended three dimensional pattern with most water molecules being connected to 4 neighboring molecules

Due to the difference in the electronegativities of the oxygen and hydrogen atom, water molecules are held together by hydrogen-bonds involving highly oriented dipole-dipole interactions. Due to the asymmetry of the position of the protons along the hydrogen bonds, these interactions are highly directed. Numerous sequential connections of these hydrogen bonds lead to the creation of water's 3-dimensional hydrogen bond network. Figure 1.1 illustrates this hydrogen bond network (HBN) for liquid water where the hydrogen bond between two water molecules is highlighted with the dashed oval. Although the HBN has been invoked to rationalize many of water's unique properties, the microscopic origins and nature of its fluctuations has been the source of numerous controversies [9, 6, 13, 14]. One of the central challenges here has been examining relevant hydrogen-bond patterns for a disordered system that occur on various spatial length scales[15, 12, 16, 17, 18, 19].

Almost a century ago, Wilhelm Roentgen provided one of the earliest attempts to describe different environments in water that could be used to explain its anomalous properties[20]. He speculated that water consists of two environments namely, a low-density form (LDL) and a high-density form (HDL). In this view, LDL is locally ordered and favoured by enthalpy, while HDL is disordered and entropically stabilized. Therefore, upon cooling, the fraction of HDL decreases while LDL increases. This simple picture allows both to explain the metastable behavior of supercooled water, whose origin will be in the fluctuations between these two states as well as the density maximum, as a shift between the equilibrium populations from a majority of HDL towards being dominated by LDL.

Numerous models of water have been put forward building on these original ideas suggested by Roentgen. These range from *mixture* models [21, 22, 13, 14, 23, 24], often interpreted as consisting of the two dominant LDL and HDL states, as well as continuous random-network models [25, 26]. In the latter, the local tetrahedral structure of water with some variation in the bond distances and angles, leads to the creation of local topological structures such as closed rings [27, 28, 29, 29, 30, 31, 32, 17, 33]. Several theoretical and simulation studies have shown that the changes in the topology of network are manifested across the phase diagram[34, 17, 15].

Four decades ago, Poole and co-workers conducted molecular dynamics simulations of a coarse grained model of water (ST2) and suggested the possibility of observing a second liquid-liquid critical point [35]. This phenomenon, has since then, been interpreted in terms of the existence of the HDL and LDL water environments [14, 12, 35] and confirming Roentgen's speculations. Very recently, this observation was confirmed by Sciortino, Debenedetti et. al using microsecond timescale atomistic molecular dynamics simulations of several realistic water models [36, 37, 38].

The existence of LDL and HDL water environments is commonly used to interpret the existence of heterogeneities in small angle scattering measurements [18, 19]. Specifically, these experiments show the presence of density fluctuations and inhomogeneities that occur on the nanometer length scale even at room temperature [18]. In this context, a common strategy that is adopted is to use molecular dynamics simulations performed at different thermodynamic state points and then conducting an inherent structure analysis where one quenches the system to zero-

Kelvin[39]. Thereafter, one examines the distribution of different types of order parameters or reaction coordinates designed to distinguish between the LDL and HDL structures [40, 23].

Underlying the techniques and approaches used to interpret liquid water in terms of the LDL and HDL states particularly at room temperature, involves two assumptions: firstly, it is not clear that the fluctuations relevant in the supercooled regime, translate to higher temperature and secondly, the directions along which these fluctuations occur may not necessarily be the same. The timescales associated with the possible heterogeneities in water has also been fervently discussed in the literature [41, 42, 43]. In this regard, there have been a plethora of chemical-intuition based coordinates that have been developed to distinguish between the LDL and HDL environments [44, 45, 46, 47]. Many of these variables build on the original notion set out by Roentgen which aim at identifying the difference between a locally more ordered vs disordered structure induced by the changes between the first and second solvation shells. Figure 1.2 reproduces a schematic illustration of the changes typically thought of in HDL and LDL environments taken from a previous work by Car and co-workers [48].

The vast majority of chemical-intuition based parameters involve projecting along a reduced set of dimensions and for the most part, require making *ad hoc* assumptions for example, on how many or which specific water molecules one uses in the first or second solvation shell (as seen in Figure 2.1). Furthermore, to really understand and quantify the fluctuations associated with the 3-dimensional HBN, the underlying free energy landscape in which it occurs needs to be examined in its correct embedding which is not known *a priori*. Indeed we will show in this thesis, that valuable information about the structural fluctuations in water can be lost when projecting in a low-dimensional space.

Besides the thermodynamics associated with the free energy landscape of liquid water, another important and complementary aspect that has also been the subject of many experimental and theoretical studies is its dynamics [49, 8]. Liquid water has a very rich dynamical spectrum as observed by Infra-Red (IR) [43], Raman[50] and TeraHertz (ThZ)[51] spectroscopies. Due to the directed interactions formed by the water dipoles as eluded to earlier, fluctuations of the HBN at the molecular scale involve hydrogen-bond breaking and formation where water molecules undergo reorientational motions [52, 49].

Until over two decades ago, the accepted interpretation for water reorientations was that it involved a slow diffusive process[53]. Laage and Hynes in 2006 used molecular dynamics simulations of an atomistic model of water and showed instead that water molecules undergo large amplitude jumps [49]. The angular jump mechanism as it is now referred to, has been used as a framework to understand changes in water re-orientational dynamics in a wide variety of contexts such as different electrolyte solutions and also near organic and inorganic interfaces[54, 55, 56]. Although this mechanism has been found in a wide variety of different water models[57, 58], a key underlying assumption has been that it is primarily a localized effect involving three water molecules as shown in the Figure 1.3 reproduced from Ref [49].

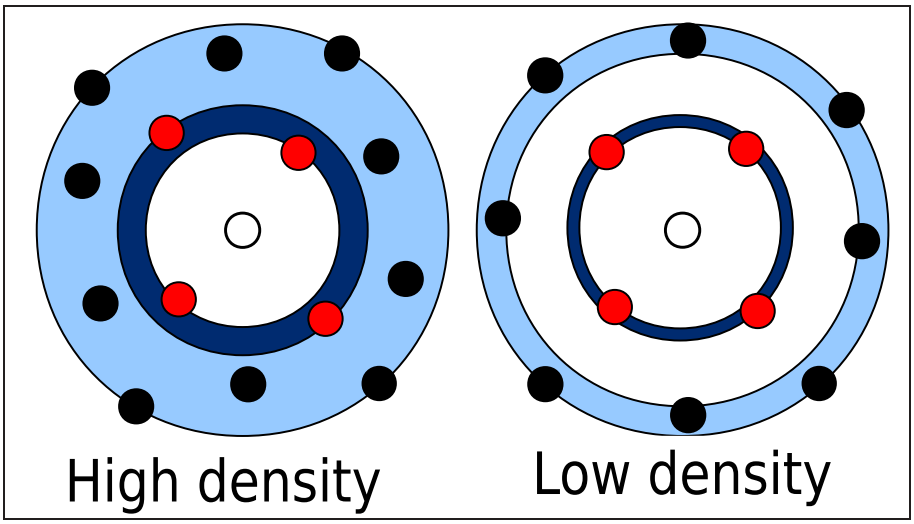


Figure 1.2: This figure illustrates high-density-like and locally disordered (left) vs. low-density-like and locally ordered (right) environments. The dark and light blue areas represent the first and second coordination shells around the central water molecule, respectively. The separation between the water molecules in the first and second coordination shells [48]

Since the water hydrogen bond network involves both density and orientational correlations that can extend to the length scale of ∼1nm [59, 19, 12], we postulate that the angular jump mechanism involves more collective behavior than previously expected. Furthermore, since water can be seen as a topological directed network with fluctuations in defects and non-defective water molecules, changes in water re-orientations would imply changes in directionality of the network. Thus, one might expect to see cooperative behavior [59].

In this thesis we revisit some of these commonly held assumptions regarding the fluctuations associated with the thermodynamical and dynamical properties of water. In particular, we bring in recently developed modern tools in data science[60, 61, 62] to examine these issues in an unsupervised manner. Specifically, a key overarching challenge that we address is to circumvent the bias of chemical intuition in describing the complexity of different environments and how this is manifested in both the thermodynamics and dynamics of liquid water.

The advent of machine learning techniques to chemistry has brought in many tools allowing for the construction of higher quality potentials as well as investigating the complex data of disordered liquids such as water with minimal chemical bias. In this thesis we re-visit some fundamental questions in the theory of aqueous systems armed with these modern tools: What are the relevant degrees of freedom needed

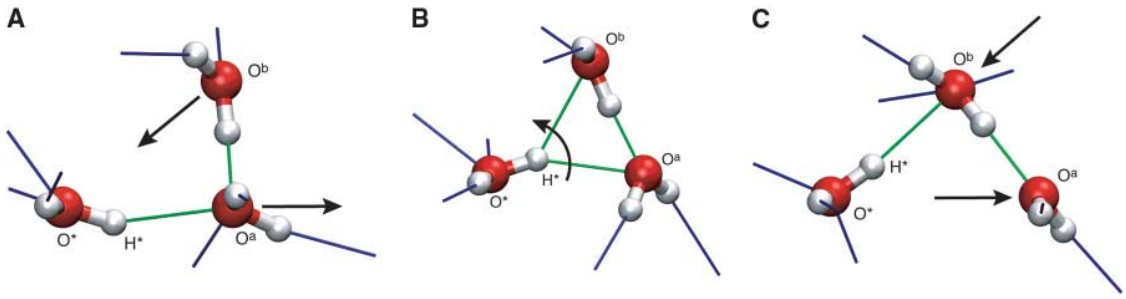


Figure 1.3: This figure from reference [49] illustrates the angular jump in which a water molecule(*) breaks a bond with neighboring molecule **a** and becomes bonded to another water **b** originally in the second shell. In panel A, the undercoordinated water is approaches the first shell of molecule(*) and while the overcoordinated water b moves away for O*. Panel B shows an intermediate stage of the flip between a and b where H* molecule is equidistant from the both both waters. In panel C the exchange has taken place and thermal fluctuations have resulted in all waters having the same hydrogen bond coordination number of 4

to describe the broad spatial and temporal range of fluctuations in the water? How much *information* can we extract from different components of the hydrogen bond network? What implications does this have on our understanding of the existence of LDL and HDL phases across the phase diagram?

In an effort to address these questions, the thesis tackles two topics: firstly, characterizing the complex high dimensional nature of the free energy landscape of water at room temperature and in the supercooled regime (Chapter 3) and secondly, revisiting the collective nature of the angular jump mechanism within this landscape (Chapter 4).

In Chapter 3, constructing the free energy landscape of water involves three important steps: i) describing local atomic environments that are not biased by chemical intuition, ii) estimating the intrinsic dimension of the water network and finally iii) clustering and free energy construction. A combination of both machine-learning (ML) and chemically inspired coordinates are used to understand the free energy landscape. Our key finding here is that liquid water is a high dimensional free energy landscape characterized by a rather broad minimum with small ripples arising from low barriers separating the different minima constituting very different structures. These results offer a much more nuanced perspective on the interpretation of water as a two-state picture which is currently invoked in many contexts. Interestingly, we find that a combination of both the ML based atomic descriptors as well as chemical-based parameters are required to accurately understand the fluctuations in the network.

Chapter 4 revisits the angular jump mechanism that was originally proposed by Laage and Hynes. We propose an unsupervised approach for automatically detecting angular swings in water at room temperature. This approach allows for probing the collective nature associated with the angular motions. We find that the fluctuations in water topology creating defects, come in waves on the timescale of several

picoseconds. By identifying correlations between the large angular jumps and both the ML based and chemical intuition based variables used in Chapter 3, we identify both density and topological fingerprints of the network that are key players in water's cooperative fluctuations.

Overall, the findings of this thesis point to a much richer physics and chemistry associated with the fluctuations of liquid water. These findings are also independent of the choice of the water model. The protocol we employ in Chapter 3 for constructing the free energy landscape of water provides a general framework by which the thermodynamics of water in different regions of the phase diagram as well as near biological interfaces, maybe explored in the future. The collective nature of the angular jump mechanism also opens up interesting questions on how this changes for example, upon supercooling and also near biologically relevant systems such as proteins, where correlations get more enhanced.

In summary, the thesis is organized in the following manner. We begin in Chapter 2 with an overview of all the methods used including a summary of molecular dynamics and the analysis techniques used to characterize water's free energy landscape. Within this Chapter we also try to highlight where possible, how the data-science methods we use go beyond the current state-of-the-art in the field. Chapter 3 discusses our results on the high-dimensional fluctuations in water while Chapter 4 we discuss both the method we developed for detecting angular fluctuations in an unsupervised manner and subsequently how it is used to extract the cooperative nature of water reorientational dynamics. Finally, we end in Chapter 5 with some conclusions and perspectives for future work.

Chapter 2

Methods

In this chapter, we review the basic theory behind the methods employed in the thesis. We will begin by briefly summarizing some of the techniques underlying classical molecular dynamics (MD) which allows for generating the water configurations. Subsequently, we will also discuss the various analysis techniques that have been employed, ranging from cataloging the chemical based parameters of the water environments, to modern protocols that allow for a multidimensional description of the underlying physics and chemistry using advanced data-science techniques.

2.1 Classical Molecular Dynamics (MD) Simulations

At the microscopic level, the physics of atoms and molecules should be studied with a quantum mechanical description of both the electrons and nuclei[63]. However, modeling this quantum nature can be prohibitively expensive and, therefore, several approximations need to be employed to describe atomistic systems[64]. A generally employed one is the Born-Oppenheimer approximation [65]; that is, due to the separation in timescales of the electrons and the nuclei, the former instantaneously adapt to the nuclear geometry, generating a force field in which the nuclei move. In most empirical classical molecular dynamics models an additional set of assumptions are made namely that the nuclei follow Newton's equations of motion and the forces between the nuclei can be parameterized in some manner thereby integrating out the electronic degrees of freedom[66].

MD has been widely used to study many different systems, ranging from biological systems [67, 68], liquids[69] and also inorganic materials[70]. In the following, we will summarize how the MD protocol is used to simulate the water systems considered in this thesis. Typically one begins with a set of initial positions of atoms and updates the positions each time step based on the forces acting on the system. As mentioned earlier, the rules governing the change of the atomic positions is dictated by Newton's equations of motion, yielding a second-order differential equation of the form:

$$\ddot{\vec{r}}_i = \frac{d\vec{v}_i}{dt} = \frac{\vec{F}_i}{m_i}, i = 1, .., N \tag{2.1}$$

where $\vec{r}_i$ is the three dimensional vector representing the position of atom i in a

system with N atoms. The force $\vec{F}_i$ on each atom in turn, depends on the coordinates of all the atoms in the system through a potential U in the following manner:

$$\vec{F}_i = -\frac{\partial U(\vec{r}_1, ..., \vec{r}_N)}{\partial \vec{r}_i} \tag{2.2}$$

In classical non-reactive force fields, the potential $U(\vec{r}_1, ..., \vec{r}_N)$ is usually split into bonded and non-bonded interactions. The bonded part takes into account the changes in the internal degrees of freedom of the molecules such as bond distances, angles, and dihedrals[66]. The non-bonded contributions are related to the electrostatic and van der Waals forces which typically involve pairwise interactions. Electrostatic forces are usually modeled by a Coulomb potential, while the van der Waals ones with a $6-12$ Lennard-Jones potential, leading to the following equation:

$$U(\vec{r}_1, ..., \vec{r}_n) = U_{bond} + \sum_{<i,j>} A_{i,j} \frac{q_i q_j}{r_{i,j}} + (\frac{C_{i,j}}{r_{i,j}^{12}} - \frac{B_{i,j}}{r_{i,j}^{6}}) \tag{2.3}$$

2.1.1 Water Models

The functional form for the force field and the respective parameters of the bonded and non-bonded interactions defines the water model. Due to the key role of water as a solvent in many atomistic systems, a lot of different water models have been developed [71, 72, 73]. Among these empirical potentials, some of the most popular potentials include TIP3P[71], SPC[72], SPC/E[74], TIP4P [71], TIP4P-Ew[75], TIP4P/ICE[36] and finally TIP4P/2005[73]. A critical distinguishing feature of these water models is the number of sites considered for computing the non-bonded interactions. Specifically, three-site models such as TIP3P, SPC and SPC/E have three interaction points corresponding to the three atoms of the water molecule, while four-site models (TIP4P, TIP4P-EW, TIP4P/ICE and TIP4P/2005) move the charge corresponding to the oxygen along the bisector of the HOH angle while maintaining four sites for the Lennard-Jones interactions. This somewhat *ad hoc* fix has been shown to improve the electrostatic potential around the water molecule which is then manifested in better reproducing thermodynamic properties of water across the phase diagram[73].

In this thesis, unless otherwise stated, most of the analysis is performed using the TIP4P/2005 model to explore the free energy landscape at room temperature in the supercooled regime. The orientational dynamics in Chapter 4 is studied using the SPC/E water model. In both cases, we use the rigid-body version of these models allowing for using larger integration time steps. The TIP4P/2005 water model on the other hand, has been shown to reproduce the condensed phase properties of water including the melting and vaporization point as well as the dielectric constant and structural properties across a wide temperature range between 100-600K[73]. Recently, the TIP4P/2005 water model was also shown to display the second critical point at $172 \pm 1\,\mathrm{K}$, $1861 \pm 9 bar$ [37] consistent with numerous previous theoretical and experimental proposals[76]. The SPC/E model consists in a modification of the original SPC water model by scaling the point charges to account for polarization in an effective manner. The SPC/E model has been shown to generate a diffusion constant of water that is consistent with experiments[74].

It is worth stressing that many-body polarization effects, known to be present in water, are not captured by pairwise simple force-field models [77] described above.

More accurate models take advantage of the many body expansion of the energy and polarizability to accurately deal with these effects[78]. Perhaps the most accurate model for neutral water is the MB-pol potential[78]. The principal idea here is that the binding energy of many water molecules can be written as a many-body expansion shown below:

$$\begin{aligned} E_{\text{bind}} &= E_{\text{1B}} + E_{\text{2B}} + E_{\text{3B}} + \cdots + E_{\text{NB}} \\ &= \sum_{n=1}^{N} E_{\text{nB}} \end{aligned} \tag{2.4}$$

In MB-pol, the one body interaction (E_{1B}) term is represented by the spectroscopically accurate potential energy surface developed by Partridge and Schwenke[79]. The two body interaction term (E_{2B}), that is, the interaction between two water molecules, is split into a short and long range contribution, with the short range term represented by a permutationally invariant polynomial that smoothly switches to zero once the separation between two water molecules becomes larger than a predetermined cutoff value. On the other hand, the long-range interactions arise from electrostatic contributions originating from interactions between permanent and induced moments as well as weak dispersion forces. The three-body term (E_{3B}) was also separated into a short and long range induction energy. The many-body interactions in the MB-pol water model was fitted to highly accurate quantum chemistry calculation at the CCSD(T) level[80]. The MB-pol force field accurately reproduces structural, dynamical and spectroscopic properties of water across the phase diagram[77] both in the bulk and also at interfaces[81]. Wherever necessary and possible, the predictions made from the empirical potentials we use (TIP4P/2005 and SPC/E) are compared with the MB-pol water model in order to validate our results.

2.1.2 Time Evolution of Molecular Dynamics

Upon obtaining the force on each atom at time t, we determine its position and velocity at some time $t + \Delta t$ by numerically integrating the Equation 2.1. The Δt is the timestep that is used in the simulation, the magnitude of which can affect the accuracy of the predicted positions and momenta. Many procedures exist to integrate Newton's equation of motion such as the Verlet[82], the velocity Verlet[83], and leapfrog algorithms[84]. For example, the Verlet integrator can be obtained by Taylor expanding the positions $r(t)$ forward and backward in time which subsequently yields the following equation allowing for predicting the positions of the particles forward in time:

$$\vec{r}(t + \Delta t) = \vec{r}(t) - \vec{r}(t - \Delta t) + \frac{\vec{a}(t)\Delta t^2}{2} + O(\Delta t^4) \tag{2.5}$$

Specific physical quantities such as kinetic energy require the accurate calculation of velocities and it is therefore useful to compute them on the fly. This is implemented both in the leapfrog and velocity-Verlet, the latter of which is shown below:

$$\vec{r}(t+\Delta t) = \vec{r}(t) + \vec{v}(t)\Delta t + \frac{\vec{a}(t)\Delta t^2}{2} \tag{2.6}$$

$$\vec{v}(t+\Delta t) = \vec{v}(t) + \frac{\vec{a}(t) + \vec{a}(t+\Delta t)}{2}\Delta t \tag{2.7}$$

$$\tag{2.8}$$

The use of rigid water models without considering the bonding interactions requires modifications to the equations of motion to maintain the internal water structure. The conservation of the interatomic distances in these models is enforced through the use of Lagrange multipliers that are exploited in procedures such as the SHAKE[85] and RATTLE[86] algorithms.

Up to now, the formalism explained allows to perform simulations in the microcanonical ensemble (NVE) in which the conserved quantities are the number of particles, the total volume of the system and the total energy. However, the thermodynamic conditions that are often of interest to compare with experiments are situations at constant temperature and pressure. There are several methods to modify the dynamics in order to satisfy these requirements and obtain configurations from the canonical (NVT) or isothermal-isobaric (NPT) ensembles. The temperature in MD simulations is usually controlled using thermostats which are numerical recipes that essentially modulate the exchange of energy with the velocities of the particles. Examples include the Anderson[87], Nose-Hoover[88, 89], or the velocity-rescaling thermostat[90]. Similarly, the pressure is maintained using a barostat such as Berendsen[91] or Parrinello-Rahman[92] where the volume of the system fluctuates. In our NVT or NPT simulations that were conducted for this thesis, we use the velocity-rescaling thermostat and Parrinello-Rahman barostat.

2.2 Descriptors for Water Environments

The final output of the molecular dynamics simulation is a trajectory, a time-ordered sequence of frames consisting of atomic coordinates. The set of atomic coordinates in a frame corresponds to a snapshot of the system at some instant in time. If one runs the simulation long enough, all relevant structural configurations will be sampled, provided there is no ergodicity breaking. At the same time, if one selects structures far apart in time, i.e., at times greater than the correlation time, we arrive at an independent identically distributed (i.i.d.) collection of configurations of the system.

This ensemble of configurations can be used to compute the thermodynamic properties of the system under study. In this work, we use the MD simulations of water to extract water configurations. More precisely, we will investigate the different *environments* around the water molecules to obtain a structure-based understanding of the macroscopic properties.

2.2.1 Chemical-Based Descriptors

The standard approach to the description of environments involves using chemical-intuition defined functions of atomic coordinates called collective variables(CV)[93].

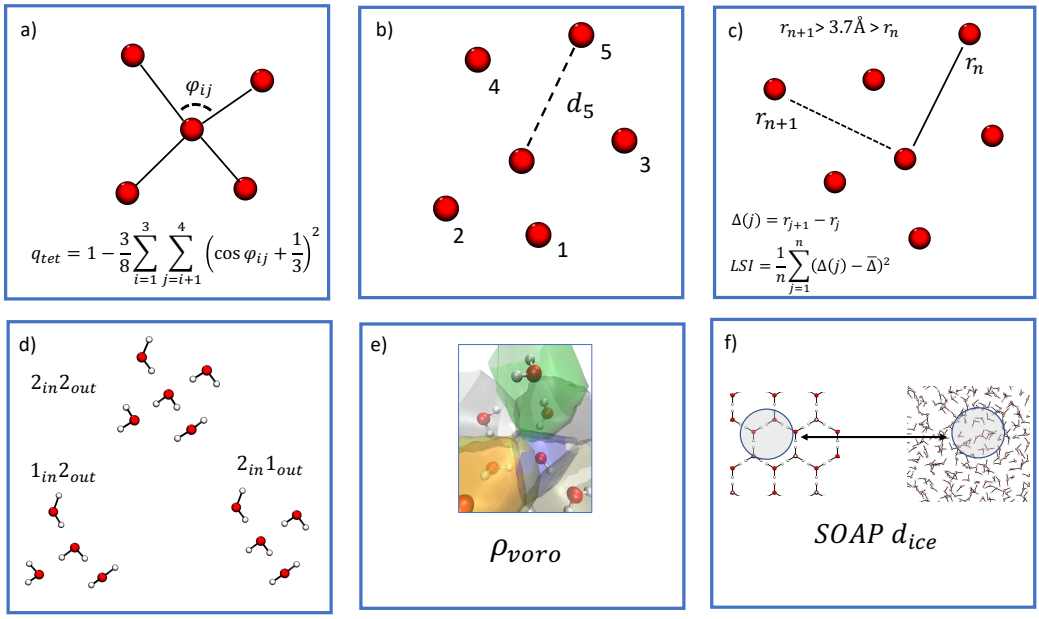


Figure 2.1: A visual schematic summarizing all the one-dimensional descriptors of water environments that were used in this work. a) q_{tet}, Tetrahedrality associated with the four nearest neighbors. b) d_5, distance from the fifth nearest neighbor. c) LSI, Local Structure Index measuring the extent or order/disorder between 1st and 2nd solvation shells. d)The number of neighbors for Topological defects in water. Shown are some examples including $2_{in}2_{out}$, $1_{in}2_{out}$ and $2_{in}1_{out}$ water molecules. The in, out subscript corresponds to the number of hydrogen bonds being accepted/donated respectively by the central water molecule which is the order parameter) ρ_{voro}, the local density estimated as the inverse of the Voronoi volumes corresponding to each water molecule. f) SOAP d_{ice} is the distance in the SOAP space from a given liquid water environment to the one present in hexagonal ice.

Their use is quite extensive in both the physics and chemistry community, for example, to study not only the structure of water under different thermodynamic conditions [47, 94, 9] but also water near different types of interfaces where the properties of water are known to change [95].

In the following, we will detail the essential CVs behind the descriptors used in this work (summarized in Figure 2.1). Tetrahedrality (q_{tet}, Fig. 2.1a) measures the similarity between the first layer environment and a tetrahedron. Its input values are the angles computed taking as vertex the oxygen atom of the central water and all the possible couples generated by the four nearest neighbors, yielding a value of 1 for a perfectly tetrahedral environment such as in hexagonal ice[96, 97]. More precisely, the q_{tet} is defined by the following equation:

$$q_{tet} = 1 - \frac{3}{8}\sum_{i=1}^{3}\sum_{j=i+1}^{4}(\cos(\phi_{i,j}) + \frac{1}{3})^2 \tag{2.9}$$

where $\phi_{i,j}$ is the angle formed by the lines joining the oxygen atom of the central water molecule its nearest neighbor oxygen atoms **i** and **j**. Consequently q_{tet} ranges [-3 1]

The d_5 parameter (Fig. 2.1b) is the distance from the fifth nearest neighbor to

the central atom and reflects the extent of separation between the first and second solvation shells. A larger value of d_5 is interpreted as being a more open and locally ordered structure[98]. The LSI parameter was designed to distinguish environments with well-separated first and second coordination shells from those that are more disordered. Consider the distances between the central water molecule (using the oxygen atoms) and the ith neighboring water molecule ordered in the following manner $r_1 < r_2 < ... < r_i < r_{i+1} < ... < r_n < 3.7\mathbf{\AA} < r_{n+1}$. The LSI is then defined as,

$$\mathbf{LSI} = \frac{1}{n}\sum_{j=1}^{n}(\Delta(j) - \bar{\Delta})^2 \tag{2.10}$$

where $\Delta(j) = r_j - r_{j+1}$ and $\bar{\Delta}$ corresponds to the difference in consecutive distances and the mean respectively. Fig. 2.1c) schematically illustrates the variables going into the LSI function, which is designed to probe the order in the first and second coordination shells by examining all neighboring water molecules within a cutoff of 3.7Å by looking at the O-O distances. One can build up an intuition of the physical meaning of this parameter by comparing two extreme cases. In the case of hexagonal ice, $\Delta(j)$ will be nearly equal to zero for all values except for the very last term which will lead to a large value of the LSI. Indeed, in this case the $LSI \simeq 0.3$. On the other hand in a random gas, $\Delta(j)$ should be much lower and the consequently lead to a very small value of LSI.

All these variables previously discussed do not explicitly include the hydrogen atoms. However, numerous previous theoretical studies have shown that there are essential correlations in the hydrogen-bond network created by the local topology, which involves directed hydrogen bonds between water molecules[99, 100, 101, 102]. Figure 2.1 d) shows some examples of topological defects that can be created from the canonical $2_{in}2_{out}$ (two hydrogen bond donors and two acceptors) water, which we also examine in our work. A geometrical definition is typically used to construct hydrogen bonds[103] In particular, to identify topological defects two molecules are hydrogen-bonded if their inter oxygen distance is less than 3.5 Å and the smallest angle between the O-O axis and the O-H axis is less than 30 degrees.

Although the previously described parameters are often interpreted in terms of high and low-density environments, this can only be inferred indirectly. Therefore, to obtain a more quantitative measure of density variations, we computed the Voronoi density (ρ_{voro}) as illustrated in Figure 2.1 e). ρ_{voro} is computed as the inverse of the Voronoi-volume associated with a water molecule which is the sum of the volume of the oxygen and two hydrogen atoms [41, 104, 105]. This volume is found by constructing surfaces equidistant between neighbouring atoms. This Voronoi partitioning is closely linked to the Wigner-Seitz cell in solid-state physics[106, 107] and provides a very elegant and robust manner in which to partition 3-d space.

Finally, Figure 2.1 f) shows the last one-dimensional descriptor that we used in this work, namely the SOAP distance from a hexagonal ice structure. This variable quantifies how different a local water environment in liquid water is from a water molecule obeying the ice rules in the ice lattice. Depending on the choice of the various SOAP parameters, one can generate a wide variety of distance measures.

While the collective variables provide a very physical understanding of atomic systems, they have two main limitations: 1) since they are low dimensional descriptors, there is no guarantee that they capture all the relevant aspects of the

complexity of environments due to possible information loss and, 2) these variables are typically chosen through some form of chemical intuition based on visualising structures and therefore involve either implicitly or explicitly, human bias. In the next section, we will provide the basis for an agnostic multidimensional description of water environments.

2.2.2 Machine Learning Based Descriptors

The second class of descriptors aims to represent the atomic environments directly from the atomic coordinates. The main challenge for obtaining this is that in order to properly reflect the physics of the system, they must be invariant to translations, rotations, and permutation of the atoms. More explicitly, if one translates the whole environment, or rotates it, or changes the order of the atoms in the coordinates, the descriptors must remain the same. The manner in which to achieve this fall broadly into two categories namely graph-based versus density-based descriptors.

Graph-based methods such as Coulomb[108] and Gaussian overlap matrices[109] encode structural features of an environment by constructing matrices whose elements are functions of pairwise distances between molecules. While such representations are translationally and rotationally invariant, one must carry out a further step to make them permutationally invariant. We can achieve permutational invariance either by sorting elements according to the magnitude or using the spectrum. The spectrum of a matrix is invariant to the permutation of indices but suffers from a problem of non-uniqueness. In other words, very different graphs can possess the same spectrum. Sorting matrix elements by magnitude resolves this problem. However, small changes in distances between molecules can result in significant changes of graph matrix elements and makes it inconvenient for systems with substantial vibrations such as liquids. At the beginning of this work, we attempted to generalize a graph-based descriptor developed by Hamm and co-workers [8] to identify states in water. However, very small vibrations of the hydrogen bonds were found to lead to large changes in the descriptor, a feature that was not desirable.

Density-based descriptors such as the smooth overlap of atomic positions (SOAP)[110], Atomic Cluster Expansion ACE[111], and Behler-Parrinello[112] symmetry functions begin by building a three-dimensional density.

In this thesis we use the SOAP descriptors to describe the fluctuations in liquid water. The SOAP descriptor has in the last couple of years been successfully applied in various contexts such as characterizing hydrogen bond networks[113], identification and prediction of inorganic crystals[114] and finally also determining fingerprints in various biological systems[115]. Very recently, Chen et al. [116] and Pavan et al. [113], have used averaged SOAP descriptors to characterise the similarity of liquid water environments to phases of ice. In the following, we describe the theoretical formalism underlying the construction of the SOAP descriptors.

Let χ be an atomic environment consisting of all molecules within a certain spherical volume of radius r_{cut} around a central water. Then for a particular atomic species Z_i, one begins by encoding the local environment of χ in density constructed as a sum of Gaussian functions with variance σ^2 centered on each of the neighbors of a central atom including the central atom itself:

$$\rho^{Z_i}(\mathbf{r}) = \sum_{j \in \chi} \exp\left(\frac{-\left|\mathbf{r} - \mathbf{r_j}\right|^2}{2\sigma^2}\right) \tag{2.11}$$

This atomic neighbour density can be expanded in a basis of radial basis functions and spherical harmonics as illustrated below:

$$\rho^{Z_i}(\mathbf{r}) \approx \sum_{n=0}^{nmax} \sum_{l=0}^{lmax} \sum_{m=-l}^{l} c_{nlm}^{Z_i} g_n(r) Y_{lm}(\theta, \phi) \tag{2.12}$$

where the $c_{nlm}^{Z_i}(\mathbf{r})$ are the coefficients. Given $\rho^{Z_i}(\mathbf{r})$ one can obtain the coefficients as

$$c_{nlm}^{Z_i} = \iiint_{\mathcal{R}^3} \mathrm{d}V g_n(r) Y_{lm}(\theta, \phi) \rho^{Z_i}(\mathbf{r}). \tag{2.13}$$

The number of coefficients of the basis functions one chooses to compute, is bounded by the number of radial basis functions $nmax$ and that of the angular basis functions $lmax$. The parameter $rcut$ identifies all molecules within some radial cutoff of the central atom. One can then define a rotationally invariant power spectrum as

$$p_{nn'l}^{Z_1 Z_2} = \pi \sqrt{\frac{8}{2l+1}} \sum_{m} {c_{nlm}^{Z_1}}^* c_{n'lm}^{Z_2} \tag{2.14}$$

By accumulating the elements of the power spectrum into a vector $\mathbf{p}$, the distance between two environments χ and χ' is related to the SOAP kernel by the following expression.

$$d(\chi, \chi') = 1 - K^{\mathrm{SOAP}}(\mathbf{p}, \mathbf{p}') \tag{2.15}$$

where,

$$K^{\mathrm{SOAP}}(\mathbf{p}, \mathbf{p}') = \left(\frac{\mathbf{p} \cdot \mathbf{p}'}{\sqrt{\mathbf{p} \cdot \mathbf{p}\ \mathbf{p}' \cdot \mathbf{p}'}}\right) \tag{2.16}$$

2.3 Data Science Protocol

The machine learning-based approaches described in the previous chapter provides a very high-dimensional representation of the atomic environment. However, making sense of these representations and extracting useful information remains a difficult task. For instance, to compute a free energy profile, one needs to perform a density estimation (traditionally a histogram) and then take the logarithm of this estimate as the free energy in k_BT units. However, as the dimension of the data increases, this procedure is destined to fail due to the increasing number of empty bins. This is one manifestation of the so-called *curse of dimensionality*[117] namely that the number of data points needed for obtaining information of similar quality increases with the power of D, where D is the dimension of the data.

As traditional methods are not able to deal with these high-dimensional descriptors, machine-learning techniques developed with this target are earning considerable attention in this field. In this thesis, we used a series of *unsupervised* machine learning techniques which aims to uncover the structure of the data associated with liquid water using minimal human intervention/chemical bias[118].

2.3.1 Intrinsic Dimension (ID)

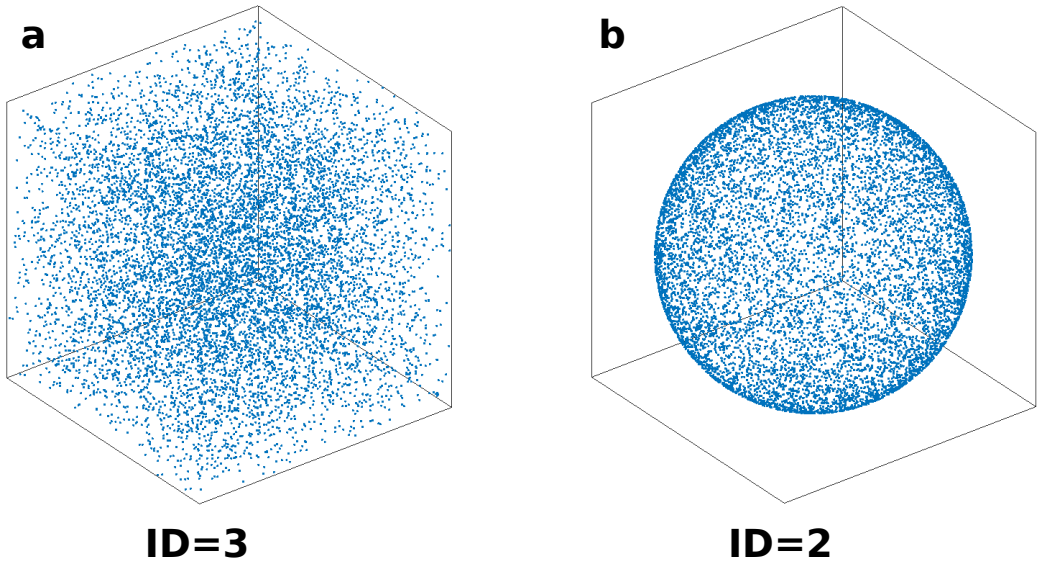


Figure 2.2: A visual schematic two three dimensional datasets with different intrinsic dimensionalities. Panel a is sampled from a uniform distribution and with an intrinsic dimension of 3. Panel b shows a dataset with points on the surface of a sphere having an intrinsic dimension of 2.

As mentioned earlier, the *curse of dimensionality* makes it very challenging to extract meaningful information in the data as the dimension of the system increases. However, machine learning techniques have been shown to be effective even when applied to high dimensional data. How is this possible? The answer is that, while the dimension of the data is formally high, correlations between the different variables describing each data point imply that the system of interest likely resides

(approximately) in a landscape with lower *intrinsic dimension* (ID). More formally, the ID corresponds to the dimension of the manifold in which the data lies which is typically much lower than the embedding dimension. To see this is in a very simple example, taking a set of points in three dimensions that are randomly distributed would yield an ID of 3. However, correlations between the coordinates could result in the data points lying only on the sphere's surface, which would, in turn, lead to an ID with reduced dimension of 2 (see Fig. 2.2).

Besides its importance in understanding the content and structure of information in data science, the ID is a key player in understanding complexity in physics. On the one side, due to its deep relationship with the extent of correlations in physical systems, it can be used to unveil both classical[119] and quantum [120] phase transitions. On the other hand, the ID is a critical unknown parameter needed for computing the free energies in high dimensions without using collective variables[61]. Later in this section we will discuss the free energy extraction techniques in more detail. We begin first with the methodology we employed to determine the ID.

The methods for estimating the intrinsic dimension fall broadly into three categories namely fractal, projection and finally nearest neighbor methods. Fractal methods estimate the ID by counting the observed points in a neighborhood around a specific point and estimating how this count scales with increasing distance [121, 122]. A pitfall of these methods is that they typically require the number of points to be exponential in the dimension to get a reasonable estimate.

Projection methods such as principal component analysis (PCA) and multidimensional scaling (MDS) search for a subspace to project the data in by minimizing a projection error[123, 124, 125]. For example, in PCA, this involves diagonalizing a covariance matrix and counting the number of dominant eigenvalues as an ID. These techniques work well when the data lie on a hyperplane so that there is typically a gap in the spectrum. However, this does not work well on twisted manifolds. Other methods, like ISOMAP[126], kernel PCA[123], t-SNE[127] or UMAP[128] are designed to deal with non-linearity in the data. However, some of these methods do not provide a manner to estimate the ID. In addition, those that do have a spectrum do not present a clear gap in many practical applications, which is needed to infer the ID.

Nearest-Neighbors like MLE[129] or DANCO[130] methods assume local uniformity and infer the ID based on statistics of nearest neighbor distances. In this work, we made use of a recently developed technique, namely the Two-NN estimator[62], which estimates the ID based on information of the first and second nearest neighbor of data points. The method has been successfully applied in studying different molecular systems[131, 132, 133]. The main ideas and derivation are shown next.

Consider a data set of **n** D-dimensional vectors in $X_1, ..., X_n$ that are independently and identically (IID) distributed and selected from some probability distribution. For a given point **i**, let $r_{i,l}$ correspond to the distance to the lth nearest neighbor from **i**. Suppose that our data lies on a lower-dimensional manifold so that the intrinsic dimensionality (d) is less than D. If the data-set is locally uniform up to the second nearest neighbor, then the volume of the shell between the first l nearest neighbor and the l-1 nearest neighbor($\Delta v_{i,l}$) for $l \in \{1, 2\}$ is given by the following:

$$\Delta v_{i,l} = \omega_d (r_l^d - r_{l-1}^d) \tag{2.17}$$

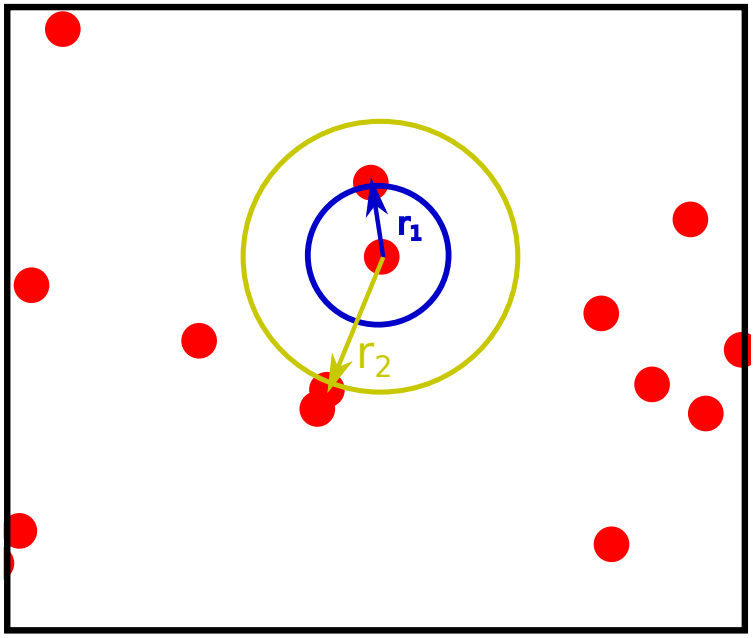


Figure 2.3: This figure shows the first and second nearest neighbors r_1 and r_2 needed for computing the ID

where $\omega_d = \frac{\pi^{\frac{d}{2}}}{\Gamma(d+\frac{1}{2})}$ is the volume of a unit d-dimensional hypersphere. If the density is locally constant around a point i, all the $\Delta v_{i,l}$ are independently drawn from an exponential distribution with rate equal to the density

$$p(v_{i,l} \in [v, v+dv]) = \rho e^{-\rho v} dv \tag{2.18}$$

Letting $R = \frac{\Delta v_{i,2}}{\Delta v_{i,1}}$, this yields the following equations,

$$P(R \in [R, R+dR]) = \rho^2 e^{-\rho(v_{i,1}+v_{i,2})} dv_1 dv_2 \delta(R - \frac{v_{i,2}}{v_{i,1}}) \tag{2.19}$$

$$p(R) = \rho^2 e^{-\rho(v_{i,1}+v_{i,2})} dv_1 dv_2 \delta(R - \frac{v_{i,2}}{v_{i,1}}) = \frac{1}{(R+1)^2} \tag{2.20}$$

The ratio of the second nearest neighbor and first nearest neighbor ($\mu = \frac{r_2}{r_1}$, see Fig. 2.3), can easily be related to R by the expression $R+1 = \mu^d$ so the probability density of μ becomes

$$p(\mu) = p(R(\mu))\frac{dR}{d\mu} = \frac{d}{\mu^{d+1}} \tag{2.21}$$

A very important ingredient in our further derivation involves the use of Log-likelihood/Maximum likelihood estimation. The goal of this is the estimation of parameters of an assumed probability distribution, given some observed data and is achieved by maximizing a likelihood function so that, under the assumed statistical model, the observed data is most probable. In practice this typically done by finding maxima in the logarithm of the likelihood function.

Considering the values of μ for **N** independent points in our dataset $\{\mu_i\}_{i=1}^N$, the probability of having observed this set of values is given by the expression below

$$p(\{\mu_i\}_{i=1}^{N}) = \frac{d^n}{\prod_{i=1}^{n} \mu_i^{d+1}} \tag{2.22}$$

The log-likelihood estimate of the intrinsic dimension can then be found by minimizing the logarithm of this probability with respect to d which finally yields the following estimator of the intrinsic dimension:

$$d = \frac{N}{\sum_i \log(\mu_i)} \tag{2.23}$$

It is worth noting that since this derivation relies on the statistics of only the first and second nearest neighbor, the only requirement is local uniformity up to the second nearest neighbor. This is a rather mild requirement that one expects to be fulfilled in many data sets. Consequently Two-NN is thus able to deal with systems characterized by density heterogeneities as well as being more robust to noise.

Using the SOAP distances, we estimated the ID of the environment around a water molecule. Ultimately, the physical significance of the ID is that it corresponds to the minimum number of independent order parameters or reaction coordinates required to describe, in our case, the environment around a water molecule. In this way, one can also quantify using the ID the amount of information that is gained or lost when including different variables[134].

2.3.2 High Dimensional Free Energy

The next step in our analysis involves computing a high dimensional probability density function in the space of atomic coordinates. This probability distribution is directly related with the free energy, which is the negative of the logarithm of this distribution in units of k_BT. The importance of the free energy in understanding of molecular systems cannot be overstated. Differences in the free energy between minima of states or meta-stable states can be directly related to equilibrium constants while the barrier heights control the dynamical rates of processes between different states. Both these quantities can be directly related to experimental observables.

Several approaches have been developed to compute the free energy [135, 136, 137]. One of the most popular approaches which has been hinted at above, involves using collective variables (CV), namely human-defined functions of the atomic coordinates. These are often constructed based on chemical intuition and thus the CVs often provide deep insight in a very efficient way (with a small amount of data). Unfortunately, the use of a single (or a few) human-designed variable(s) has two main drawbacks: 1) it corresponds to an uncontrolled dimensionality reduction, often lower than the intrinsic dimension, therefore leading to some information loss and 2), constructing a physically correct CV is highly non-trivial and subject to a lot of human intervention. The projection methods described earlier provide a possible alternative to limit the human intervention needed in constructing appropriate CVs[93]. However, these methods require both careful selection of various parameters and involve dimensionality reduction.

As stated above, whether one employs a human designed collective variable or an automatic projection method, the probability density needs to be estimated in order to extract the free energy. Density estimation methods can be divided into parametric and non-parametric methods. Parametric methods such as Gaussian

mixture models (GMM)[138] make assumptions on the functional form of the underlying probability distribution and attempt to fit parameters to the data. However, these methods are often criticized as the number of clusters and the initial parameters affect the final analysis. Non-parametric methods do not make assumptions on the functional form of the density. A classical example of this approach, is the construction of histograms[139], in which the data space is divided into bins and the probability density function at each point is estimated by counting the number of data points within its corresponding bin.

More elaborated non-parametric methods are the k-NN estimator[140], in which the density is estimated as k times the inverse of the volume occupied by the k-nearest neighbors of this point. An alternative to this is the kernel density estimation, in which a kernel function (usually a Gaussian although other functions can be used) is placed at each data point and the total density at a given point is the average value of all these kernel functions. Although these methods are referred to as non-parametric, each point requires a choice of a threshold of influence. In the case of the kernel density estimator, this is the bandwidth, while in k-NN it is the choice of the k value. Furthermore, these techniques do not work well for data sets with large heterogeneities in the density. All these methods however, suffer from the curse of dimensionality mentioned earlier and therefore the quality of the probability estimates significantly decreases with increased dimension.

In this work, we employ a recently developed Point Adaptive K-nearest estimator(PAK)[61] that avoids the need for any projection and has been used to study a wide variety of complex molecular systems[132, 133, 141]. In brief, the method uses the ID as a parameter to construct a point-dependent density (ρ_i). The derivation for this method is outlined below.

The log-likelihood function of the ρ given the observation of the k-nearest neighbor distances from point i is:

$$\ell(\rho|\{v_{i,l}\}_{l<k}) = k\log(\rho) - \rho\sum_{l=1}^{k} v_{i,l} = k\log(\rho) - \rho V_{i,K} \tag{2.24}$$

where $v_{i,l}$ is the volume of the hyper-spherical shell centered on i and enclosed between neighbors $l-1$ and l. Therefore, $V_{i,k} = \sum_{l=1}^{k} v_{i,l}$ is the total volume of the hyper-sphere with a radius equal to the distance from i to its k nearest neighbor. By maximizing this function with respect to ρ, we arrive at the log-likelihood estimate of the density $\rho = \frac{k}{V_{i,k}}$ which is that of the k-NN estimator already mentioned above. The asymptotic standard deviation of the parameter ρ given by $\epsilon = \frac{\rho}{\sqrt{k}}$, provides us with an estimate of the error.

This form of the error suggests that increasing the value k decreases the error. However, increasing k would change the density and affect one of the main assumptions. Consequently, the PAK density estimator resolves this problem by choosing for each point i, the largest k_i where the density is constant within a certain confidence interval. Practically, this density is computed by adding a linear correction to the standard k-nearest neighbor estimator of the density, and the k_i's are chosen to minimize the errors in the density.

2.3.3 High Dimensional Clustering

The PAK method provides an estimation of the point-dependent free energies which allows for determining the free energy basins on the system of interest. More specifically, to do this task we employed a modified form of the density peak clustering algorithm (DPA)[142].

Clustering techniques aim for the detection of natural groups within data sets. Therefore, they are well suited for discovering the states that appear in molecular dynamics simulations. Indeed, over three decades ago, Elber & Karplus employed k-means clustering for assessing the conformers of a protein in an MD trajectory[143]. While a complete review of all the clustering techniques is out of the scope of this thesis, we will briefly discuss both the original density peaks clustering algorithm[60] and its modified version in order to demonstrate its suitability for determining the free energy minima.

In the original algorithm, the density at a data point is roughly estimated as the number of neighbours within a radius $\mathbf{d_c}$. Then, for each data point a new quantity δ is computed as the minimum distance from a point with higher density. Cluster peaks are then detected as outliers in the so-called *decision graph*, which represents the δ as a function of the density. This simple procedure relies on the observation that density peaks are characterized by points that have a relative high density and are typically far away than other points with higher density. The non centers are then assigned in order of decreasing density to the same cluster as its nearest neighbor with higher density.

This algorithm described above already has some of the characteristics we need for identifying the clusters as density maxima, that is, the free energy basins that we are looking for. However, it suffers from several limitations. Firstly, the density estimation is very rough and depends on a prudent choice of the parameter d_c. Secondly, its effectiveness relies on visually inspecting the decision graph for outliers, a process that becomes difficult when statistical fluctuations due to sampling are significant. Finally, it cannot deal with hierarchical relationships between clusters which is a critical component of understanding free energy landscapes. Therefore, a modified version was used to address these challenges.

The first limitation is addressed by employing PAK (see section above) as a density estimator. PAK provides a more stable estimate of the density allowing for a rigorous quantification of the error. Rather than relying on visual inspection of the decision graph, a set of heuristics are implemented that first identify putative centers and then distinguish between real centers and those coming from statistical fluctuations of the sampling. More specifically, point i is a putative center if its density is higher than all the other points in the neighborhood employed for computing the density. Subsequently, the rest of the points are then assigned in order of decreasing density to the same cluster as its nearest neighbor with higher density. Finally, after the saddle points between clusters are determined, the significance between the difference of two clusters is established if the difference between the value of the density at the maximum and the density at the saddle is smaller than the sum of the error times a parameter Z. In this case, two clusters would be merged. This free parameter Z reflects the statistical confidence of the stability of the clusters found and, unless otherwise stated, the z-value chosen for the analysis presented in this thesis is set to 2.5.

Finally, for understanding the presence of hierarchies among the clusters de-

tected, the method makes direct use of the saddles detected in the last step. Indeed, the lower the barrier between clusters, the more likely that they should be considered as part of a bigger merged cluster. These aspects of the cluster analysis can be understood more intuitively in the form of a dendrogram constructed in a manner where the extent of the similarity between the cluster determines their relative distance. Figure 2.4 shows an example of how applying DPA to a set of points sampled from the distribution shown in A leads to the dendrogram in panel B which illustrates the structure of the peaks of the underlying probability distribution function.

DPA has the additional advantage that one has a more rigorous way of assessing the statistical confidence of a cluster. Thus it is even possible to find data that is composed by only one cluster if that is the case. This feature does not naturally emerge in non-density based clustering techniques.

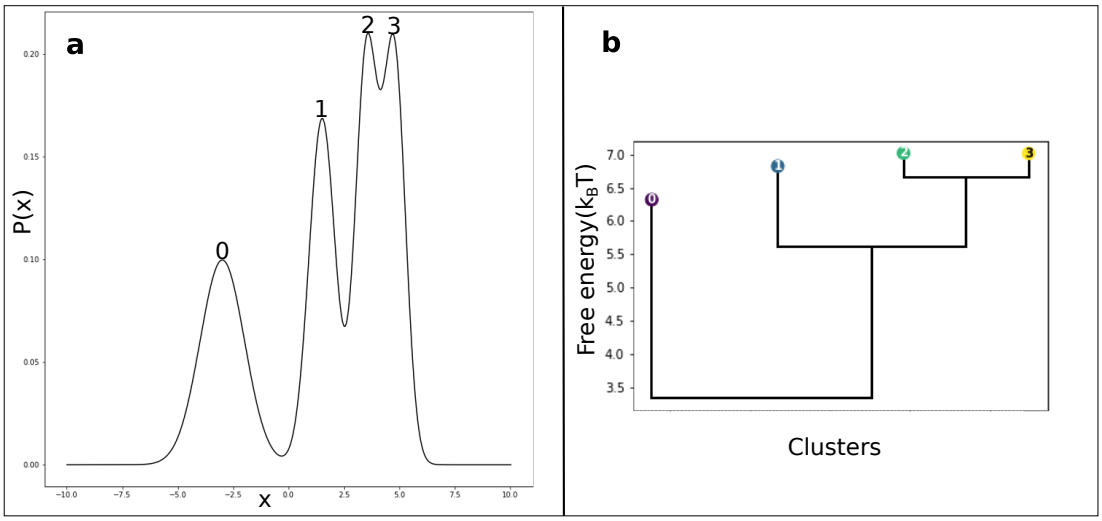


Figure 2.4: Panel a of this figure is a one-dimensional probability density from which a data set is sampled. Panel b, obtained by applying DPA to this dataset, shows the important structure of the peaks and saddle-points of the underlying probability distribution function in a dendrogram.

Chapter 3

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

A version of this chapter is at present being reviewed for publication and has been uploaded on the arxiv as:
Offei-Danso, Adu, Ali Hassanali, and Alex Rodriguez. "High Dimensional Fluctuations in Liquid Water: Combining Chemical Intuition with Unsupervised Learning." arXiv preprint arXiv:2112.11894 (2021).

3.1 Introduction

Having established the theoretical protocol and methods that we will be using in the thesis, we next move on in this chapter to examine the insights that we obtain on the free energy landscape of water. As was mentioned in the Introduction, there have been numerous theoretical and computational studies devoted to rationalizing the possibility of LDL and HDL phases in liquid water at room temperature water by examining the inherent structure of the liquid at zero-K[144, 40, 14]. This has been used to interpret several experimental observations from scattering measurements[11, 7]. On the other hand, most molecular dynamics simulations of water at room temperature show that it is a homogeneous liquid [9, 41] and that any heterogeneities in its structure arise from transient short-lived fluctuations[145, 41, 146]. At the heart of understanding this problem lies the question of the discovery of molecular probes of the local environments in a disordered liquid medium and subsequently, how to capture highly complex patterns in the hydrogen bond network on different length-scales.

Over the last three decades, since the advancement in the use of computer simulations, a wide plethora of different reaction coordinates or order parameters have been constructed to interrogate local environments in water. These include to name a few, the tetrahedrality (q_{tet})[44, 147], local-structure index (LSI)[45], the distance of the 5th closest water, molecule (d_5) to a central water[148, 47, 98] and local coordination defects[100, 149, 46, 99]. Besides variables that quantify correlations in the water network, there have also been measures to quantify local density using varia-

tions of the Voronoi volumes which probes the free-space in the network and hence the density[104, 41, 105, 150, 131, 102]. Details on how these various parameters were constructed was elaborated on in Chapter 2. All these various quantities are inspired by chemical intuition and are used to infer differences between ordered and disordered water environments. The vast majority of these quantities are designed using cutoffs in the number of water molecules, such as in q_{tet} or d_5 or radially defined thresholds in the case of the LSI. If water is seen as a percolating directed liquid network with medium-to-long range correlations beyond the first solvation shell, the ability of these chemically inspired parameters to capture all the complexities of the water network, remains an open question.

In this chapter, we employ the techniques discussed in the Methods section to investigate the fluctuations underlying the free energy landscape of the TIP4P/2005 [73] model of liquid water at room temperature and also close to the critical point. Our strategy is implemented in three steps and lays the ground-work for a general framework through which fluctuations of water in different contexts may be studied. Firstly, we encode the information of water environments using a recently developed atomic-descriptor, the smooth-overlap of atomic positions (SOAP) which has the power of preserving rotational, permutational and translational invariances [110]. This is used to compare water environments on different length scales and topologies to important milestones such as ice.

In the second step, the dimension of the manifold in which the SOAP descriptors lie is estimated using the two nearest neighbors intrinsic dimension estimator (TWO-NN) [62]. This quantity, known as the intrinsic dimension (ID), has been successfully used to characterize changes in the conformation of proteins[133, 141] as well as phase transitions in simple classical and quantum Hamiltonians[119, 120]. The ID also feeds into the third and final step of our procedure in which the minima and transition states of the high-dimensional free energy landscape are located by using a density peaks clustering algorithm[60, 61, 142].

The chapter is organized as follows. We begin in Section 1 with a summary of some specific computational aspects for this chapter. In Section 2, we report on all our results where we discuss: our findings of the intrinsic dimensionality of the hydrogen bond network, the free energy landscape of liquid water at room temperature, and the molecular origins of the associated high dimensional fluctuations. Within the results, we also elucidate the behavior of the high dimensional fluctuations upon supercooling and also close to the critical point.

3.2 Methods

3.2.1 Molecular Dynamics Simulations

All-atom molecular dynamics simulations (MD) of 1019 water molecules were performed using the GROMACS 5.0 package[151]. For most of the results we report in this paper, we use the TIP4P/2005[73] rigid water model. Energy minimization was first carried out to relax the system, followed by an NVT and NPT equilibration at 300K and 1 atmosphere for 10ns each. A timestep of 2fs was used for all the simulations. The NVT simulations were performed using the velocity-rescaling thermostat[90] with a time constant of 2ps, while the NPT runs were conducted using the Parrinello-Rahman[70] barostat using a pressure coupling time constant of 2ps. The production run at 300K was carried out for 50 ns[152] in the NPT ensemble.

We also extend some of our analysis of both the chemical-intuition and SOAP-based descriptors to their evolution upon supercooling. To generate the supercooled trajectories, we used a linear temperature ramp procedure as done in previous studies[153] where the temperature was decreased in steps of 10K from 300K down to 250K at ambient pressure. In particular at each of the temperature, after an equilibration of 10 ns, a production run of 50ns was carried out within the NPT ensemble to obtain the supercooled trajectories.

Besides these simulations, we also analyzed molecular dynamics trajectories of water reported recently by Debenedetti and Sciortino which showed for the first time, that atomistic models such as TIP4P/2005 and TIP4P/ICE [36] also display a second critical point. In these simulations, one observes fluctuations between high and low density phases of water. The work by Debenedetti and Sciortino to generate supercooled trajectories close to the critical point is a tremendous feat as they require numerous initial conditions as well as very long simulation times on the order of 10s of microseconds. In summary, these authors perform exhaustive NVT and NPT simulations of bulk water boxes ranging between 300-35000 water molecules covering timescales of up to 100 microseconds. For more details on how these trajectories were generated, the reader is referred to the original manuscript[38].

3.2.2 SOAP Descriptors for Water

For our simulations of bulk water, the SOAP descriptor for a water molecule is constructed involving different combinations of the oxygen and hydrogen atoms and their environments. The first descriptor ($\vec{\mathbf{O}}$) is formed by computing the power spectrum of the density constructed by placing Gaussian functions on only the oxygen atoms within a certain radius centered about the position of the oxygen of the central water molecule. The other two descriptors include the hydrogen atoms of the water molecule. Since a water molecule contains two hydrogen atoms, it is necessary to choose the centers to make a new descriptor invariant to the permutation of the two indices. We achieved this by averaging the power spectra generated with centers on each of the hydrogen atoms ($\vec{\mathbf{H}}_{ave}$). In order to preserve information about the possible asymmetries present in the environment of the two hydrogen atoms, the absolute value of the difference in the descriptors was also considered ($\vec{\mathbf{H}}_{dif}$).

The SOAP descriptors were constructed using the Dscribe package[154]. In practice 10 randomly chosen water molecules were selected from each frame with a sam-

pling frequency of 4ps. This was done to ensure independence of points by reducing the effects of spatial and temporal correlations between sampled environments. In total, 120000 points were extracted for a radial cutoff of 3.7 and 6.0 Å. These cutoffs were chosen to enclose the first and second hydration shell of water respectively. Our analysis was done using 8 (**nmax**) radial and 6 (**lmax**) angular basis functions. These values are similar to those used in previous studies[116, 155]. With the power spectrum of the SOAP-based environments in hand, one can compute distances between the different local environments in water and other milestone structures for example ice.

In this work, for most of our analysis, we focus on comparing the local environments in water to those in hexagonal ice. In the rest of the manuscript, this distance is referred to as d_{ice}. One can also select many different milestone structures to compare with. For example, Pettersson and co-workers have recently proposed the possibility of low-density liquid environments in water arising from fused dodecahedron structures[156]. SOAP distances can then be constructed with respect to this structure. When this is done, we will make reference to this distance as d_{dod}.

Using these soap descriptors, we focused the ensuing analysis on three variations: $\vec{\mathbf{O}}$, ($\vec{\mathbf{O}}$, $\vec{\mathbf{H}}_{ave}$) and finally, ($\vec{\mathbf{O}}$, $\vec{\mathbf{H}}_{ave}$,$\vec{\mathbf{H}}_{dif}$). Extracting the SOAP descriptors as outlined previously results in high dimensional power spectra. Specifically, the dimensions for the three SOAP descriptors outlined earlier, are 252, 1904 and 2856 for $\vec{\mathbf{O}}$, ($\vec{\mathbf{O}}$, $\vec{\mathbf{H}}_{ave}$) and ($\vec{\mathbf{O}}, \vec{\mathbf{H}}_{av}\vec{\mathbf{H}}_{dif}$) respectively. The high dimensionality of these spectra implies the need to use advanced techniques to extract meaningful information. As described in the Methods section, the SOAP descriptors above then serve as input for extracting both the intrinsic dimension and subsequently the clustering and free energy construction.

Finally, as a way for understanding the results obtained from this analysis, we projected the SOAP descriptors in two dimensions using the uniform manifold approximation and projection (UMAP)[128] method. UMAP provides a convenient way of visualizing the high dimensional free energy in two dimensions, as has been done in several recent applications[157].

3.3 Results

3.3.1 ID Analysis of Local Water Configurations

Using the Two-NN estimator, we extracted the ID of the hydrogen bond network using the SOAP descriptors described earlier. In the context of water fluctuations, the ID provides a quantitative measure of the changes in the information content on adding several features when describing the local water environment. Specifically, in our analysis, we systematically examine how the ID changes when increasing the cutoff from layer 3.7 Å to the second 6.0 Å and when adding hydrogen atoms to the descriptors.

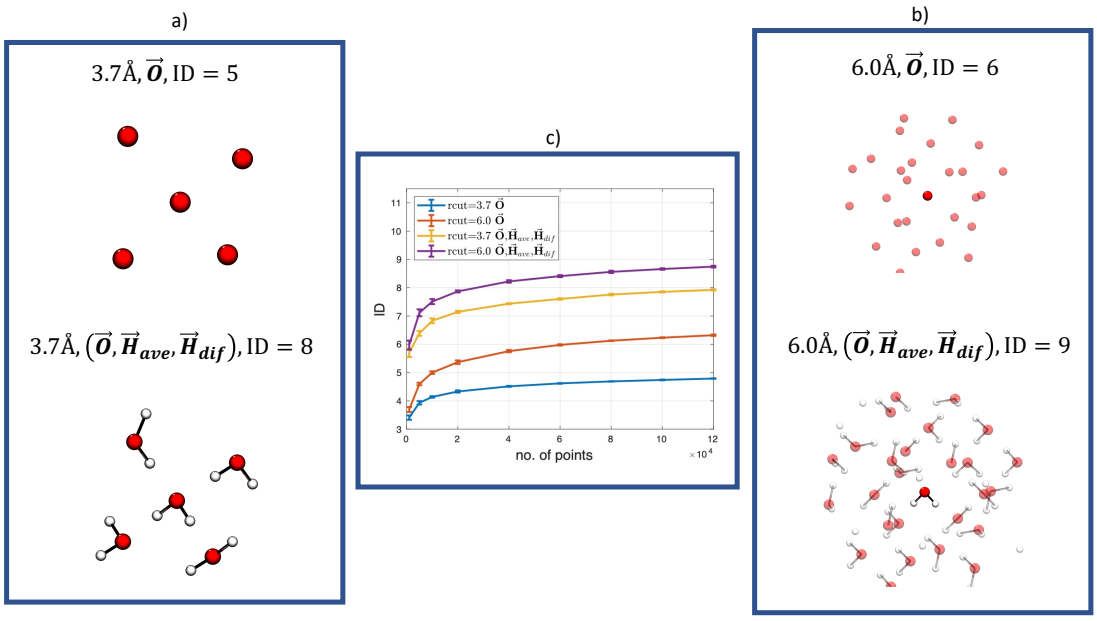


Figure 3.1: Panel a) shows 3.7 Å environment involving only oxygen atoms (top) and the same environment with hydrogen atoms (bottom). The intrinsic dimensionality is found to increase by 3 when hydrogens are included. b) Shows 6.0 Å environment involving only oxygen atoms (top) and the same environment with hydrogen atoms (below). The intrinsic dimensionality is found to increase by 3 when hydrogens are included. c) Shows the scaling of the intrinsic dimensionality with the number of points of the data set.

Figure 3.1 summarizes the results obtained from the ID analysis. Panels a) and b) schematically illustrate the local environments that are included with the corresponding inferred IDs. Panel c) shows the convergence of the ID as a function of the number of data points. Essentially, the ID can be increased in two possible ways: firstly by including or excluding the chemical species in a water molecule namely oxygen or hydrogen atoms, and secondly, by increasing the size of the solvation shell of the local water environment.

Interestingly, the ID analysis shows there is a much bigger change in the importance of including hydrogen atoms into the descriptors compared to expanding the radial cutoff. For both the 3.7 Å and 6 Å radial cutoffs, the ID increases by ≈ 3 upon the inclusion of the hydrogen atoms. On the other hand, moving from the smaller to larger radial cutoff increases the ID by a unit value. Several of the

chemical-based order parameters described earlier in Figure 3.1, for example, q_{tet}, d_5 and the LSI do not explicitly include the hydrogen atoms and therefore are very likely to miss out important coordinates needed to characterize water environments.

To understand better the molecular origins of these differences in the ID, one can examine the effect of the inclusion of the hydrogen atoms when comparing different water environments. For instance, we can take the two defects that are shown in Figure 2.1 panel d namely, $1_{in}2_{out}$ and $2_{in}1_{out}$ and compute the distribution of the SOAP distances within each class of defect type and between the two defects. In Figure 3.2 it can be seen that when using only the oxygen atoms ($\vec{\mathbf{O}}$), the distributions within and across different defects are almost identical. However, upon adding the hydrogen atoms contributions ($\vec{\mathbf{O}}$, $\vec{\mathbf{H}}_{ave}$,$\vec{\mathbf{H}}_{dif}$), the distributions are different with a slight bias towards higher values in the case of the inter-group distance distributions. Although there is clearly significant overlap in all these distributions, our analysis shows that by only using ($\vec{\mathbf{O}}$), there is important information about the hydrogen bond network that is lost, i.e. orientations of neighboring hydrogen atoms with respect to the neighboring waters.

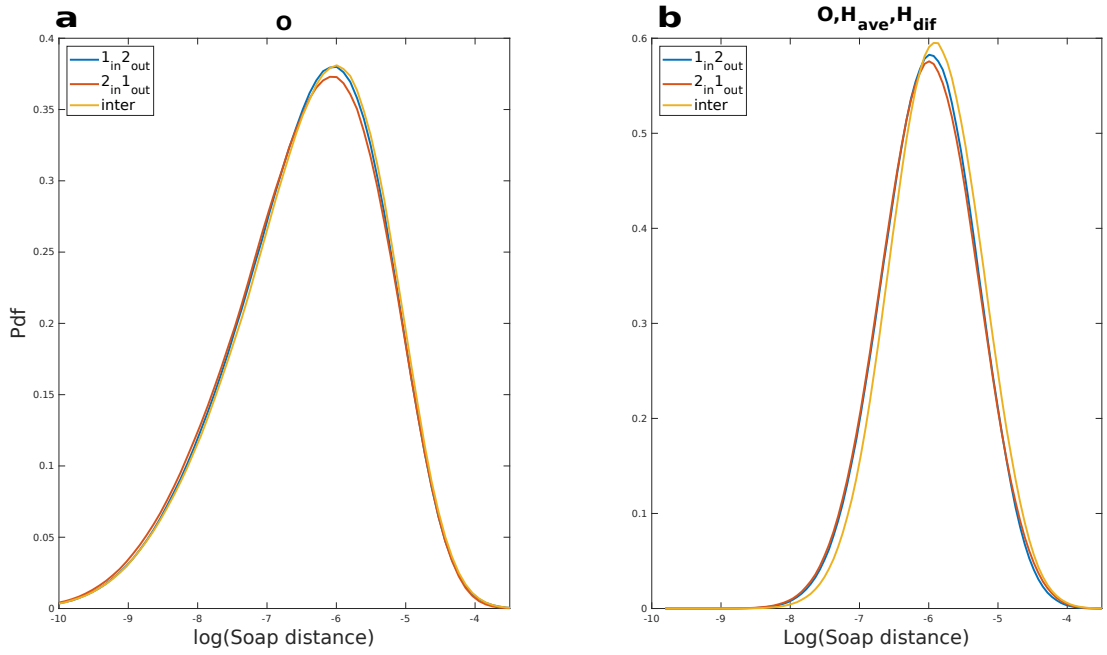


Figure 3.2: This figure shows the probability density estimates of the logarithm of SOAP distances between $1_{in}2_{out}$ environments(blue) , $2_{in}1_{out}$ environments(red) as well as distances between $1_{in}2_{out}$ and $2_{in}1_{out}$ environments(yellow). Panel a shows that there is a complete 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.

The changes in the ID has important implications on our understanding of the free energy landscape of water. Firstly, the hydrogen bonds between water molecules involve directed dipole-dipole interactions which arise from the asymmetry in the position of the hydrogen atoms. This feature of the chemistry is clearly reflected in the change of the ID upon including the hydrogen atoms. At the same time, the

presence of medium-to-longer range structural and orientational correlations in the network are manifested also in the increase in the ID albeit to a smaller extent than the role of directionality[158]. In some sense, the enhancement in the ID from the oriented hydrogen bonds within 3.7Å is strongly coupled to the order or disorder at longer distances.

In addition to the effect of the ID on the combination of the variables shown in Figure 3.1, we also examined the effect of the two hydrogen atom based SOAP descriptors namely $\vec{\mathbf{H}}_{ave}$ and $\vec{\mathbf{H}}_{dif}$. For both the 3.7 Å and 6.0 Å environments, eliminating $\vec{\mathbf{H}}_{dif}$ reduces the ID by one unit to 7 and 8 respectively. This effect on the ID, indicates that there are important asymmetries involving the environments of the two hydrogen atoms that can donate hydrogen bonds. A clear example of this, is shown in Figure 2.1 d) illustrating the creation of different types of topological defects.

3.3.2 Free Energy Landscape of Liquid Water

Having computed the ID, we are now in a position to generate the high-dimensional free energies using the PAK-Nearest density estimator. One of the challenges in constructing the point-free energies is that the error grows with the ID[61] which implies that using both a large cutoff and the hydrogen atoms would enhance the errors. We thus begin by focusing our analysis on using the SOAP descriptor environments consisting of only the oxygen atoms with a radial cutoff of 3.7Å.

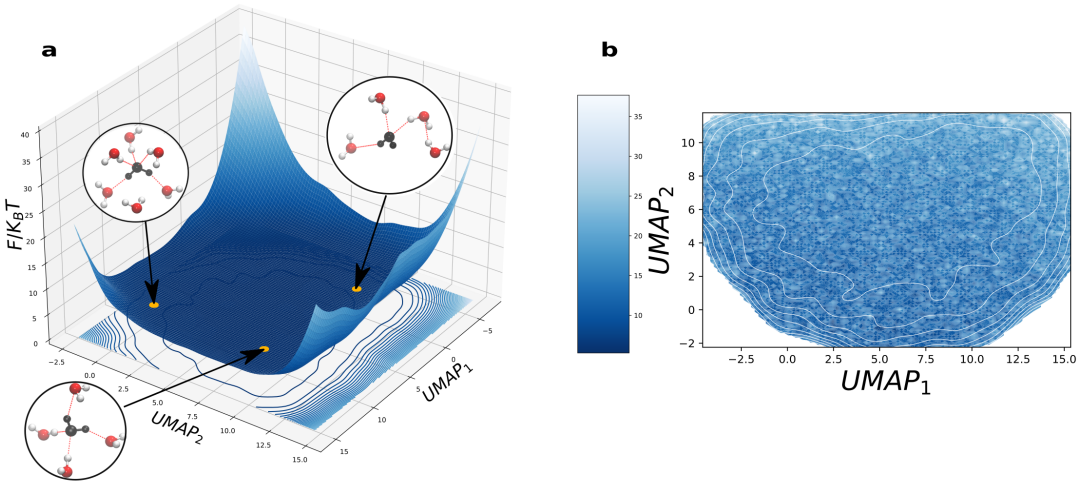


Figure 3.3: a)Free energy surface constructed in 2D UMAP manifold reveals a single basin. Also shown are three different water environments which are separated by no significant barriers indicating that fluctuations between different structures occurs within a flat free energy landscape. b) Contour plot of the 2D UMAP manifold colored by the actual free energy values is also consistent with a rough but rather flat basin.

Visualizing the high-dimensional free energies is extremely challenging. With the PAK free energies, we performed the modified density peak clustering using a confidence interval of z=2.5 which indicates the presence of one big cluster. The presence of one cluster at this z value is reproduced across several different water molecule environments suggesting that this is not an artifact of statistical fluctuations.

In order to gain a more visual inspection of the free energy landscape, we project the SOAP coordinates in two dimensions using the UMAP method[128]. Although reducing the dimensionality below the ID leads to an unavoidable information loss, the UMAP method has shown to fairly preserving the global structure of the data[159] providing a convenient way to visualize the free energies. The left panel of Figure 3.3 shows the free energy surface obtained with UMAP. Interestingly, the landscape is characterized by a very broad and rather flat free energy with small barriers (k_BT) separating shallow minima. These minima are characterized by water environments that are quite diverse as seen in the three snapshots taken at various points in the basin. These fluctuations between defective and non-defective environments without deep minima, is consistent with the presence of short-lived (between fs-ps) heterogeneities in water[41, 99].

The right panel of Figure 3.3 shows a 2d-contour map along the UMAP coordinates colored by the free energy which more clearly illustrates these features and confirms that at these temperatures, liquid water is indeed a homogeneous liquid[9, 6]. The UMAP manifold for two other datasets corresponding to different choices of water molecule environments, were found to be essentially the same suggesting the main features of the landscape are not artefacts of statistical fluctuations. Figure 3.4 shows the UMAP projection obtained for these three different water molecules which essentially provide a consistent picture on the topography of the free energy landscape.

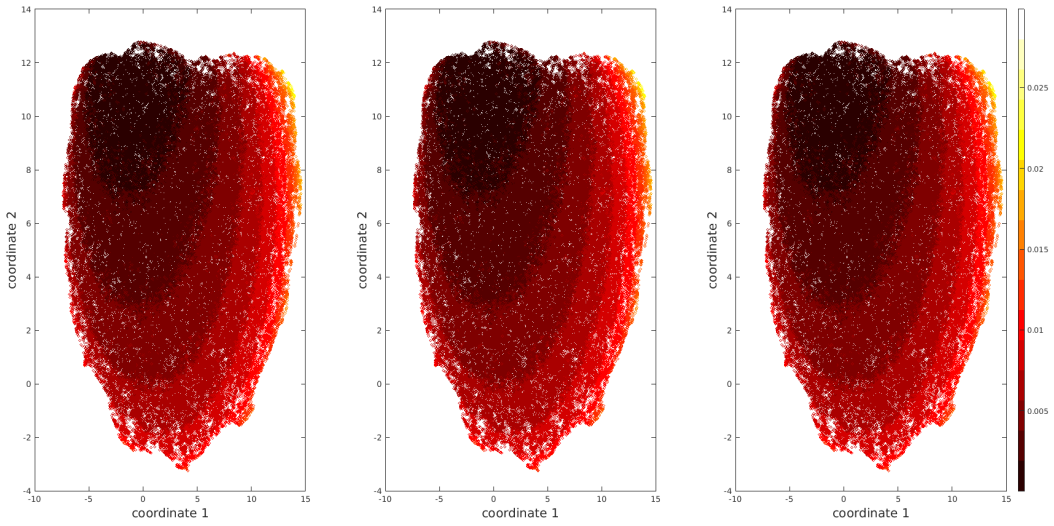


Figure 3.4: 2D UMAP projection of the environments for three datasets colored by d_{ice}.

The preceding analysis is performed on the SOAP descriptors involving only oxygen atoms within 3.7Å. Since the ID changes quite significantly when including the hydrogen atoms we repeated the PAK and UMAP analysis with the other SOAP variable combinations. Expanding the solvation environment does not lead to any significant changes in the free energy landscape. More quantitatively, Figure 3.5 shows a scatter plot of the free energies comparing the results using $\vec{\mathbf{O}}$, and $(\vec{\mathbf{O}}, \vec{\mathbf{H}}_{av}\vec{\mathbf{H}}_{dif})$ both with a radial cutoff 3.7Å using the same water coordinates. The two free energies are very well correlated with each other and shows that while the

hydrogen atoms expand the dimensionality of the free energy landscape, the key physical features remain very similar. As pointed out earlier, the full descriptor including the hydrogens is important for distinguishing different defect environments and therefore, the effects on the underlying free energy may become more pronounced in regimes where these defects are enhanced such as the air-water interface [101].

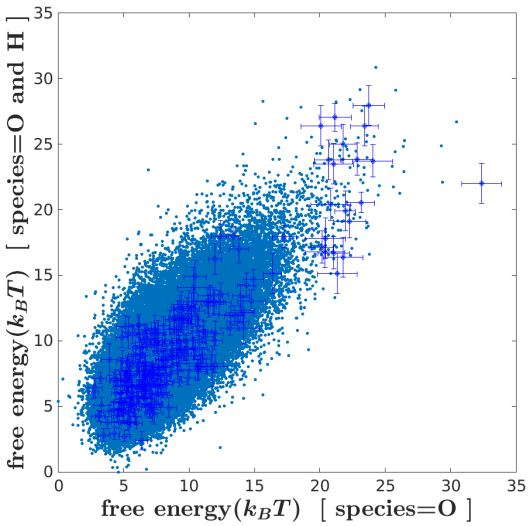


Figure 3.5: This figure which shows the scatter plot of the free energy values of $\vec{\mathbf{O}}$ environments versus ($\vec{\mathbf{O}}$, $\vec{\mathbf{H}}_{ave}$,$\vec{\mathbf{H}}_{dif}$). There is close to a linear relationship with a correlation coefficient of 0.7 and an RMSE between the two free energies of $\sim 2k_BT$. In this figure some points have been randomly selected and their errors displayed.

The current results have been extracted using the TIP4P/2005 water model which neglects many-body polarization effects. To validate our results, we repeated our procedure of extracting the ID, PAK and UMAP projection on trajectories of the MB-pol water model at room temperature[77]. As stated in Chapter 2 MB-pol is a many-body interaction potential which is the most accurate in-silico model for neutral and non-dissociative water across the phase diagram[77, 78]. In this model, we also observe the same features, namely the presence of a broad and flat free energy landscape. Figure 3.6 confirms that even the more accurate MB-pol water model presents a very similar free energy landscape giving us more confidence in our results.

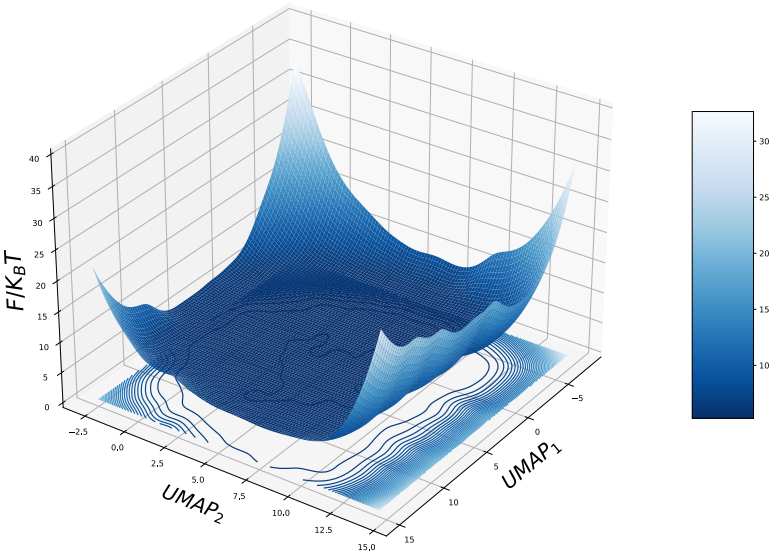


Figure 3.6: Free energy surface of MB-pol constructed in 2D UMAP manifold reveals a single basin without an appreciable barrier consistent with the results from TIP4P/2005 shown earlier.

3.3.3 Molecular Origins of High Dimensional Fluctuations

The preceding analysis of the ID shows that the fluctuations involving the hydrogen bond network involve a rather larger number of solvent degrees of freedom moving in different directions. In the following, we will examine the correlations that exist between the various chemically inspired coordinates such as q_{tet}, d_5, LSI and Voronoi density (ρ_{vor}), as well as the new SOAP-based descriptor, d_{ice} that was described earlier. Note that in this analysis a radial cutoff of 3.7 Å is used. Also in this discussion below we restrict our analysis to results in which d_{ice} was computed with only oxygen atoms. However, our findings are consistent with using a larger cutoff as well as the inclusion of the hydrogens (see Appendix Figure 6.1 and Figure 6.2).

Tetrahedrality and d_{ice}

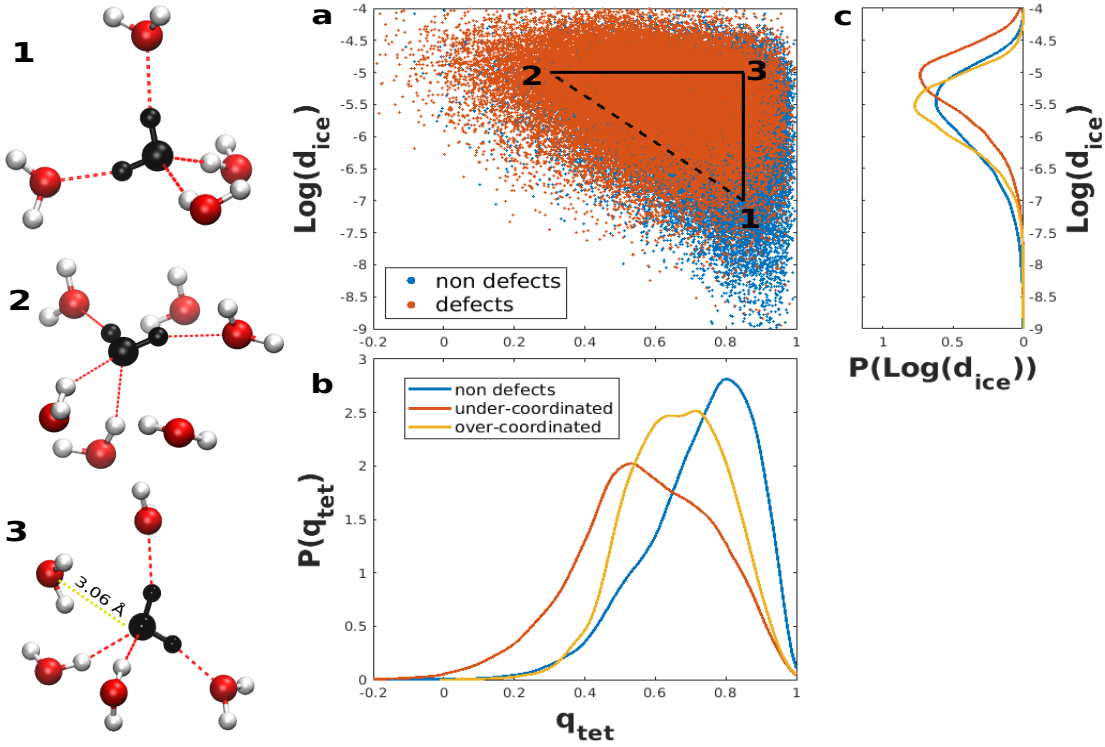


Figure 3.7: Figure a) shows the scatter plot of q_{tet} versus Log(d_{ice}) with 3 points which correspond to the environments shown on the left. The points are colored to differentiate between defects and non defects. A large overlap is found between defects and non defects. Panels b) shows the probability density distributions along q_{tet} and Log(d_{ice}) for non-defects, and, under and over coordinated defects as defined in the main text.

We begin by showing in Figure 3.7 a scatter plot of q_{tet} and Log(d_{ice}). Also highlighted are 3 points (1-3) which are illustrated in the panels to the left of the Figure. The q_{tet} and Log(d_{ice}) values refer to that of the central water colored in black. The dashed line connecting points 1 and 2 corresponds to the regime where these two variables are strongly correlated with each other. Specifically, point 1 is a water environment that has a high tetrahedrality and a large negative value of Log(d_{ice}) implying that it is closer in distance to a locally-ice like environment.

Point 2 on the other hand has a low tetrahedrality and is non ice-like. The origin of the difference between these two structures is seen more clearly in the bottom panels where we observe that in point 2, there is an asymmetry in the angles between the donating and accepting side that are used to compute the tetrahedrality.

Perhaps the more surprising aspect of Figure 3.7 are all the points that lie above the dashed line. One of these limiting cases is shown by point 3 which has a high tetrahedrality but a large distance from ice using the SOAP coordinates. This environment illustrated in the bottom panel of Figure 3.7 (Point 3) shows that there is a water molecule that is within the first-shell ($\sim$ 3.23 Å) that is not hydrogen bonded to the central water. Nonetheless, the angles that are used to extract q_{tet} involving only the nearest four neighbours do not include this water molecule and thus the central water is flagged incorrectly as a tetrahedral environment.

Also shown in Figure 3.7 are the overlapped scatter plots for the defective and non-defect water molecules. Recall that defect waters are those which break the ice rules of accepting and donating 2 hydrogen bonds. Figure 3.7 b) and c) show the 1-d distributions for Log(d_{ice}) and q_{tet} for non-defects, under-coordinated (defined as when the sum of the number of donating and accepting hydrogen bonds is less than 4) and over-coordinated defects (where the sum of the number of donating and accepting hydrogen bonds is greater than 4). In this case, we see that while q_{tet} and Log(d_{ice}) are both characterized by differences in their average values for defect and non-defect populations, the former appears to show larger variation across the different environments.

d_5 and d_{ice}

Figure 3.8 shows the analysis performed on d_5 and Log(d_{ice}). The d_5 parameter was designed in order to quantify fluctuations that occur between the first and second hydration shell[148]. Specifically, a larger d_5 has been interpreted as a water environment that is more open and low-density like, while smaller values of d_5 as compact and high-density like.

Similar to that analysis, we illustrate three landmark points in the scatter plot which are illustrated to the left of Figure 3.8. The water molecules referred to by the blue arrow correspond to waters that satisfy the d_5th criterion. The fluctuations along the line connecting points 1 and 3 reflect changes where the two parameters are well correlated: point 1 is a locally tetrahedral environment where the d_5 water resides in the second shell and separated by two hydrogen bonds from the central water, while in point 3, the d_5 water undergoes a large fluctuation bringing it from 3.7 Å to within 3.1 Å of the central water.

The fluctuations along the points 1-2 and 2-3 are more non-trivial as it shows that both d_{ice} and d_5 play an important role in characterizing the local environments independently. Although point 2 has a high d_5 of approximately 3.7 Å, asymmetries in the hydrogen bonds between the donating and accepting side of the first shell, renders it with a local configuration that is non-ice like. Examining the constrained distributions of the d_5 for the non-defects and under/over coordinated defects as before, shows that unlike q_{tet}, d_5 is much less sensitive in distinguishing these different environments. Intuitively, this is because the q_{tet} is a parameter that uses the 4 nearest neighbours while d_5 focuses on just a single water molecule that fluctuates between the first and second hydration shell.

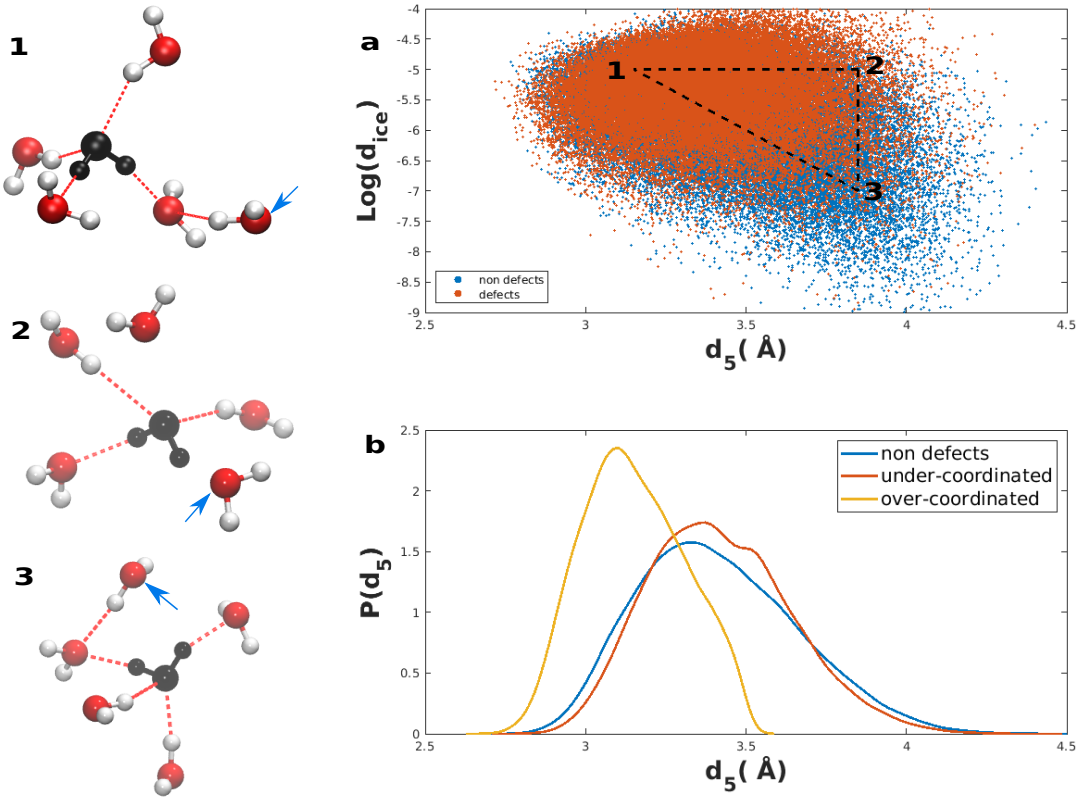


Figure 3.8: Panel a) shows the scatter plot of d_5 versus Log(d_{ice}) with the 3 numbered configurations corresponding to the environments shown pictorially on the left. The blue arrow points to the water molecule that satisfies the d_5th criterion. Panel b shows the probability densities obtained along d_5 for the defective and non-defective water molecules.

Voronoi density(ρ) and d_{ice}

Figure 3.9 shows the analysis performed on the Voronoi density ρ and Log(d_{ice}). The ρ parameter has been used previously in the literature [41, 104, 105] to quantify local density fluctuations in liquid water at different thermodynamic conditions. As done in the previous analysis, a series of landmark points are illustrated to aid with the discussion. The fluctuations along the line connecting points 1 and 2 reflect changes where the two parameters are well correlated: point 1 is a low density open and ice-like environment, while point 2 corresponds to a higher density environment with 8 neighboring waters within 3.7 Å. As expected, this high density fluctuation leads to the creation of an environment with a low value of Log (d_{ice}).

Moving along points 1-3 and 2-3 confirms again, the importance of understanding the fluctuations of the network using a combination of several different variables. Point 2 is a high density environment created by a water molecule that participates in a six-membered ring and maintains a local tetrahedral order and therefore has a low value of Log(d_{ice}). Point 3 on the other hand, corresponds to a low density environment but the orientations of the nearby water molecules do not have a local tetrahedral structure therefore leading to a higher value of Log(d_{ice}). Figure 3.9 c) shows the ρ parameter performs rather well compared to the d_5 at distinguishing over and undercoordinated defect water molecules. It is interesting to note however, that there is significant overlap in the densities for water molecules that accept and

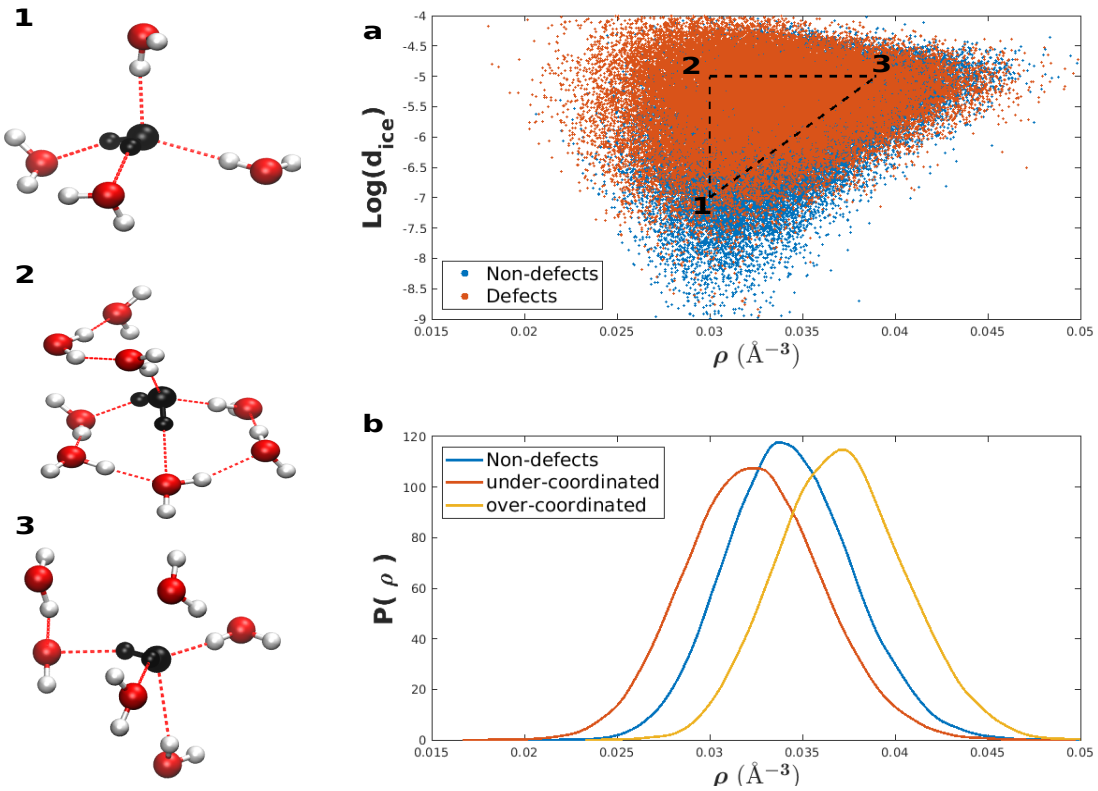


Figure 3.9: Panel a) shows the scatter plot of ρ versus Log(d_{ice}) with the 3 numbered configurations corresponding to the environments shown pictorially on the left. Panels b shows the probability densities obtained along ρ for the defective and non-defective water molecules.

donate 2 hydrogen bonds and both under/over coordinated waters.

LSI and d_{ice}

Finally, we conclude this section with a comparison of the LSI and SOAP based parameter d_{ice}. The LSI variable was designed in order to quantify fluctuations between more ordered and disordered environments due to fluctuations at the boundary between the first and second solvation shell [45, 160]. Specifically, a larger value of LSI has been interpreted as a water environment that is more open and characterized by a separation between first and second shell while smaller values of LSI, correspond to environments without a well separated first and second shell owing to the presence of interstitial waters.

The dashed line connecting points 1 and 3 corresponds to the regime where these two variables are well correlated with each other. Specifically, point 1 is a water environment that has a large LSI and a large negative value of Log(d_{ice}) implying that it looks locally like an ice-like environment. Point 3 on the other hand has a low LSI and is non-ice like. The origins of the difference between these two structures is seen more clearly in the leftmost panels where we observe in point 3 the presence of several interstitial water molecules.

The region connecting the points 2 and 3 illustrates the challenge in interpreting the LSI coordinate in terms of the local order/disorder. Point 2, which has a small distance from ice but smaller separation between first and second shell is found to

have the same LSI value as point 3 which corresponds to a rather compressed and high density unicy environment.The LSI is able to distinguish appreciably between defects and non-defects, while the effect is more subtle distinguishing undercoordinated from overcoordinated environments.

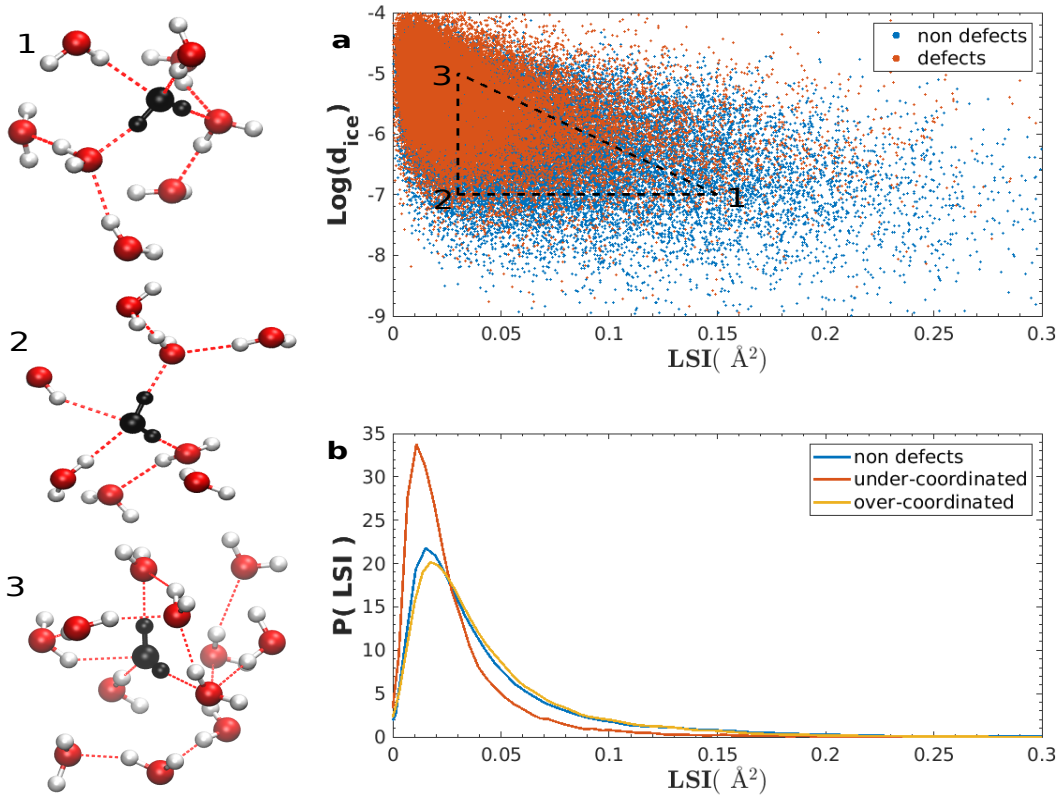


Figure 3.10: Panel a) shows the scatter plot of LSI versus Log(d_{ice}) with the 3 numbered configurations corresponding to the environments shown pictorially on the left. Panels b show the probability densities obtained along LSI for the defective and non-defective water molecules.

3.3.4 Evolution of Molecular Descriptors with Free Energy

In the preceding sections we have shown that the free energy landscape of liquid water at room temperature is best characterized as having one broad basin with small barriers separating the different minima. Furthermore, we have also seen as anticipated by the ID analysis, that the fluctuations within this landscape involve the coupling of several different molecular descriptors. In this section, we explore how these quantities change as a function of the free energy at room temperature. In addition, we also examine the behavior of some of the descriptors close to the critical point of supercooled water based on an analysis of microsecond long trajectories by Debenedetti and Sciortino[38].

Room Temperature Liquid Water

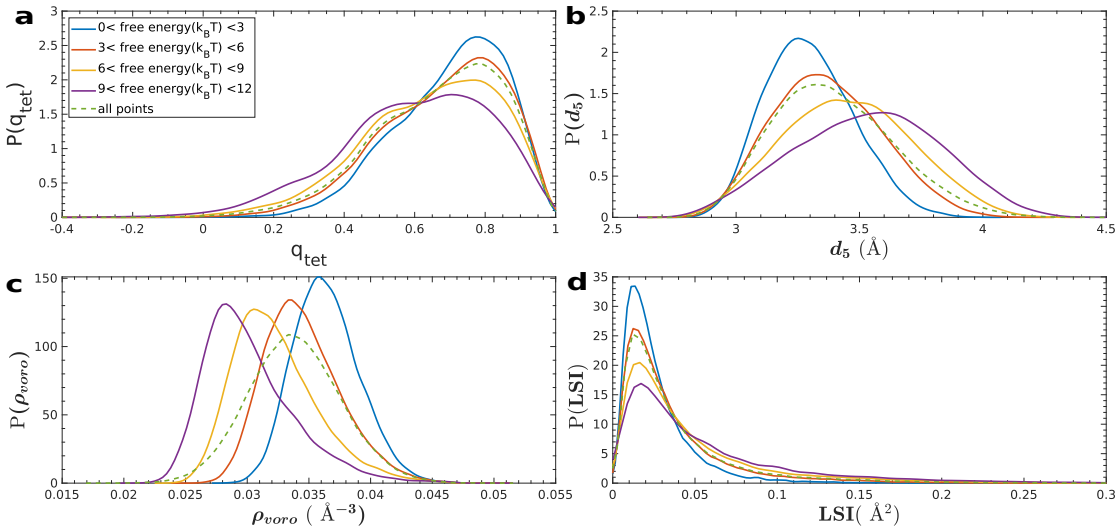


Figure 3.11: Panels a)-d) show the evolution of the chemical-based parameters as a function of different cuts along the free energy (shown with the various colored curves). The dashed curve in each panel corresponds to the distribution obtained by averaging over all the water molecules regardless of its free energy.

Using the free energies of the points extracted earlier, we examined how the various descriptors, evolve as a function of being on different regions of the actual free energy surface(different from the UMAP free energies). Figure 3.11 and Figure 3.12 show distributions of q_{tet}, d_5, LSI, ρ_{vor} and Log(d_{ice}) in slices of the free energy ranging between the minimum and 10 k_BT. Also shown in each panel, is the distribution of the respective variable obtained from all points independent of its position on the free energy surface (FES).

Starting with q_{tet}, we observe that the water tetrahedrality reduces as one moves higher in FES. Interestingly, the shoulder at lower values of $q_{tet} \sim 0.5$ becomes much more pronounced for the points higher up in the FES. In order to better understand the origin of this shoulder in the tetrahedrality, a feature which has been reported in numerous previous studies[7, 13], we show in the left and right panels of Figure 3.13 the fraction of defects as a function of free energy cuts and the q_{tet} distributions for

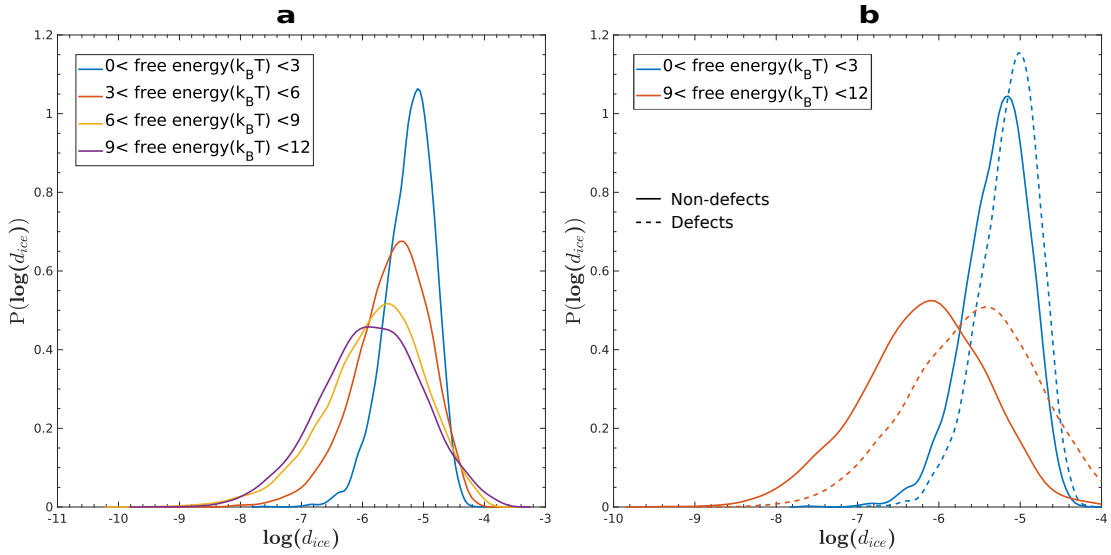


Figure 3.12: Panel a shows the evolution of Log(d_{ice}) as a function of different free energy cuts while panel b illustrates the Log(d_{ice}) for a low and high free energy region specifically for defective and non-defective water

defects and non-defects respectively. Moving up higher in free energy increases the fraction of non-tetrahedral water molecules (left panel of Figure 3.13). Furthermore, the shoulder in q_{tet} arises from these defective water molecules in the network (right panel of Figure 3.13).

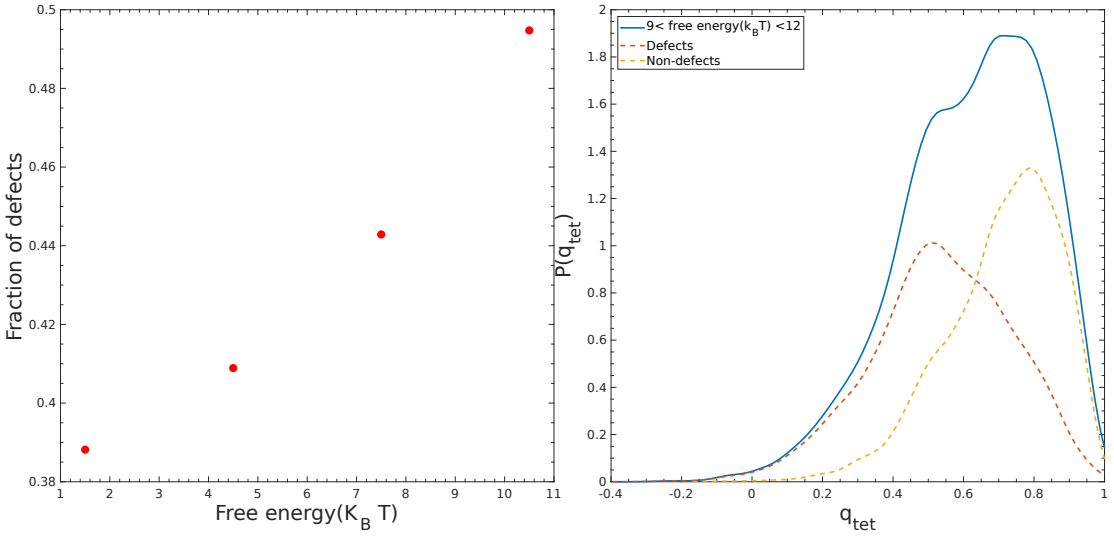


Figure 3.13: Left panel is the fraction of defects for $3k_BT$ cuts of the free energy. Right panel shows the q_{tet} distribution of points high in free energy. Also shown are the weighted q_{tet} distributions of points high in free energy restricted to defective(red) environments and non-defects(orange).

Defective and non-tetrahedral water molecules which break the ice-rules can either be undercoordinated or overcoordinated. To dissect the connection between

the defect type, q_{tet} and free energy, in Figure 3.14 we show the distribution for the coordination number of water molecules residing in low and high free energy cuts. Interestingly, we see that it is mostly the undercoordinated water molecules with coordination number less than or equal to 3 which contribute to the shoulder in q_{tet} and therefore also the high free energy regions of the landscape.

The evolution of the variables such as d_5 and ρ_{vor} reflect other changes in the hydrogen bond network. In particular, d_5 increases from 3.3 to 3.5Å moving above the minimum in free energy while ρ_{vor} decreases from 0.037 to 0.03 $Å^{-3}$. These changes correspond to water environments that become more open and less tetrahedral. It also worth noting that the points near the minima correspond to densities that are 12% larger than the average bulk density. The LSI distributions in Figure 3.11 shows more subtle changes toward higher values as a function of free energy again consistent with the formation of a more open local structure. It is clear however, that there is significant overlap along all these variables across the entire free energy landscape.

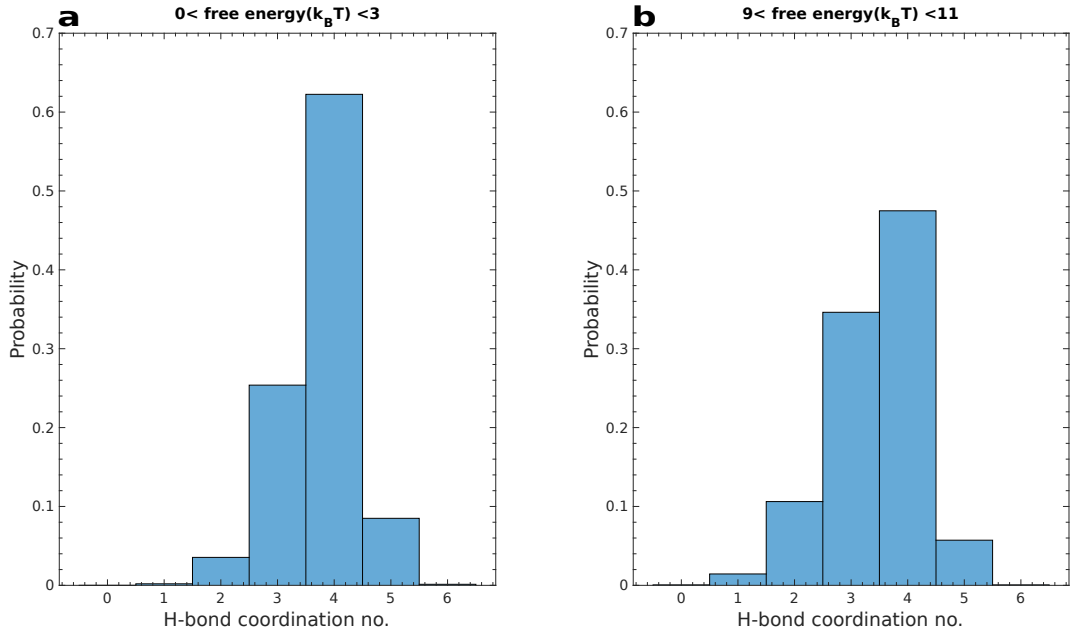


Figure 3.14: Panels a and b show the probability distributions of the no. of hydrogen bonds of the central water molecule for different cuts low and high in free energy

The left panel of Figure 3.12 shows the distributions of $\mathrm{Log}(d_{ice})$ for different free energy cuts. Interestingly, as one moves to regions of the FES that are higher in free energy, the environments look more ice-like which is consistent with the lower free energy structures in ambient temperature water being dominated by high density and more disordered environments. In the right panel of Figure 3.12, the changes in $\mathrm{Log}(d_{ice})$ as a function of free energy for both defective and non-defective water molecules are shown. Firstly, we note that the free energy minimum is characterized by the presence of both defective and non-defect water molecules consistent with the earlier analysis on the presence of a broad free energy basin characterized by low barriers separating different water structures. Secondly, we observe that fluctuations in the hydrogen bond network away from the free energy minimum results in the creation of both more ice-like or less-ice like environments as revealed by the changes

in Log(d_{ice}). For non-defects which accept and donate two hydrogen bonds, the higher lying free energy structures arise from more-ice like environments which are energetically stabilized but entropically disfavored.

Having observed the relevance of defects on the q_{tet} distribution for regions high in free energy, we examined the behavior of both defecting and non-defective water molecules as a function of free energy for the different collective variables. For q_{tet} (Figure 3.15 a)) the separation between the defects and non-defects increases as one goes higher in free energy while in the case of d_5 (Figure 3.15 b)) the distributions remain very similar. In the case of ρ_{vor}, it is quite interesting to note that defective environments higher in free energy, have significantly lower densities. In the ensuing section, we will show that this feature has important implications for temperature dependent structural evolution. Finally, Figure 3.15 d) shows the constrained distribution for the LSI. In both high and low FES regions, the LSI probability distributions for defects are found to be peaked close to zero and with less fat tails. This is indicative of smaller separation between first and second shells of the central water for defective environments. Four illustrative examples reflecting the differences in defect/non-defect water molecule environments are shown in the figure 3.17. The environments high in free energy (a and b) differ from environments low in free energy in having fewer interstitial waters.

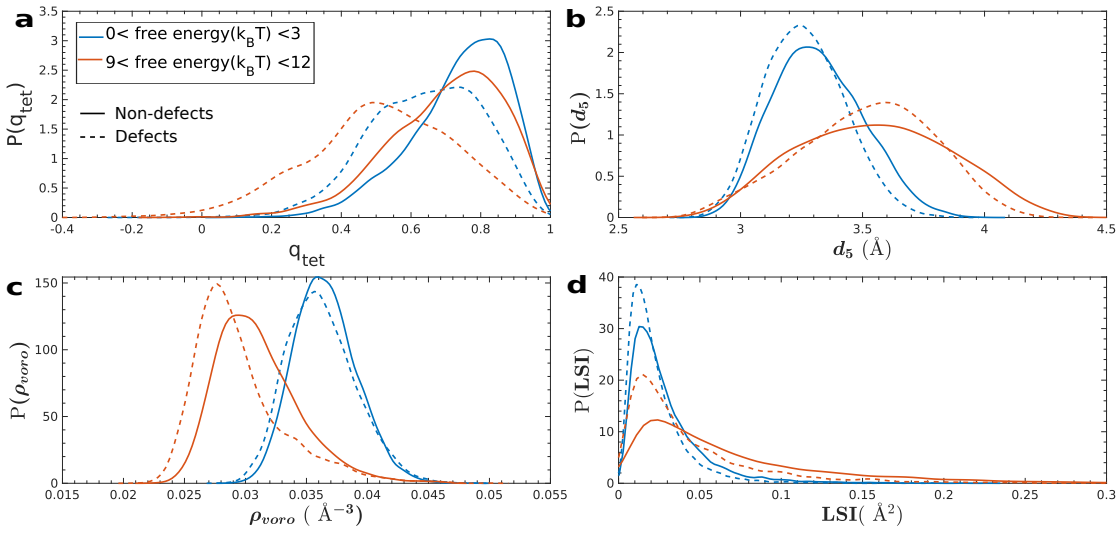


Figure 3.15: Panels a b c d illustrate the probability distributions of the $q_{tet}, d_5, \rho_{voro}, LSI$ for a low and high free energy region specifically for defective and non-defective waters

Although the Log (d_{ice}) we have used in the preceding analysis only uses the oxygen atoms, we have shown earlier that the inclusion of the hydrogen atoms contains important information as seen in the sensitivity of the magnitude of the ID (see Figure 3.1) and SOAP-distance analysis (Figure 3.2). Since there have been several recent studies using SOAP descriptors with only the oxygen atoms to characterize water environments[113, 116], we thought it was prudent to quantify better with the Log(d_{ice}) parameter the effect of including the various SOAP descriptors described earlier.

We compared the distribution of Log(d_{ice}) for defects and non-defects comparing $\vec{\mathbf{O}}$, ($\vec{\mathbf{O}}$, $\vec{\mathbf{H}}_{ave}$) and ($\vec{\mathbf{O}}$, $\vec{\mathbf{H}}_{ave}$,$\vec{\mathbf{H}}_{dif}$). Figure 6.3 shows the distributions of Log(d_{ice}) for non-defective, under and over coordinated defects for the various SOAP descriptor combinations. While there are expected shifts in the absolute values of Log(d_{ice}) with the different prescriptions, the differences are not striking. However, examining the difference in the probability densities of the defects and non defects as seen in the Appendix (Figure 6.4) shows that the use of ($\vec{\mathbf{O}}$, $\vec{\mathbf{H}}_{ave}$,$\vec{\mathbf{H}}_{dif}$) allows for the largest difference in distinguishing between defective and non-defective water molecules. Although this may not be so critical in our understanding of the free energy landscape of bulk water, the use of the hydrogen atoms will likely play a more important role in understanding the structure of water at interfaces [95].

The strategy we have adopted here using hexagonal ice as reference milestone to compare the environments in water with serves as one of many possible references. For example, recent theoretical studies proposed the possibility of fused dodecahedron structures as a possible source of a low density environment in water[156]. In order to assess this possibility, we also examined the d_{dod} distance as described in the methods section where environments in liquid water are compared with the proposed fused dodecahedron. We find that in the models of liquid water examined in this work, the environments in room temperature water yield larger values for d_{dod} compared to d_{ice} as seen in Figure 3.16.

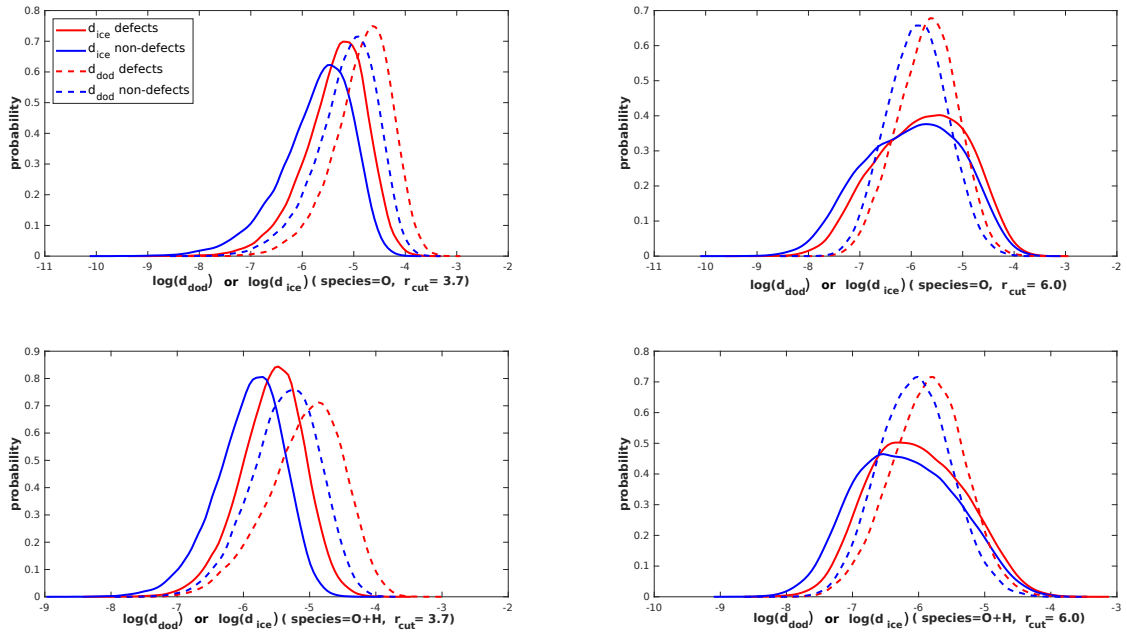


Figure 3.16: Probability density estimates of Log(d_{ice}) and Log(d_{dod}) restricted to defects and non-defects for radial cutoffs of 3.7 Å and 6.0 Å. The top panels are constructed using only oxygen atoms ($\mathbf{O}$) while bottom panels include the hydrogen atoms in computing the distance ($\mathbf{O}, \mathbf{H}_{ave}, \mathbf{H}_{dif}$).

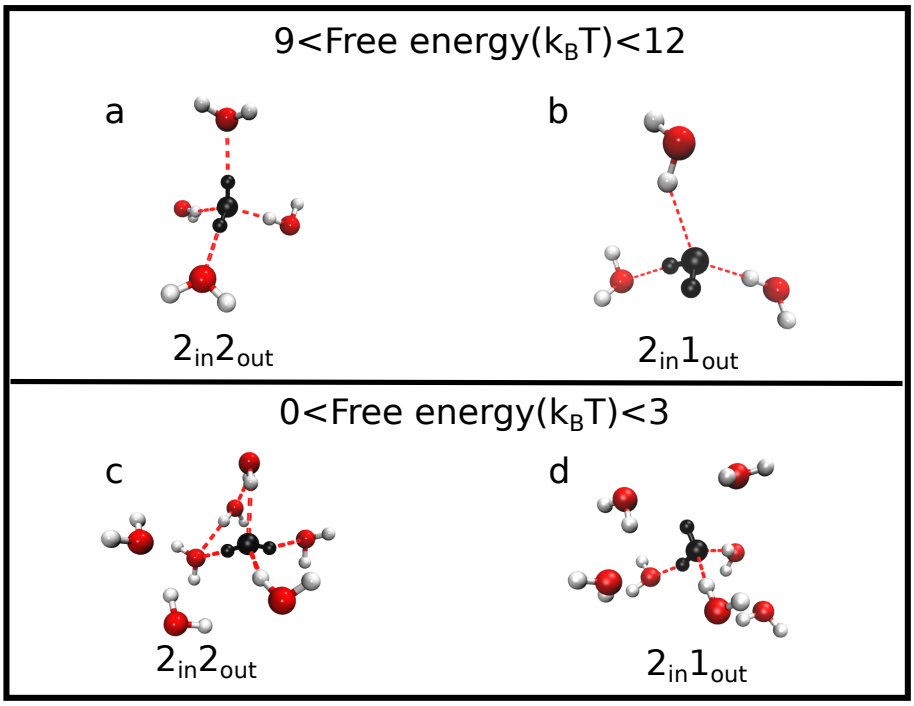


Figure 3.17: This figure shows water molecules within 3.7 Å of a central water(black) for representative environments in different regions of the free energy landscape. Panels a and b show examples a non-defective and defective($2_{in}2_{1out}$) high in free energy. Panels c and d show the corresponding local topologies for environments for low in free energy.

3.3.5 Supercooled Water and Origins of Density Maximum

With a large set of new tools in hand that allow us to examine the fluctuations in water at room temperature in a much more nuanced way, we are in a position to examine the behavior of water upon cooling. In particular, we will tackle two issues: firstly how the free energy landscape and the variables discussed previously evolve up to 250K and secondly exploring the origins of the density maximum in terms of our PAK free energies. We note that cooling up to 250K still keeps the liquid very far from the critical point. Later in this chapter we will also examine the behaviour of water near this point analyzed from previous microsecond simulations [38].

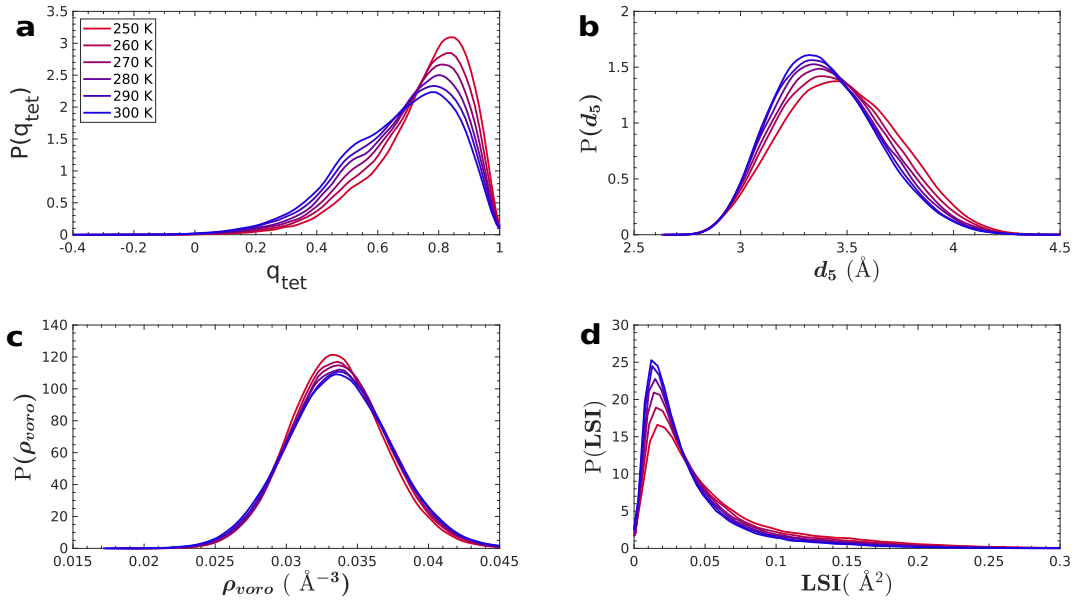


Figure 3.18: Panels a b c d show the probability distributions of the q_{tet} ,d_5,ρ_{voro},LSI within the temperature ranges of 250-300K

Figures 3.18 and 3.20 show probability density estimates of q_{tet} ,d_5,ρ_{voro}, LSI and Log (d_{ice}) as a function of temperature. In the case of q_{tet}, the shoulder in distribution is found to reduce upon supercooling. This is consistent with the decrease in fraction of defects when the temperature is decreased (see Figure 3.19). The peak of the position of d_5 increasing upon supercooling and the fattening of the tails in the LSI is also consistent with a well separated first and second shell. In Figure 3.20 a, the peaks of probability density estimates of Log(d_{ice}) shift towards more negative values indicating that the local-environments in water become more ice-like consistent with what is expected. This is seen more clearly in the right panel of Figure 3.20 which shows the average value of Log(d_{ice}) across as a function of temperature. For both defective and non-defective environments, the evolution of this parameter as a function of temperature is essentially indistinguishable.

Examining the global average of the various quantities like the way we have done for Log(d_{ice}) provides a probe into the behavior of macroscopic quantities relevant to thermodynamics, for example the density. In Figure 3.21 we illustrate the behavior as a function of temperature for the various different quantities such as q_{tet}, LSI, d_5 and ρ_{voro}. Figure 3.21 shows that the average value of q_{tet}, LSI, d_5 all increase as a function of cooling. Furthermore, while the difference between these variables for

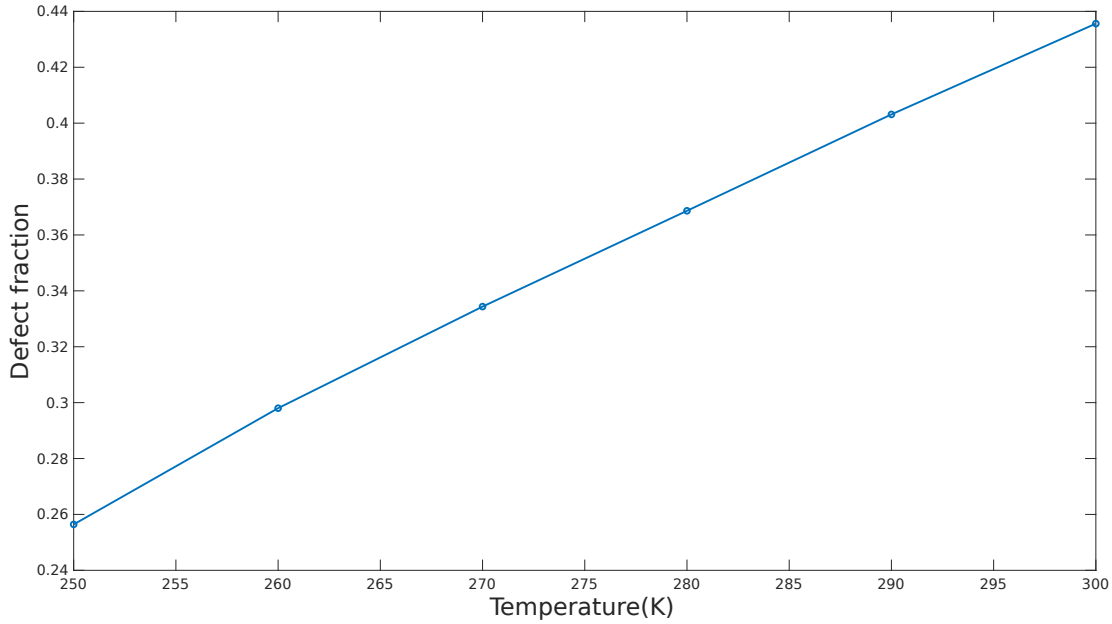


Figure 3.19: Fraction of defects as a function of temperature from room temperature to supercooled water.

defective and non-defective environments is more pronounced than for the average Log(d_{ice}) (Figure 3.20 a), both defective and non-defective environments exhibit the same trends.

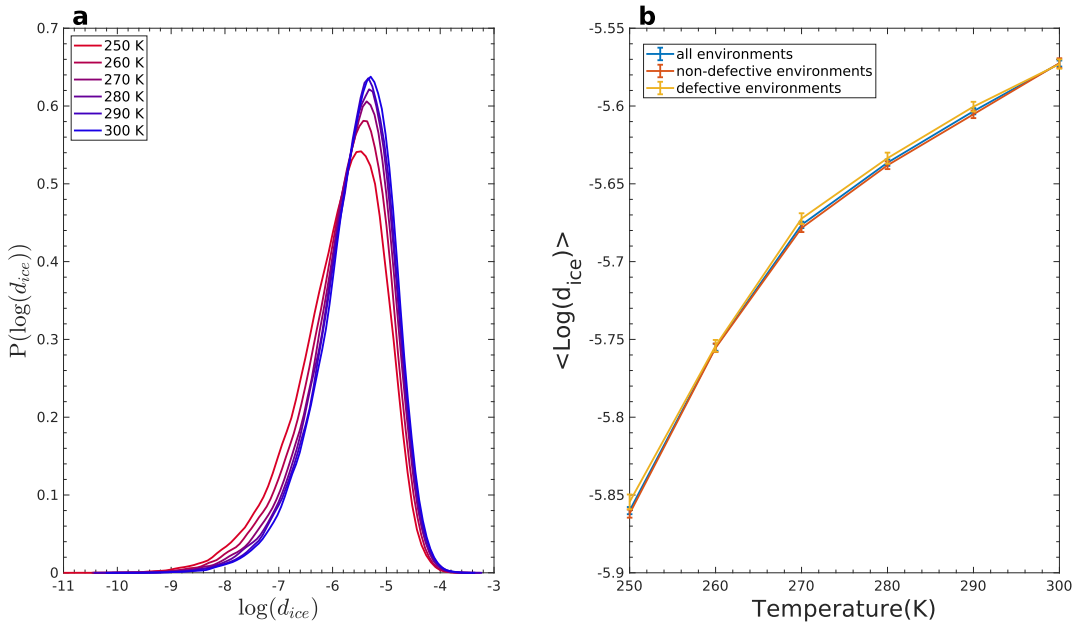


Figure 3.20: Panel a shows the probability distributions of the Log(d_{ice}) within the temperature ranges of 250-300K. Panel b shows the average values of Log(d_{ice}) for defective and non-defective environments in the same temperature range.

More interesting perhaps is the case of the average ρ_{voro} where we observe that the Voronoi density associated with all environments which clearly shows a maxi-

mum at around 280K close to the maximum of density of the TIP4P/2005 water model[161]. The density maximum of water has previously been rationalized in terms of the competition between LDL and HDL water environments [162, 7]. The LDL structures have been further decomposed into partially and non-hydrogen bonded environments[163].

If one computes the mean Voronoi density of defects as a function temperature(see Figure 3.21c), we observe, that it increases upon cooling. This effect on the overall average density is however, only important at higher temperatures, since the fraction of defects is small at low temperatures (see Figure 3.19). The net effect of the presence of defects is to decrease the overall average density at room temperature. As earlier pointed out, non-defective environments with comparatively lower densities are found higher up in free energy and it is these defects in particular that are responsible for the low density at high temperatures. Upon cooling, fluctuations in the direction of these under-coordinated yet vacuous defects from the minimum, become less probable (see the left tails of the defect Voronoi probability distribution SI Figure 6.5e and f). Rather differently from reference[163] we do not interpret these defects in terms of a HDL and LDL environments since these are not in any sense well defined minima or possessing a barrier.

On the whole, the dominant contribution to the anomalous expansion of water is from the non-defective environments which increase as the temperature is cooled. For such environments, there is a clear increase in average d_5 as the temperature is reduced. This is indicative of the formation of a well separated first and second shell and is manifest in the decrease of interstitial waters.

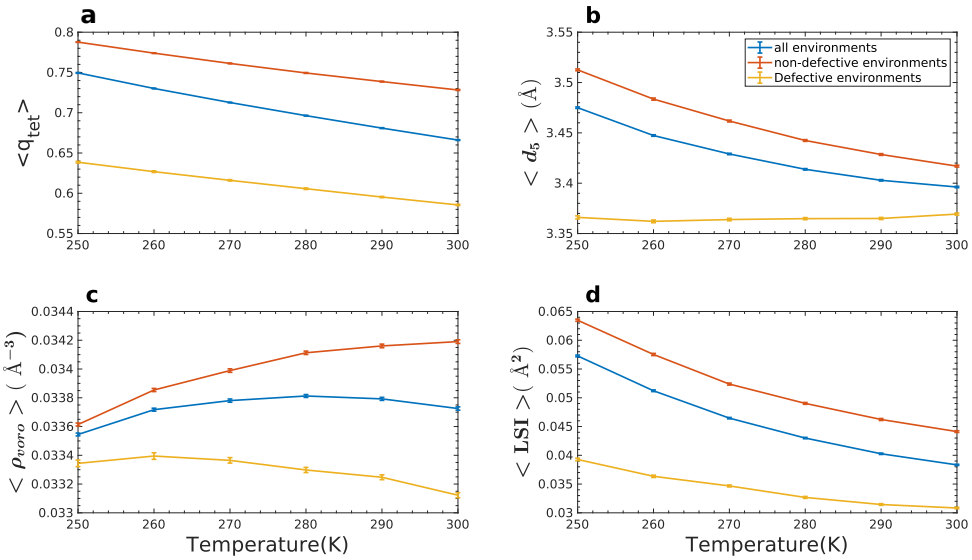


Figure 3.21: Panels a b c d show the average values of the q_{tet} ,d_5,ρ_{voro}, LSI within the temperature ranges of 250-300K . The blue lines correspond to average values computed with all points in the data sets, the red corresponds to the average value computed with non defects and green involves changes in the defective population.

3.3.6 Liquid Water near the Critical Point

The analysis of supercooled of water discussed in the preceding section do not give any direct evidence of the existence of the HDL and LDL phases due the fact that the simulations are short and the temperatures simulated are rather far from the putative critical point. In a recent work, the second critical point of water was studied from long microsecond simulations of realistic point-charge water models[38]. In these simulations close to the critical point of water at $\sim 171K, 1861bars$, fluctuations between a high density (HD) phase at 1064 kgm^{-3} and a low density (LD) phase at 977 kgm^{-3} were observed. These transitions occurring over the course of several tens of microseconds are illustrated in Figure 3.22a . The local water environments of the HD and LD phases have typically been rationalized using chemical based descriptors such as q_{tet}, LSI and the bond-order Steinhardt order parameters[45, 164, 165, 23, 166, 7, 148].

We have seen that a combination of the chemical-based and SOAP variables provide a more nuanced perspective on the nature of the fluctuations in the hydrogen bonded network. In Figure 3.22 b, we show the Log(d_{ice}) distributions for water environments extracted from the HD and LD regions of the trajectory in panel a. Specifically, SOAP environments were determined for all water molecules in the frames where the density was within $1kg/m^3$ of the minimum in the HD phase and LD phase separately. Additionally, the distributions for water at 300K and supercooled water at 230K are also shown. The LD phase is characterized by water environments that are more ice-like by 4-5 orders of magnitude compared to those in the HD phase. As expected, the environments observed in bulk water at 300K are much more similar to those in the HD phase close to the critical point. Interestingly, even though the global densities are quite different, there is a significant region of overlap in the environments observed in the HD and LD phases. This suggests the existence of heterogeneities within each of the liquid phases near the critical point.

Figure 3.22 c)-h) shows the behavior of the coupling between the d_5, q_{tet} and LSI parameters as a function of Log(d_{ice}) for the HD and LD phases in the left and right panels respectively. For both d_5 and q_{tet} in panels c)-f), the extent of the correlation between these variables and Log(d_{ice}) changes rather significantly when comparing the HD and LD phases. For the LD phase there appear to be a significant number of environments that have a large d_5 and high tetrahedrality, but cover a broad spread of Log(d_{ice}) values. The LSI for the LD phase is the only variable that shows the presence of a bimodal character. However in the HD phase, most of the local environments have the signatures of a closed, non-tetrahedral and higher local density (see Appendix Figure 6.6). A visual inspection of the trajectories suggests that the LD phase is characterized by the presence of larger domains that are built up of both low (where there are connected regions with lower d_{ice}) and smaller intermediately high density regions (with larger values of d_{ice}. Details of this analysis will be the subject of a forthcoming study[167].

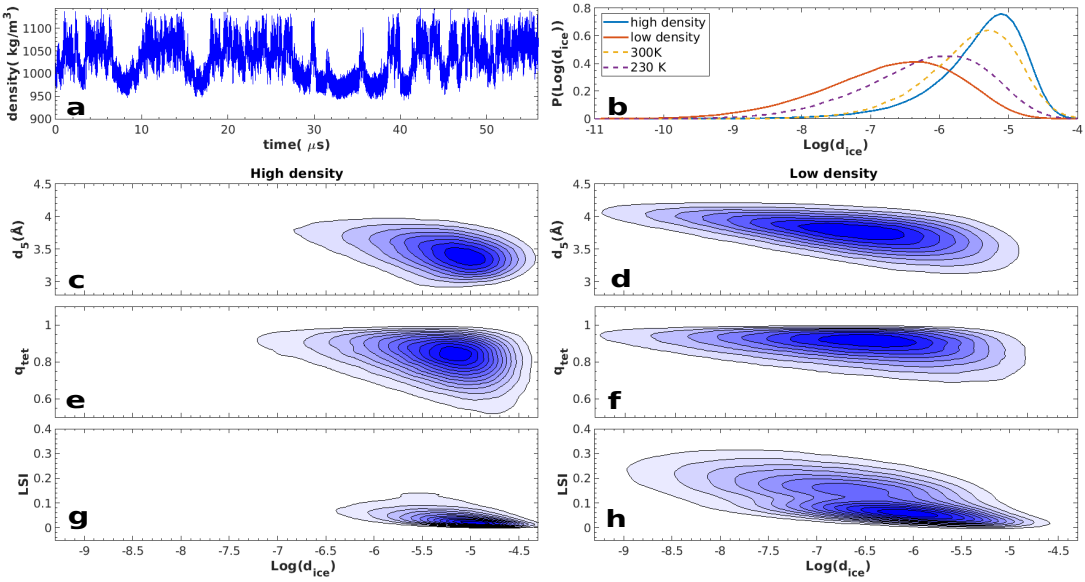


Figure 3.22: Panel a is a plot of the time series of the global density of the trajectory at 171K,1751 bars. Panel b shows the comparison of the Log(d_{ice}) at 3.7 Å between environments in HD and LD regions contrasted with Log(d_{ice}) at 230K, 300K. Panels c-h show the 2d kernel density estimates of the Log d_{ice} versus chemical-based parameters. Panels c and d, show the plots of d_5 and Log d_{ice} for HD and LD regions respectively while Panel's e and f show those q_{tet} and Log d_{ice}. Figures g and h show the plots of the LSI and Log d_{ice}. The growth of a shoulder is observed in the LD phase more clearly when using the LSI parameter.

3.4 Conclusions

There is currently tremendous growth in the development and application of machine learning methods to understand complex molecular systems. These approaches are aimed at circumventing the intervention of human or chemical bias as well as allowing for helping interpreting physical models that are beyond chemical imagination.

In this chapter we used a series of advanced unsupervised learning techniques to study the fluctuations in simulated liquid water at room temperature. This procedure involves the use of state-of-the-art local atomic descriptors (SOAP) to describe water environments, followed by the extraction of the intrinsic dimension of the water network and then finally determining the topography of the free energy landscape. We also complement this analysis by studying the behavior of various chemically inspired coordinates that have been used to study water structure. We believe that this establishes a rigorous theoretical protocol for studying fluctuations in liquids and aqueous solutions in general.

Our analysis confirms previous theoretical and experimental observations, that room temperature water is a homogeneous liquid. However, the picture that emerges is much more nuanced. Fluctuations in the hydrogen bond network occur on a rather high dimensional free energy landscape that is broad and rather flat with small ripples separated by small free energy barriers. These features are found both in TIP4P/2005 water and in the many-body potential MB-pol water[78]. This implies the presence of short-lived heterogeneities on the fs-to-ps timescale. The use of SOAP descriptors, by revealing the intrinsically multidimensional character of the local environment, leads to an additional conclusion: individually, variables such as q_{tet}, d_5, LSI or for that matter d_{ice}, cannot be used to infer the existence of LDL or HDL like environments. In the next chapter we will explore how all these different variables are manifested in the collective orientational fluctuations of water.

Finally, we also examined the evolution of all these variables within the supercooled regime by analyzing trajectories from recent work by Sciortino and Debenedetti[38]. While the HDL phase in supercooled water resembles the majority of local water environments in room temperature water, the situation is more complicated with LDL. Here there appear to be larger domains involving water environments that are more ice-like as well as high density-like but lower than the density of the HDL phase. The possibility of creating these domains is consistent with an earlier study in our group showing that upon supercooling, a network of connected branched-voids develop surrounded by smaller spherical cavities[102].

Chapter 4

Unsupervised Detection of the Collective Nature of Angular Swings in Liquid Water

The contents of the chapter are at present being prepared for publication by Adu Offei-Danso, Uriel Morzan, Alex Rodriguez, Ali Hassanali, and Asja Jelic.

4.1 Introduction

Hydrogen-bond network fluctuations in water are at the heart of a wide range of physical, chemical, and biological processes, ranging from proton transfer in the ionization of water [168, 169] to the folding of proteins and aggregation of molecules in solution [133, 56]. Since water molecules are characterized by a rather large dipole moment which in turn leads to directed interactions between water molecules, the re-orientational dynamics underlying the network reorganization has attracted the interest of both experimentalists and theoreticians alike[41, 170, 171, 172, 54].

The complex reorientational dynamics of water can be probed through various experimental techniques. In particular, the frequency dependent dielectric spectrum of water at room temperature covers a wide range of frequencies up to approximately 20 THz[173]. While the main dielectric relaxation peak reaching several tens of gigahertz is well described by a Debye relaxation process, the high frequency regime between 0.1-1 THz deviated significantly from the Debye law[173]. Numerous theoretical models have been invoked to rationalize the microscopic origins of the dielectric spectrum including the flickering cluster model by Frank and Wen[21], Pople's continuum random network model [25], and finally, the jump and wait diffusive model[53, 174].

The flickering cluster model is a statistical-physics based description of water where the liquid is thought to consist of non-hydrogen bonded monomers and clusters of hydrogen-bonded waters in equilibrium with each other. On the other hand, in Pople's continuum random network model, water consists of an extensive three-dimensional network with distorted hydrogen bonds of varying degrees of strength. Although these two models have been successful at reproducing some thermodynamic and static properties, they do not capture the above mentioned fast (sub-THz) dielectric relaxation of liquid water[25, 175]. Since the jump and wait diffusive model forms a central part of our story, it will be elaborated on in more detail later.

Since the early days of the development of molecular dynamics techniques to simulate water, atomistic simulations have played a critical role in dissecting the complex reorientational motions of water[176, 29]. Even in these early studies, Stillinger and Rahman hinted at the importance of cooperative or collective water reorganization of the water network without specifying the mechanisms.

One of the enormous hurdles in pinpointing the microscopic origins of the collective behavior from numerical simulations is the difficulty in disentangling fluctuations occurring over a wide spread of both length and timescales that create and form labile hydrogen bonds with a broad spectrum of patterns. For this reason, several early theoretical studies by Ohmine and his collaborators [52, 177] focused on inherent structures at zero-K and pointed to the importance of larger numbers of water molecules in water reorganization [52]. However, as mentioned in the third chapter in reference to the inherent structure analysis, while certainly instructive, the collective nature of the water dynamics at room temperature occurs on a free energy landscape (see Figure 3.3 in Chapter 3) that is fundamentally different from the potential energy surface.

Until over a decade ago, the primary mechanism by which water molecules reorient was thought to be a jump and wait diffusive model. In this framework, after a hydrogen bond is broken, the water molecule undergoes a period of diffusive motion remaining bonded to another water molecule during a so-called *waiting period* [175, 178, 103]. This model was found to adequately describe elastic neutron scattering experiments[179, 180]. However, similar to the flickering cluster and continuum random network models, the jump and wait diffusion does not adequately capture the dielectric spectrum in the sub-THz range.

In a seminal work by Laage and Hynes [49], it was demonstrated through the use of computer simulations that water rotations do not occur solely via small diffusive steps but involve large-amplitude angular jumps. This mechanism describes water molecules' large and quick rotations as a localized event in the hydrogen bond network. A schematic picture of the mechanism is shown in Figure 1.3 in the Introduction). While the angular jump mechanism eludes to the importance of cooperative or collective fluctuations in triggering the angular jumps, the details associated with this are essentially swept under the rug. This mechanism has now become the standard manner in which to view reorientational dynamics of aqueous solutions in both the bulk and interfaces[181].

In the last several years, important advances in time-dependent spectroscopy have opened up a new window into probing dynamical processes in water on the fs-ps timescale [172, 171, 182]. Ultrafast 2D IR anisotropy measurements have infact suggested that hydrogen bond switches in the water network are a concerted process involving large reorientations [171, 182]. However, the microscopic origins of these collective fluctuations have remained unknown.

In this chapter, we unravel a mechanism that elucidates the collective nature by which water molecules reorient. Using classical molecular dynamics simulations of the SPC/E water model [74], and by automatizing the detection of angular motions, we demonstrate that there is a heterogeneity in the types of angular motions that occur and that large reorientations are facilitated by a highly orchestrated motion of dozens of water molecules. The heterogeneity in the fast reorientational dynamics is in turn associated with transitions involving different types of defective and non-defective water molecules discussed in Chapter 3. We assert that these features are

a generic property of the fluctuations in the topology of the water hydrogen-bond network on the TeraHertz timescale[183] which are facilitated by density fluctuations. These effects are akin to previous ideas suggesting the role of defects in the molecular mobility of water[184, 175].

4.2 Methods

4.2.1 Molecular Dynamics Simulations

We performed a molecular dynamics (MD) simulation of 1019 water molecules using the GROMACS 5.0 package [151] with the SPC/E rigid water model [74]. We also compare some of our results with the MB-pol potential which is the most accurate in-silico potential reproducing both structural and dynamical properties of water across the phase diagram [185, 77]. Energy minimization was first carried out to relax the system, followed by an NVT and NPT equilibration at 300K and 1 atmosphere for 10ns each. A timestep of 1fs was used for all the simulations with a sampling time of 4fs. The NVT simulations were performed using the velocity-rescaling thermostat [90] with a time constant of 2ps, while the NPT runs were conducted using the Parrinello-Rahman [70] barostat using a pressure coupling time constant of 2ps. The production run at 300K was carried out for 2ns in the NVT ensemble [152].

4.2.2 Angular Swing Detection Protocol

Water reorientation dynamics include various processes happening at different time scales, from swift librational motions causing limited angular changes to slower reorientation through sudden large-amplitude angular jumps. It remains still an open question how and to which extent each of these processes is involved in the underlying collective hydrogen bond rearrangements that was invoked in previous studies. In order to elucidate the mechanisms behind the collective reorganization of water, we developed an automatized protocol for detecting all the various angular changes in water reorientation, which we term angular swings. In this section, we describe all the steps of the protocol depicted in Fig.4.1, before getting into detailed analysis of the diverse angular motions identified in the following section.

To track down angular changes in water orientation, we rely on two body-fixed vectors, the HH vector and the dipole moment (see Fig.4.1(a)). From the MD simulation, we first extracted the HH and dipole vector time series for each water molecule, herein referred to as $\vec{v}(t)$, with t being time. A new times series, $\vec{v}_{\mathrm{F}}(t)$, was constructed by filtering $\vec{v}(t)$ with a second-order low-pass digital butterworth filter [186] with the cutoff frequency of 25 THz, after which a mean filter of 100fs was applied. In this way we remove fast fluctuations arising from high frequency librational modes at 20 THz. The original unfiltered and the filtered time series are shown in blue and red, respectively, in Fig.4.1(b). In these time series of one of the vector component we examine, we can observe very clear sudden changes, such is the one at around time 1500 fs, that should be detected by the automatized protocol.

Next, we compute the derivative of $\vec{v}_{\mathrm{F}}(t)$ using a finite difference method, and calculate the cross product of this vector with $\vec{v}_{\mathrm{F}}(t)$ in order to obtain a new vector

$$\vec{n}(t) = \vec{v}_{\mathrm{F}}(t) \times \frac{d\vec{v}_{\mathrm{F}}(t)}{dt}, \tag{4.1}$$

that corresponds to the vector perpendicular to the plane of rotation of the body-fixed vector $\vec{v}_{\mathrm{F}}(t)$.

Our protocol defines the angular swing to be the process that does not change the plane of rotation of the body-fixed vector $\vec{v}_{\mathrm{F}}(t)$. This implies that, over the time of one angular swing, the direction of $\vec{n}(t)$ does not change. The start and the

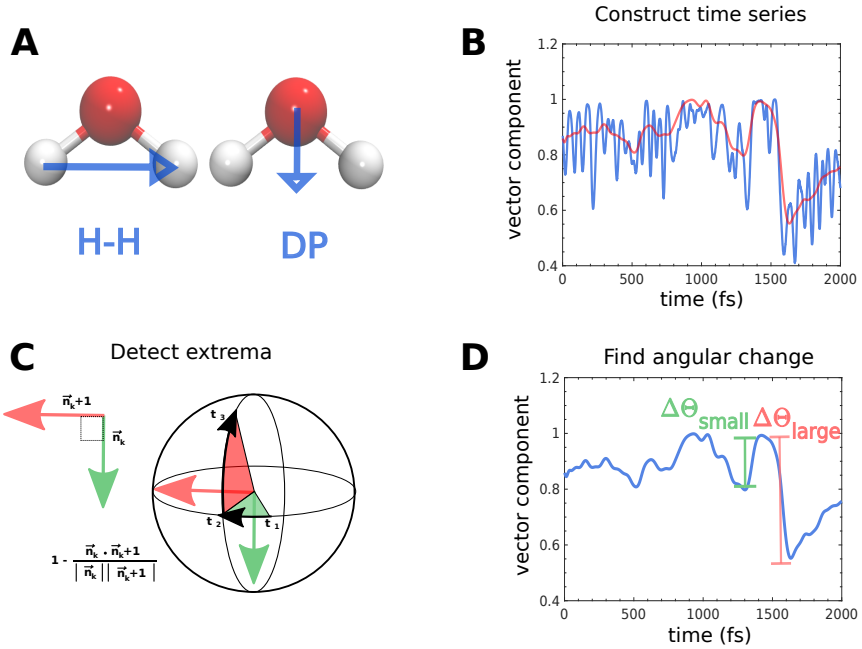


Figure 4.1: Summary of the protocol for angular swing detection. (a) Definition of the HH and dipole (DP) vectors extracted for every molecule. (b) The second step in our protocol involves the construction of the filtered time series (red) of each of the three components of the H-H (DP) vector from the original time series (blue). (c) Illustration of the detection of two successive swing events by identifying the start and the endpoints of an angular swing as an instantaneous change in the direction of the vector n_k and n_{k+1} perpendicular to the plane of rotation of the H-H (DP) vector. This example shows our protocol for two successive swing events, $E_k = [t_1 t_2]$ (green) and $E_{k+1} = [t_2 t_3]$ (red), detected from the filtered time series of the DP vector of one molecule. The direction of the vector normal to plane of rotation (green arrow for E_k and red arrow E_{k+1}) is found to change only when transitioning from E_k to E_{k+1} (d) Angular swings E_k (small green swing) followed by E_{k+1}(large red swing) indicated on the filtered DP vector component time series. The start and end points correspond to extrema in the time series.

end points of swing events are then identified as large instantaneous changes in the direction of $\vec{n}(t)$. More precisely, we look at the following quantity

$$q(t) = 1 - \frac{\vec{n}(t) \cdot \vec{n}(t+\delta t)}{|\vec{n}(t)||\vec{n}(t+\delta t)|}, \tag{4.2}$$

which is equal to 0 during the swing. At the start and at the end of the swing, this quantity is found to be non-zero, corresponding to the change in the plane of rotation of the body-fixed vector $\vec{v}_{\mathrm{F}}(t)$. Consequently, start and end points of angular swings can be identified as maxima in $q(t)$.

Indeed, we find that these points, where the plane of rotation changes, correspond to extrema in the filtered time series, $\vec{v}_{\mathrm{F}}(t)$, therefore identifying angular swings as shown in Fig.4.1(d). In order to show more precisely that the extrema of $q(t)$ determine the start and end points of angular swings, in Fig.4.2(a) we look at the time series of one of the components of the HH-vector of one water molecule. The dashed blue line corresponds to the unfiltered time series $\vec{v}(t)$, while the connected

line is the filtered time series $\vec{v}_{\mathrm{F}}(t)$. The four vertical lines correspond to the start and end points of two selected swings detected using the automatized protocol. The first small swing occurs between the two green lines, while the larger swing is between the red lines. Panel (b) in Fig.4.2 shows the value of $q(t)$ for the region in which the selected swings take place, the start and end points of the swings being the points at which $q(t)$ peaks. We see that the swings detected through our automatic procedure indeed correspond to strong angular changes in the HH-vector as seen from one of its component plotted in panel (a), whose extrema correspond well to the non-zero values of $q(t)$.

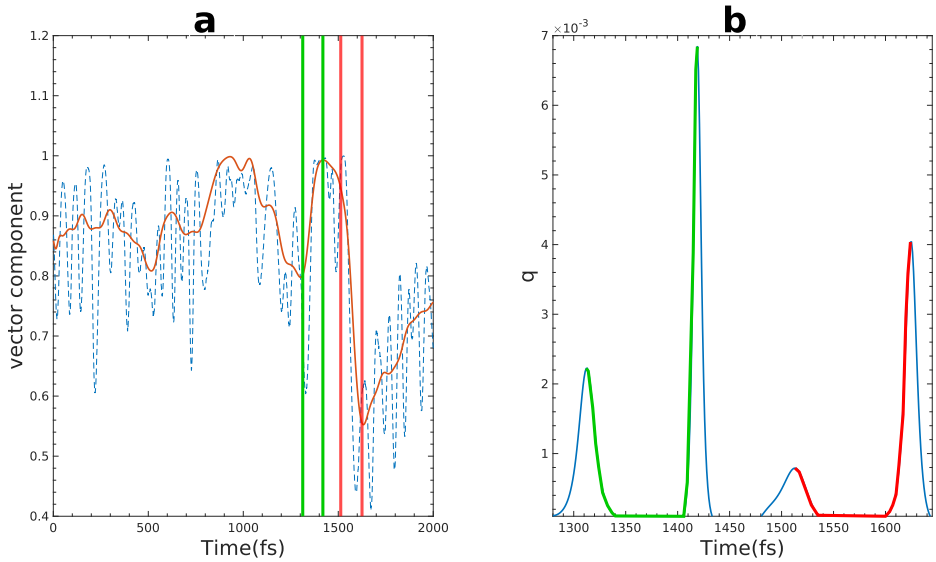


Figure 4.2: (a) Time series of one of the HH-vector components of one molecule in an interval of 2ps. The unfiltered and filtered time series are the dashed and full lines, respectively. The four vertical lines correspond to the start and end points of two selected swings. The first small swing occurs between the two vertical green lines, while the larger swing occurs between the red vertical lines. (b) Time series of q(t), between 1580 and 1625 fs. The start and end points of the swings are points at which q(t) peaks, with green and red lines corresponding to the swings observed in panel (a).

Finally, having identified the start and endpoints of the swings, the duration of the swings are taken to be the times between two peaks of $q(t)$, and the magnitude is found by computing the angle between the unfiltered HH or dipole vector at the start and at the end point of the swing. The final output of the protocol is the start time (t), duration (Δt), and magnitude ($\Delta\Theta$), for each angular swing detected. We performed the procedure both for the HH vectors and the dipole vectors, as some angular fluctuations of water molecules can be better captured through one or the other vector. In the next section, we will show the results for both vectors.

4.3 Results

4.3.1 Angular swings and changes in the local environment

We have applied an automatized protocol that identifies angular fluctuations of each molecule in the system on the trajectories obtained from the molecular dynamics simulation, as elaborated in the previous section. As a result, a total of around 10^8 angular swings were found for a system of 1019 water molecules over a time interval of 2 ns.

As we see in Fig.4.3, the angular changes we identified through the protocol have a broad range of amplitudes and duration. Therefore, we term all detected molecular reorientations as angular swings, rather than angular jumps. The latter is the predominant term in the literature, but it only refers to large-amplitude angular swings accompanied by hydrogen-bond breaking [49]. Here, we also analyze small angular fluctuations which would essentially correspond to small angular diffusive steps, that don't necessarily involve H-bond breaking.

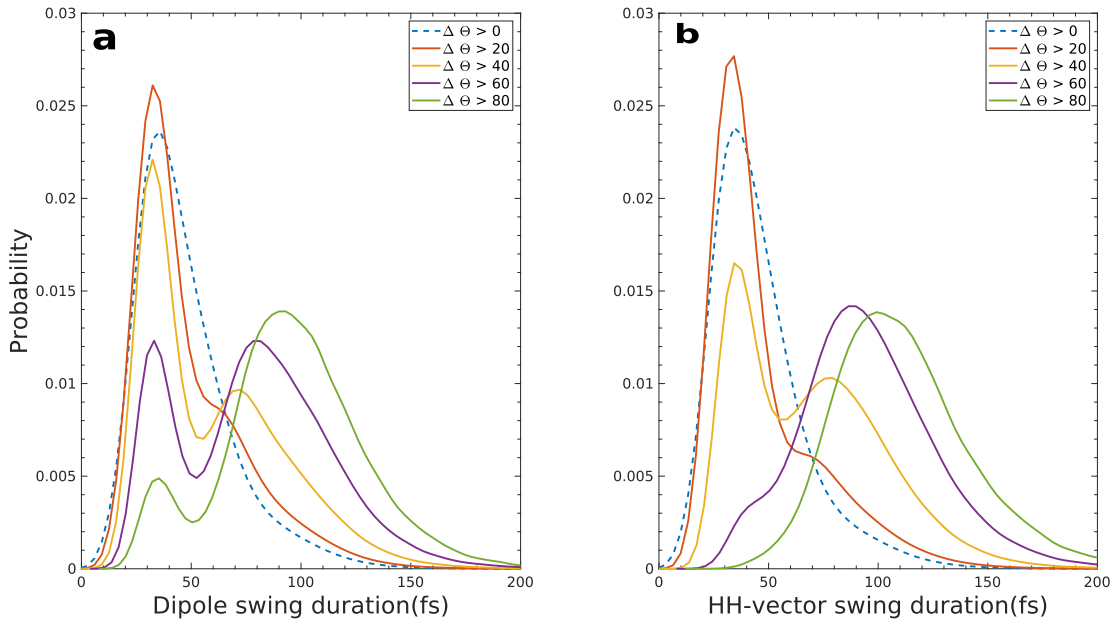


Figure 4.3: Panels (a) and (b) show the probability distributions of angular swing duration Δt for swings with the angular magnitude $\Delta\Theta$ greater than a certain threshold, both for dipole and HH vectors, respectively. For low angular thresholds, we observe bimodality in the probability distributions and a trend towards unimodality as the angle threshold increases. In particular, the surviving peak for large angles is found to be around 100 fs.

In Fig.4.3, we show the probability distributions of the duration of swings Δt, both for swings detected by looking at the time series of the dipole vector and those detected from the HH vector, panels (a) and (b), respectively. We select the swings depending on their magnitude and plot the distribution of the swing duration only for those swings with an amplitude $\Delta\Theta$ larger than a certain angular threshold. For small angular threshold, i.e. when the selected swings are predominantly of small magnitude, the swing duration peaks at roughly 30 fs. This corresponds to a fast

hindered rotational mode. As we increase the angular magnitude threshold, we find that the probability distributions change, both for the dipole and the HH vector.

For intermediate values of the angular amplitude, a second peak in the probability distribution appears and the distribution becomes bimodal. A new characteristic time of around 100 fs emerges as the angle threshold increases to 60°, i.e. when looking only at angular fluctuations with the magnitude larger than 60°. This is the characteristic time of large angular swings. Indeed, the angular value at which the second peak in the bimodal distribution becomes prominent, provides us with a criterion for what we can call large angular swings. In a total number of 10^8 swings that were detected in our simulation, 1% percent of them were found to be large swings. Based on our detection protocol, we find that on average, 50 out of 1019 water molecules undergo large angular swings (detected either through its dipole or HH vector) in a period of 100 fs.

Let us now look at how many of the detected angular swings involve hydrogen bond breaking. In Fig.4.4(a), the fraction of swing events that break hydrogen bonds for different angle thresholds is shown. This monotonically increasing function shows that large swings (with the angular magnitude larger than 60°) break hydrogen-bonds 90% of the time. Panel (b) in Fig.4.4 shows a more detailed analysis by distinguishing the swings that break either out hydrogen-bonds i.e. interactions in which the central water donates a hydrogen atom to a neighboring acceptor, or in hydrogen-bonds i.e. interactions in which the central water accepts a hydrogen from a neighboring donor.

Our study reveals that HH vector swings break out hydrogen-bonds roughly 90% of the time, while in hydrogen-bonds are broken 40% of the time. In the case of the dipole vector, this asymmetry persists, although to a lesser degree (70% for out and 40% in). To understand the nature of this assymetry, it is worth noting that for the molecule undergoing a swing, the oxygen atom is originally hydrogen bonded to two neighboring hydrogen atoms, while the each hydrogen atom is bonded to one oxygen of a nearest neighbor. Consequently a fluctuation originating from this molecule more easily breaks its proton donating (out) interaction rather than its proton accepting (in) interaction.

It is important to clarify, that the timescales we discuss here associated with the swings consider all the orientational fluctuations that occur independent of the hydrogen bonding interactions. For small angular swings, these motions correspond to small angular diffusive steps of water molecules which typically will not change the local topology of the water. On the other hand, for the large amplitude swings, the swing duration is most akin to the timescales associated with passing through the transition state during successful angular jumps, in the Laage and Hynes mechanism[56].

In the following, we will restrict ourselves to swings that break hydrogen bonds in order to analyze the local topology of the swinging molecules by calculating the hydrogen-bond network at the start and end points of the swings. This analysis is summarized in Fig.4.5. Panels (a) and (c) show changes in the topology of molecules undergoing small magnitude HH and dipole swings, respectively, while large magnitude swings are shown in panels (b) and (d). Small swings ($\Delta\Theta < 20°$) were found to involve transitions between non-defective topology $2_{in}2_{out}$ and an under-coordinated defect $1_{in}2_{out}$ or between non-defective environments and an over-coordinated defect $3_{in}2_{out}$ (see Fig.4.5(a) and (c)).

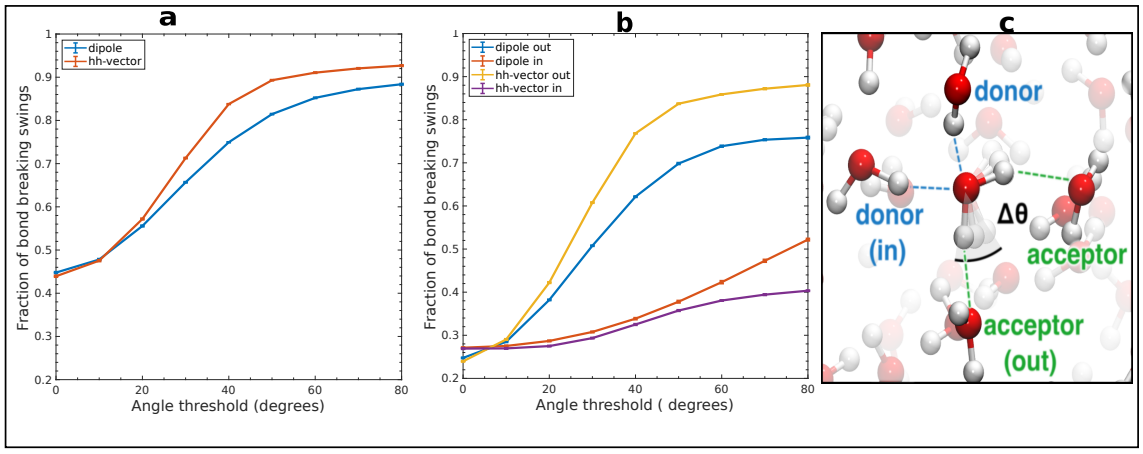


Figure 4.4: Angular changes that lead to hydrogen bond breaking and change in local water network topology. (a) The plot of the fraction of events in which we detect hydrogen bond breaking, depending on the angular swing amplitude. The x-axis represents a threshold angle for the HH (red) and dipole (blue) vector so that we count all the events with an amplitude larger than the threshold value. For swings with large amplitudes, most events involve hydrogen bond breaking. (b) We show the fraction of bonds that break during angular swing events depending on whether they are with proton donors (out, blue and yellow) and proton acceptors (in, red and purple) as the angle threshold is increased. We see that the hydrogen bond associated with the hydrogen atom of the swinging molecule is affected the most by the large angular swing of the water. (c) Sketch of a non-defective water (central molecule) with its proton donating bond (out) and its proton accepting h-bond (in) highlighted.

For large swings ($\Delta\Theta > 60°$) dominant transitions occur between $2_{in}2_{out}$ and $1_{in}2_{out}$, as well as $2_{in}1_{out}$ and $2_{in}1_{out}$ for the HH vector. For the dipole vector the dominant contributions are between two $2_{in}1_{out}$ and $2_{in}1_{out}$. Transitions between under-coordinated environments such as $2_{in}1_{out}$ and $1_{in}2_{out}$ become prominent as well in both large dipole and HH vector fluctuations.

In order to investigate in more depth the origin of angular swings, in the following analysis we will try to obtain more information on the changes in the local environment of the swinging water molecules. For that purpose, we will characterize the local environment beyond the binary geometric hydrogen bond criteria, by looking at the Voronoi density and the SOAP built logarithm of the distance from ice $\mathrm{Log}(d_{ice})$ (for the definition see Section 2.2). We compute these two quantities at the mid-point of every detected swing. We then proceed by constructing the probability distributions of the Voronoi density and $\mathrm{Log}(d_{ice})$ for swings with an amplitude larger than certain angular threshold $\Delta\theta$. The results are shown in Fig.4.6.

As we increase the threshold of the magnitude of the angular swings we observe that the peak in the distribution of the Voronoi density is found to shift towards smaller values, indicating that large angular swings occur in low density environments (see Fig.4.6 panels (a) and (b)). Furthermore, the behavior of the distance from ice, shown in panels (c) and (d) in Fig.4.6, also suggests that large angular swings occur in more disordered environments, since the peak of the distribution shifts towards larger values of $\mathrm{Log}(d_{ice})$ as the threshold for the swing magnitude is increased. However, there is a significant overlap between the probability dis-

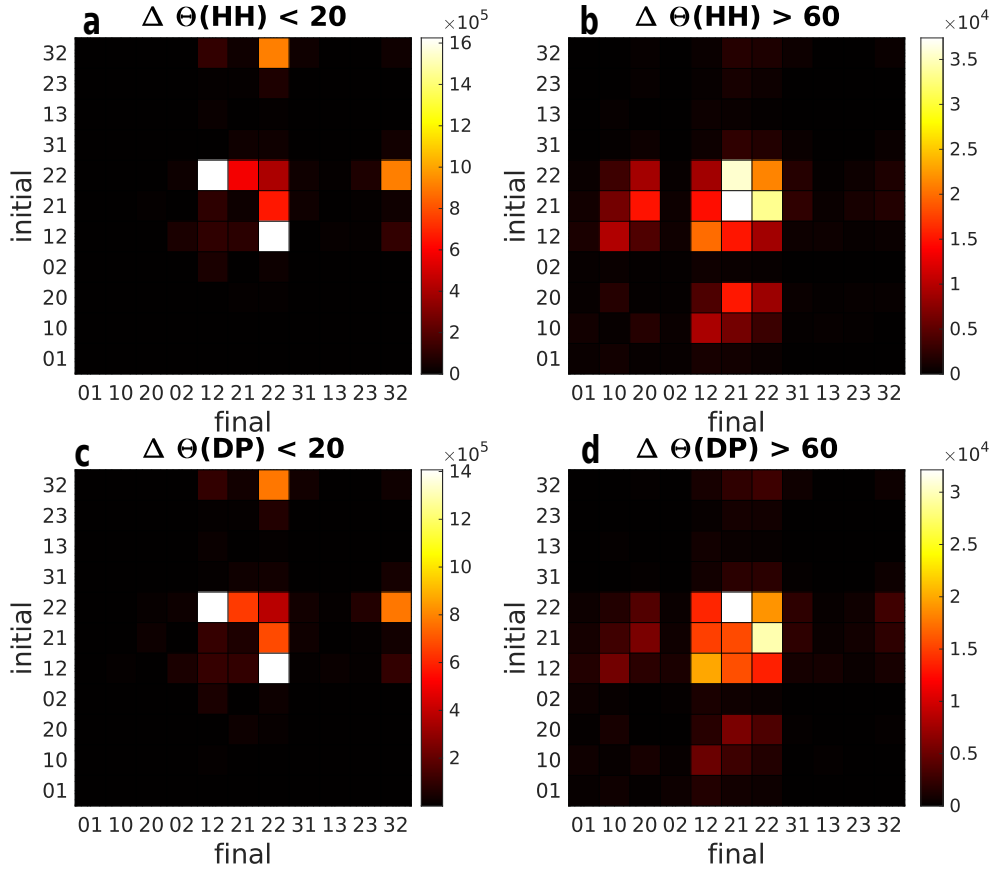


Figure 4.5: Changes in topology during HH and dipole (DP) vectors detected swings for events that break hydrogen bonds. The y-axis corresponds to the initial local topology at the start of the swing, while the x-axis shows the topology at the end of the jump, where the tick label (x,y) corresponds to $x_{in}y_{out}$. Panels (a) and (c) show the transition matrices for small HH and DP swings ($\Delta\Theta < 20°$), while (b) and (d) show transitions for large swings ($\Delta\Theta > 60°$).

tributions for the large and small swings. For example, if a small swing occurs in close vicinity to a large swing of one of the neighboring molecules, this event may not necessarily be a simple diffusive motion but rather occurring in a disordered environment triggered by the large swing.

As seen from the previous analysis on the changes in the local environment involving the local topology of the hydrogen bond network, large swings must naturally result in disruption of the hydrogen bond network of at least one of the nearest neighbors. Furthermore, large angular swings also tend to occur in low density and more disordered environments. Since density fluctuations involve collective reorganizational processes within the hydrogen bond network, large swings could presumably either *lead* to consequent angular jumps of the near by molecules, as suggested in the literature [187, 188, 177] or alternatively, be part of several large angular swings that occur simultaneously. In the next section, we will therefore study in more detail the collective nature of angular jumps.

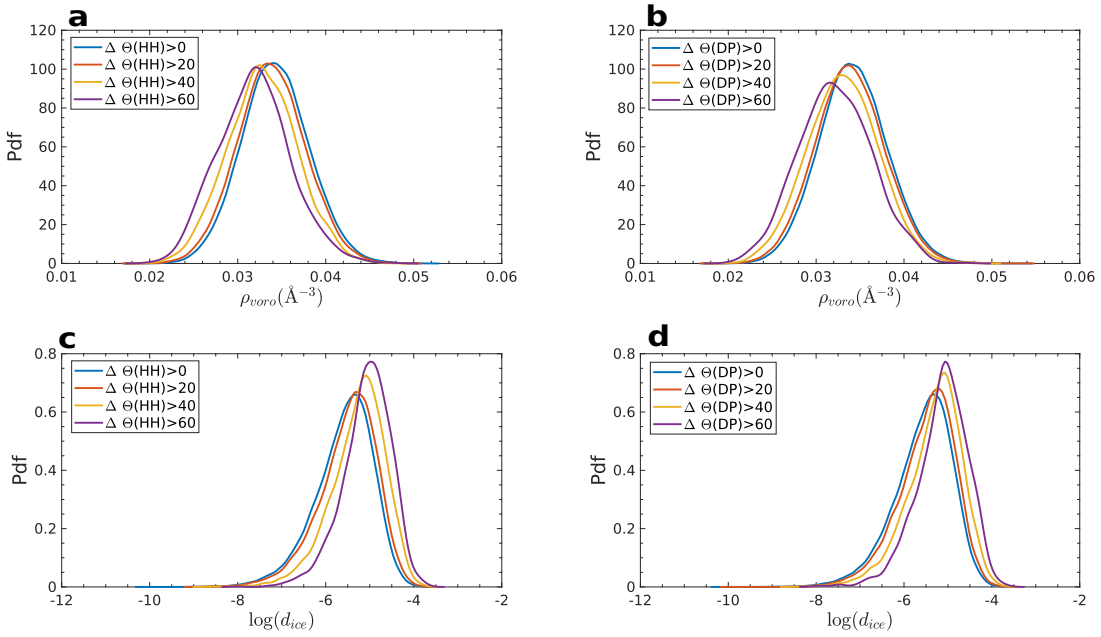


Figure 4.6: Panels (a) and (b) show the magnitude constrained probability distributions of the Voronoi density, ρ_{voro}, for the water molecules undergoing angular swings as detected from the HH and dipole (DP) vectors, respectively. Panels (c) and (d) show the corresponding probability distributions for Log(d_{ice}). The probability distributions shifts towards lower local densities and more disordered environments as we restrict ourselves to swings with the larger angular magnitude.

4.3.2 Collective Nature of Angular Swings

To build our intuition on the collective nature of angular jumps, in Fig.4.7(a), we highlight all water molecules in the system that perform large-amplitude swings of the dipole vector within a selected time interval of 350 fs. The round panels (b) and (c) in the middle of Fig.4.7 show a close-up of several of these molecules before and after the angular swing, as seen through their dipole vector orientations. Also shown in the background are all the other water molecules in close vicinity to this event. Finally, for the selected group of molecules, the plot in Fig. 4.7(d) follows the change of the dipole vector in time through the time evolution of the angle it forms for one of the axes of the laboratory coordinate system (this is a proxy of the angular change, used here for simplicity; a more precise definition described in the swing detection protocol in section 4.2.2 is used in the analysis below).

We see that the angular change of the dipole vectors of the eight molecules involved in this event ranges between 60–120 degrees within the time interval of 350 fs that we are observing. This type of angular reorientation modifies the direction in which the dipoles of these eight water molecules point and, as we will see, requires a collective reorganization of the topology of the hydrogen-bond network.

Underpinning the large angular jumps in the HB network are fluctuations in the topology of water molecules. As we have seen in the previous section, large angular motions usually create coordination defects that affect the hydrogen bonding patterns. This, in turn, affects the surrounding molecules' local topology, leading to rearrangements of nearby water molecules and possibly other large reorientation

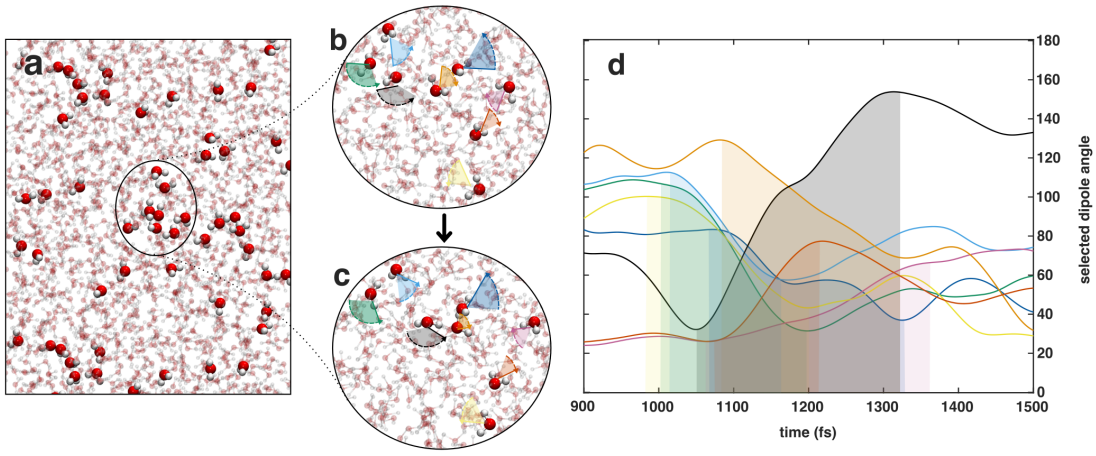


Figure 4.7: Collective nature of angular jumps. (a) Highlighted are all molecules undergoing large angular swings of magnitude greater than 60 degrees in a box of 3nm within the time interval of 350fs (which spans between time steps 1000fs and 1350fs in the MD simulation). (b,c) Close-ups of 8 of these molecules in a small box region at the start (panel (b)) and the end (panel (c)) of a large angular swing are observed from the changes in their dipole vectors. The colored arcs outline the angular motion carried by the dipole vectors in the direction of the dashed arrow. Positions of the molecules in (b) and (c) are slightly different due to translational motion during the observed time interval. (c) For each of the selected molecules, we show the change of their dipole vector in time through the time evolution of the angle it forms with respect to one of the axes of the laboratory coordinate system (for each molecule, we show the component which changes most in this time interval). The regions between the start and the end of the swings are shaded by the colors of the corresponding molecules in panels (b) and (c).

events. Here we examine the connection between local topology and angular swings in more detail by quantifying the occurrence of these events in time for the entire ensemble of water molecules.

First, at every time step of the simulation, we calculate the number of waters in the system that are non-defective, i.e. those that accept two and donate two hydrogen bonds, and the number of all the other water molecules, which we refer to as defects. The time series of the fractions of these two quantities with respect to the total number of water molecules in the system is shown in Fig.4.8(a). Interestingly, we observe that the fluctuations in non-defective and defective water molecules occur in waves. The apparent anti-correlation between the two-time series is due to the definition of the defective and non-defective topologies, which is equal to the total number of molecules in the system. The oscillations shown in Fig.4.8(a) reflect processes in the network which on a picosecond timescale, for example, lead to the creation or annihilation of up to 10-20 defective water molecules in the network. Many of these bursts then appear to accumulate over a longer timescale leading to a slower process occurring on 10s of picoseconds. Our findings bare some similarities with recent work by Liu and co-workers [187, 189] where they show that an angular jump of a given water molecule could enhance the subsequent jump motions of the same water molecule and surrounding water molecules up to the 2nd coordination shell.

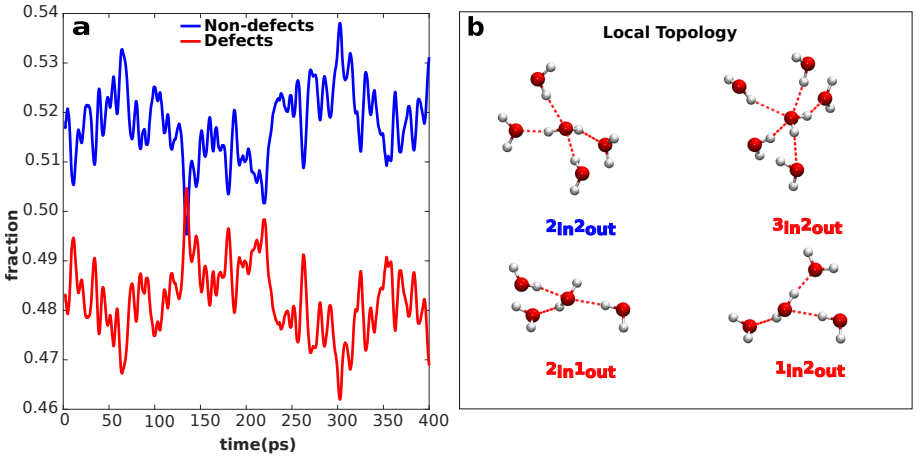


Figure 4.8: Fluctuations in the topology of the water HB network. (a) Time series of the fraction of the total number of water molecules with a non-defective topology, i.e., those that accept and donate two hydrogen bonds (blue), and the fraction of the defective ones (red). Mean-filter with the time window of 1ps was applied to the time series. The two-time series are found to undergo fluctuations of the order of tens of picoseconds. Trivial anti-correlation in time is an artifact of the definition of the defective and non-defective topologies. (b) Examples of a non-defective environment ($2_{\text{in}}2_{\text{out}}$), a defective over-coordinated environment ($3_{\text{in}}2_{\text{out}}$), and defective under-coordinated environments ($2_{\text{in}}1_{\text{out}}$ and $1_{\text{in}}1_{\text{out}}$)

In Fig.4.8(b), several examples of under and over-coordinated defects are shown. In all cases, we observe that these defects form throughout the network in waves and can exhibit different magnitudes of correlations with the non-defective waters. An examination of the power spectrum of these time series shows that these fluctuations occur on the timescale of several THz (see SI 7.1 and 7.2). These dynamics are consistent with several experimental spectroscopies [190]. The wide variety of different topologies undergoing angular fluctuations imply that there is a heterogeneity in the jumping mechanisms consistent with the interpretations made by Tokmakoff and co-workers from 2D-IR ultrafast anisotropy measurements [190].

To quantify the collective nature of angular swings in the water network, we examine how many angular swings occur simultaneously in the system. We assume that the rearrangements of the local HB network can give rise to causality between two swings occurring close by in time. Therefore, at every time step, we calculate how many molecules perform angular swings within a specific time interval around it. Previous studies report that the water reorientation happens on a time scale of about 1ps that includes not only the angular jump itself but also the breakage and forming of the HB before and after the jump. Thus, we look at a time window of 1ps, within which we calculate all the angular swings happening in the system.

In Fig.4.9(a), we plot the number of swings with an amplitude larger than 60°, detected either from the angular motion of the dipole or the HH vector. We note that the number of concurrent large swings in the system fluctuates with the same frequency as the number of defected waters, of the order of dozen of picoseconds. Moreover, the oscillations seem to be correlated in time: the larger the number of the defective water molecules, the more large-amplitude swings simultaneously happen in the system. In fact, in panels (b) and (c) of Fig.4.9, we show that when the

local topology in the HB network is more defective, there are less small-amplitude angular changes in the system and more of the large-amplitude ones. This implies a connection between the underlying topology of the HB network and the type and amount of angular swings occurring simultaneously in the system. Large-amplitude angular jumps cause rearrangements in the HB network, as they typically involve hydrogen bond breaking (as seen in the previous section), and those changes in the local topology facilitate further angular swings and reorientations of water molecules. These results reinforce a picture of water reorientations being an outcome of highly coordinated dynamics of water molecules, rooted in the collective fluctuations of the network's topology. It is worth noting that the correlation between the number of

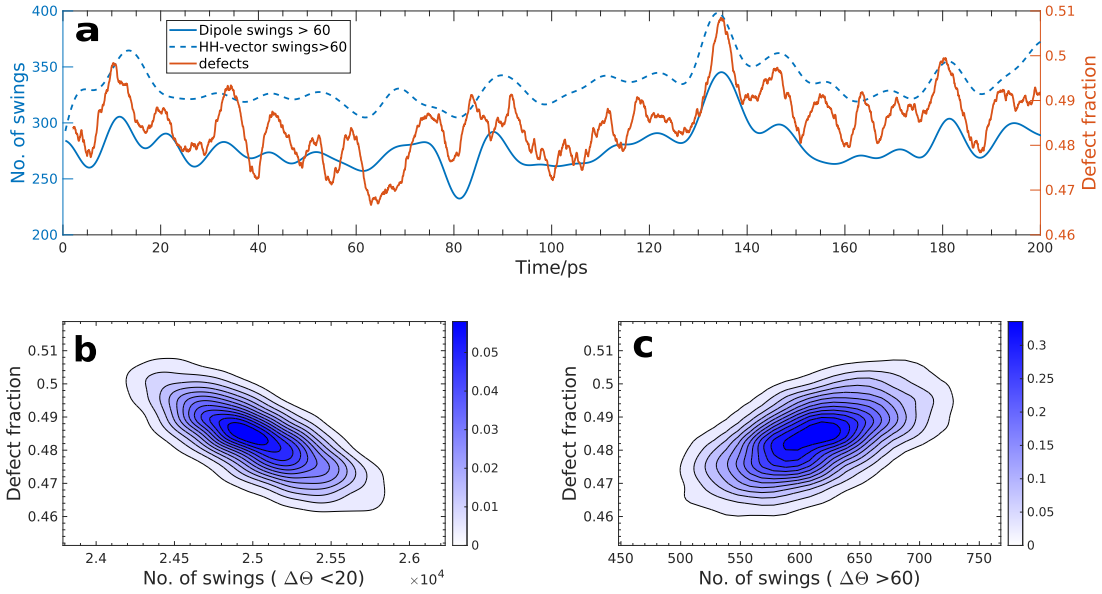


Figure 4.9: Correlation between the number of molecules performing large angular swings simultaneously and the fluctuations in the local topology of the water HB network. (a) Time series of the number of molecules in the HB network performing large angular swings (amplitude $\Delta\Theta > 60$) at each moment as detected from the observation of the dipole vector (full blue line) and HH vector (blue dashed line). At each moment, we count the number of swings happening in the system within a time window of 1ps around it. We superimpose these time series with the fraction of molecules in the HB network that are defective (red). We observe fluctuations of the order of tens of picoseconds in all three curves that often appear to be correlated in time. (b) Density plot of the fraction of defects in the HB network with respect to the number of molecules in the network performing small angular swings ($\Delta\Theta < 20$) within 1ps. Anti-correlation between these two quantities means that when there are more molecules with defective local topology, the less small-amplitude angular swings occur in the HB network. We find the correlation coefficient to be -0.7390 ± 0.0089, with $p < 0.01$. (c) Density plot of the fraction of defects in the HB network with respect to the number of molecules making large-amplitude angular swings ($\Delta\Theta > 60$) within 1ps. Correlation between these two quantities indicates that the more the local topology in the HB network is defective, the larger is the number of molecules that perform large-amplitude angular swings. The correlation coefficient found is 0.5604 ± 0.0151, with $p < 0.01$.

swings and defects were found to be stronger for small swings. This arises because large swings also involve local rearrangements from defective topologies to other defective topologies, as seen in the previous section.

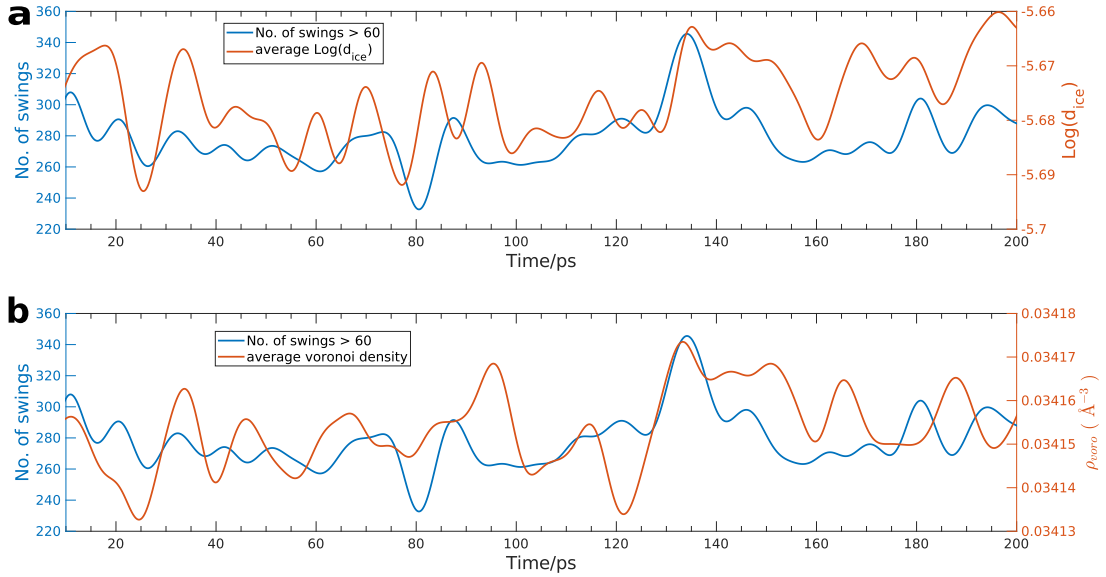


Figure 4.10: Correlation between the number of molecules performing large angular swings simultaneously and the descriptors of the local environment of the swinging molecule. (a) We superimpose the filtered time series of the average Log(d_{ice}) in 1ps as a function of time (red) with the filtered time series of the number of molecules in the HB network performing large angular swings (amplitude $\Delta\theta > 60°$) as detected from the observation of the dipole and HH vector (blue). As before, at each moment, we count the number of swings happening in the system within a time window of 1ps around it. (b) The number of large swings occurring concurrently in the system (blue) is now superimposed with the filtered time series of the average Voronoi density ρ_{voro} (red).

As pointed out in the third chapter, fluctuations in the local environment of a water molecule involve fluctuations which cannot be captured solely in terms of local network topology extracted by means of the geometric criteria. Let us now examine how the total number of angular swings occurring concurrently in the system correlates with other descriptors of the local environment, in order to gain better understanding of the collective nature of large jumps.

As in the previous section, we will look at the logarithm of the average distance from ice and the average Voronoi density of those molecules that perform angular swings. In Fig.4.10, we examine the correlation between these quantities with the number of simultaneous large-amplitude angular swings happening in the system. Both time series, of the Voronoi density and $Log(d_{ice})$, show fluctuations of the order of tens of picoseconds and often seem to correlate well with the total number of large swings occurring simultaneously in the system. However, calculation of the correlation coefficients for the unfiltered time series, did not find large correlation with the number of large swings, for neither the Voronoi density nor $Log(d_{ice})$. Instead, the large wavelength components were correlated with a Pearson coefficient of 0.42 ± 0.03 and 0.59 ± 0.02 for the Voronoi density and distance from ice, respectively.

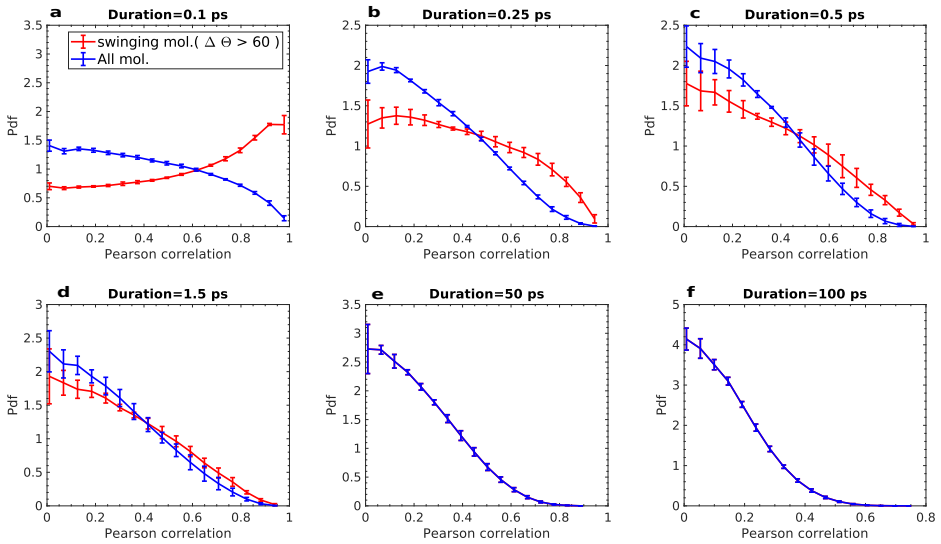
We will now concentrate on the large-amplitude swings that happen concurrently in the system and examine whether they are correlated in time. For that purpose, for the molecules that perform angular swings, we look at how correlated the parts of the H-H and dipole vector trajectories are around the times of these swings. Our analysis shows that an average duration of a large-amplitude angular swing is around 0.1ps (see Fig.4.3). Note that this swing duration does not include times needed for the HB breakage and forming, as previously considered in the literature, but only the angular motion of the water molecule. We, therefore, calculate the correlation for parts of the trajectories of length 0.1ps and longer.

To understand better the nature of large collective angular swings, we computed the Pearson coefficients between the HH and DP vector time series of different molecules over a certain time interval of length dT. It is worth noting that while calculating the Pearson coefficient between two single component time series is straightforward, the manner of computing correlation between n-component time series (in our case the motion is 3-dimensional, therefore n=3) is somewhat ambiguous. To this end, several possible prescriptions have been suggested []. In our case, restricting ourselves to some time interval of consideration dT, for every molecule, we select the time series of the component of the HH or dipole vector with the largest change in magnitude during that time interval dT. In that way, we approximate the angular motion of each water molecule by looking at the direction with the most significant change within the time interval of consideration. We believe that the obtained Pearson coefficient, constructed from a single component of the molecules' HH or dipole vector time series, is a good measure of how correlated their angular motions are.

For varying time interval widths dT, ranging from 0.1ps to 50ps, we extracted the Pearson correlation coefficients between molecules undergoing large angular swings greater than 60°, as described above. We then constructed probability density estimates from the Pearson correlation coefficients. On average, 50 out of 1019 waters are found to undergo large angular fluctuations within 0.1ps. Therefore, for these results to be statistically significant, we considered different time intervals of length dT in the trajectory, by shifting the initial time of the window of consideration by dT, in order to avoid overlapping statistics. Each probability density estimate was constructed using 10,000 Pearson coefficients. Finally, the probability distributions for molecules undergoing large angular swings were then contrasted with distributions generated by computing the Pearson correlation coefficients between all water molecules over time interval dT in a similar manner. The results of this analysis are shown in Fig.4.3.2.

In the top panels of Fig.4.3.2, we show that when looking at the short trajectories, the Pearson correlation coefficient for the molecules performing large-amplitude angular swings is typically high. In particular, for the shortest time interval dT that is just of the order of the average duration of the large-amplitude angular swings ($\tilde{1}$00fs), the probability distribution of the correlation coefficient between the trajectories of the molecules performing the large jumps has a peak close to 1 (red curves). On the contrary, when we look at the same length trajectories for any other two molecules, they are typically much less correlated (blue curves), as expected since motion of any two water molecules in the system is not expected to be correlated.

In the bottom panels of Fig.4.3.2, as we increase the time interval dT over which we compare the angular motion of the molecules, we observe that the trajectories



(d,e,f) The bottom panels show the distributions of the Pearson correlation coefficient for the parts of the trajectories of length 1.5, 50, and 100ps, respectively. While the distributions for large-amplitude swings overlap with the ones of all the molecules, the fat tail that we find for intermediate times becomes less and less pronounced, and at the order of 10ps, the time series become less and less correlated.

(d,e,f) The bottom panels show the distributions of the Pearson correlation coefficient for the parts of the trajectories of length 1.5, 50, and 100ps, respectively. While the distributions for large-amplitude swings overlap with the ones of all the molecules, the fat tail that we find for intermediate times becomes less and less pronounced, and at the order of 10ps, the time series become less and less correlated.

Figure 4.11: Correlation of reorientation dynamics of the water molecules. (a,b,c) In the top three panels, we plot the probability distribution functions of the Pearson correlation coefficient between the H-H vector trajectories of length 0.1, 0.25, and 0.5ps. We show the difference in the correlations between the time series of the molecules with large-amplitude swings (red) for those of all molecules that perform angular swings (blue) within the time window of interest. While for the short time intervals, the correlation between the large swings is typically high (peak of the curve is close to 1), as the time interval increases, the time series become more and more uncorrelated, and we find that the two distributions almost overlap for 0.5ps. (d,e,f) The bottom panels show the distributions of the Pearson correlation coefficient for the parts of the trajectories of length 1.5, 50, and 100ps, respectively. While the distributions for large-amplitude swings overlap with the ones of all the molecules, the fat tail that we find for intermediate times becomes less and less pronounced, and at the order of 10ps, the time series become less and less correlated.

on average become less and less correlated, even for the waters that reorient with the large-amplitude swing. This is also as expected, since we are now looking at the time window much larger than the duration of the angular swing and the two molecules that at some point in time performed angular jump simultaneously, do not in general have correlated motion over longer times. Moreover, as we increase the time interval dT, the two probability distributions for large swinging molecules

and all molecules start to overlap, and the fat tail of the distribution that we observe for intermediate times becomes less and less prominent. The results on the overall correlation between the strongly jumping molecules reinforce the assumption that water reorientation is not just a local phenomenon, as often studied until now, but a result of highly correlated dynamics of dozens of water molecules.

4.4 Conclusions

Recent advances in ultrafast infrared spectroscopy renewed the discussion about the collective rearrangements of the hydrogen bond network and the role that the large angle reorientations play in it [172, 171, 182]. Indeed, since the seminal work by Laage and Hynes [49], the large jump mechanism was considered primarily a localized event, with not much attention given to the possible collective effects in the reorganization of water.

In this chapter, we investigated the coordinated dynamics of water molecules and its relation to the underlying changes in the hydrogen bond network. For that purpose, we developed an automatized protocol that identifies large and small angular swings from the time evolution of the HH and dipole vector. This procedure relies on the filtering of the time series of the HH and dipole vectors, minimizing the effects of librations in order to identify instantaneous changes in the vectors. The unsupervised detection of angular changes beyond fast librational modes, independent of what their effect on the underlying hydrogen bond network might be, enabled investigation of large and small angular fluctuations occurring over varied time scales.

We have first analyzed changes in the local environment that are associated with small and large angular swings. In particular, we found that large-amplitude swings change local topology of the swinging molecule, as they usually entail hydrogen bond breaking, often involving transitions between under-coordinated defects. We further demonstrated that large swings occur in more disordered and low density environments. All this implies a strong possibility of an orchestrated dynamics of water molecules through collective rearrangements of the underlying hydrogen bond network and density fluctuations.

The collective and cooperative nature of angular motion was explored by quantifying the total number of large swings happening simultaneously in the system and by looking at how this corresponds to the overall changes in the local environment – total number of defects, as well as average Voronoi density, and logarithm of distance from ice. Time series of the total number of large swings, as well as the descriptors of the local environment in the system, show a strong component in the TeraHertz and sub TeraHertz domain, which coincides with the experimental results [173], and is an effect of the changes in the topology of the underlying hydrogen-bond network and density fluctuations.

Furthermore, the low frequency component of the times series of the number of large swings was found to be correlated with the time series of the fraction of defects, average Voronoi density and Logarithm of the distances from ice. We further found that molecules undergoing large angular swings close by in time were more strongly correlated in their vector components, with respect to any other pair of molecules in the system. This reinforces a picture of large reorientations being a collective and cooperative phenomenon.

Our simulations provide important details on the origins of collective fluctuations in water often swept under the rug, with broad implications for water at different thermodynamic conditions such as in supercooled water [5], as well as water near hydrophobic and hydrophilic interfaces in chemical and biological contexts [56, 191].

Chapter 5

Conclusions

The implications of the data-science and machine learning revolution has built significant momentum and its importance in atomistic simulations is likely to become an integral part of physics, chemistry and biology. Some of these approaches circumvent the need for human intervention as well as enabling the interpretation of physical models in a parameter-free manner. While data-driven techniques have become almost standard protocols in the development of new potentials to describe the dynamics of complex systems in atomistic settings as well as in analyzing data of biological systems, their application to understanding the thermodynamic and dynamical landscape of liquids, remains unexplored. In this thesis we take an important step to fill this knowledge gap.

Specifically, we investigated the structural and dynamical properties of liquid water using unsupervised learning techniques. One of the core challenges in understanding the complex landscape of water is detecting hydrogen bonding patterns on various time and length scales. The techniques that were used to codify this are discussed in detail in Chapter 2.

Chapter 3 re-visits using these state-of-the-art data-science techniques, the notion that liquid water consists of two co-existing states. Using the TIP4P/2005 water model we encode the local environment of water molecules using local-atomic descriptors that preserve important symmetries in the system. We show that water's hydrogen bond network resides in a high-dimensional space (an intrinsic dimension greater than 5) where the fluctuations are tuned by various interactions in both the first and second solvation shell. Extracting the free energy and clustering shows the landscape at room temperature cannot be characterized by a two-state liquid but instead is a broad and flat surface with small barriers on the order of thermal energy that separate different types of environments. While high density liquid (HDL) and low density liquid (LDL) form some of these structures, the free energy landscape is populated with a continuum of different states. Our analysis is thus consistent with several theoretical observations that water at room temperature is a homogeneous liquid with transient short-lived heterogeneities. Our findings are fully consistent with the analysis of one of the best water models of neutral water, namely the MB-pol water model [78].

Within this free energy landscape, we also provide clues into the various collective variables or order parameters that underlie the complex fluctuations of the network. We demonstrate the combination of *both* chemical-intuition inspired and machine-learning based atomic descriptors must be used in concert to characterize

the structural motions of the network. Armored with these new insights at room temperature, we provide a more nuanced perspective on the microscopic origin of the density maximum of water in terms of the local environments. We further examine the behavior of water under supercooling which has evoked intense debate in the literature[11]. Analyzing microsecond trajectories of water near the critical point generated recently by Sciortino and Debenedetti[37] where transitions between an HD and LD macroscopic phase are observed. While the HDL phase in supercooled water resembles the majority of local water environments in room temperature water, the situation is more complicated with LDL. Here there appear to be larger domains involving water environments that are more ice-like as well as high density-like but lower than the density of the HDL phase.

One of the key outcomes of Chapter 3 is that it establishes a rigorous theoretical protocol for studying fluctuations in liquids and aqueous solutions in general. Future directions in the development of this protocol, would be to automatize the choice of collective variables with which we understand fluctuations of the free energy landscape. This could possibly be done by constructing measures from information theory such as the Kullback–Leibler divergence [192] or the mutual information [193] that inform us about which CVs out of a large set of collective variables are most insightful about fluctuations in liquid environments. Another aspect is that constructing free energies in high dimensions naturally suffers from larger errors and therefore, coming up with more optimized ways to map the changes point free energies to atomic enironments. In this regard, it would be interesting to explore new avenues in predicting the free energies using neural networks[194].

Chapter 4 tackles the problem of the re-orientational hydrogen-bond dynamics in liquid water within the framework of the fluctuations occurring within the free energy landscape that was elucidated in Chapter 3. Specifically, we revisit the textbook accepted picture of large angular jumps in water which is currently depicted as a primarily localized event involving three active water molecules that coarse-grains away any cooperative or collective nature of the process. We develop an unsupervised protocol for detecting large changes in the orientational motion of water that does not depend on any a priori criteria of hydrogen bond interactions and also lends itself to the identification of collective orientational *swings*.

The nature of molecular environments during these swings were explored by bringing in our understanding of several of the parameters that were analyzed in Chapter 3 specifically, the local topology and density. In particular, large swings were found to occur in more disordered and low density environments. Leveraging on this connection, we demonstrate further, that large angular swings occur in a rather orchestrated fashion involving approximately 5% of the total population of water molecules. These angular swings, are manifested also in the collective nature of the creation of defects in the network and average density of water molecules which occur on the sub-to-ThZ timescale. Although these dynamics have been observed in several previous spectroscopy based experiments[59, 195, 196], pinpointing the collective origin of these modes has remained poorly understood.

If the re-orientational dynamics of water at room temperature, involves highly cooperative phenomena, how does this effect change upon supercooling or for that matter, near organic and in-organic interfaces? The spatial extent of the correlations and coupling between the topology and density in these contexts would be interesting to explore in the future. At the moment, all the studies to our knowledge

have used only variations of the localized angular jump model to examine the reorientational dynamics[49, 191, 181]. Another aspect that warrants attention is how the large angular swings are reflected if at all, in dielectric spectroscopy within the sub-ThZ regime[173]. Finally, our numerical results and molecular insights should motivate the creation of theoretical models to describe the cooperative dynamics in hydrogen bonded liquids[197] which may play an important role in tuning chemical reactions[198]. Our large angular swings are akin to the tunneling dynamic pathways of water clusters at low temperature[199] which might be an interesting starting point for building such models.

From water clusters to condensed phased liquid water in the bulk and near interfaces, there have been close to 4500 papers referenced on the broader topic of the structure, dynamics and spectroscopy of water on Martin Chaplin's resourceful website[200]. Despite long study, aqueous science continues to offer to the scientific community both theoretical and experimental challenges. In this thesis, we took an agnostic approach to investigate both the thermodynamics and dynamics of liquid water using state-of-the-art unsupervised approaches. These techniques offer the possibility of providing significantly new insights in to complexity of liquids in a wide variety of interdisciplinary contexts.

Chapter 6

Appendix A

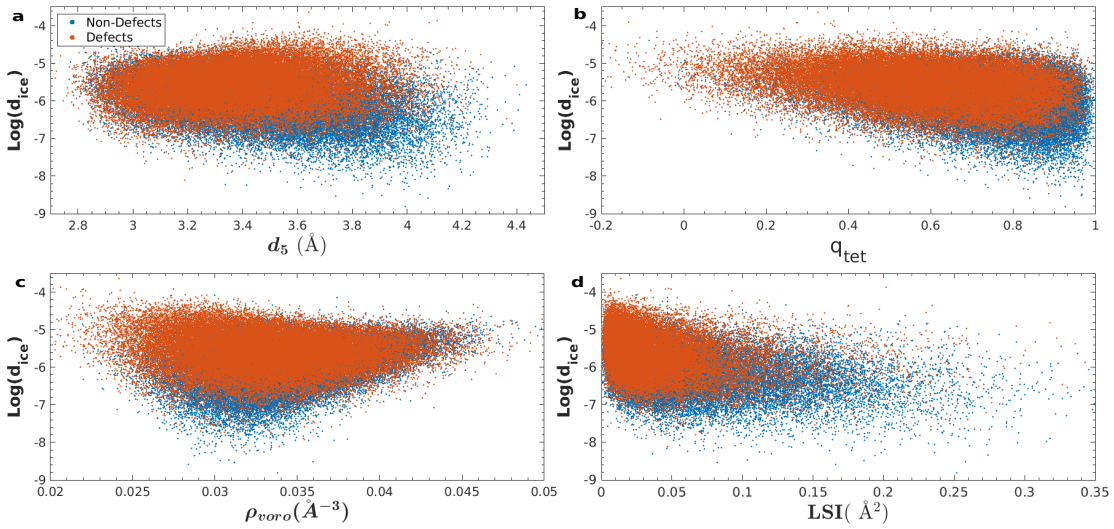


Figure 6.1: The panels a)-d) show the scatter plots of Log(d_{ice}) at 3.7 Å (including the hydrogen atom SOAP descriptors) versus the chemical-based collective variables for $q_{tet}, d_5, \rho_{voro}$ and LSI.

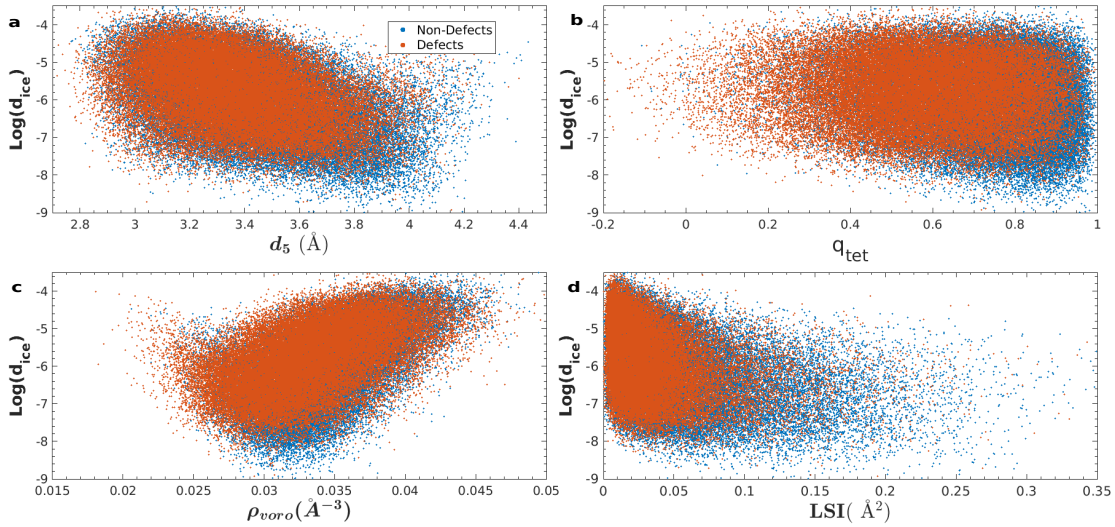


Figure 6.2: The panels a)-d) show the scatter plots of Log(d_{ice}) at 6.0 Å (including only the oxygen atoms for the SOAP descriptor) versus collective variables for $q_{tet}, d_5, \rho_{voro}$ and LSI.

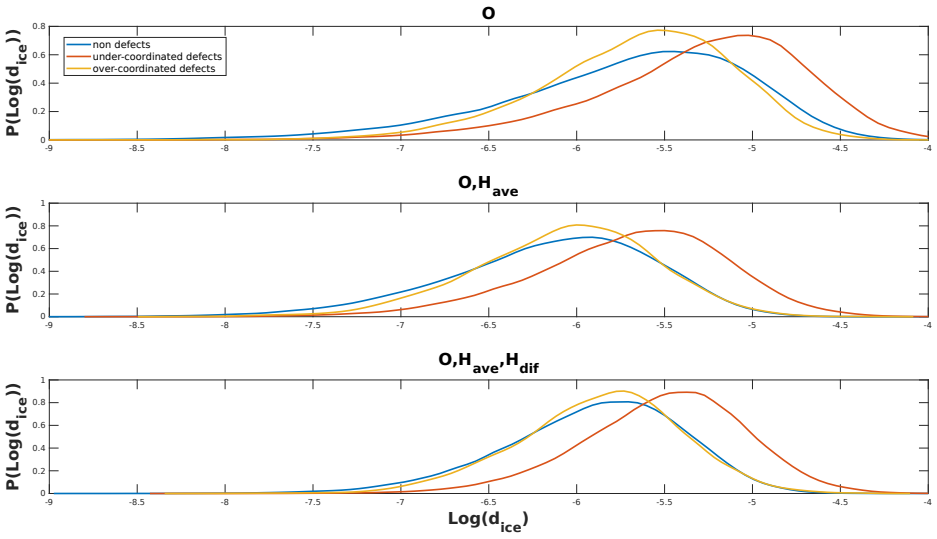


Figure 6.3: Probability density estimate of Log(d_{ice}) for non-defects, under-coordinated defects and over-coordinated defects.

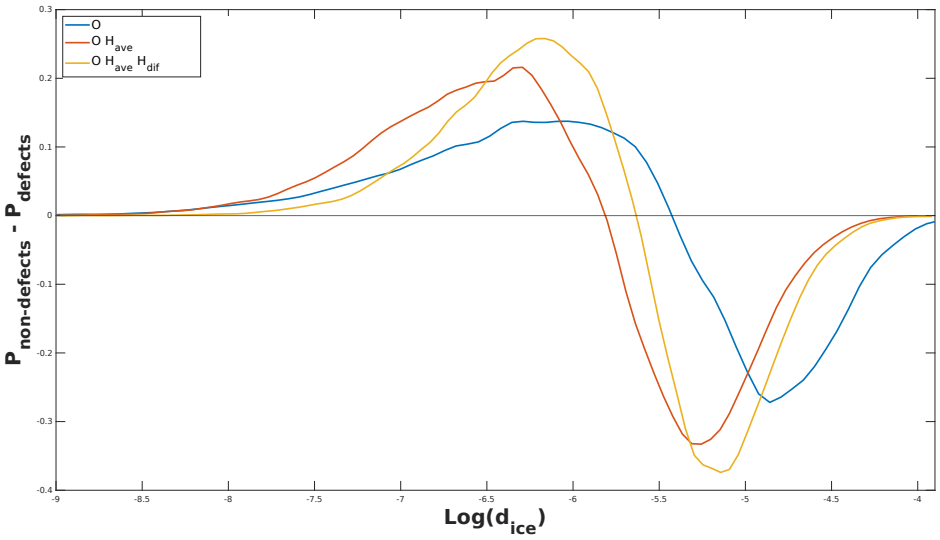


Figure 6.4: Figure shows the difference in distribution of Log(d_{ice}) of non-defects and defects for the three variations of the SOAP descriptors: ($\mathbf{O}$), ($\mathbf{OH}_{ave}$), ($\mathbf{O}, \mathbf{H}_{ave}, \mathbf{H}_{dif}$). The descriptor including both $\mathbf{H}_{ave}$ and $\mathbf{H}_{dif}$ is found to have the greatest difference between defects and non-defects.

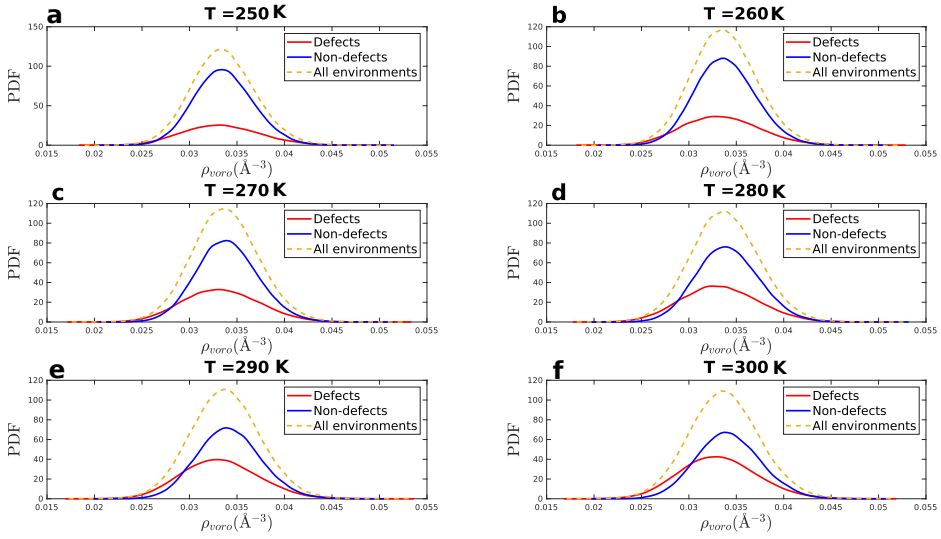


Figure 6.5: This Figures a-f show the probability distributions of the Voronoi density as a function of temperature for 250K,260K,270K,280K,290K,300K respectively. The full probability distribution(dashed yellow) is decomposed into the defect Voronoi distribution(red) and non-defect distributions(blue). The latter two distributions were reweighted by population so that the sum is the full Voronoi density

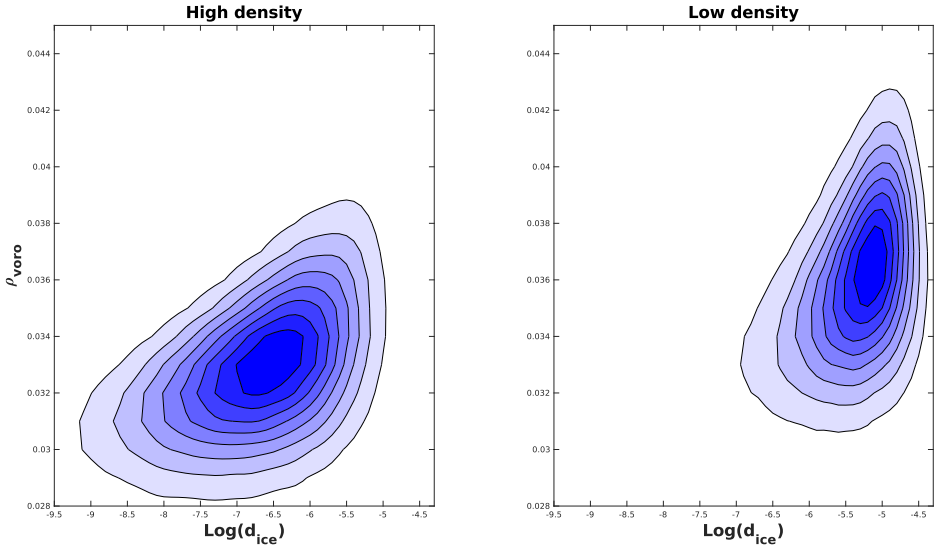


Figure 6.6: Density plot of $\log(d_{ice})$ versus ρ_{voro} for HD and LD environments of supercooled water.

Chapter 7

Appendix B

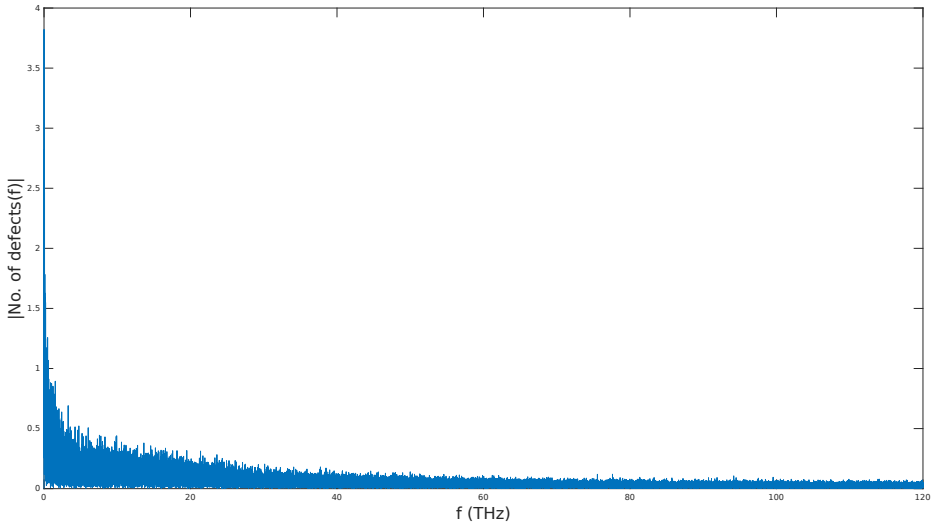


Figure 7.1: This figure shows the amplitude spectrum of the number of defects as a function of time.

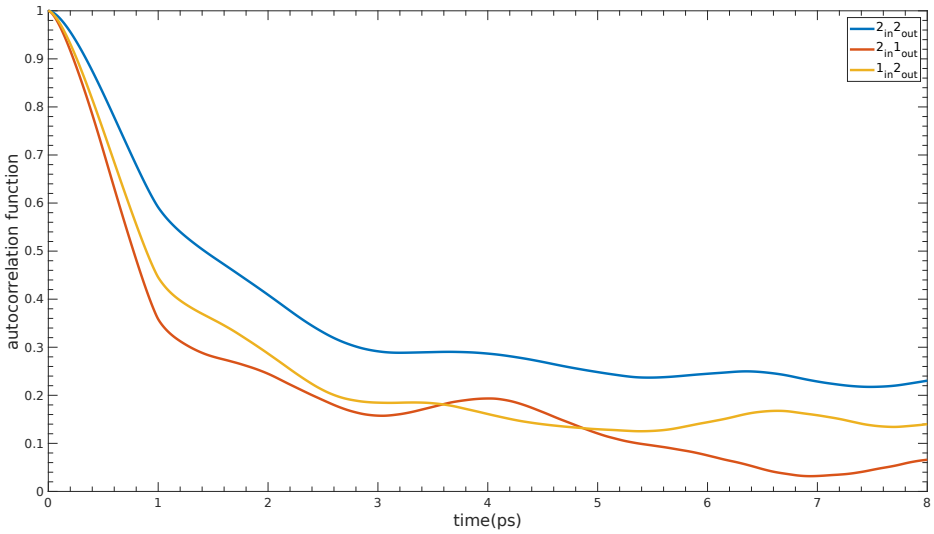


Figure 7.2: This figure shows the autocorrelation function of time non-defects and two types of defects $1_{in}2_{out}$ and $2_{in}1_{out}$.

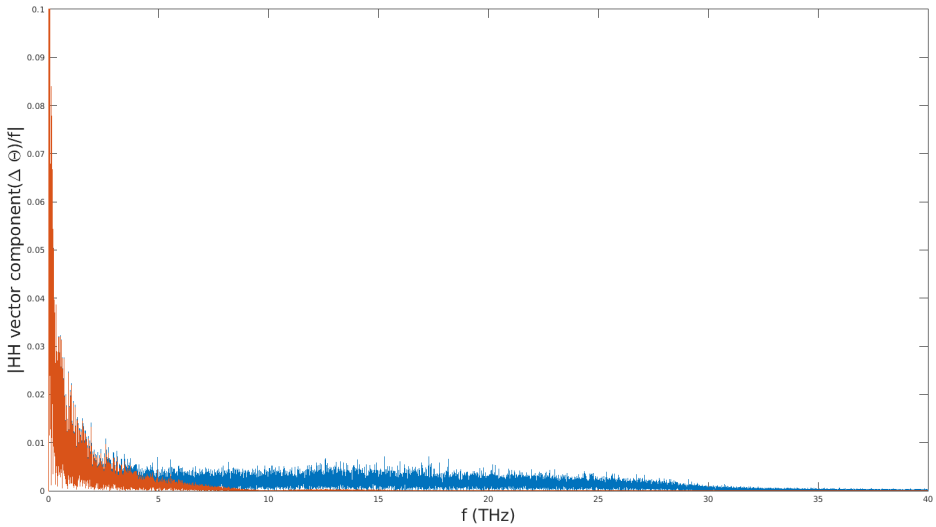


Figure 7.3: This figure shows the amplitude spectrum of the HH vector. The blue corresponds to the unfiltered time series, while the red is that of the filtered

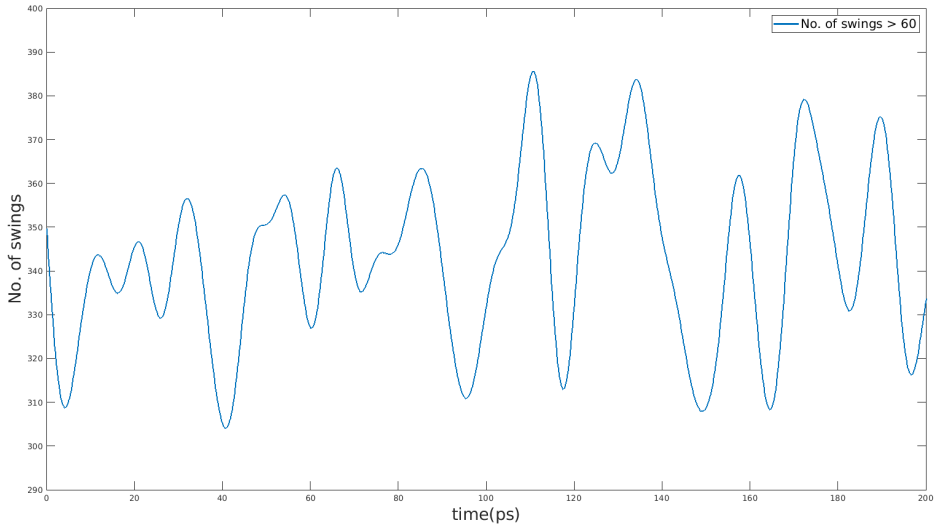


Figure 7.4: Number of swings $> 60°$ for mb-pol water model

Bibliography

[1] Felix Franks. *Water: a matrix of life*. Vol. 21. Royal Society of Chemistry, 2000.

[2] Philip Ball. "Water as an active constituent in cell biology". In: *Chemical reviews* 108.1 (2008), pp. 74–108.

[3] Myron W Evans et al. *Water in biology, chemistry and physics: experimental overviews and computational methodologies*. Vol. 9. World Scientific, 1996.

[4] RJ Speedy and CA Angell. "Isothermal compressibility of supercooled water and evidence for a thermodynamic singularity at- 45 C". In: *The Journal of Chemical Physics* 65.3 (1976), pp. 851–858.

[5] Charles A Angell. "Supercooled water". In: *Annual Review of Physical Chemistry* 34.1 (1983), pp. 593–630.

[6] AK Soper. "Is water one liquid or two?" In: *The Journal of chemical physics* 150.23 (2019), p. 234503.

[7] Paola Gallo et al. "Water: A tale of two liquids". In: *Chemical reviews* 116.13 (2016), pp. 7463–7500.

[8] Francesco Rao, Sean Garrett-Roe, and Peter Hamm. "Structural inhomogeneity of water by complex network analysis". In: *The Journal of Physical Chemistry B* 114.47 (2010), pp. 15598–15604.

[9] Elise Duboué-Dijon and Damien Laage. "Characterization of the Local Structure in Liquid Water by Various Order Parameters". In: *The Journal of Physical Chemistry B* 119.26 (June 2015), pp. 8406–8418. DOI: 10.1021/acs.jpcb.5b02936. URL: https://doi.org/10.1021/acs.jpcb.5b02936.

[10] S Myneni et al. "Spectroscopic probing of local hydrogen-bonding structures in liquid water". In: *Journal of Physics: Condensed Matter* 14.8 (2002), p. L213.

[11] Anders Nilsson and Lars GM Pettersson. "Perspective on the structure of liquid water". In: *Chemical Physics* 389.1-3 (2011), pp. 1–34.

[12] Osamu Mishima. "Volume of supercooled water under pressure and the liquid-liquid critical point". In: *The Journal of chemical physics* 133.14 (2010), p. 144503.

[13] Peter Hamm. "Markov state model of the two-state behaviour of water". In: *The Journal of chemical physics* 145.13 (2016), p. 134501.

[14] Lars GM Pettersson. "A Two-State Picture of Water and the Funnel of Life". In: *International Conference Physics of Liquid Matter: Modern Problems*. Springer. 2018, pp. 3–39.

[15] A Geiger et al. “Structure and dynamics of the hydrogen bond network in water by computer simulations”. In: *Le Journal de Physique Colloques* 45.C7 (1984), pp. C7–13.

[16] Alfons Geiger and H Eugene Stanley. “Low-Density” Patches” in the Hydrogen-Bond Network of Liquid Water: Evidence from Molecular-Dynamics Computer Simulations”. In: *Physical Review Letters* 49.24 (1982), p. 1749.

[17] Masakazu Matsumoto and Iwao Ohmine. “A new approach to the dynamics of hydrogen bond network in liquid water”. In: *The Journal of chemical physics* 104.7 (1996), pp. 2705–2712.

[18] Congcong Huang et al. “The inhomogeneous structure of water at ambient conditions”. In: *Proceedings of the National Academy of Sciences* 106.36 (2009), pp. 15214–15218.

[19] Congcong Huang et al. “Increasing correlation length in bulk supercooled H 2 O, D 2 O, and NaCl solution determined from small angle x-ray scattering”. In: *The Journal of chemical physics* 133.13 (2010), p. 134504.

[20] Wilhelm Conrad Röntgen. “Ueber die constitution des flüssigen wassers”. In: *Annalen der Physik* 281.1 (1892), pp. 91–97.

[21] Henry S Frank and Wen-Yang Wen. “Ion-solvent interaction. Structural aspects of ion-solvent interaction in aqueous solutions: a suggested picture of water structure”. In: *Discussions of the Faraday Society* 24 (1957), pp. 133–140.

[22] Arnold T Hagler, Harold A Scheraga, and George Nemethy. “Structure of liquid water. Statistical thermodynamic theory”. In: *The Journal of Physical Chemistry* 76.22 (1972), pp. 3229–3243.

[23] Vincent Holten et al. “Two-state thermodynamics of the ST2 model for supercooled water”. In: *The Journal of chemical physics* 140.10 (2014), p. 104502.

[24] Paul F McMillan. “Polyamorphic transformations in liquids and glasses”. In: *Journal of Materials Chemistry* 14.10 (2004), pp. 1506–1512.

[25] John Anthony Pople. “The molecular orbital theory of chemical valency. V. The structure of water and similar molecules”. In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 202.1070 (1950), pp. 323–336.

[26] Jichen Li and DK Ross. “Evidence for two kinds of hydrogen bond in ice”. In: *Nature* 365.6444 (1993), pp. 327–329.

[27] Frank H Stillinger and Aneesur Rahman. “Molecular dynamics study of temperature effects on water structure and kinetics”. In: *The Journal of chemical physics* 57.3 (1972), pp. 1281–1292.

[28] Frank H Stillinger and Aneesur Rahman. “Improved simulation of liquid water by molecular dynamics”. In: *The Journal of Chemical Physics* 60.4 (1974), pp. 1545–1557.

[29] Aneesur Rahman and Frank H Stillinger. “Hydrogen-bond patterns in liquid water”. In: *Journal of the American Chemical Society* 95.24 (1973), pp. 7943–7948.

[30] A Geiger, FH Stillinger, and A Rahman. "Aspects of the percolation process for hydrogen-bond networks in water". In: *The Journal of Chemical Physics* 70.9 (1979), pp. 4185–4193.

[31] Jer-Lai Kuo et al. "On the use of graph invariants for efficiently generating hydrogen bond topologies and predicting physical properties of water clusters and ice". In: *The Journal of Chemical Physics* 114.6 (2001), pp. 2527–2540.

[32] Rahul Dandekar and Ali A Hassanali. "Hierarchical lattice models of hydrogen-bond networks in water". In: *Physical Review E* 97.6 (2018), p. 062113.

[33] Ali Khosravi et al. "Ring population statistics in an ice lattice model". In: *The Journal of Chemical Physics* 155.22 (2021), p. 224502.

[34] Riccardo Foffi, John Russo, and Francesco Sciortino. "Structural and topological changes across the liquid–liquid transition in water". In: *The Journal of Chemical Physics* 154.18 (2021), p. 184506.

[35] Peter H Poole et al. "Phase behaviour of metastable water". In: *Nature* 360.6402 (1992), pp. 324–328.

[36] JLF Abascal et al. "A potential model for the study of ices and amorphous water: TIP4P/Ice". In: *The Journal of chemical physics* 122.23 (2005), p. 234511.

[37] Pablo G. Debenedetti, Francesco Sciortino, and Gül H. Zerze. "Second critical point in two realistic models of water". In: *Science* 369.6501 (2020), pp. 289–292. DOI: 10.1126/science.abb9796. eprint: https://www.science.org/doi/pdf/10.1126/science.abb9796. URL: https://www.science.org/doi/abs/10.1126/science.abb9796.

[38] Pablo G Debenedetti, Francesco Sciortino, and Gül H Zerze. "Second critical point in two realistic models of water". In: *Science* 369.6501 (2020), pp. 289–292.

[39] Frank H Stillinger. *Energy landscapes, inherent structures, and condensed-matter phenomena*. Princeton University Press, 2015.

[40] KT Wikfeldt, Anders Nilsson, and Lars GM Pettersson. "Spatially inhomogeneous bimodal inherent structure of simulated liquid water". In: *Physical Chemistry Chemical Physics* 13.44 (2011), pp. 19918–19924.

[41] Guillaume Stirnemann and Damien Laage. *Communication: On the origin of the non-Arrhenius behavior in water reorientation dynamics*. 2012.

[42] R Schulz et al. "Collective hydrogen-bond rearrangement dynamics in liquid water". In: *The Journal of chemical physics* 149.24 (2018), p. 244504.

[43] CJ Fecko et al. "Ultrafast hydrogen-bond dynamics in the infrared spectroscopy of water". In: *Science* 301.5640 (2003), pp. 1698–1702.

[44] Nicolas Giovambattista et al. "Structural order in glassy water". In: *Physical Review E* 71.6 (2005), p. 061505.

[45] Gustavo Adrian Appignanesi, JA Rodriguez Fris, and Francesco Sciortino. "Evidence of a two-state picture for supercooled water and its connections with glassy dynamics". In: *The European Physical Journal E* 29.3 (2009), pp. 305–310.

[46] Francesco Sciortino and SL Fornili. "Hydrogen bond cooperativity in simulated water: Time dependence analysis of pair interactions". In: *The Journal of chemical physics* 90.5 (1989), pp. 2786–2792.

[47] Hajime Tanaka et al. "Revealing key structural features hidden in liquids and glasses". In: *Nature Reviews Physics* 1.5 (2019), pp. 333–348.

[48] Biswajit Santra et al. "Local structure analysis in ab initio liquid water". In: *Molecular Physics* 113.17-18 (2015), pp. 2829–2841.

[49] Damien Laage and James T. Hynes. "On the Molecular Mechanism of Water Reorientation". In: *The Journal of Physical Chemistry B* 112.45 (2008). PMID: 18942871, pp. 14230–14242. DOI: 10.1021/jp805217u. eprint: https://doi.org/10.1021/jp805217u. URL: https://doi.org/10.1021/jp805217u.

[50] Yasunori Tominaga, Aiko Fujiwara, and Yuko Amo. "Dynamical structure of water by Raman spectroscopy". In: *Fluid Phase Equilibria* 144.1-2 (1998), pp. 323–330.

[51] Nikita Penkov et al. "Terahertz spectroscopy applied for investigation of water structure". In: *The Journal of Physical Chemistry B* 119.39 (2015), pp. 12664–12670.

[52] Iwao Ohmine and Hideki Tanaka. "Fluctuation, relaxations, and hydration in liquid water. Hydrogen-bond rearrangement dynamics". In: *Chemical reviews* 93.7 (1993), pp. 2545–2566.

[53] P Debye. "Polar molecules, the chemical catalog company". In: *Inc., New York* (1929), pp. 77–108.

[54] Minbiao Ji, Michael Odelius, and KJ Gaffney. "Large angular jump mechanism observed for hydrogen bond exchange in aqueous perchlorate solution". In: *Science* 328.5981 (2010), pp. 1003–1005.

[55] Damien Laage et al. "Water jump reorientation: from theoretical prediction to experimental observation". In: *Accounts of chemical research* 45.1 (2012), pp. 53–62.

[56] Damien Laage, Guillaume Stirnemann, and James T Hynes. "Why water reorientation slows without iceberg formation around hydrophobic solutes". In: *The Journal of Physical Chemistry B* 113.8 (2009), pp. 2428–2435.

[57] Francesco Paesani, Satoru Iuchi, and Gregory A Voth. "Quantum effects in liquid water from an ab initio-based polarizable force field". In: *The Journal of chemical physics* 127.7 (2007), p. 074506.

[58] Gaia Camisasca et al. "Translational and rotational dynamics of high and low density TIP4P/2005 water". In: *The Journal of chemical physics* 150.22 (2019), p. 224507.

[59] Christopher J Fecko et al. "Local hydrogen bonding dynamics and collective reorganization in water: Ultrafast infrared spectroscopy of HOD/D 2 O". In: *The Journal of chemical physics* 122.5 (2005), p. 054506.

[60] Alex Rodriguez and Alessandro Laio. "Clustering by fast search and find of density peaks". In: *science* 344.6191 (2014), pp. 1492–1496.

[61] Alex Rodriguez et al. “Computing the free energy without collective variables”. In: *Journal of chemical theory and computation* 14.3 (2018), pp. 1206–1215.

[62] Elena Facco et al. “Estimating the intrinsic dimension of datasets by a minimal neighborhood information”. In: *Scientific reports* 7.1 (2017), pp. 1–8.

[63] Paul Adrien Maurice Dirac. *The principles of quantum mechanics*. 27. Oxford university press, 1981.

[64] Attila Szabo and Neil S Ostlund. *Modern quantum chemistry: introduction to advanced electronic structure theory*. Courier Corporation, 2012.

[65] Max Born and Robert Oppenheimer. “Zur quantentheorie der molekeln”. In: *Annalen der physik* 389.20 (1927), pp. 457–484.

[66] Michael P Allen and Dominic J Tildesley. *Computer simulation of liquids*. Oxford university press, 2017.

[67] J Andrew McCammon, Bruce R Gelin, and Martin Karplus. “Dynamics of folded proteins”. In: *Nature* 267.5612 (1977), pp. 585–590.

[68] Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. “Protein folding kinetics and thermodynamics from atomistic simulation”. In: *Proceedings of the National Academy of Sciences* 109.44 (2012), pp. 17845–17850.

[69] Aneesur Rahman. “Correlations in the motion of atoms in liquid argon”. In: *Physical review* 136.2A (1964), A405.

[70] Michele Parrinello and Aneesur Rahman. “Crystal structure and pair potentials: A molecular-dynamics study”. In: *Physical review letters* 45.14 (1980), p. 1196.

[71] William L Jorgensen et al. “Comparison of simple potential functions for simulating liquid water”. In: *The Journal of chemical physics* 79.2 (1983), pp. 926–935.

[72] Herman JC Berendsen et al. “Interaction models for water in relation to protein hydration”. In: *Intermolecular forces*. Springer, 1981, pp. 331–342.

[73] Jose LF Abascal and Carlos Vega. “A general purpose model for the condensed phases of water: TIP4P/2005”. In: *The Journal of chemical physics* 123.23 (2005), p. 234505.

[74] HJC Berendsen, JR Grigera, and TP Straatsma. “The missing term in effective pair potentials”. In: *Journal of Physical Chemistry* 91.24 (1987), pp. 6269–6271.

[75] Hans W Horn et al. “Development of an improved four-site water model for biomolecular simulations: TIP4P-Ew”. In: *The Journal of chemical physics* 120.20 (2004), pp. 9665–9678.

[76] Stephen Harrington et al. “Liquid-liquid phase transition: Evidence from simulations”. In: *Physical Review Letters* 78.12 (1997), p. 2409.

[77] Sandeep K Reddy et al. “On the accuracy of the MB-pol many-body potential for water: Interaction energies, vibrational frequencies, and classical thermodynamic and dynamical properties from clusters to liquid water and ice”. In: *The Journal of chemical physics* 145.19 (2016), p. 194504.

[78] Sandeep K. Reddy et al. “On the accuracy of the MB-pol many-body potential for water: Interaction energies, vibrational frequencies, and classical thermodynamic and dynamical properties from clusters to liquid water and ice”. In: *The Journal of Chemical Physics* 145.19 (2016), p. 194504. DOI: 10.1063/1.4967719. eprint: https://doi.org/10.1063/1.4967719. URL: https://doi.org/10.1063/1.4967719.

[79] Harry Partridge and David W Schwenke. “The determination of an accurate isotope dependent potential energy surface for water from extensive ab initio calculations and experimental data”. In: *The Journal of Chemical Physics* 106.11 (1997), pp. 4618–4639.

[80] Rodney J Bartlett. “Many-body perturbation theory and coupled cluster theory for electron correlation in molecules”. In: *Annual Review of Physical Chemistry* 32.1 (1981), pp. 359–401.

[81] Maria Carolina Muniz et al. “Vapor–liquid equilibrium of water with the MB-pol many-body potential”. In: *The Journal of Chemical Physics* 154.21 (2021), p. 211103.

[82] Loup Verlet. “Computer” experiments” on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules”. In: *Physical review* 159.1 (1967), p. 98.

[83] William C Swope et al. “A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters”. In: *The Journal of chemical physics* 76.1 (1982), pp. 637–649.

[84] Charles K Birdsall and A Bruce Langdon. *Plasma physics via computer simulation*. CRC press, 2018.

[85] Jean-Paul Ryckaert, Giovanni Ciccotti, and Herman JC Berendsen. “Numerical integration of the cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes”. In: *Journal of computational physics* 23.3 (1977), pp. 327–341.

[86] Hans C Andersen. “Rattle: A “velocity” version of the shake algorithm for molecular dynamics calculations”. In: *Journal of computational Physics* 52.1 (1983), pp. 24–34.

[87] Hans C Andersen. “Molecular dynamics simulations at constant pressure and/or temperature”. In: *The Journal of chemical physics* 72.4 (1980), pp. 2384–2393.

[88] Shuichi Nosé. “A unified formulation of the constant temperature molecular dynamics methods”. In: *The Journal of chemical physics* 81.1 (1984), pp. 511–519.

[89] William G Hoover. “Canonical dynamics: Equilibrium phase-space distributions”. In: *Physical review A* 31.3 (1985), p. 1695.

[90] Giovanni Bussi, Davide Donadio, and Michele Parrinello. “Canonical sampling through velocity rescaling”. In: *The Journal of chemical physics* 126.1 (2007), p. 014101.

[91] Herman JC Berendsen et al. “Molecular dynamics with coupling to an external bath”. In: *The Journal of chemical physics* 81.8 (1984), pp. 3684–3690.

[92] Michele Parrinello and Aneesur Rahman. “Polymorphic transitions in single crystals: A new molecular dynamics method”. In: *Journal of Applied physics* 52.12 (1981), pp. 7182–7190.

[93] Giovanni Bussi and Alessandro Laio. “Using metadynamics to explore complex free-energy landscapes”. In: *Nature Reviews Physics* 2.4 (Apr. 2020), pp. 200–212. ISSN: 2522-5820. DOI: 10.1038/s42254-020-0153-0. URL: https://doi.org/10.1038/s42254-020-0153-0.

[94] John Russo and Hajime Tanaka. “Understanding water’s anomalies with locally favoured structures”. In: *Nature communications* 5.1 (2014), pp. 1–11.

[95] Olle Björneholm et al. “Water at interfaces”. In: *Chemical reviews* 116.13 (2016), pp. 7698–7726.

[96] P-L Chau and AJ Hardwick. “A new order parameter for tetrahedral configurations”. In: *Molecular Physics* 93.3 (1998), pp. 511–518.

[97] RM Lynden-Bell and Pablo G Debenedetti. “Computational investigation of order, structure, and dynamics in modified water models”. In: *The Journal of Physical Chemistry B* 109.14 (2005), pp. 6527–6534.

[98] Ivan Saika-Voivod, Francesco Sciortino, and Peter H Poole. “Computer simulations of liquid silica: Equation of state and liquid–liquid phase transition”. In: *Physical Review E* 63.1 (2000), p. 011202.

[99] Richard H Henchman and Sheeba Jem Irudayam. “Topological hydrogen-bond definition to characterize the structure and dynamics of liquid water”. In: *The Journal of Physical Chemistry B* 114.50 (2010), pp. 16792–16810.

[100] Piero Gasparotto, Ali A Hassanali, and Michele Ceriotti. “Probing defects and correlations in the hydrogen-bond network of ab initio water”. In: *Journal of chemical theory and computation* 12.4 (2016), pp. 1953–1964.

[101] Emiliano Poli, Kwang H. Jong, and Ali Hassanali. “Charge transfer as a ubiquitous mechanism in determining the negative charge at hydrophobic interfaces”. In: *Nature Communications* 11.1 (Feb. 2020), p. 901. ISSN: 2041-1723. DOI: 10.1038/s41467-020-14659-5. URL: https://doi.org/10.1038/s41467-020-14659-5.

[102] Narjes Ansari et al. “Insights into the Emerging Networks of Voids in Simulated Supercooled Water”. In: *The Journal of Physical Chemistry B* 124.11 (2020). PMID: 32032486, pp. 2180–2190. DOI: 10.1021/acs.jpcb.9b10144. eprint: https://doi.org/10.1021/acs.jpcb.9b10144. URL: https://doi.org/10.1021/acs.jpcb.9b10144.

[103] Alenka Luzar and David Chandler. “Hydrogen-bond kinetics in liquid water”. In: *Nature* 379.6560 (1996), pp. 55–57.

[104] Yu-ling Yeh and Chung-Yuan Mou. “Orientational relaxation dynamics of liquid water studied by molecular dynamics simulation”. In: *The Journal of Physical Chemistry B* 103.18 (1999), pp. 3699–3705.

[105] John D Bernal. “A geometrical approach to the structure of liquids”. In: *Nature* 183.4655 (1959), pp. 141–147.

[106] Eugene Wigner and Frederick Seitz. “On the constitution of metallic sodium”. In: *Physical Review* 43.10 (1933), p. 804.

[107] J Bohm, RB Heimann, and M Bohm. “Voronoi polyhedra: A useful tool to determine the symmetry and Bravais class of crystal lattices”. In: *Crystal Research and Technology* 31.8 (1996), pp. 1069–1075.

[108] Matthias Rupp et al. “Fast and accurate modeling of molecular atomization energies with machine learning”. In: *Physical review letters* 108.5 (2012), p. 058301.

[109] Ali Sadeghi et al. “Metrics for measuring distances in configuration spaces”. In: *The Journal of chemical physics* 139.18 (2013), p. 184118.

[110] Albert P Bartók, Risi Kondor, and Gábor Csányi. “On representing chemical environments”. In: *Physical Review B* 87.18 (2013), p. 184115.

[111] Ralf Drautz. “Atomic cluster expansion for accurate and transferable interatomic potentials”. In: *Physical Review B* 99.1 (2019), p. 014104.

[112] Jörg Behler and Michele Parrinello. “Generalized neural-network representation of high-dimensional potential-energy surfaces”. In: *Physical review letters* 98.14 (2007), p. 146401.

[113] Riccardo Capelli, Francesco Muniz-Miranda, and Giovanni M. Pavan. “Ephemeral ice-like local environments in classical models of liquid water”. In: (2021). arXiv: `2112.02022 [physics.chem-ph]`.

[114] Volker L Deringer and Gábor Csányi. “Machine learning based interatomic potential for amorphous carbon”. In: *Physical Review B* 95.9 (2017), p. 094203.

[115] Dmitrii Maksimov, Carsten Baldauf, and Mariana Rossi. “The conformational space of a flexible amino acid at metallic surfaces”. In: *International Journal of Quantum Chemistry* 121.3 (2021), e26369.

[116] Bartomeu Monserrat et al. “Liquid water contains the building blocks of diverse ice phases”. In: *Nature communications* 11.1 (2020), pp. 1–8.

[117] RICHARD Bellman. “Dynamic programming, princeton univ”. In: *Press Princeton, New Jersey* (1957).

[118] Wei Bin How et al. “Significance of the Chemical Environment of an Element in Nonadiabatic Molecular Dynamics: Feature Selection and Dimensionality Reduction with Machine Learning”. In: *The Journal of Physical Chemistry Letters* 12 (2021), pp. 12026–12032.

[119] T Mendes-Santos et al. “Unsupervised learning universal critical behavior via the intrinsic dimension”. In: *Physical Review X* 11.1 (2021), p. 011040.

[120] T. Mendes-Santos et al. “Intrinsic Dimension of Path Integrals: Data-Mining Quantum Criticality and Emergent Simplicity”. In: *PRX Quantum* 2 (3 Aug. 2021), p. 030332. DOI: `10.1103/PRXQuantum.2.030332`. URL: `https://link.aps.org/doi/10.1103/PRXQuantum.2.030332`.

[121] Peter Grassberger and Itamar Procaccia. “Measuring the strangeness of strange attractors”. In: *The theory of chaotic attractors*. Springer, 2004, pp. 170–189.

[122] Francesco Camastra and Alessandro Vinciarelli. “Estimating the intrinsic dimension of data with a fractal-based method”. In: *IEEE Transactions on pattern analysis and machine intelligence* 24.10 (2002), pp. 1404–1407.

[123] Sebastian Mika et al. “Kernel PCA and De-noising in feature spaces.” In: *NIPS*. Vol. 11. 1998, pp. 536–542.

[124] Ian Jolliffe. “Principal component analysis”. In: *Encyclopedia of statistics in behavioral science* (2005).

[125] Joseph B Kruskal. *Multidimensional scaling*. 11. Sage, 1978.

[126] Joshua B Tenenbaum, Vin De Silva, and John C Langford. “A global geometric framework for nonlinear dimensionality reduction”. In: *science* 290.5500 (2000), pp. 2319–2323.

[127] Laurens van der Maaten and Geoffrey Hinton. “Visualizing Data using t-SNE”. In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605. URL: `http://jmlr.org/papers/v9/vandermaaten08a.html`.

[128] Leland McInnes, John Healy, and James Melville. “Umap: Uniform manifold approximation and projection for dimension reduction”. In: *arXiv preprint arXiv:1802.03426* (2018).

[129] Elizaveta Levina and Peter J Bickel. “Maximum likelihood estimation of intrinsic dimension”. In: *Advances in neural information processing systems*. 2005, pp. 777–784.

[130] Claudio Ceruti et al. “Danco: An intrinsic dimensionality estimator exploiting angle and norm concentration”. In: *Pattern recognition* 47.8 (2014), pp. 2569–2581.

[131] Narjes Ansari, Alessandro Laio, and Ali Hassanali. “Spontaneously forming dendritic voids in liquid water can host small polymers”. In: *The journal of physical chemistry letters* 10.18 (2019), pp. 5585–5591.

[132] Matteo Carli et al. “Candidate binding sites for allosteric inhibition of the SARS-CoV-2 main protease from the analysis of large-scale molecular dynamics simulations”. In: *The journal of physical chemistry letters* 12.1 (2020), pp. 65–72.

[133] KwangHyok Jong and Ali A Hassanali. “A data science approach to understanding water networks around biomolecules: the case of tri-alanine in liquid water”. In: *The Journal of Physical Chemistry B* 122.32 (2018), pp. 7895–7906.

[134] Francesco Camastra and Antonino Staiano. “Intrinsic dimension estimation: Advances and open problems”. In: *Information Sciences* 328 (2016), pp. 26–41.

[135] Shankar Kumar, Philip W Payne, and Maximiliano Vásquez. “Method for free-energy calculations using iterative techniques”. In: *Journal of computational chemistry* 17.10 (1996), pp. 1269–1275.

[136] EA Carter et al. “Constrained reaction coordinate dynamics for the simulation of rare events”. In: *Chemical Physics Letters* 156.5 (1989), pp. 472–477.

[137] Shankar Kumar et al. “Multidimensional free-energy calculations using the weighted histogram analysis method”. In: *Journal of Computational Chemistry* 16.11 (1995), pp. 1339–1350.

[138] Miguel A. Carreira-Perpinan. "Mode-finding for mixtures of Gaussian distributions". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22.11 (2000), pp. 1318–1323.

[139] Karl Pearson. "X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material". In: *Philosophical Transactions of the Royal Society of London.(A.)* 186 (1895), pp. 343–414.

[140] YP Mack and Murray Rosenblatt. "Multivariate k-nearest neighbor density estimates". In: *Journal of Multivariate Analysis* 9.1 (1979), pp. 1–15.

[141] Giulia Sormani, Alex Rodriguez, and Alessandro Laio. "Explicit Characterization of the Free-Energy Landscape of a Protein in the Space of All Its Cα Carbons". In: *Journal of chemical theory and computation* 16.1 (2019), pp. 80–87.

[142] Maria d'Errico et al. "Automatic topography of high-dimensional data sets by non-parametric Density Peak clustering". In: *arXiv preprint arXiv:1802.10549* (2018).

[143] R Elber and Martin Karplus. "Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin". In: *Science* 235.4786 (1987), pp. 318–321.

[144] SR Accordino et al. "Quantitative investigation of the two-state picture for water in the normal liquid and the supercooled regime". In: *The European Physical Journal E* 34.5 (2011), pp. 1–7.

[145] Richard H Henchman. "Water's dual nature and its continuously changing hydrogen bonds". In: *Journal of Physics: Condensed Matter* 28.38 (2016), p. 384001.

[146] Thomas D. Kühne and Rustam Z. Khaliullin. "Electronic signature of the instantaneous asymmetry in the first coordination shell of liquid water". In: *Nature Communications* 4.1 (Feb. 2013), p. 1450. ISSN: 2041-1723. DOI: 10.1038/ncomms2459. URL: https://doi.org/10.1038/ncomms2459.

[147] Zhenyu Yan et al. "Structure of the first-and second-neighbor shells of simulated water: Quantitative relation to translational and orientational order". In: *Physical Review E* 76.5 (2007), p. 051201.

[148] Megan J Cuthbertson and Peter H Poole. "Mixturelike behavior near a liquid-liquid phase transition in simulations of supercooled water". In: *Physical review letters* 106.11 (2011), p. 115706.

[149] Noam Agmon. "Liquid water: From symmetry distortions to diffusive motion". In: *Accounts of chemical research* 45.1 (2012), pp. 63–73.

[150] N. Ansari et al. "High and low density patches in simulated liquid water". In: *The Journal of Chemical Physics* 149.20 (2018), p. 204507. DOI: 10.1063/1.5053559. eprint: https://doi.org/10.1063/1.5053559. URL: https://doi.org/10.1063/1.5053559.

[151] Henk Bekker et al. "Gromacs-a parallel computer for molecular-dynamics simulations". In: *4th International Conference on Computational Physics (PC 92)*. World Scientific Publishing. 1993, pp. 252–256.

[152] R Hockney and J Eastwood. “Computer simulations using particles mcgraw-hill”. In: *New York* (1981).

[153] Narjes Ansari et al. “Insights into the Emerging Networks of Voids in Simulated Supercooled Water”. In: *The Journal of Physical Chemistry B* 124.11 (2020). PMID: 32032486, pp. 2180–2190. DOI: 10.1021/acs.jpcb.9b10144. eprint: https://doi.org/10.1021/acs.jpcb.9b10144. URL: https://doi.org/10.1021/acs.jpcb.9b10144.

[154] Lauri Himanen et al. “DScribe: Library of descriptors for machine learning in materials science”. In: *Computer Physics Communications* 247 (2020), p. 106949.

[155] Albert P Bartók et al. “Machine learning a general-purpose interatomic potential for silicon”. In: *Physical Review X* 8.4 (2018), p. 041048.

[156] Gaia Camisasca et al. “A proposal for the structure of high- and low-density fluctuations in liquid water”. In: *The Journal of Chemical Physics* 151.3 (2019), p. 034508. DOI: 10.1063/1.5100875. eprint: https://doi.org/10.1063/1.5100875. URL: https://doi.org/10.1063/1.5100875.

[157] Etienne Becht et al. “Dimensionality reduction for visualizing single-cell data using UMAP”. In: *Nature Biotechnology* 37.1 (Jan. 2019), pp. 38–44. ISSN: 1546-1696. DOI: 10.1038/nbt.4314. URL: https://doi.org/10.1038/nbt.4314.

[158] Cui Zhang and Giulia Galli. “Dipolar correlations in liquid water”. In: *The Journal of Chemical Physics* 141.8 (2014), p. 084504. DOI: 10.1063/1.4893638. eprint: https://doi.org/10.1063/1.4893638. URL: https://doi.org/10.1063/1.4893638.

[159] Alex Diaz-Papkovich et al. “UMAP reveals cryptic population structure and phenotype heterogeneity in large genomic cohorts”. In: *PLoS genetics* 15.11 (2019), e1008432.

[160] Eli Shiratani and Masaki Sasai. “Molecular scale precursor of the liquid–liquid phase transition of water”. In: *The Journal of chemical physics* 108.8 (1998), pp. 3264–3276.

[161] Helena L Pi et al. “Anomalies in water as obtained from computer simulations of the TIP4P/2005 model: density maxima, and density, isothermal compressibility and heat capacity minima”. In: *Molecular Physics* 107.4-6 (2009), pp. 365–374.

[162] Anders Nilsson and Lars GM Pettersson. “The structural origin of anomalous properties of liquid water”. In: *Nature communications* 6.1 (2015), pp. 1–11.

[163] Francesco Mallamace et al. “The anomalous behavior of the density of water in the range 30 K¡ T¡ 373 K”. In: *Proceedings of the National Academy of Sciences* 104.47 (2007), pp. 18387–18391.

[164] Pablo G Debenedetti. “Supercooled and glassy water”. In: *Journal of Physics: Condensed Matter* 15.45 (2003), R1669.

[165] Rakesh S Singh et al. “Two-state thermodynamics and the possibility of a liquid-liquid phase transition in supercooled TIP4P/2005 water”. In: *The Journal of chemical physics* 144.14 (2016), p. 144504.

[166] A Taschin et al. “Evidence of two distinct local structures of water from ambient to supercooled conditions”. In: *Nature communications* 4.1 (2013), pp. 1–8.

[167] Ali Hassanali Adu-Offei Danso and Alex Rodriguez. “Manuscript in preparation”. In: ().

[168] Koji Ando and James T Hynes. “HF acid ionization in water: the first step”. In: *Faraday discussions* 102 (1995), pp. 435–441.

[169] Ali Hassanali et al. “Proton transfer through the water gossamer”. In: *Proceedings of the National Academy of Sciences* 110.34 (2013), pp. 13723–13728.

[170] Rajib Biswas et al. “IR spectral assignments for the hydrated excess proton in liquid water”. In: *The Journal of chemical physics* 146.15 (2017), p. 154507.

[171] Krupa Ramasesha et al. “Ultrafast 2D IR anisotropy of water reveals reorientation during hydrogen-bond switching”. In: *The Journal of chemical physics* 135.5 (2011), p. 054509.

[172] Rebecca A Nicodemus et al. “Collective hydrogen bond reorganization in water studied with temperature-dependent ultrafast infrared spectroscopy”. In: *The Journal of Physical Chemistry B* 115.18 (2011), pp. 5604–5616.

[173] Ivan Popov et al. “The mechanism of the dielectric relaxation in water”. In: *Phys. Chem. Chem. Phys.* 18 (20 2016), pp. 13941–13953. DOI: 10.1039/C6CP02195F. URL: http://dx.doi.org/10.1039/C6CP02195F.

[174] KS Singwi and Alf Sjölander. “Diffusive motions in water and cold neutron scattering”. In: *Physical Review* 119.3 (1960), p. 863.

[175] David Eisenberg and Walter Kauzmann. *The structure and properties of water*. OUP Oxford, 2005.

[176] Aneesur Rahman and Frank H Stillinger. “Molecular dynamics study of liquid water”. In: *The Journal of Chemical Physics* 55.7 (1971), pp. 3336–3359.

[177] Iwao Ohmine, Hideki Tanaka, and Peter G Wolynes. “Large local energy fluctuations in water. II. Cooperative motions and fluctuations”. In: *The Journal of chemical physics* 89.9 (1988), pp. 5852–5860.

[178] Biman Bagchi. “Water dynamics in the hydration layer around proteins and micelles”. In: *Chemical Reviews* 105.9 (2005), pp. 3197–3219.

[179] KE Larsson. “Rotational and translational diffusion in complex liquids”. In: *Physical Review* 167.1 (1968), p. 171.

[180] K-E Larsson. “Liquid dynamics”. In: *Inelastic Scattering of Neutrons. Vol. II. Proceedings of the Symposium on Inelastic Scattering of Neutrons*. 1965.

[181] Damien Laage et al. “Water Jump Reorientation: From Theoretical Prediction to Experimental Observation”. In: *Accounts of Chemical Research* 45.1 (2012). PMID: 21749157, pp. 53–62. DOI: 10.1021/ar200075u. eprint: https://doi.org/10.1021/ar200075u. URL: https://doi.org/10.1021/ar200075u.

[182] David E Moilanen et al. “Ion–water hydrogen-bond switching observed with 2D IR vibrational echo chemical exchange spectroscopy”. In: *Proceedings of the National Academy of Sciences* 106.2 (2009), pp. 375–380.

[183] Matthias Heyden et al. “Dissecting the THz spectrum of liquid water from first principles via correlations in time and space”. In: *Proceedings of the National Academy of Sciences* 107.27 (2010). Publisher: National Academy of Sciences _eprint: https://www.pnas.org/content/107/27/12068.full.pdf, pp. 12068–12073. ISSN: 0027-8424. DOI: 10.1073/pnas.0914885107. URL: https://www.pnas.org/content/107/27/12068.

[184] R Podeszwa and V Buch. “Structure and dynamics of orientational defects in ice”. In: *Physical review letters* 83.22 (1999), p. 4570.

[185] Volodymyr Babin, Claude Leforestier, and Francesco Paesani. “Development of a “first principles” water potential with flexible monomers: Dimer potential energy surface, VRT spectrum, and second virial coefficient”. In: *Journal of chemical theory and computation* 9.12 (2013), pp. 5395–5403.

[186] Stephen Butterworth et al. “On the theory of filter amplifiers”. In: *Wireless Engineer* 7.6 (1930), pp. 536–541.

[187] Chao Liu, Wenfei Li, and Wei Wang. “Correlation of reorientational jumps of water molecules in bulk water”. In: *Physical Review E* 87.5 (2013), p. 052309.

[188] Marco G Mazza et al. “Relation between rotational and translational dynamic heterogeneities in water”. In: *Physical review letters* 96.5 (2006), p. 057803.

[189] Chao Liu et al. “Interplay between translational diffusion and large-amplitude angular jumps of water molecules”. In: *The Journal of Chemical Physics* 148.18 (2018), p. 184502.

[190] JD Eaves et al. “Hydrogen bonds in liquid water are broken only fleetingly”. In: *Proceedings of the National Academy of Sciences* 102.37 (2005), pp. 13019–13022.

[191] Damien Laage, Thomas Elsaesser, and James T Hynes. “Water dynamics in the hydration shells of biomolecules”. In: *Chemical Reviews* 117.16 (2017), pp. 10694–10725.

[192] S Kullback and RA Leibler. “10.1214/aoms/1177729694”. In: *Ann. Math. Stat* 22 (1951), pp. 79–86.

[193] Claude Elwood Shannon. “A mathematical theory of communication”. In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[194] Tobias Lemke and Christine Peter. “Neural network based prediction of conformational free energies-a new route toward coarse-grained simulation models”. In: *Journal of chemical theory and computation* 13.12 (2017), pp. 6213–6221.

[195] Joseph J Loparo, Sean T Roberts, and Andrei Tokmakoff. “Multidimensional infrared spectroscopy of water. II. Hydrogen bond switching dynamics”. In: *The Journal of chemical physics* 125.19 (2006), p. 194522.

[196] Robert Laenen, Konstantinos Simeonidis, and Alfred Laubereau. “Time resolved spectroscopy of water in the infrared: New data and discussion”. In: *Bulletin of the Chemical Society of Japan* 75.5 (2002), pp. 925–932.

[197] S. Perticaroli et al. “Water-like Behavior of Formamide: Jump Reorientation Probed by Extended Depolarized Light Scattering”. In: *The Journal of Physical Chemistry Letters* 9.1 (2018). PMID: 29243934, pp. 120–125. DOI: 10.1021/acs.jpclett.7b02943. eprint: https://doi.org/10.1021/acs.jpclett.7b02943. URL: https://doi.org/10.1021/acs.jpclett.7b02943.

[198] Manuel F. Ruiz-Lopez et al. “Molecular reactions at aqueous interfaces”. In: *Nature Reviews Chemistry* 4.9 (Sept. 2020), pp. 459–475. ISSN: 2397-3358. DOI: 10.1038/s41570-020-0203-2. URL: https://doi.org/10.1038/s41570-020-0203-2.

[199] Jeremy O. Richardson et al. “Concerted hydrogen-bond breaking by quantum tunneling in the water hexamer prism”. In: *Science* 351.6279 (2016), pp. 1310–1313. DOI: 10.1126/science.aae0012. eprint: https://www.science.org/doi/pdf/10.1126/science.aae0012. URL: https://www.science.org/doi/abs/10.1126/science.aae0012.

[200] Martin Chaplin. *Water Structure and Science*. URL: https://water.lsbu.ac.uk/water/water_structure_science.html.