


Beyond Local Structures in Critical Supercooled Water through Unsupervised Learning

Edward Danquah Donkor, Adu Offei-Danso, Alex Rodriguez,* Francesco Sciortino, and Ali Hassanali*

 Cite This: *J. Phys. Chem. Lett.* 2024, 15, 3996–4005

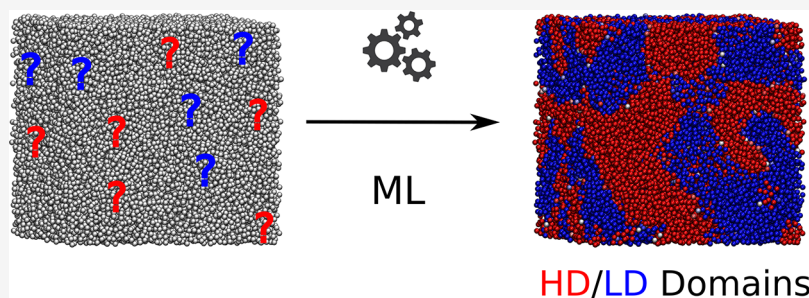
 Read Online

ACCESS |

 Metrics & More

 Article Recommendations

 Supporting Information



ABSTRACT: The presence of a second critical point in water has been a topic of intense investigation for the last few decades. The molecular origins underlying this phenomenon are typically rationalized in terms of the competition between local high-density (HD) and low-density (LD) structures. Their identification often requires designing parameters that are subject to human intervention. Herein, we use unsupervised learning to discover structures in atomistic simulations of water close to the liquid–liquid critical point (LLCP). Encoding the information on the environment using local descriptors, we do not find evidence for two distinct thermodynamic structures. In contrast, when we deploy *nonlocal* descriptors that probe instead heterogeneities on the nanometer length scale, this leads to the emergence of LD and HD domains rationalizing the microscopic origins of the density fluctuations close to criticality.

The physics of the critical behavior of matter close to phase transitions remains one of the most cherished areas of study in both experimental and theoretical physics.^{1–4} One of the most lively areas of discussion in this regard pertains to the microscopic origins of the complex phase diagram of water.^{5–8} Besides the rather well characterized liquid–gas critical point, a series of theoretical predictions over the last few decades have proposed the existence of another critical point: the Liquid Liquid Critical Point (LLCP) of water in the supercooled regime.^{9–11} The physics underlying this criticality is thought to be one of the essential ingredients for understanding the anomalies of water.

Probing the molecular origins of this second critical point has been dominated by theoretical and numerical predictions due to the challenge of spontaneous nucleation of ice at supercooled conditions.^{12–14} Just over three decades ago, Poole and co-workers demonstrated using the ST2 water model,¹⁵ that deeply supercooled water showed the presence of two distinct liquid phases, with fluctuations between the two phases terminating at the LLCP.⁹ Several groups have also shown in the past decade, using advanced sampling free energy calculations, that the ST2 model exhibits two distinct liquid phases.^{10,16–19} More recently, this has been bolstered by *tour-de-force* microsecond simulations of realistic classical models of liquid water,¹¹ as well as ab initio neural network models of

liquid water²⁰ that give further evidence for a LLCP scenario at least, on the numerical front. On the experimental side, work on supercooled water under elevated pressures as well as pioneering sound velocity measurements appear to be breaking the boundary of the so-called *no-mans land* giving strong indications of the existence of a LLCP.^{21–23}

One of the central holy grails of understanding the possible polymorphic nature of liquid water, has been the use of locally stable structures^{7,24–28} which are thought to be rooted in water’s unique hydrogen bond network. Many of the anomalies in water have been rationalized in terms of a competition between two types of hydrogen bonding structures.^{6,25} One is said to have a more ordered tetrahedral and therefore open structure, often referred to as a low-density (LD) local configuration, and the other is disordered due to the presence of interstitial water molecules, which is referred to as a high-density (HD) local configuration.

Received: February 5, 2024

Revised: March 29, 2024

Accepted: April 1, 2024

Published: April 4, 2024



Numerous order parameters have been constructed in an attempt to identify and distinguish these two local environments.^{7,25,28–34} These order parameters have also been shown to be tightly coupled to the macroscopic density fluctuations that occur close to the critical point (CP³⁵). Although they provide a manner in which to physically interpret the simulation data, these order parameters often require significant human intervention, which necessarily involves chemical bias (and often some arbitrary cutoff in their definition). Furthermore, it is also not clear a priori whether the interpretations made through these parameters are transferable across different regions of the phase diagram.

Recently, some of us proposed a protocol that streamlines an unsupervised learning procedure for liquids and applied it to study the structure of water at room temperature³⁶ as well as the study of the excess proton in hydrochloric acid.³⁷ In brief, the method involves a three-step process. We begin by encoding the information on local environments using local atomic descriptors computed from the Smooth Overlap of Atomic Positions (SOAP³⁸), which preserve important symmetries when comparing different molecular structures.³⁹ In the second step, these high-dimensional descriptors are subsequently processed through an algorithm that extracts the Intrinsic Dimension (ID)⁴⁰ which is crucial to understand the embedding manifold of the data. In the final step, the ID is used to extract the high-dimensional free energy of the system⁴¹ and identify the minima.^{42,43} For room temperature liquid water, we found a rather broad and rough landscape separated by small barriers on the order of thermal energy where the shallow minima arise from a continuum of local molecular structures that continuously connect the canonical low- or high-density local environments.³⁶

In this work, we apply this protocol to understand fluctuations in supercooled water. Specifically, we uncover the molecular origins of critical-like fluctuations using unsupervised learning, analyzing trajectories recently reported in ref 11 connected to the presence of a second critical point in atomistic water models. The free energy landscape constructed using local SOAP descriptors results in a single minimum despite there being macroscopic fluctuations in the global density. By systematically expanding the SOAP descriptor to include fluctuations on a length scale up to 1 nm, we uncover nonlocal domains relevant to critical-like fluctuations in supercooled water. The free energy landscape close to the critical point evolves between the high- and low-density macroscopic phases, through a complex topography which we link to collective fluctuations of chemical-based order parameters that include nonlocal information of the water network.

We begin by summarizing the methods we employ in our work. The trajectory used for our analysis obtained from ref 11 is a 40 μ s long *NPT* trajectory of 300 TIP4P/2005 water molecules produced using the Gromacs 5.1.4 software close to critical conditions (177 K, 1751 bar). More information on the simulation conditions is detailed in the main and supplementary text of ref 11. Additionally, we apply our analysis protocol to a large system (36424 TIP4P/2005 water molecules) in the *NVT* ensemble at a temperature and density of (180 K, 1011.83 kg/m³) in order to explore the larger length-scale fluctuations in the density.

Our unsupervised learning protocol developed in ref 36 involves encoding local environments of molecules in a local atomic descriptor, extracting the intrinsic dimension, and

constructing a high-dimensional point-dependent probability density function from which the thermodynamic information can be inferred. The details of this procedure are outlined in the following paragraphs.

As indicated earlier, the first step in our analysis is to encode the water molecular environments with a local atomic descriptor. To this end, we use the Smooth Overlap of Atomic Positions (SOAP) descriptor,^{38,39} which preserves rotational, translational, and permutational symmetries of our molecular environments. In brief, given an atomic environment χ around a central atom, one characterizes the local density as a sum of Gaussian functions with variance σ^2 centered on each of the neighbors of the central atom, including the central atom itself:

$$\rho_{\chi}(\mathbf{r}) = \sum_{j \in \chi} \exp\left(\frac{-|\mathbf{r} - \mathbf{r}_j|^2}{2\sigma^2}\right) \quad (1)$$

This atomic neighbor density can be expanded in terms of radial basis functions and spherical harmonics Y_{lm} such that

$$\rho_{\chi}(\mathbf{r}) \approx \sum_{n=0}^{n_{\max}} \sum_{l=0}^{l_{\max}} \sum_{m=-l}^l c_{nlm} g_n(r) Y_{lm}(\theta, \phi) \quad (2)$$

where the c_{nlm} are the expansion coefficients.

The number of expansion coefficients one chooses to compute is bound by the number of radial and angular basis functions (n_{\max} , l_{\max}). In practice, one defines a cutoff radius (r_{cut}) for the atomic environment being considered. One can then define a rotationally invariant power spectrum (\mathbf{p}), whose elements are

$$p_{n'l'} = \pi \sqrt{\frac{8}{2l+1}} \sum_m (c_{nlm})^{\dagger} c_{n'lm} \quad (3)$$

Thus, the distance between two environments χ and χ' is related to the SOAP kernel by the following expression:

$$d(\chi, \chi') = 1 - K^{\text{SOAP}}(\mathbf{p}, \mathbf{p}') \quad (4)$$

where

$$K^{\text{SOAP}}(\mathbf{p}, \mathbf{p}') = \left(\frac{\mathbf{p} \cdot \mathbf{p}'}{\sqrt{\mathbf{p} \cdot \mathbf{p} \mathbf{p}' \cdot \mathbf{p}'}} \right) \quad (5)$$

Using the *Dscribe* package,⁴⁴ the local SOAP descriptor for a water molecule $\mathbf{p}(i)$ is formed by computing the power spectrum on only oxygen species within a cutoff radius ($r_{\text{cut}} = 3.7$ Å) centered about each oxygen atom. The local SOAP descriptors encode fluctuations on the length scales around the first coordination shell. To explore nonlocal fluctuations, we form a *glocal* SOAP descriptor by taking an average of the SOAP descriptor for each molecule and its neighbors within a distance r_{gloc} given as

$$\mathbf{p}_{r_{\text{gloc}}}(i) = \frac{1}{n} \sum_{j=1}^n \mathbf{p}(j) \quad (6)$$

Here, n is the number of neighbors within a distance r_{gloc} of the water molecule i .

We also consider $\mathbf{p}_{\text{Global}}(i)$ as the SOAP descriptor obtained by averaging the descriptors of all molecules within a snapshot. Similar types of descriptors have been previously used by Lechner and Dellago⁴⁵ as a means to accurately include

nonlocal structural information in crystalline solid-state systems.

The quality, size, and accuracy of the SOAP descriptors depend on the parameters that go into its definition. In particular, one needs to have a balance between the level of detail the descriptors encode and also the computational management of the data sets one uses. In this work, we compute the SOAP descriptors considering only oxygen species and with the following parameters: $n_{\max} = 8$, $l_{\max} = 6$, and $\sigma = 1.0 \text{ \AA}$ since it offers a good balance between the level of detail of the molecular environment encoded and the size of the descriptors. In section S1 of the [Supporting Information](#) we explain how the descriptors used in the ensuing analysis and the ones that are built to include hydrogens, as well as the use of smaller σ , encode similar information.

Besides the SOAP based local atomic descriptors, we are also interested in examining if and how well chemically based order parameters capture the relevant fluctuations in liquid water. Of the many order parameters, the ones of interest to us in this context were q_{tet} ,^{30,31} LSI,^{46–48} d_s ,⁴⁹ ρ_{voroi} ,⁵⁰ ψ ,³⁵ and ζ .⁶ More details on these chemical order parameters are provided in section S6 of the [Supporting Information](#) as well as in the main text of the referenced material.

In data sets with numerous dimensions, the presence of correlations among variables describing each data point suggests that the system of interest likely lies on a manifold whose dimension (the Intrinsic Dimensionality of the data set) is much lower than the embedding dimension of the data. To illustrate this, consider a set of points in three dimensions; if distributed randomly, the Intrinsic Dimensionality (ID) would be three. However, correlations between coordinates could restrict data points to lie only on the surface of a sphere, resulting in an ID of 2.

Computing the ID is closely tied to dimensionality reduction techniques,^{51–53} where the data set is projected into a lower-dimensional space for analysis, visualization, and interpretation. The ID denotes the minimum dimensionality in which the data can be projected by applying such techniques without a significant information loss. A proper understanding of the ID guides the selection of the space used to analyze system fluctuations. In our study, the ID is crucial for estimating a point-dependent density function, influencing the extraction of free energy, as elaborated later.

In this work, we employed the Two-NN estimator,⁴⁰ a recently developed technique estimating the ID based on information from the first and second nearest neighbors of data points. This method, successfully applied to various molecular systems,^{54–56} operates on the assumption that the density of a data point can be considered approximately uniform within the distance to the second nearest neighbor of a data point, demonstrating that the ratio of the second to the first nearest neighbor distances ($\mu = r_2/r_1$) follows a specific distribution:

$$P(\mu) = \frac{d}{\mu^{d+1}} \quad (7)$$

Here, d is the ID. Assuming independence of sampled ratios μ_i , the ID can be estimated by maximum likelihood (other estimators are also possible) as

$$d = \frac{N}{\sum_{i=1}^N \log(\mu_i)} \quad (8)$$

where N is the total number of samples in the data set.

Using SOAP distances, we estimated the ID of the water molecule environment. The ID represents the minimum number of independent order parameters needed to describe the environment, aiding in quantifying information gained or lost with different variables.⁵⁷

The considerations of the previous section have a direct impact on the reconstruction of the free energy landscape of water. To this end, understanding relevant variables characterizing structural fluctuations is essential. A common strategy is to examine probability densities along chemically inspired variables like q_{tet} , LSI, and d_s .^{7,47,58} However, this assumes no information loss in the projection (something that cannot be strictly true if the number of variables employed is smaller than the ID of the data) and that the variable correctly encodes the process of interest. Recent techniques automatically identify important degrees of freedom^{59,60} and construct free energies in high dimensions.^{41,61–64} For a detailed discussion, refer to a recent review.⁶⁵

In this work, we employed the Point Adaptive k -nearest neighbor estimator (PAK),⁴¹ avoiding the need for projection and used successfully in studying complex molecular systems.^{55,56,66} The method uses the ID as a parameter to construct a point-dependent density (ρ_i). This density is computed by adding a linear correction to the standard k -nearest neighbor estimator, where the density is $\rho_i = \frac{k_i}{r_{k_i}^d}$, and k_i 's are computed for each data point as the larger neighborhood for which the density can be considered approximately constant. The rationale is that, at constant density, the variance of the density estimation scales with $\frac{1}{\sqrt{k_i}}$

while the inclusion of regions with different densities introduces a bias term to the error; therefore, the procedure controls the bias-variance trade-off. The point-dependent free energy is $-\text{Log}(\rho_i)$. Previous work shows this method accurately estimates free energy errors up to dimensions as large as 8.⁴¹

With point-dependent free energies, independent minima in the free energy landscape (clusters) are determined using a modified density peak clustering algorithm (DPA),⁴³ an extension of the original density peak clustering.⁴² In this procedure, cluster center candidates are chosen as those whose density is maximum within their k_i neighbors. Then, the saddle points between these free energy basins are computed, and the clusters are considered as coming from statistical fluctuations (and therefore merged in one) if the free energy difference between the basin minima and the saddle point is lower than Z times the sum of the errors associated with these free energy estimates. The parameter Z is the only free parameter in DPA clustering and can be interpreted as a measure of the statistical confidence of the clustering partition. The higher its value, the more one can be sure that the clusters are not coming from statistical fluctuations but, at the same time, the higher the probability of losing real clusters whose statistical confidence is low due to the limited number of data points. In this work, the choice of Z was made by varying it in two independently generated data sets until the clusters were consistent.

Finally, PAK and DPA results are visualized and interpreted using the uniform manifold approximation and projection (UMAP),⁶⁷ providing a convenient way to visualize high-dimensional free energies in two dimensions.⁶⁸

To unravel the relationships between various order parameters and the macroscopic density, we used a statistical

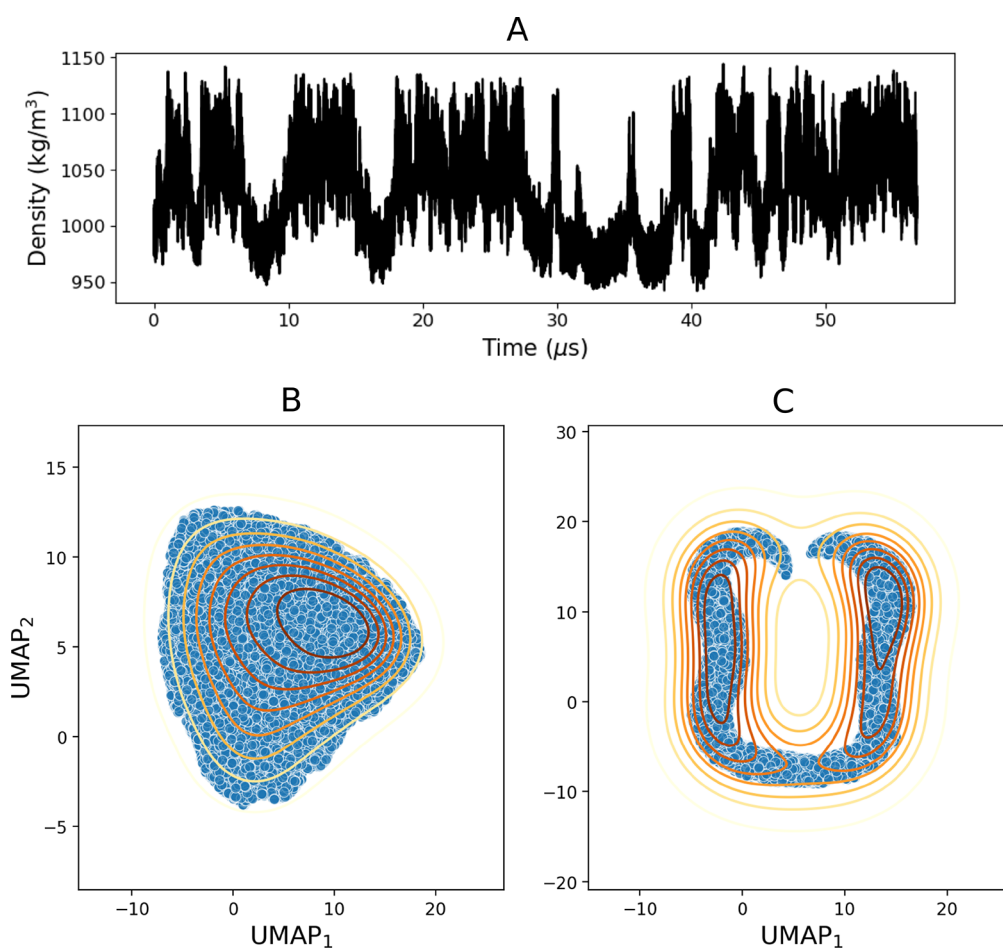


Figure 1. (A) Critical macroscopic density fluctuations close to the critical point as obtained in ref 11. (B) 2D UMAP representation of the data manifold obtained from the local SOAP descriptors; there is no clear separation between the two phases at the local level, as obtained from our clustering procedure and also seen from the unimodal nature of the free energy surface. (C) 2D UMAP representation of the data manifold obtained from the global average SOAP descriptors; the two minima correspond to the low and high macroscopic density phases obtained from our clustering procedure.

test called the Information Imbalance (IB). More details on the method is provided in ref 69. In brief, given a data set with N data points and F features, one can construct different distance measures A and B using any subset of the feature space of choice, and the IB is then defined as

$$\Delta(A \rightarrow B) = \frac{2}{N} \langle R^B | R^A = 1 \rangle = \frac{2}{N^2} \sum_{i,j: R_{ij}^A=1} R_{ij}^B \quad (9)$$

where R_{ij}^A and R_{ij}^B are the rank matrices obtained from distances A and B , respectively. Thus, $R_{ij}^A = 1$ if point j is the first neighbor of point i in space A . With this definition, if $\Delta(A \rightarrow B) \sim 0$, then space A is predictive of B and, if $\Delta(A \rightarrow B) \sim 1$, then the two spaces are unrelated.

The IB is by definition asymmetric, in the sense that, if $\Delta(A \rightarrow B) \sim 0$ and $\Delta(B \rightarrow A) \sim 1$, then it means that distance measure A can be used to predict B with more reliability than the reverse.

Having described the methods we employed in this work, we move next to discussing our results. We estimate the ID of the environment around a water molecule with $\mathbf{p}(i)$, $\mathbf{p}_{r_{\text{gloc}}}(i)$ for all $r_{\text{gloc}} \in [3.7 \text{ \AA}, 6.0 \text{ \AA}, 10.0 \text{ \AA}]$ and $\mathbf{p}_{\text{Global}}(i)$. We find an ID of 5 with the purely local SOAP descriptors ($\mathbf{p}(i)$) and this decreases to 4 as we increase the radial threshold to include molecules in the whole frame, indicating that the averaging

enhances the correlations in the descriptor. Figure S2 in the Supporting Information shows how the ID scales as a function of the number of data points sampled from the trajectory.

With the ID computed, we are now in a position to analyze the free energy landscape. Panel A of Figure 1 shows the time series of the simulation trajectory as reported in ref 11, where critical-like fluctuations between the HD and LD phases are observed. When using the global density as an order parameter, the underlying free energy landscape is clearly bimodal. Clear two-peak distributions have been observed for several geometric and energetic order parameters, when averaged over all molecules in the system.³⁵ It has also been shown that the distributions of the same descriptors, if evaluated at particle level, do not show a clear bimodal character. A notable exception is ψ , an indicator based on the topology of the hydrogen bond network surrounding each molecule.^{28,35} We address the question of the onset of bimodality on crossing from local to global descriptors by performing DPA clustering using the purely local SOAP descriptors $\mathbf{p}(i)$. Our clustering analysis reveals one cluster despite the pronounced macroscopic density fluctuations. In panel B, we show the UMAP projection of the local SOAP descriptors in two dimensions. Confirming our clustering results, we see that the local environments coming from both LD and HD snapshots lie in one free energy basin with no clear separation between local

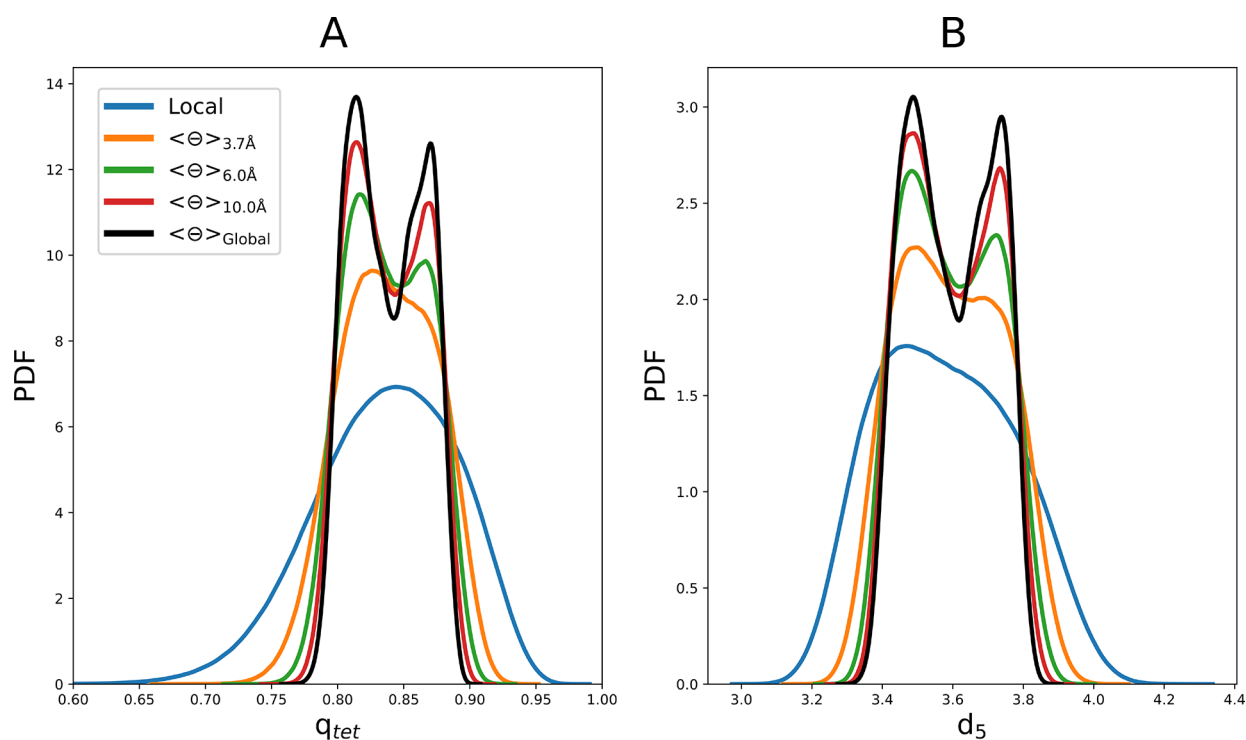


Figure 2. Evolution of the order parameters as a function of averaging. the average in the legend denotes the average within some cutoff until the global average. One can see the emergence of bimodality upon averaging beyond the second coordination shell.

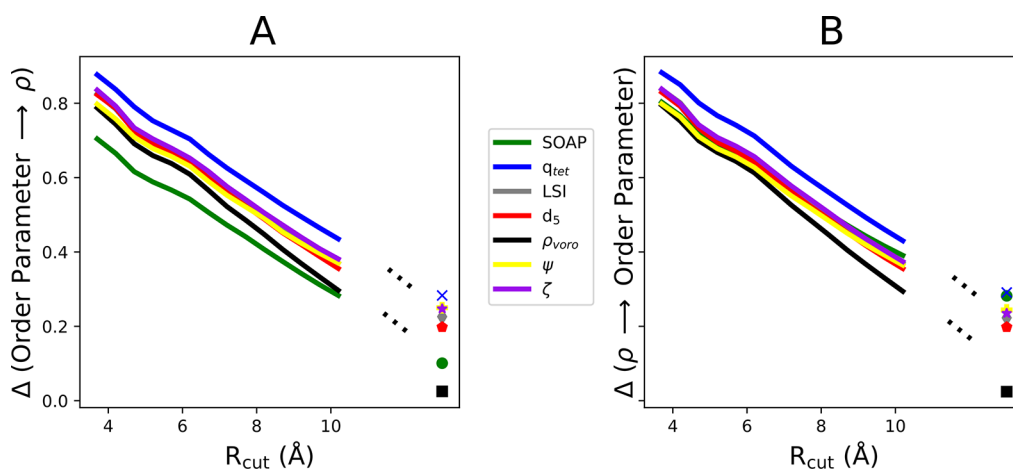


Figure 3. (A) Information imbalance between the descriptors and the macroscopic density (ρ) as a function of radial averaging. We see a consistent reduction in the IB as we increase the cutoff radius for the averaging. It is noteworthy that the different descriptors reach different IB values with the whole box average, with the ρ_{voro} being the most predictive of ρ , followed by the SOAP descriptor. On the flip side (panel B), which is the IB between ρ and the descriptors, one observes a symmetry between ρ and the other descriptors except SOAP, indicating that SOAP contains some information about the global average structure which ρ misses.

LD and HD environments akin to what is observed in water at ambient conditions.³⁶

The microscopic origins of what we observe are likely rooted in the large heterogeneity of the local environments in both phases. Our local descriptor, however, only directly encodes information on the water hydrogen bond network on the length scale of ~ 3.7 Å. Indeed, several previous studies have pointed to the important role of structural information beyond the second coordination shell that may be essential in understanding the differences between an HD and LD phase.^{35,70,71} With this in mind, we performed clustering with $\mathbf{p}_{r_{\text{gloc}}}(i)$. As shown in Figure S1 of the Supporting

Information, by increasing r_{gloc} the topology of the UMAP manifold starts to change, and we see the emergence of two clusters (confirmed by the DPA clustering) when we average beyond the second solvation shell. Panel C of Figure 1 shows the UMAP projection of the $\mathbf{p}_{\text{Global}}(i)$ SOAP descriptor. The clustering analysis using the global descriptors reveals two clusters that are consistent with the macroscopic HD and LD phases, further indicating that there is structural information beyond the second solvation shell that is important in distinguishing the LD and HD phases.

The density plots shown in the bottom panels of Figure 1 involve a projection of the high-dimensional SOAP features at both the local and global scale onto two UMAP coordinates

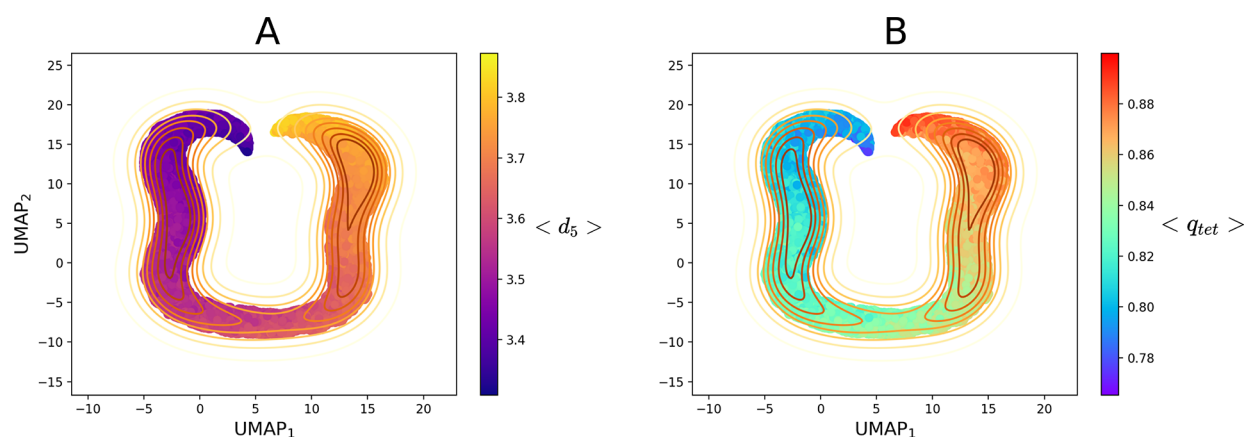


Figure 4. (A) 2D UMAP representation of the SOAP data manifold colored with the average d_5 . The peak on the left half of this panel corresponds to low average d_5 while the peak on the right half corresponds to relatively high average d_5 . (B) 2D UMAP representation of the SOAP data manifold colored with the average q_{tet} . The peak on the left half of this panel corresponds to low average q_{tet} while the peak on the right half corresponds to relatively high average q_{tet} .

which are rather difficult to interpret physically. We thus turn to examining how chemical descriptors such as tetrahedrality (q_{tet}) and the distance to the fifth water molecule (d_5) evolve as a function of the nonlocal averaging. Figure 2 shows the distributions of the q_{tet} (panel A) and d_5 (panel B) order parameters computed from the critical point trajectory in Figure 1. We observe that the distribution for the local order parameters is essentially unimodal with no characteristic peaks. However, upon averaging the descriptors within radial cut-offs ($\langle \Theta \rangle_{r_{glob}}$) we observe the emergence of a bimodal structure in the distributions. Specifically, beyond 6 Å, one peak grows at relatively low q_{tet} values (~ 0.8) and hence low d_5 values (~ 3.4 Å), corresponding to environments sampled from the HD phase. The other peak is located at high q_{tet} values (~ 0.9) and hence higher d_5 values (~ 3.8 Å) corresponding to average environments sampled from the LD phase. For q_{tet} , there is a larger proportion of environments in the LD phase consistent with what is observed with the macroscopic density¹¹ whereas this feature is much less pronounced with d_5 .

The emergence of bimodality in the distribution of the order parameters on longer length scales signifies that there is some correlation between them and macroscopic density. However, the manner in which this bimodal structure develops and how strongly this reflects the density (LD vs HD) are rather sensitive to the choice of the chemical order parameter that are used. In a recent work, some of us have shown that for liquid water at room temperature, a full description of the fluctuations in the water hydrogen bond network involves the coupling of several different order parameters together.⁷² In the context of this work, we wanted to explore which order parameters and on what length-scales best probe the HD-LD density fluctuations.

To address this question, we applied the IB method to investigate the coupling between several chemically based order parameters and the density. The IB provides a quantitative measure of how well variables such as q_{tet} or d_5 averaged over different length scales can predict the global density and vice versa. Figure 3 illustrates the behavior of the IB as a function of radial averaging. We observe in panel A that the information about the macroscopic density contained in all the descriptors starts increasing (corresponding to low IB values) as we increase the radial cutoff for the averaging. Up to a cutoff distance of ~ 1 nm, which is approximately half of the

whole box size, the IB reduces significantly, reaching values smaller than ~ 0.3 for the SOAP descriptor and ρ_{voro} .

In the bottom right corners of each panel in Figure 3, one can see the IB value obtained for the average descriptors ($\langle \Theta \rangle_{Global}$) which reaches values of ~ 0.1 and ~ 0.03 for the SOAP descriptor and ρ_{voro} , respectively. We note that this tight coupling between the descriptors and the macroscopic density is strictly a feature observed in subcritical supercooled water (as we show in Figures S3 and S4 in the Supporting Information) and it is only achieved upon including structural information on up to ~ 1 nm length scale, as has also been discussed in ref 35.

The asymmetric nature of the IB allows us to also compare the information contained in the macroscopic density with the different descriptors. As seen in panel B, there is symmetric information shared between the macroscopic density and all other descriptors, except for the SOAP descriptor. This is not surprising since the SOAP descriptor by nature is complete and contains information about the molecular orientations, which the macroscopic density does not contain.

All in all, the preceding results build strong evidence to a picture where the LD-HD density fluctuations cannot be described in terms of local competing structures but instead involve clusters of at least 100 water molecules. Thus, according to our unsupervised learning protocol, the density fluctuations underlying LD-HD transitions cannot be associated with properties assigned at the single molecule level. With this picture in mind, we can revisit the UMAP projections, providing more chemical interpretability. In panels A and B of Figure 4, we show the UMAP projection of the SOAP data in 2 dimensions, now colored with the corresponding average d_5 ($\langle d_5 \rangle_{Global}$) and q_{tet} ($\langle q_{tet} \rangle_{Global}$), respectively. We confirm from this that the two density peaks (or free energy minima) emerging from our clustering correspond to the HD and LD phases since one of the peaks overlaps with $\langle d_5 \rangle_{Global} \sim 3.4$ Å hence $\langle q_{tet} \rangle_{Global} \sim 0.8$ (high density), and the other peak overlaps with $\langle d_5 \rangle_{Global} \sim 3.8$ Å and thus $\langle q_{tet} \rangle_{Global} \sim 0.9$ (low density).

One of the important signatures of critical behavior is the divergence in the structure factor in the low $|k|$ limit, which ultimately translates into an enhancement in long-range density fluctuations. To investigate this anomalous scattering behavior of water close to the critical point, Debenedetti and

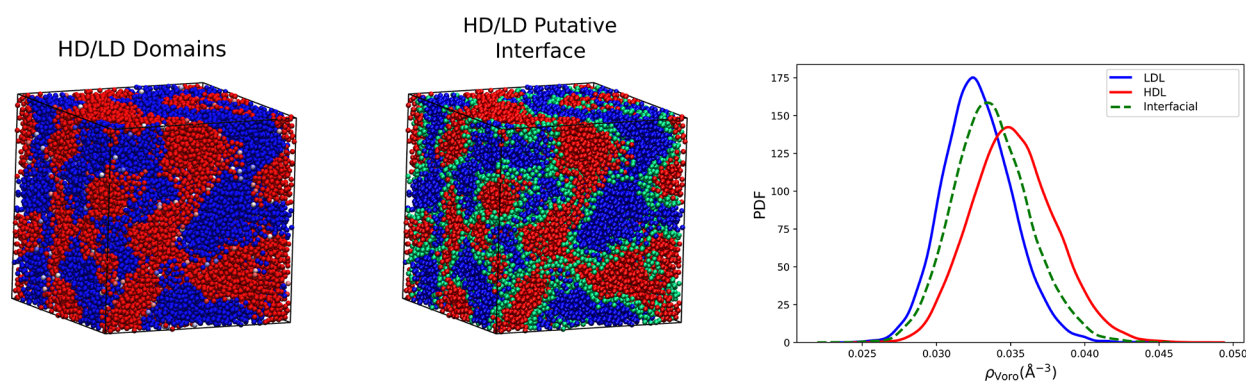


Figure 5. (Left) Snapshot of the large system colored by which density phase it was assigned to blue for LD and red for HD. We observe the LD and HD domains extend over 1 nm spatial distance. (Middle) Same snapshot now coloring molecules that are found in the boundary between LD and HD domains in green. (Right) PDF of the Voronoi density for all LD assigned molecules (blue full line), PDF of the Voronoi density for all HD assigned water molecules (red full line) and the PDF of Voronoi density values for all molecules assigned as interfacial molecules (green dashed line). We note how the distributions are peaked toward low density, high density, and intermediate density, respectively, albeit with a huge overlap.

co-workers explored the properties of the static structure factor using a large system (36424 TIP4P/2005 water molecules at subcritical conditions run in the *NVT* ensemble¹¹). Their analysis indeed shows the signature of the critical behavior. A question that remains, however, is how exactly one rationalizes the relationship between the nonlocal structures that emerge from our preceding analysis and the long-range density fluctuations.

The two clusters that have been automatically identified from the SOAP descriptors averaged on the nanometer length scale (see Figure 1 earlier) provide a protocol for classifying water molecule environments in other contexts such as those used to construct the structure factor previously described. Using the *k*-nearest neighbor classifier (see details in section S7 of the Supporting Information), we assign water molecules in the large box simulation to either LD or HD type depending on their respective similarities.

Applying this procedure leads to automatic identification of LD and HD domains. In the left-most panel of Figure 5, we show one snapshot with only oxygen atoms (for clarity) colored by the phase they have been assigned to: blue spheres represent molecules assigned as LD-like, while red spheres are molecules assigned as HD-like. By visual inspection, one can see a tendency for the LD and HD water molecules to cluster together, forming LD-like and HD-like domains. These domains extend over spatial distances of several nanometers and essentially percolate throughout the periodic box, and to the best of our knowledge, this is the first instance where the LD and HD domains have been identified in a completely unsupervised manner.

If the density fluctuations close to the critical point are indeed creating LD and HD domains, then this implies that there should be some signature of an interfacial region forming at the boundary of the domains. One signature of this would be that water molecules close to the boundary would not be classified as pure LD or HD environments. A manner in which this can be quantified is to measure the probability of identifying either an LD or HD environment and subsequently identifying pure LD environments as those with $p_{\text{LD}} > 0.7$ and the pure HD environments as those with $p_{\text{HD}} > 0.7$. Water molecules with $0.7 > p_{\text{LD}} > 0.4$ or $0.7 > p_{\text{HD}} > 0.4$ are then identified as those that are putatively assigned as *boundary* or *interfacial* points.

In the middle panel of Figure 5, we show the same simulation snapshot but now also coloring points that have been identified to be so-called *boundary* water molecules. From visual inspection, we can see how the green molecules are typically located between LD and HD domains. These findings nicely demonstrate that our procedure of agnostically identifying environments with appropriately averaged SOAP descriptors on the nanometer length scale leads to the emergence of LD and HD domains, which are identified at the same thermodynamic state point.

In the right-most panel of Figure 5, we plot the PDF of the ρ_{Voro} order parameter constrained to the LD (blue full line), HD (red full line), and interfacial molecules (green dashed line). We note that the full distribution of ρ_{Voro} is unimodal and broad. However, by restricting the distribution to the identified domains separately, we find that the peaks in the distributions are consistent with those associated with the LD and HD phases in the smaller box. It is also curious to observe interfacial molecules, which have ρ_{Voro} values peaked between the peaks of the LDL/HDL ρ_{Voro} distributions.

In conclusion, in this work, we have used unsupervised machine learning techniques to analyze data coming from molecular dynamics simulations of liquid water close to the second critical point, where one observes pronounced fluctuations of the global density between a high-density (HD) and low-density (LD) liquid phase. We show that the free energy landscape in the space of local descriptors consists of one minimum despite pronounced density fluctuations. This is rooted in the large heterogeneity of the local configurations sampled by the water molecules in both phases. However, by using descriptors that account for *nonlocal* information of the water network, bimodality in the free energy landscape emerges.

We further confirm the importance of *nonlocal* information by deploying a statistical test that allows us to evaluate the strength of the mapping between different descriptors and the macroscopic density fluctuations. We find that the mapping is strongest when the descriptors are constructed to include information on approximately a nanometer length scale. Finally, armed with the *nonlocal* HD and LD structures that emerge from our analysis, we characterize the formation of HD and LD domains that is manifested in the anomalous scattering behavior of water close to the second critical point.

Our results bring forward important challenges in assigning and interpreting fluctuations of the hydrogen bond network in terms of single particle properties where longer range structural correlations are clearly more important. These findings should motivate more work³² in trying to understand the relationships between local atomic descriptors and local molecular chemically inspired parameters and how they change our understanding of fluctuations across the phase diagram of water. We believe our work provides a general framework for understanding water's structural and dynamic properties in other scenarios where long-range correlations may be important, such as at interfaces^{73–75} as well as under confinement.^{76,77}

■ ASSOCIATED CONTENT

SI Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jpcllett.4c00383>.

Information imbalance between SOAP descriptors computed from different hyper-parameters, clustering labels; from local to global descriptors, intrinsic dimension scaling; from local to global descriptors, descriptor-density coupling; comparison between subcritical supercooled water and water at ambient conditions, SOAP descriptor-order parameter coupling; comparison between subcritical supercooled water and water at ambient conditions, description of chemical order parameters used in the main text, test scores for the KNN model used in Figure 5 of the main text (PDF)

Transparent Peer Review report available (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Alex Rodriguez – *The Abdus Salam International Center for Theoretical Physics (ICTP), 34151 Trieste, Italy; Dipartimento di Matematica, Informatica e Geoscienze, Università degli studi di Trieste, 34127 Trieste, Italy;* orcid.org/0000-0002-0213-6695; Email: alejandro.rodriguezgarcia@units.it

Ali Hassanali – *The Abdus Salam International Center for Theoretical Physics (ICTP), 34151 Trieste, Italy;* orcid.org/0000-0002-3208-1488; Email: ahassana@ictp.it

Authors

Edward Danquah Donkor – *The Abdus Salam International Center for Theoretical Physics (ICTP), 34151 Trieste, Italy; Scuola Internazionale Superiore di Studi Avanzati (SISSA), 34136 Trieste, Italy;* orcid.org/0000-0001-7565-7603

Adu Offei-Danso – *The Abdus Salam International Center for Theoretical Physics (ICTP), 34151 Trieste, Italy; Scuola Internazionale Superiore di Studi Avanzati (SISSA), 34136 Trieste, Italy;* orcid.org/0000-0002-8035-0769

Francesco Sciortino – *Dipartimento di Fisica, Sapienza Università di Roma, 00185 Rome, Italy;* orcid.org/0000-0002-2418-2713

Complete contact information is available at: <https://pubs.acs.org/doi/10.1021/acs.jpcllett.4c00383>

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

A.H. thanks the European Commission for funding on the ERC Grant HyBOP 101043272. E.D.D. thanks Alex Chen Yi Zhang for very fruitful discussions during this work.

■ REFERENCES

- (1) Corró, M.; El Hichou, A.; Cesari, E.; Kustov, S. Study of magnetic transitions in Dy by means of reversible Villari effect. *J. Phys. D: Appl. Phys.* **2016**, *49*, No. 015001.
- (2) Shamba, P.; Wang, J. L.; Debnath, J. C.; Kennedy, S. J.; Zeng, R.; Md Din, M. F.; Hong, F.; Cheng, Z. X.; Studer, A. J.; Dou, S. X. The magnetocaloric effect and critical behaviour of the Mn_{0.94}Ti_{0.06}CoGe alloy. *J. Phys.: Condens. Matter* **2013**, *25*, No. 056001.
- (3) Ekimov, E.; Sadykov, R.; Mel'nik, N.; Presz, A.; Tat'yanin, E.; Slesarev, V.; Kuzin, N. Diamond crystallization in the system B 4 C–C. *Inorganic materials* **2004**, *40*, 932–936.
- (4) Noguera, C.; Mackrodt, W. A theoretical study of magnetic phase transitions in ultra-thin films: Application to NiO. *Surface science* **2006**, *600*, 861–872.
- (5) Shi, R.; Russo, J.; Tanaka, H. Common microscopic structural origin for water's thermodynamic and dynamic anomalies. *J. Chem. Phys.* **2018**, *149*, No. 224502.
- (6) Russo, J.; Tanaka, H. Understanding water's anomalies with locally favoured structures. *Nat. Commun.* **2014**, *5*, 3556.
- (7) Cuthbertson, M. J.; Poole, P. H. Mixturelike behavior near a liquid-liquid phase transition in simulations of supercooled water. *Physical review letters* **2011**, *106*, No. 115706.
- (8) Pipolo, S.; Salanne, M.; Ferlat, G.; Klotz, S.; Saitta, A. M.; Pietrucci, F. Navigating at will on the water phase diagram. *Physical review letters* **2017**, *119*, No. 245701.
- (9) Poole, P. H.; Sciortino, F.; Essmann, U.; Stanley, H. E. Phase behaviour of metastable water. *Nature* **1992**, *360*, 324–328.
- (10) Palmer, J. C.; Martelli, F.; Liu, Y.; Car, R.; Panagiotopoulos, A. Z.; Debenedetti, P. G. Metastable liquid–liquid transition in a molecular model of water. *Nature* **2014**, *510*, 385–388.
- (11) Debenedetti, P. G.; Sciortino, F.; Zerze, G. H. Second critical point in two realistic models of water. *Science* **2020**, *369*, 289–292.
- (12) Hou, Y.; Yu, M.; Shang, Y.; Zhou, P.; Song, R.; Xu, X.; Chen, X.; Wang, Z.; Yao, S. Suppressing ice nucleation of supercooled condensate with biphilic topography. *Physical review letters* **2018**, *120*, No. 075902.
- (13) Murray, B.; O'sullivan, D.; Atkinson, J.; Webb, M. Ice nucleation by particles immersed in supercooled cloud droplets. *Chem. Soc. Rev.* **2012**, *41*, 6519–6554.
- (14) Mishima, O. Liquid-liquid critical point in heavy water. *Physical review letters* **2000**, *85*, 334.
- (15) Stillinger, F. H.; Rahman, A. Improved simulation of liquid water by molecular dynamics. *J. Chem. Phys.* **1974**, *60*, 1545–1557.
- (16) Liu, Y.; Panagiotopoulos, A. Z.; Debenedetti, P. G. Low-temperature fluid-phase behavior of ST2 water. *J. Chem. Phys.* **2009**, *131*, 104508.
- (17) Sciortino, F.; Saika-Voivod, I.; Poole, P. H. Study of the ST2 model of water close to the liquid–liquid critical point. *Phys. Chem. Chem. Phys.* **2011**, *13*, 19759–19764.
- (18) Kesselring, T. A.; Franzese, G.; Buldyrev, S. V.; Herrmann, H. J.; Stanley, H. E. Nanoscale dynamics of phase flipping in water near its hypothesized liquid-liquid critical point. *Sci. Rep.* **2012**, *2*, 474.
- (19) Poole, P. H.; Bowles, R. K.; Saika-Voivod, I.; Sciortino, F. Free energy surface of ST2 water near the liquid-liquid phase transition. *J. Chem. Phys.* **2013**, *138*, 034505.
- (20) Gartner III, T. E.; Piaggi, P. M.; Car, R.; Panagiotopoulos, A. Z.; Debenedetti, P. G. Liquid-liquid transition in water from first principles. *Physical review letters* **2022**, *129*, No. 255702.
- (21) Amann-Winkel, K.; Gainaru, C.; Handle, P. H.; Seidl, M.; Nelson, H.; Böhmmer, R.; Loerting, T. Water's second glass transition. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 17720–17725.
- (22) Kim, K. H.; Amann-Winkel, K.; Giovambattista, N.; Späh, A.; Perakis, F.; Pathak, H.; Parada, M. L.; Yang, C.; Mariedahl, D.;

- Eklund, T.; et al. Experimental observation of the liquid-liquid transition in bulk supercooled water under pressure. *Science* **2020**, *370*, 978–982.
- (23) Amann-Winkel, K.; Kim, K. H.; Giovambattista, N.; Ladd-Parada, M.; Späh, A.; Perakis, F.; Pathak, H.; Yang, C.; Eklund, T.; Lane, T. J.; et al. Liquid-liquid phase separation in supercooled water from ultrafast heating of low-density amorphous ice. *Nat. Commun.* **2023**, *14*, 442.
- (24) Hamm, P. Markov state model of the two-state behaviour of water. *J. Chem. Phys.* **2016**, *145*, No. 134501.
- (25) Tanaka, H.; Tong, H.; Shi, R.; Russo, J. Revealing key structural features hidden in liquids and glasses. *Nature Reviews Physics* **2019**, *1*, 333–348.
- (26) Tanaka, H. Liquid–liquid transition and polyamorphism. *J. Chem. Phys.* **2020**, *153*, 130901.
- (27) Skarmoutsos, I.; Franzese, G.; Guardia, E. Using Car-Parrinello simulations and microscopic order descriptors to reveal two locally favored structures with distinct molecular dipole moments and dynamics in ambient liquid water. *J. Mol. Liq.* **2022**, *364*, No. 119936.
- (28) Foffi, R.; Sciortino, F. Identification of local structures in water from supercooled to ambient conditions. *J. Chem. Phys.* **2024**, *160*, 094504.
- (29) Errington, J. R.; Debenedetti, P. G. Relationship between structural order and the anomalies of liquid water. *Nature* **2001**, *409*, 318–321.
- (30) Chau, P.-L.; Hardwick, A. A new order parameter for tetrahedral configurations. *Mol. Phys.* **1998**, *93*, 511–518.
- (31) Lynden-Bell, R.; Debenedetti, P. G. Computational investigation of order, structure, and dynamics in modified water models. *J. Phys. Chem. B* **2005**, *109*, 6527–6534.
- (32) Martelli, F.; Leoni, F.; Sciortino, F.; Russo, J. Connection between liquid and non-crystalline solid phases in water. *J. Chem. Phys.* **2020**, *153*, 104503.
- (33) Faccio, C.; Benzi, M.; Zanetti-Polzi, L.; Daidone, I. Low- and high-density forms of liquid water revealed by a new medium-range order descriptor. *J. Mol. Liq.* **2022**, *355*, No. 118922.
- (34) Montes de Oca, J. M.; Sciortino, F.; Appignanesi, G. A structural indicator for water built upon potential energy considerations. *J. Chem. Phys.* **2020**, *152*, No. 244503.
- (35) Foffi, R.; Sciortino, F. Correlated fluctuations of structural indicators close to the liquid–liquid transition in supercooled water. *J. Phys. Chem. B* **2023**, *127*, 378–386.
- (36) Offei-Danso, A.; Hassanali, A.; Rodriguez, A. High-dimensional fluctuations in liquid water: Combining chemical intuition with unsupervised learning. *J. Chem. Theory Comput.* **2022**, *18*, 3136–3150.
- (37) Di Pino, S.; Donkor, E. D.; Sánchez, V. M.; Rodriguez, A.; Cassone, G.; Scherlis, D.; Hassanali, A. ZundEig: The Structure of the Proton in Liquid Water from Unsupervised Learning. *J. Phys. Chem. B* **2023**, *127*, 9822–9832.
- (38) Bartók, A. P.; Gillan, M. J.; Manby, F. R.; Csányi, G. Machine-learning approach for one-and two-body corrections to density functional theory: Applications to molecular and condensed water. *Phys. Rev. B* **2013**, *88*, No. 054104.
- (39) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769.
- (40) Facco, E.; d’Errico, M.; Rodriguez, A.; Laio, A. Estimating the intrinsic dimension of datasets by a minimal neighborhood information. *Sci. Rep.* **2017**, *7*, No. 12140.
- (41) Rodriguez, A.; d’Errico, M.; Facco, E.; Laio, A. Computing the free energy without collective variables. *J. Chem. Theory Comput.* **2018**, *14*, 1206–1215.
- (42) Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. *Science* **2014**, *344*, 1492–1496.
- (43) d’Errico, M.; Facco, E.; Laio, A.; Rodriguez, A. Automatic topography of high-dimensional data sets by non-parametric density peak clustering. *Information Sciences* **2021**, *560*, 476–492.
- (44) Himanen, L.; Jager, M. O.J.; Morooka, E. V.; Federici Canova, F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; Foster, A. S. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **2020**, *247*, No. 106949.
- (45) Lechner, W.; Dellago, C. Accurate determination of crystal structures based on averaged local bond order parameters. *J. Chem. Phys.* **2008**, *129*, 114707.
- (46) Shiratani, E.; Sasai, M. Growth and collapse of structural patterns in the hydrogen bond network in liquid water. *J. Chem. Phys.* **1996**, *104*, 7671–7680.
- (47) Appignanesi, G. A.; Rodriguez Fris, J. A.; Sciortino, F. Evidence of a two-state picture for supercooled water and its connections with glassy dynamics. *Eur. Phys. J. E* **2009**, *29*, 305–310.
- (48) Malaspina, D.; Schulz, E.; Alarcón, L.; Frechero, M.; Appignanesi, G. Structural and dynamical aspects of water in contact with a hydrophobic surface. *Eur. Phys. J. E* **2010**, *32*, 35–42.
- (49) Saika-Voivod, I.; Sciortino, F.; Poole, P. H. Computer simulations of liquid silica: Equation of state and liquid–liquid phase transition. *Phys. Rev. E* **2000**, *63*, No. 011202.
- (50) Rycroft, C. H.; Grest, G. S.; Landry, J. W.; Bazant, M. Z. Analysis of granular flow in a pebble-bed nuclear reactor. *Phys. Rev. E* **2006**, *74*, No. 021306.
- (51) Mika, S.; Schölkopf, B.; Smola, A.; Müller, K.-R.; Scholz, M.; Rätsch, G. Kernel PCA and de-noising in feature spaces. *Advances in neural information processing systems* **1998**, *11*, 147.
- (52) Jolliffe, I. T. Principal component analysis: a beginner’s guide—I. Introduction and application. *Weather* **1990**, *45*, 375–382.
- (53) Kruskal, J. B.; Wish, M. *Multidimensional scaling*; Sage, 1978; Vol. 11.
- (54) Ansari, N.; Laio, A.; Hassanali, A. Spontaneously forming dendritic voids in liquid water can host small polymers. *Journal of physical chemistry letters* **2019**, *10*, 5585–5591.
- (55) Carli, M.; Sormani, G.; Rodriguez, A.; Laio, A. Candidate binding sites for allosteric inhibition of the SARS-CoV-2 main protease from the analysis of large-scale molecular dynamics simulations. *Journal of physical chemistry letters* **2021**, *12*, 65–72.
- (56) Jong, K.; Hassanali, A. A data science approach to understanding water networks around biomolecules: the case of tri-alanine in liquid water. *J. Phys. Chem. B* **2018**, *122*, 7895–7906.
- (57) Camastra, F.; Staiano, A. Intrinsic dimension estimation: Advances and open problems. *Information Sciences* **2016**, *328*, 26–41.
- (58) Wikfeldt, K.; Nilsson, A.; Pettersson, L. G. Spatially inhomogeneous bimodal inherent structure of simulated liquid water. *Phys. Chem. Chem. Phys.* **2011**, *13*, 19918–19924.
- (59) Tenenbaum, J. B.; Silva, V. d.; Langford, J. C. A global geometric framework for nonlinear dimensionality reduction. *science* **2000**, *290*, 2319–2323.
- (60) Roweis, S. T.; Saul, L. K. Nonlinear dimensionality reduction by locally linear embedding. *science* **2000**, *290*, 2323–2326.
- (61) Carreira-Perpinan, M. A. Mode-finding for mixtures of Gaussian distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2000**, *22*, 1318–1323.
- (62) Gasparotto, P.; Ceriotti, M. Recognizing molecular patterns by machine learning: An agnostic structural definition of the hydrogen bond. *J. Chem. Phys.* **2014**, *141*, No. 174110.
- (63) Gasparotto, P.; Meißner, R. H.; Ceriotti, M. Recognizing local and global structural motifs at the atomic scale. *J. Chem. Theory Comput.* **2018**, *14*, 486–498.
- (64) Geissler, P. L.; Dellago, C.; Chandler, D. Kinetic pathways of ion pair dissociation in water. *J. Phys. Chem. B* **1999**, *103*, 3706–3710.
- (65) Glielmo, A.; Husic, B. E.; Rodriguez, A.; Clementi, C.; Noé, F.; Laio, A. Unsupervised learning methods for molecular simulation data. *Chem. Rev.* **2021**, *121*, 9722–9758.
- (66) Sormani, G.; Rodriguez, A.; Laio, A. Explicit characterization of the free-energy landscape of a protein in the space of all its *ca* carbons. *J. Chem. Theory Comput.* **2020**, *16*, 80–87.
- (67) McInnes, L.; Healy, J.; Saul, N.; Großberger, L. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* **2018**, *3*, 861.

(68) Becht, E.; McInnes, L.; Healy, J.; Dutertre, C.-A.; Kwok, I. W.; Ng, L. G.; Ginhoux, F.; Newell, E. W. Dimensionality reduction for visualizing single-cell data using UMAP. *Nature biotechnology* **2019**, *37*, 38–44.

(69) Glielmo, A.; Zeni, C.; Cheng, B.; Csányi, G.; Laio, A. Ranking the information content of distance measures. *PNAS Nexus* **2022**, *1*, No. pgac039.

(70) Soper, A. Is water one liquid or two? *J. Chem. Phys.* **2019**, *150*, 234503.

(71) Daidone, I.; Foffi, R.; Amadei, A.; Zanetti-Polzi, L. A statistical mechanical model of supercooled water based on minimal clusters of correlated molecules. *J. Chem. Phys.* **2023**, *159*, No. 094502.

(72) Donkor, E. D.; Laio, A.; Hassanali, A. Do machine-learning atomic descriptors and order parameters tell the same story? The case of liquid water. *J. Chem. Theory Comput.* **2023**, *19*, 4596–4605.

(73) Björneholm, O.; Hansen, M. H.; Hodgson, A.; Liu, L.-M.; Limmer, D. T.; Michaelides, A.; Pedevilla, P.; Rossmeisl, J.; Shen, H.; Tocci, G.; et al. Water at interfaces. *Chem. Rev.* **2016**, *116*, 7698–7726.

(74) Das, S.; Imoto, S.; Sun, S.; Nagata, Y.; Backus, E. H.; Bonn, M. Nature of excess hydrated proton at the water–air interface. *J. Am. Chem. Soc.* **2020**, *142*, 945–952.

(75) Creazzo, F.; Pezzotti, S.; Bougueroua, S.; Serva, A.; Spöner, J.; Saija, F.; Cassone, G.; Gaigeot, M.-P. Enhanced conductivity of water at the electrified air–water interface: a DFT-MD characterization. *Phys. Chem. Chem. Phys.* **2020**, *22*, 10438–10446.

(76) Di Pino, S.; Perez Sirkin, Y. A.; Morzan, U. N.; Sánchez, V. M.; Hassanali, A.; Scherlis, D. A. Water Self-Dissociation is Insensitive to Nanoscale Environments. *Angew. Chem.* **2023**, *135*, No. e202306526.

(77) Tielrooij, K. J.; Cox, M. J.; Bakker, H. J. Effect of Confinement on Proton-Transfer Reactions in Water Nanopools. *Chem. Phys. Chem.* **2009**, *10*, 245–251.