IMA Journal of Complex Networks (2023) Page 1 of 25
doi:10.1093/comnet/xxx000

Structural analysis of water networks

MICHELE BENZI
Scuola Normale Superiore, Piazza dei Cavalieri, 7, 56126, Pisa, Italy

ISABELLA DAIDONE
Department of Physical and Chemical Sciences, University of L'Aquila, via Vetoio (Coppito 1), L'Aquila, 67010, Italy

CHIARA FACCIO*
Scuola Normale Superiore, Piazza dei Cavalieri, 7, 56126, Pisa, Italy
*Corresponding author: chiara.faccio@sns.it

AND

LAURA ZANETTI-POLZI
Center S3, CNR-Institute of Nanoscience, Via Campi 213/A, Modena, 41125, Italy

[Received on 14 January 2023]

Liquid water, besides being fundamental for life on Earth, has long fascinated scientists due to several anomalies. Different hypotheses have been put forward to explain these peculiarities. The most accredited one foresees the presence in the supercooled region of two phases at different densities: the low-density liquid phase and the high-density liquid phase. In our previous work [Faccio et al., J. Mol. Liq. 355 (2022): 118922], we showed that it is possible to identify these two forms in water networks through a computational approach based on molecular dynamics simulation and on the calculation of the total communicability of the associated graph, in which the nodes correspond to water molecules and the edges represent the connections (interactions) between molecules. In this paper, we present a more in-depth investigation of the application of graph-theory based approaches to the analysis of the structure of water networks. In particular, we investigate different connectivity and centrality measures and we report on the use of a variety of global metrics aimed at giving a topological and geometrical characterization of liquid water.

Keywords: liquid water, LDL/HDL phases, graph structure, centrality measures, global metrics.

2010 AMS Subject Classification: 05C50, 05C90.

1. Introduction

The mathematical theory of graphs has long found important applications in chemistry. In its most basic form it has been used, for example, to construct mathematical models of molecules and their interactions. A host of graph-theoretic concepts and results can be applied to analyze chemical compounds and understand their properties, and thousands of papers have been published in this field; see, e.g., [7], or [66] for a more recent survey.

Only recently, concepts and techniques from the field of network science have been introduced to chemical physics with the goal of gaining a deeper understanding of collective phenomena involving molecular networks. This is in contrast with other fields, like biology, where network science tools have been in heavy use for a number of years, for example in the study of protein and gene networks,

see [8, 38]. In the present paper we apply concepts from network science, together with molecular dynamics (MD) tools, to advance our understanding of the structure of liquid water through the analysis of networks of water molecules under different conditions. The paper can be seen as a follow-up to our previous work [28], but it is self-contained and can be read independently of it.

In this paper we are interested, primarily, in the problem of characterizing a liquid phase in a molecular simulation. In particular, we focus here on the differentiation between two forms of liquid water differing in density: the low-density liquid (LDL) and high-density liquid (HDL). The existence of a first-order liquidliquid phase transition (LLPT) between these two forms in the deeply supercooled region at high pressure has been hypothesized on the basis of computational studies [19, 36, 37, 52, 55, 59]. The LDL and HDL phases feature different structural and topological properties (see section 2) and several studies in the literature proposed a variety of chemical order parameters to investigate these two forms, [17, 20, 31, 32, 47–49, 49, 50, 56–59, 61]. We also mention works of Bakó et al, [5, 6], in which the authors use a spectral clustering method, based on the analysis of the eigenvalues of the graph Laplacian, to study the global properties of an H-bonded network.

In [28], we presented a new order parameter based on a centrality measure, the total communicability (TC), see [10], which was found to be effective in differentiating the molecules in the LDL phase from the molecules in the HDL phase. We used for our analysis a water model that was previously shown to exhibit a metastable liquidliquid critical point in deeply supercooled conditions and we chose four temperatures along an isobar that, in this water model, crosses the coexistence line, i.e., along which a first order phase transition between LDL and HDL occurs. We compared the distributions of the total communicability with the distributions of the eigenvector centrality and the degree, and we found that the last two are not well-suited to the task of characterizing molecules in the two phases. Regarding the eigenvector centrality, this can be explained by observing that the adjacency matrices associated with water networks exhibit a small spectral gap, hence using only the dominant eigenvector to analyze the network leads to a loss of structural information. The degree, on the other hand, being a local measure of connectivity, is unable to capture indirect interactions between non-neighboring water molecules and thus to account for medium to long range effects. We also computed the fraction of LDL population at four different reference temperatures of physical interest and found a sudden and marked decrease in the LDL fraction upon raising the temperature, consistently with a transition between the two pure phases, as expected by crossing the coexistence line. We compared the above populations with those obtained from well-known parameters in the physical chemistry literature. The data allowed us to conclude that TC differentiates well the two phases of liquid water.

In the present work we investigate the performance of additional centrality measures when used to identify if water is in the LDL or HDL phase and to characterize the internal organization of the two liquid forms. We consider the closeness, betweenness, Katz, and subgraph centrality, and we compare them with the results obtained with the total communicability. Furthermore, we also analyze the structural and connectivity properties of the two phases using other well-established metrics, such as clustering coefficients, a bipartivity measure, the algebraic connectivity, the graph energy, and the relative number of closed cycles of certain lengths.

The paper is organized as follows. In section 2, we briefly describe the LDL and HDL phase; in section 3, we recall some basic concepts from graph theory and specify the centrality measures and global metrics used in this work. In section 4 we describe the used methods of molecular dynamics simulations, in section 5, we explain how we construct the graph given a box with water molecules, while in section 6 we present the results of our numerical experiments. Conclusions and ideas for future works are given in section 7.

2. Low- and high-density liquid phases

Water is a complex liquid with anomalous properties: for example, to cite a few, liquid water is denser than solid water (ice), and this implies that the ice floats instead of sinking; it has very high specific heat, meaning that the water temperature rises more slowly than the temperature of almost any other substance; liquid water at ambient pressure shows a maximum in density at around 4 C. One of the most popular hypotheses for explaining many water anomalies is based on the existence of a transition between two liquid phases, referred to as low-density liquid (LDL) and high-density liquid (HDL), [19, 30, 36, 37, 52, 55, 59, 68]. The existence of a liquid-liquid phase transition (LLPT) between these two liquids was hypothesized in the *supercooled region*, in which the temperature is lowered below the water freezing point, without the liquid becoming a solid. The measurements in this region are challenging since supercooled liquids quickly crystallize. In contrast, glassy water at these low-temperatures is stable, and hence well characterized, and presents two structural forms: the low-density amorphous (LDA) form and the high-density amorphous (HDA) one. For this reason, it was conjectured that a counterpart of these two forms exists in the supercooled region.

The structural characterization of liquids is extremely difficult because, differently from crystalline solids, liquids lack long-range order, while retaining short- to medium-range order. At a local level, the low-density liquid form of water exhibits tetrahedral order, while the high-density liquid one is more disordered and less tetrahedral, with a high probability of having interstitial water molecules within the first hydration shell. Nevertheless, their characterization on a medium-to-long length-scale is very challenging and is still a matter of debate [30, 47, 56].

3. Graph theory background and notation

Most of the material in this section is standard, but we include it for the sake of setting the notation and terminology used in the paper. A graph $G = (V,E)$ consists of a finite set $V = \{v_1, ..., v_N\}$ of nodes (or vertices) and a set $E \subseteq V \times V$ whose elements are called the edges (or links) of G. If the edges have an orientation, i.e., (v_i, v_j) is a directed edge from the node v_i to the node v_j, then the graph is *directed*. Otherwise, it is *undirected*, and we do not distinguish between (v_i, v_j) and (v_j, v_i). A positive number (the *weight*) is associated with each edge. If these weights are equal to one for all the edges, then the network is called *unweighted*, otherwise it is *weighted*.

The structural information of a graph is contained in the associated adjacency matrix A. It is a matrix of size $N \times N$, where N is the number of nodes of G, and the entry $a_{i,j}$ is equal to the weight of the edge (v_i, v_j). If such an edge does not exist, $a_{i,j}$ is set to zero. A is symmetric if and only if the associated graph is undirected. In this work we consider undirected, unweighted graphs without self-loops, i.e., without edges of the form (v_i, v_i). Hence, the adjacency matrix A is symmetric, binary, and with all zeros on the main diagonal.

The *degree* of a node v_i is the number of links incident upon the node. It is equal to the sum of the entries in the ith row of the matrix A. A graph is *regular* if all the nodes in it have the same degree. A *walk* between the nodes v_i and v_j is a sequence of nodes that starts from the node v_i, follows a sequence of edges in the graph G, and ends at node v_j. A *closed walk* is a walk where the final node coincides with the initial node. A *path* is a walk that never visits the same vertex twice, and a *cycle* is a closed walk where each node is visited only once, except the initial node (which is equal to the final one). A *shortest path* between two nodes is a path of minimal length between such nodes; the number of edges comprising a shortest path defines the *geodesic distance* between the two nodes (if no such path exists, we say that the distance between the two nodes is infinity). A graph G is *connected* if for any pair $v_i, v_j \in V$, there exists a walk starting at node v_i and ending at node v_j. In this case, the adjacency matrix

A is *irreducible*, i.e., it cannot be brought to block diagonal form by a symmetric permutation of its rows and columns. A *connected component* of a graph G is a connected subgraph $G' = (V', E')$ of $G = (V, E)$ with $V' \subseteq V$ and $E' \subseteq E$.

Using the adjacency matrix A, we can obtain the number of walks in the associated graph. Indeed, for any positive integer k, the entry $[A^k]_{i,j}$ is the number of walks of length k between nodes v_i and v_j, where A^k denotes the kth power of A. In particular, the diagonal entries of A^k count the number of closed walks of length k in G.

Let K be the *degree matrix* of G, defined as the diagonal matrix whose ith diagonal entry is the degree of the ith node. The *Laplacian matrix* of G is defined as

$$L = K - A\,. \tag{3.1}$$

Note that L is a singular, symmetric positive semidefinite matrix with $L\mathbb{1} = 0$, where $\mathbb{1}$ denotes the vector of all ones. The multiplicity of 0 as an eigenvalue of L is equal to the number of connected components of G.

The *density* $\delta(G)$ of a graph G is the ratio between the number $|E|$ of its edges and the number of possible edges, which is $\frac{N(N-1)}{2}$ for an undirected, loopless graph with N nodes. The *internal density* of a subgraph $G' = (V', E')$ of G is the ratio between the number of internal edges of V' and the number of all possible internal links, while the *external density* is the ratio between the sum external degree of V', defined as the total number of edges connecting nodes in V' with nodes in $V \setminus V'$, and the number of all possible such edges. Finally, a *cluster* of nodes is a set of connected nodes whose internal density (as a subgraph of G) is larger than its external one.

3.1 *Centrality measures*

Loosely speaking, a centrality measure (CM) is a quantity that measures the importance of a node in a graph. From a mathematical point of view, we can think of a CM as a function from the set of nodes to the set of positive real numbers, $CM : V \longrightarrow \mathbb{R}_{\geqslant 0}$. In recent years, centrality measures have found several applications in chemistry. For example, eigenvector centrality has been applied to the study of certain molecular substructures in [53], whereas closeness and betweenness centrality have been used to show that osmolytes act as hubs in the hydrogen-bond network of the solution [60].

In this section we describe some of the most important centrality measures in use. We are not attempting here to present an exhaustive list of centrality measures, but we select only the most important ones from our point of view. We refer to [21] or [51] for a more complete description.

3.1.1 *Degree centrality* The simplest centrality measure is the degree centrality, defined simply as the degree of a node. Since it considers only the nearest neighbors, it is a purely local notion and cannot capture interactions between non-neighboring nodes.

3.1.2 *Closeness centrality* Closeness centrality (CL) measures how close a node is to the rest of the nodes in the graph, see [35]. In the case of a connected graph, it is computed as the average inverse of the sum of the distances between the node and the other nodes in the graph:

$$CL(v_i) = \frac{N-1}{s(v_i)}, \tag{3.2}$$

where $s(v_i) = \sum_{v_j \in V} \{\text{length of the shortest path between } v_i \text{ and } v_j,\ j \neq i\}$ and N is the number of nodes in the graph G. Conventionally, if a vertex v_i is an isolated node (there are no walks that reach the node), then we set $CL(v_i) = 0$.

If the graph G is not connected, we consider for each node v_i the connected components $G' = (V', E')$ that contain such node. Let $N' \leqslant N$ be the number of vertices in G'. Then, see [63],

$$CL(v_i) = \frac{(N'-1)^2}{s(v_i)(N-1)}\,. \tag{3.3}$$

We note that if $N' = N$, we obtain equation (3.2).

3.1.3 *Betweenness centrality* Like the closeness centrality, the betweenness centrality (BC) is based on shortest paths in the network, see [34]. In this case, the importance of a node depends on its capacity to connect different parts in the graph. Hence, a node has a high value of BC if it enables and eases the passage of information between the other nodes in the network. In the simplest version, it considers the shortest path between the vertices, and it is defined as

$$BC(v_i) = \sum_{\substack{k,l=1 \\ k\neq i\neq l}}^{N} \frac{\sigma(v_k, v_i, v_l)}{\sigma(v_k, v_l)}\,, \tag{3.4}$$

where $\sigma(v_k, v_i, v_l)$ is the number of shortest paths between v_k and v_l that go through the node v_i, and $\sigma(v_k, v_l)$ is the number of all shortest paths connecting v_k with v_l.

Some variations of the betweenness centrality exist, where maximum flows or random walks are used instead of the shortest paths. In this work, we use only the classical version, and we refer to [21] for more information.

3.1.4 *Katz centrality* In many situations, in order to better understand the roles of nodes in a graph, it is advantageous to consider the effect of all walks (not just the shortest paths). Intuitively, we want to give more weight to the shorter walks, which can be achieved by means of an attenuation factor $\alpha > 0$. We recall that the powers of the associated adjacency matrix give information about the number of walks on the graph. The Katz centrality (KC) is the first walk-based measure that we consider, see [45]. It is defined as

$$KC(v_i) = [(I_N + \alpha A + \alpha^2 A^2 + \cdots + \alpha^k A^k + \cdots)\mathbb{1}]_i\,, \tag{3.5}$$

with I_N the $N \times N$ identity matrix and $\mathbb{1}$ the vector of all ones. In order for the matrix series in parentheses to converge, the parameter α must be such that $0 < \alpha < 1/\rho(A)$, where $\rho(A)$ is the spectral radius of the adjacency matrix. This guarantees that the series converges to the nonnegative matrix $(I_N - \alpha A)^{-1}$ and that we can rewrite the expression as

$$KC(v_i) = [(I - \alpha A)^{-1}\mathbb{1}]_i\,. \tag{3.6}$$

Given that here we consider undirected graphs only, the adjacency matrix A is symmetric. Hence, using the Spectral Theorem we can also express this centrality measure in terms of the eigenvectors and eigenvalues of A. In particular, for the Katz centrality we have

$$KC(v_i) = \sum_{k=1}^{N} \frac{1}{1-\alpha\lambda_k}(\mathbf{p}_k^T\mathbb{1})\mathbf{p}_k(i)\,, \tag{3.7}$$

where $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_N$ are the eigenvalues of A and $\mathbf{p}_k = [\mathbf{p}_k(1), \ldots, \mathbf{p}_k(N)]^T$ is the eigenvector associated with λ_k, normalized with respect to the Euclidean norm. It is easy to see (cf.[11]) that for $\alpha \to 0+$, the ranking obtained by Katz centrality is the same as the one obtained using degree centrality.

3.1.5 *Eigenvector centrality* The Eigenvector Centrality (EC) is a popular centrality measure which can also be interpreted in terms of walks on the graph [14]. Given that we are assuming that the graph G is connected, the adjacency matrix is irreducible. It is also a symmetric non-negative matrix, therefore by the Perron-Frobenius Theorem [42] the maximum eigenvalue λ_1 is positive and simple, and it coincides with the spectral radius $\rho(A)$. Furthermore, there exists a unique, up to normalization, eigenvector $\mathbf{p}_1$ associated to λ_1, such that $\mathbf{p}_1 > \mathbf{0}$ (entry-wise). Hence, the principal eigenvector can be used to define a centrality measure. We define $EC(v_i) = \mathbf{p}_1(i)$, where $\mathbf{p}_1(i)$ indicates the ith entry of the vector $\mathbf{p}_1$. It is not difficult to see that

$$EC(v_i) = \lim_{k\to\infty} \frac{\#\text{walks of length } k \text{ through } v_i}{\#\text{walks of length } k \text{ in } G}, \tag{3.8}$$

showing that a node has high eigenvector centrality if it keeps being visited "often" by walks on G, as the length of the walks tends to infinity. Moreover, it can be shown that eigenvector centrality is a limiting case of Katz centrality, in the sense that the ranking of the nodes obtained by Katz centrality as $\alpha \to 1/\rho(A)$ tends to the ranking given by eigenvector centrality, see [11].

3.1.6 *Subgraph centrality* The subgraph centrality (SUB) quantifies the importance of a node by considering the number of closed walks in the graph that pass through such node, again giving more weight to shorter walks [27]. Let $\beta > 0$, then the subgraph centrality of node v_i is defined in terms of the matrix exponential as

$$SUB(v_i) = \sum_{k=0}^{\infty} \frac{\beta^k}{k!} [A^k]_{i,i} = [e^{\beta A}]_{i,i}. \tag{3.9}$$

Longer walks are penalized through the use of the rapidly decreasing weights, $\frac{\beta^k}{k!}$ for a walk of length k. We note that the constant β is a tuning parameter that can be used to give more or less weight to longer walks. The default value of β is 1. The subgraph centrality can be obtained from the spectral decomposition of the adjacency matrix A:

$$SUB(v_i) = \sum_{k=1}^{N} e^{\beta\lambda_k} \mathbf{p}_k^2(i)\,. \tag{3.10}$$

The rankings obtained with SUB reduce to the degree and eigenvector centrality rankings for $\beta \to 0+$ and $\beta \to \infty$, respectively [11].

3.1.7 *Total communicability* The Total Communicability (TC) is closely related to the subgraph centrality. Unlike SUB, the TC does not consider only closed walks, but all the walks between a node v_i and every vertex in the graph, including v_i itself [10, 25]. The same weighting is used as in SUB, i.e, each walk of length k is weighted by $\frac{\beta^k}{k!}$. The TC centrality of node v_i is defined as

$$TC(v_i) = [e^{\beta A}\mathbb{1}]_i = \sum_{k=0}^{\infty} \frac{\beta^k}{k!} [A^k\mathbb{1}]_i\,, \tag{3.11}$$

where β is a positive parameter. Unlike the subgraph centrality, the TC can be computed efficiently even for very large graphs using algorithms for evaluating the action of a matrix function on a vector [9]. Like the other walk-based measures, it can be expressed in terms of the eigenvalues and eigenvectors of the matrix A:

$$TC(v_i) = \sum_{k=1}^{N} e^{\beta\lambda_k} (\mathbf{p}_k^T \mathbb{1}) \mathbf{p}_k(i) \,. \tag{3.12}$$

The limiting behavior for $\beta \to 0+$ and for $\beta \to \infty$ of the ranking obtained using TC is the same as that of the subgraph centrality, see again [11].

3.2 *Other metrics*

A number of global measures exist that can be used to characterize the geometrical organization of a graph. In the context of water molecules, for example, it is often useful to investigate the presence of cycles of various lengths, see [30]. Moreover, as is the case for other types of networks, it is of interest to study the clustering coefficient and the average shortest path lengths as indicators of possible small-world patterns [51].

3.2.1 *Counting the cycles in a graph* Some analytic formulae based on the adjacency matrix's spectral moments exist to compute the number of cycles in a graph. We recall that the kth moment is defined as $\mu_k = \sum_{j=1}^{N} \lambda_j^k = \mathrm{Tr}(A^k)$, and that $\mathrm{Tr}(A^k)$ is the number of closed walks of length k in the graph. In particular, let $|S_k|$ be the number of cycles of length k, then

$$\begin{aligned} |S_3| &= \frac{1}{6}\mu_3 \,, \\ |S_4| &= \frac{1}{8}\left(\mu_4 - 2\sum_{i=1}^{N} k_i(k_i-1) - 2m\right), \\ |S_5| &= \frac{1}{10}\left(\mu_5 - 30|S_3| - 10\sum_{k_i>2} \frac{[A^3]_{i,i}}{2}(k_i-2)\right), \end{aligned} \tag{3.13}$$

where k_i is the degree of node v_i, m is the number of edges in the graph and $[A^3]_{i,i}$ is the entry (i,i) of the third power of A. Note that the quantity $\frac{[A^3]_{i,i}}{2}$ is equal to the number of triangles attached to the node v_i. See [3] and [21] for details in obtaining the formulae.

3.2.2 *Average Watts-Strogatz clustering coefficient and transitivity index* The clustering coefficient and the transitivity index give information about how clustered a network is locally and globally, respectively [21, 51]. The clustering coefficient of node v_i, denoted C_i, is defined as

$$C_i = \frac{2t_i}{k_i(k_i-1)} \,,$$

where t_i is the number of triangles that pass through the node v_i of degree k_i. It represents the fraction of actual triangles over all potential triangles passing through node v_i. The average Watts-Strogatz clustering coefficient is just the mean of the clustering coefficients over all nodes:

$$\bar{C} = \frac{1}{N}\sum_{i=1}^{N} C_i \,.$$

The transitivity index (or Newman clustering) is a global measure of the frequency of triangles in a network. It is defined as the number of triangles in the graph divided by the total number of walks of length 2, $|P_2|$:

$$C = \frac{3|S_3|}{|P_2|}.$$

Note that $|P_2| = \sum_{i=1}^{N} k_i(k_i-1)/2$. Both these global indexes take values between 0 and 1; values close to 0 indicate a low density of triangles in the graph, while values close to 1 represent a high density. Generally speaking, the two clustering measures are in a good correlation, but there are graphs where they behave differently, see [22, 23].

3.2.3 *Bipartivity measure* A graph is bipartite if we can partition the node-set into two disjoint non-empty sets V_1 and V_2, such that every edge of the graph connects a vertex in V_1 to one in V_2. This means that in a bipartite graph, the closed walks of odd length are not present. A bipartivity measure quantifies how much a graph is close to being bipartite. Estrada has introduced the following measure:

$$B = \frac{\mathrm{Tr}(\cosh(A))}{\mathrm{Tr}(\exp(A))}, \tag{3.14}$$

where $\cosh(A) = \frac{\exp(A)+\exp(-A)}{2}$ denotes the hyperbolic cosine of the adjacency matrix A [26]. We note that $0.5 < B \leqslant 1$, and that $B = 1$ if and only if the graph is bipartite. Values of B close to 1 indicate that the graph is nearly bipartite (i.e, it can be made into a bipartite graph by rewiring a small number of edges).

3.2.4 *Average shortest path length* Recall that the distance $d(v_i, v_j)$ between the nodes v_i and v_j is the number of edges in a shortest path between such vertices. If a walk between v_i and v_j does not exist, their distance is set to $+\infty$. The diameter of a graph is defined as the maximum distance between any two vertices in the graph:

$$\mathrm{diam}\,(G) = \max_{v_i, v_j \in V} \{d(v_i, v_j)\}. \tag{3.15}$$

We can define the distance matrix D as the symmetric matrix whose (i, j) entry is equal to $d(v_i, v_j)$, for all $i, j = 1, \ldots, N$. Then the average shortest path length (ASPL) is defined as

$$\bar{l} = \frac{\mathbb{1}^T D \mathbb{1}}{N(N-1)}, \tag{3.16}$$

where we consider only pairs of nodes for which a path connecting them exists (such a path always exists if G is connected). A relatively large clustering coefficient and an average path length that grows slowly (say, logarithmically) with the number of nodes N characterize the class of small world (or Watts–Strogatz) networks [51]. In contrast, very small clustering coefficients and a fast growing (with N) ASPL are observed for grid-like graphs and for classical random graph models.

3.2.5 *Energy of a graph* Let $\lambda_1, \lambda_2, \ldots, \lambda_N$ be the eigenvalues of the adjacency matrix A, then the energy of the associated graph G is defined as

$$E(G) = \sum_{i=1}^{N} |\lambda_i|.$$

The energy is an important graph invariant inspired by chemistry (Hückel Molecular Orbital Theory) and that has been extensively studied; see, e.g, the monograph [46], as well as [24] for an interpretation of the graph energy in terms of walks on G and for applications to molecular structure. It is interesting to note that adding edges to an existing graph does not always lead to an increase in $E(G)$; see, e.g., [4].

3.2.6 *Algebraic connectivity* Let us denote the eigenvalues of the Laplacian matrix of G as follows:

$$0 = \mu_1 \leqslant \mu_2 \leqslant \ldots \leqslant \mu_n \,. \tag{3.17}$$

The second eigenvalue of the Laplacian matrix, $\alpha(G) = \mu_2$, is called the *algebraic connectivity* of the graph G, and provides a quantitative measure of "how well-connected the graph is [29]; note that $\alpha(G) > 0$ if and only if G is connected. Larger values of $\alpha(G)$ correspond to graphs that are not easily disconnected by removing a small number of edges. It is known [29] that the algebraic connectivity is edge-monotone, i.e., for two graphs with the same vertex set, $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$, with $E_1 \subseteq E_2$, it is $\alpha(G_1) \leqslant \alpha(G_2)$.

3.2.7 *Degree assortativity coefficient* This quantity measures the tendency of a node to connect with nodes that have equal, or almost equal, degrees [51]. If this quantity is positive, then the network is assortative. Otherwise, it is called disassortative. It is defined as

$$r = \frac{|P_2|(|P_{3/2}| + C - |P_{2/1}|)}{3|S_{1,3}| + |P_2|(1 - |P_{2/1}|)}\,, \tag{3.18}$$

where $|P_i|$ is the number of paths of length i, $|P_{i/j}| := |P_i|/|P_j|$, $|S_{1,3}|$ is the number of star fragments of four nodes and C is the transitivity index defined earlier. The denominator is always grater than or equal to zero, therefore $r > 0$ (the graph is assortative) if, and only if, $|P_{2/1}| < |P_{3/2}| + C$. The denominator is equal to zero in the case of a regular graph; in this case, the assortativity coefficient is not defined.

4. Molecular dynamics simulations of liquid water

Classical molecular dynamics (MD) simulation is a powerful method to sample the configurational space of a liquid (or, more generally, of complex solutions, such as biological systems) [64, 67]. Here, we use previously performed [68] MD simulations of pure water; in particular, we use MD simulations on the nanosecond timescale of TIP4P/2005 water at 1950 bar and at four different temperatures, namely 170 K (200 ns), 180 K (150 ns), 200 K (100 ns) and 240 K (100 ns). The TIP4P/2005 water model [1], which is widely used to simulate both pure water and aqueous solutions [13, 68, 69], exhibits a metastable liquidliquid critical point in deeply supercooled conditions. Given the simulation parameters, which are consistent with those of Biddle et al. [12], the liquid-liquid critical point is estimated to be at 1700 bar and 182 K. A rectangular box containing 710 water molecules was used and the MD simulations were performed in the NPT ensemble with the 5.1.2 version of the GROMACS software [2]. Temperature and pressure were kept constant by using the velocity rescaling temperature coupling [16] and the ParrinelloRahman barostat with 2 ps relaxation times [54]. Periodic boundary conditions were used, long range electrostatic interactions were treated with the particle mesh Ewald method [18] with a real space cutoff of 0.9 nm and for short range interactions a cut-off radius of 0.9 nm was employed. All bonds and intramolecular distances were constrained using the LINCS algorithm [41] and a 2 fs time step was utilized.

5. Liquid water as a network

Given a MD trajectory, the coordinates of a given set of atoms extracted at a given time frame can be used to build a network. Since in the simulations periodic boundary conditions are imposed, meaning that the box is replicated along the three directions of space, each molecule interacts with the images of the molecules which are on the opposite sides of the box.

We form the graph $G = (V,E)$ where the water molecules represent the nodes of G while links between the molecules are the edges. There are different ways of defining a link between two particles, [44]. In this work, using similar ideas and arguments to our previous paper [28], we consider two molecules as connected by an edge if the distance between their oxygen atoms is less than or equal to 0.35 nm (considering their replications along the three directions). In order to evaluate the distance, we compute the matrix of the physical distances D_{phy}. In practice, for each pair of particles v_i and v_j in the box, we calculate their distance along each direction. For example, let $dist_x(v_i, v_j)$ be their distance along the x-axis. If this quantity is larger than half of the box size in that direction L_x, we replace it with the length of the box minus this value: $dist_x(v_i, v_j) = L_x - dist_x(v_i, v_j)$. We do the same for each direction, and we determine the physical distance matrix:

$$D_{phy}(v_i, v_j) = \sqrt{dist_x(v_i, v_j)^2 + dist_y(v_i, v_j)^2 + dist_z(v_i, v_j)^2}\,.$$

At this point, if $D_{phy}(v_i, v_j) \leqslant 0.35$, there exists an edge between the nodes v_i and v_j. The corresponding adjacency matrix A is symmetric, undirected, unweighted, and of dimension $N \times N$, where N is the number of molecules (in our experiments $N = 710$).

The code we used to run all network analyses, together with instructions on how to install and use it, has been made available in the GitHub repository (https://github.com/ChiaraFaccio/WaterNetworks). The code makes heavy use of the NetworkX 2.6.3 module [39] in Python 3.7, the Python package NetworkSNS [15], the code [65] for imposing the periodic boundary conditions, and the open-source packages NumPy [40], SciPy [62], and Matplotlib [43]. All the computations were performed on a laptop with a 4 Intel core i7-8565U CPU @ 1.80GHz - 1.99 GHz and 16.0GB RAM.

6. Results

We analyze here four MD simulations on the nanosecond timescale of neat water using the TIP4P/2005 water model [59] along the 1950 bar isobar [28, 68], which is above the critical pressure (1700 bar) and along which the liquidliquid coexistence line is crossed at a temperature of around 175 K. We chose temperatures below (170K) and above (180, 200 and 240 K) this liquid-liquid transition temperature in order to maximize changes in water structure on going from the lowest temperature, at which the water is in the LDL-form, to the highest ones, at which the HDL-form prevails. For each MD trajectory, we extract 100 frames equally-spaced in time and analyze the corresponding networks using the previously described graph- and network-theoretic metrics. We note explicitly that while the number of nodes in each network remains constant, the number of edges varies along the MD trajectory as well as with the temperature, since the spatial arrangement of the oxygens varies.

6.1 *Analysis of water networks via centrality measures*

We begin by comparing the distributions for the four temperatures using different centrality measures to see how well they can distinguish the LDL phase from the HDL phase. Degree and eigenvector

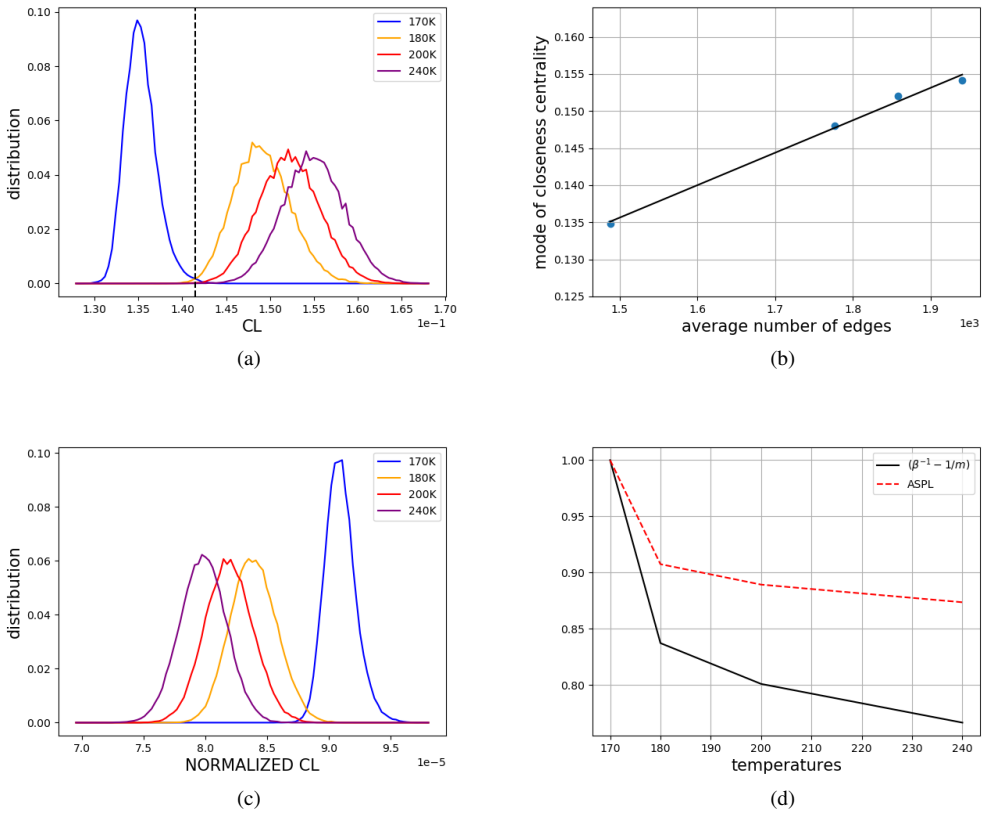


FIG. 1: Figure[1a]: Distribution of closeness centrality values. The black dashed line corresponds to the intersection between the distribution at 170K and the distribution at 180K. Figure[1b]: Linear correlation between the average number of edges in the four temperatures and the mode of CL for the four distributions. Figure[1c]: Distribution of the closeness centrality values normalized by the number of edges. Figure[1d]: Average shortest path length and average $(\beta(G)^{-1} - 1/m)$ coefficient for the four temperatures. The data are obtained from the MD simulations at 1950 bar at 170K (blue), 180K (orange), 200K (red) and 240K (violet).

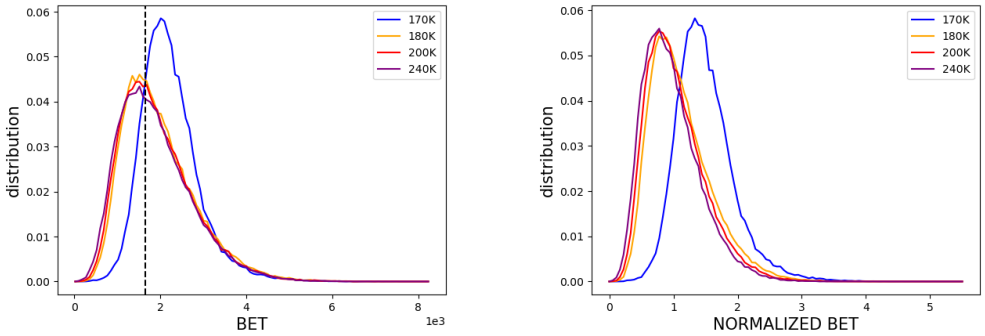
centrality, as mentioned in the Introduction (see also [28]), are not effective for this task and therefore we will not show the corresponding results, while the distributions of total communicability are reported in Figure[5a]. In Figure[1a], we report the distributions of closeness centrality at 1950 bar. This centrality measure is able to differentiate the LDL phase from the HDL one: the distribution at 170K is well-separated from the distributions at higher temperatures, while there is a significant overlap between the plots at 180K, 200K, and 240K. Closeness centrality values follow the same trend as the number of links in the network. In Figure[1b], we can see that there is, in fact, a linear correlation between the modes of the distributions at the four temperatures and the average number of edges in the graphs. When the temperature increases, the distributions are shifted towards higher values of CL. Since we

construct the adjacency matrices using a threshold on the physical distances between the oxygen atoms, the distributions indicate that in the HDL phase, the average distance between the water molecules is smaller than in the LDL phase.

Some comments on the effect of normalization of the computed centrality values are in order, since we are comparing networks with the same number of nodes but different number of links. If we normalize by the number of edges, we obtain the distributions in Figure[1c]. The normalized closeness values now show an opposite behavior, i.e., when the temperature increases, the modes of the distributions are shifted towards smaller values. This apparently surprising behavior can be attributed to the difference between the rates of change of the number of edges and of the quantity $s(v_i)$ in equation (3.2) for the four temperatures. Let m be the number of edges in the network, then we have

$$\frac{CL(v_i)}{m} = \frac{N-1}{m}\frac{1}{s(v_i)} = \left(\frac{N}{m} - \frac{1}{m}\right)\frac{1}{s(v_i)} = \left(\beta(G)^{-1} - \frac{1}{m}\right)\frac{1}{s(v_i)}, \tag{6.1}$$

where $\beta(G) = m/N$ is the *beta index* of the graph G (not to be confused with the parameter β used in the definition of sugraph and total communicability centrality), an indicator of the level of connectivity of a graph. In our case, the average quantity $(\beta(G)^{-1} - 1/m)$, reporting on the increase of the number of edges upon raising the temperature, is equal to 0.48, 0.40, 0.38, and 0.37, on going from 170 K to 240 K. As a global indicator for the behavior of $s(v_i)$, we can consider the average shortest path length (ASPL). We obtain that ASPL is 7.38, 6.70, 6.57, and 6.45 at 170 K, 180 K, 200 K and 240 K, respectively. In Figure[1d] we plot these values, normalized to be between 0 and 1, and we observe that the rate of decrease of $(\beta(G)^{-1} - 1/m)$ is greater than the rate of the ASPL. For this reason, normalizing by the number of edges we obtain the distributions of Figure[1c].

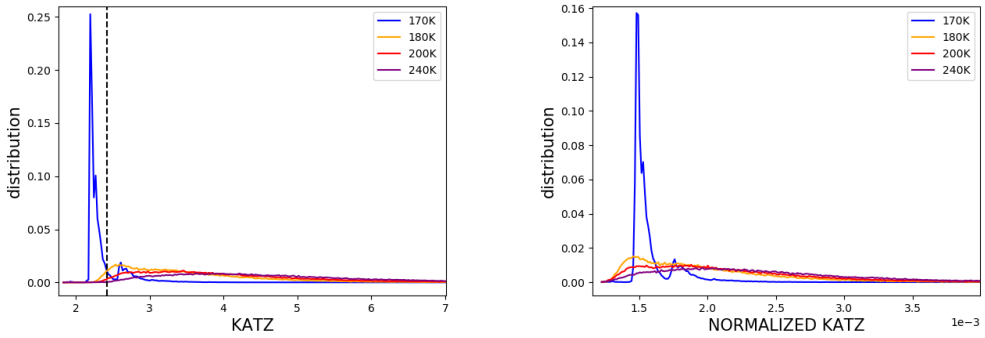


(a) (b)

FIG. 2: Figure[2a]: Distribution of betweenness centrality values. The black dashed line corresponds to the intersection between the distribution at 170K and the distribution at 180K. Figure[2b]: Distribution of the betweenness centrality normalized by the number of edges. The data are obtained from the MD simulations at 1950 bar at 170K (blue), 180K (orange), 200K (red) and 240K (violet).

The distributions of the betweenness centrality values are reported in Figure[2a]. We observe that the BC values in the LDL phase tend to be somewhat higher than for the HDL phase. This can be explained

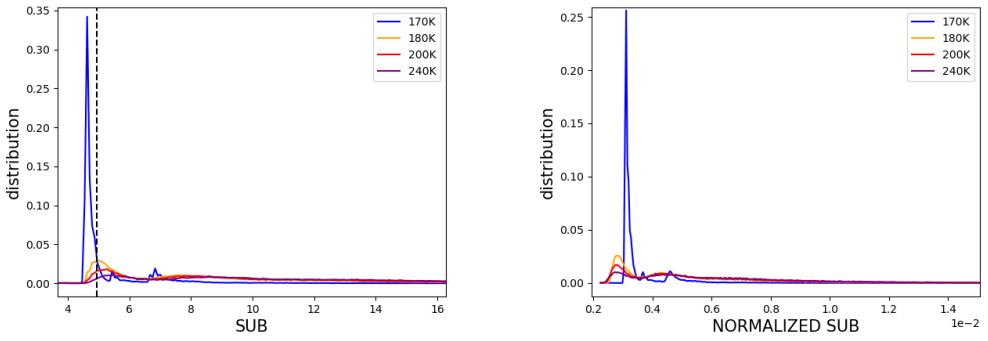
by noting that the BC of a node is proportional to the number of shortest paths that pass through such vertices. Since in the high-density liquid water the molecules are closer, and thus the number of edges increases, there is a rise in the availability of shortest paths, and the fraction of shortest paths through each vertex tends to decrease. However, there is a significant overlap between the two regimes, and betweenness centrality does not appear to be able to clearly identify the two density forms of liquid water. The normalization by the number of links (Figure[2b]) determines only minor changes (i.e, the distributions at higher temperatures become tall like at 170K), and the BC is still unable to well separate between the LDL and HDL phases. We conclude that the betweenness centrality is not an useful order parameter to identify the two density forms of liquid water.



(a) (b)

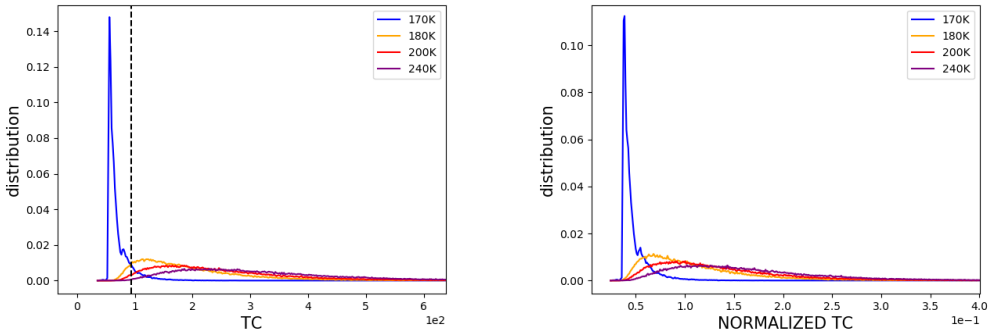
FIG. 3: Figure[3a]: Distribution of Katz centrality values using $\alpha = 0.1362$. The black dashed line corresponds to the intersection between the distribution at 170K and the distribution at 180K. Figure[3b]: Distribution of the Katz centrality normalized by the number of edges. The data are obtained from the MD simulations at 1950 bar at 170K (blue), 180K (orange), 200K (red) and 240K (violet).

The distributions of the Katz, subgraph, and total communicability centrality values are reported in Figures [3a,4a,5a], respectively. Recall that the KC of a node, like the TC, considers all the walks through that node, while SUB counts only the closed walks. All these centralities use a parameter to give less weight to the long walks. These measures have similar distributions, and they well identify the LDL phase from the HDL one. In particular, for Katz centrality we chose the penalty factor $\alpha = 0.1362$, which we determined as follows. Recall that α is subject to the condition $0 < \alpha < 1/\rho(A)$, where $\rho(A)$ is the spectral radius of the adjacency matrix, and that if $\alpha \to 1/\rho(A)$, the node ranking of KC converges to the ranking of EC, while, for $\alpha \to 0$, the ranking reduces to the ranking of the degree centrality. The spectral radii in the LDL phase are lower than in the HDL phase, so if we compute the KC of each graph using a penalty factor that depends on the spectral radius of the associated adjacency matrix A_i, $\alpha = 1/(\gamma \cdot \rho(A_i))$ (with $\gamma > 1$ a constant), then, as the temperature increases, we give less weight to the walks. On the other hand, in the HDL phase, there are more walks of length k than in the LDL phase for $k \geqslant 1$, but since α is smaller, we tend to lose this information. For this reason, we set the α parameter to be the same for all four trajectories. The maximum spectral radius for all the graphs



(a) (b)

FIG. 4: Figure[4a]: Distribution of subgraph centrality values using $\beta = 1$. The black dashed line corresponds to the intersection between the distribution at 170K and the distribution at 180K. Figure[4b]: Distribution of the subgraph centrality normalized by the number of edges. The data are obtained from the MD simulations at 1950 bar at 170K (blue), 180K (orange), 200K (red) and 240K (violet).



(a) (b)

FIG. 5: Figure[5a]: Distribution of total communicability values using $\beta = 1$. The black dashed line corresponds to the intersection between the distribution at 170K and the distribution at 180K. Figure[5b]: Distribution of the total communicability normalized by the number of edges. The data are obtained from the MD simulations at 1950 bar at 170K (blue), 180K (orange), 200K (red) and 240K (violet).

is 6.67. As a consequence, the α factor must be at least less than $1/6.67 \approx 0.15$, and far enough from this quantity to avoid reducing to the eigenvector centrality. We obtain that $\alpha = 1/(1.1 \cdot 6.67) = 0.1362$ is a good choice, in fact the centrality measure allows to identify the two densities of the liquid water. Nevertheless, at 170 K (LDL phase) there is a small secondary peak overlapping with the distributions obtained at the higher temperatures (i.e., in the HDL phase). This peak is also present in the subgraph centrality distribution at 170K with the choice $\beta = 1$. The molecules belonging to the two peaks are

almost the same, as we can observe from Figure[6a], and they correspond, in part, to the molecules of the small peak in the TC distribution at 170K with TC values between 77 and 93.

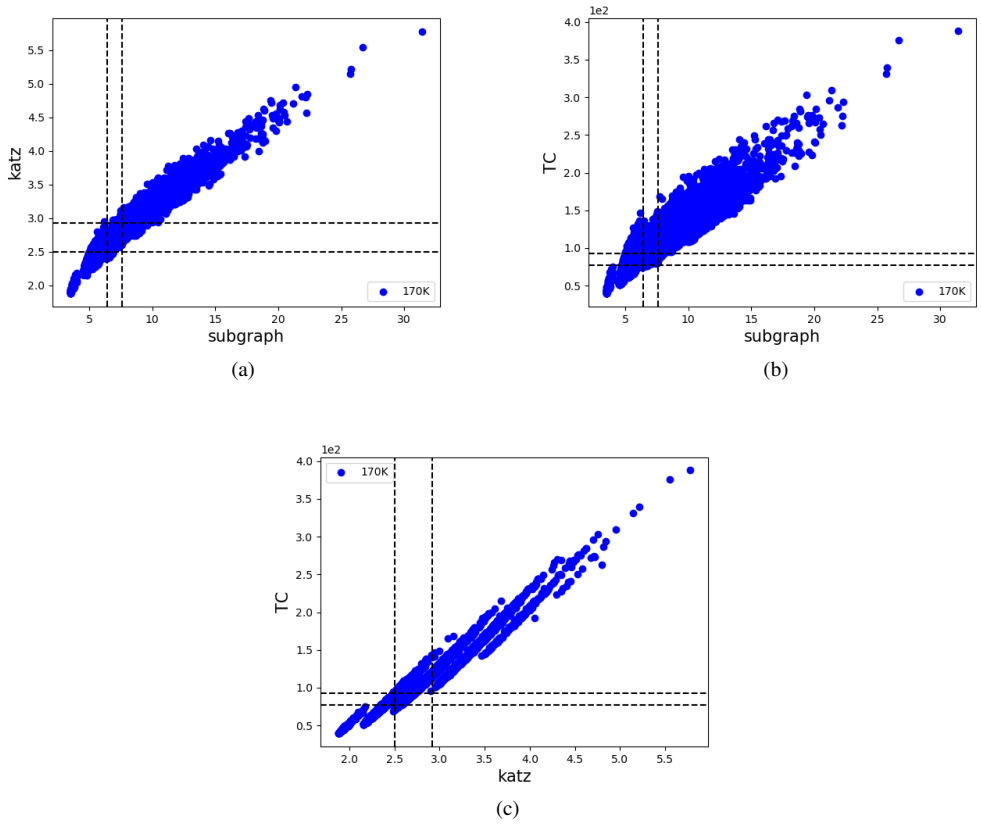


FIG. 6: Figure[6a]: Correlation between the values of SUB and KC for the temperature at 170K. Figure[6b]: Correlation between the values of SUB and TC for the temperature at 170K. Figure[6c]: Correlation between the values of TC and KC for the temperature at 170K. The black dashed lines correspond to the bounds for the peaks.

For the subgraph and total communicability centrality, we employ $\beta = 1$ for two reasons: first of all, it is widely used as a default value and, secondly, following the results in [28] for the total communicability, this choice leads to including the contributions of molecules within a 1 nm distance (the contribution of molecules at a greater distance being numerically negligible), which is the typical correlation length in the pair radial distribution function for the $O \cdots O$ contacts, see [28].

If we normalize these centralities by the number of edges, KC is no longer able to identify the two density phases, subgraph centrality distributions show an opposite behavior (similar to closeness), while for TC, the overlap between the distributions increases slightly. In particular, TC is less influenced by this normalization than Katz and Subgraph. This is probably due to the fact that the three normal-

ized centrality measures give different weights to the walks: for example, at 170K with an average of $m = 1487.61$ edges, given $\alpha = 0.1362$ and $\beta = 1$, the penalty factor referred to the degree is $9.16 \cdot 10^{-05}$, $3.36 \cdot 10^{-04}$, $6.72 \cdot 10^{-04}$ for Katz, Subgraph and Total Communicability respectively. In practice, normalizing by the number of edges leads to a loss of information about an essential feature of the two density forms, namely, the fact that water molecules in the HDL phase display a higher connectivity.

Next, we calculate the LDL fraction at the four temperatures using the previous values without normalization. As in [28], we use the crossing point between the distributions at 170K and 180K at 1950 bar to discriminate the two phases. Let x_* be such point, then v_i is selected in the LDL phase if $CM(v_i) \leqslant x_*$, otherwise it is assigned to the HDL one.

Table 1: Fraction of LDL population at 1950 bar for the four temperatures as obtained using different centrality measures.

	170K	180K	200K	240K
CL	99.53 %	0.27 %	0.03 %	0.009 %
BC	84.18 %	61.22 %	59.16 %	57.18 %
KC	79.81 %	3.15 %	0.79 %	0.18 %
SUB	71.27 %	8.51 %	3.80 %	1.49 %
TC	89.43 %	6.48 %	1.74 %	0.20 %

In Table[1], the results are summarized. It can be observed that, with the exception of the betweenness, all centrality measures well discriminate the two liquid phases. In the case of the betweenness, as previously recognized in Figure[2a], there is a significant overlap between the distributions at the four temperatures, and thus BC overestimates the fraction of LDL water molecules in the pure HDL regime.

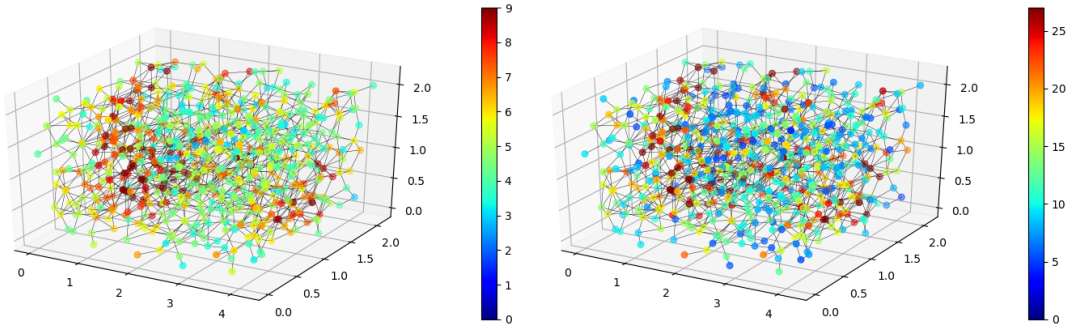
With respect to CL and TC, the Katz and subgraph centrality provide a lower estimate of the LDL fraction at 170 K. This is due to the presence of the already commented secondary peaks. In the case of the total communicability, the secondary peak is located before the crossing point used to define the two regimes and does not overlap with the distributions obtained in the HDL phase.

A strong point of centrality measures is that they provide information on each node of the graph. Therefore, as described in [28], they work at the molecular level and permit to obtain information about the internal organization of the network.

To this aim, we focus on Katz, subgraph, and total communicability centrality (we exclude the betweenness centrality because it does not identify the two liquids and the closeness because the nodes with high centrality values are all in the center of the box. Its definition introduces in fact a bias towards central nodes and therefore it cannot be used to investigate the structural organization of the water network). When the temperature increases, the distributions of Katz, subgraph and TC become broader, meaning that the values of the centrality measures tend to be higher. We previously observed [28] that in the HDL phase, the nodes with high TC are not isolated, but appear to organize themselves into patches.

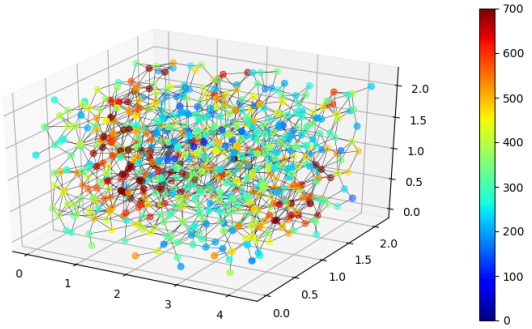
We consider the trajectory at 240K, where almost all the water molecules are in the HDL phase. From Figure[7a] and Figure[7b], we observe the same behavior that we noted in [28] also for Katz and subgraph centrality, i.e., molecules with high centrality measures are assembled in specific regions of the box.

As shown in Figure[8], there is a big patch of molecules with centrality values above the average (colored in silver in the Figure) for all three centrality measures. The red nodes are the HDL molecules, while the blue ones are LDL molecules. Only the subgraph centrality identifies some of the molecules



(a) Katz centrality

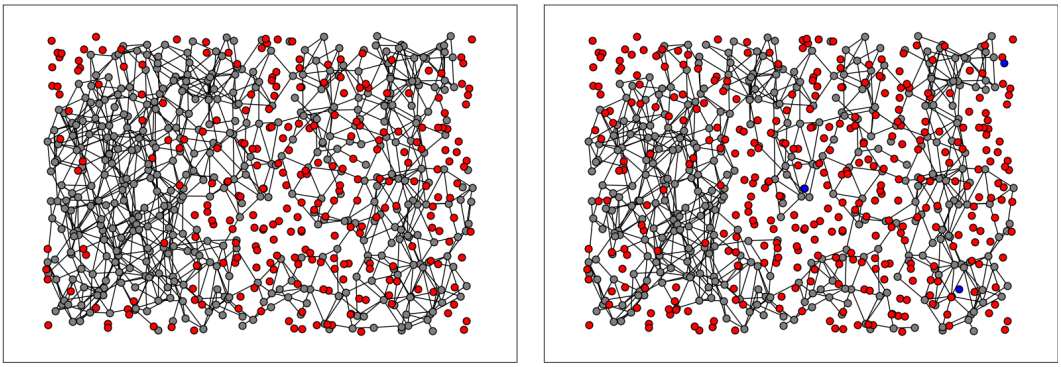
(b) Subgraph centrality



(c) Total communicability

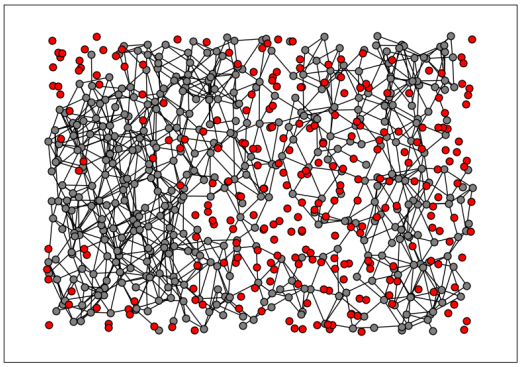
FIG. 7: 3D plot of the distribution of oxygen atoms with different Katz, subgraph and TC centrality values for frame 7 at 240K.

as being in the low-density phase, according to results in Table[1]. The regions are composed of 447 molecules in the TC case, 448 molecules in the Katz case, and 360 for SUB. We argue that the patch size obtained with SUB differs from those found with KC and TC because SUB does not consider all the walks of all the lengths (like KC and TC), but only the closed walks. The average internal density of the patches obtained with the three centrality measures is 0.01, the average external one is $4.61 \cdot 10^{-3}$, while the density $\delta(G)$ of the entire graph is $8.05 \cdot 10^{-3}$. We underline that the above values refer to the graph density, which is not necessarily correlated to the physical density of the water network. We also mention that we tried to apply standard community detection techniques to our networks, such as the Girvan–Newman algorithm or the greedy modularity technique (see [33] for details), but we obtained different clusters which, however, do not appear to be amenable to chemical-physical interpretation. We plan to further investigate this aspect in future works.



(a) Katz centrality

(b) Subgraph centrality



(c) Total communicability

FIG. 8: Snapshots of the internal organization of the water at 240K, frame 7. Blue nodes represents the LDL molecules, red nodes the HDL molecules and the silver ones are the molecules with CM above the average. The edges highlight the connections among high-CM nodes.

6.2 *Computational aspects*

In a typical simulation, a large number of water networks need to be analyzed, and computational efficiency is paramount. In Table[2], we report the average execution times for computing the centrality of every node for a single frame (network). For the closeness and betweenness centrality, the computation of all the shortest paths is by far the more expensive part. It requires a breadth-first search (BFS), which needs $O(|V|+|E|)$ time for each node of the graph. The Katz centrality can be evaluated using a direct solver (sparse Cholesky factorization) for the equation (3.6). The subgraph centrality looks at the diagonal entries of the exponential of the adjacency matrix, and it can be computed using quadrature rules and the Lanczos algorithm (see [9] for details). On the other hand, the total communicability corresponds to the product of the exponential of the adjacency matrix times the vector of all ones, and it can be

computed efficiently using a Lanczos-based algorithm for evaluating the action of a matrix function on a vector [9].

Table 2: Average of the execution times of a frame at different temperatures for the graph with 710 nodes using different centrality measures. The values are in seconds.

	CL	BC	KC	SUB	TC
170K	0.72	2.64	0.0063	13.64	0.0062
180K	0.78	2.84	0.0082	16.79	0.0054
200K	0.79	2.64	0.0085	16.74	0.0054
240K	0.81	2.72	0.0080	16.10	0.0060

From the table, we can see that the TC centrality is the fastest measure to compute, followed by the Katz centrality, while SUB is the most time-consuming.

6.3 *Results using global metrics*

Here we analyze the two liquid phases using some global metrics from graph and network theory. Water networks are 3D grid-like networks which in the LDL phase are near to being regular graphs (with common degree 4). As the temperature increases, this regular structure is perturbed as some degree of disorder sets in. However, even in the HDL phase the networks remain fairly regular, and do not display the emergence of the small world phenomenon. This is easily explained by noting that, because of the threshold we use to define an edge, the appearance of new links is limited to pairs of molecules that are not too far in space, hence the connectivity tends to remain short-ranged. In Table[3], we display the average shortest path length (ASPL) and diameter for the four different temperatures. We observe that both ASPL and diameter exhibit a moderate reduction for increasing temperature. In the third column of the table we report the (average) value of the algebraic connectivity of the networks for the four different temperatures. As expected, the algebraic connectivity increases monotonically with the temperature.

Next, we compare our water networks with the classical Erdös–Rényi random graph model, in which each pair of nodes is connected with a given probability $p > 0$. In an Erdös–Rényi random graph, the density of G is equal to its clustering coefficient, see [21]. Since for both the LDL and HDL phases the densities $\delta(G)$ are low while the average Watts–Strogatz indexes (see Table[5]) are relatively high, their clustering is not similar to that of an Erdös-Rényi random graph, reflecting the much higher transitivity of water networks. Note that as the temperature increases the graph density also increases, reflecting the creation of new links between pair of molecules as the network becomes less ordered.

Table 3: Mean values of the average shortest path length, diameter, algebraic connectivity, density, bipartivity measure, and energy of water network for the four temperatures.

	ASPL	diam(G)	$\alpha(G)$	$\delta(G)$	bipart.	$E(G)$
170K	7.38	13.20	0.101	0.0059	0.94	1271.14
180K	6.70	12.04	0.132	0.0071	0.80	1356.80
200K	6.57	12.00	0.138	0.0074	0.77	1379.63
240K	6.45	11.94	0.150	0.0077	0.73	1402.80

In ice, oxygen atoms form a hexagonal crystalline lattice, where each water molecule is bonded to four other water molecules, in a tetrahedral arrangement, through hydrogen bonds. (For a 4-regular

graph, the number of edges is $m = N \cdot 4/2$; if $N = 710$, then $m = 1420$. The average number of links at 170K is 1487.61, very close to this value). Such a regular lattice is easily shown to be bipartite. Since liquid water at 170K behaves similarly to ice, we expect that for water networks at this low temperature the bipartivity measure defined in (3.14) is close to 1 (nearly bipartite graph). At higher temperatures, when this ordered structure is perturbed by the appearance of new links (see first column of Table[6]), the (near) bipartivity of the network is lost, and we expect the bipartivity measure to decrease. This intuition is confirmed by the results reported in Table[3]. This behaviour is confirmed by inspecting the adjacency spectrum at different temperatures. Let $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_N$ be the eigenvalues of A. It is a basic fact of spectral graph theory that a graph is bipartite if and only if $\lambda_N = -\rho(A) = -\lambda_1$. In Table[4], we report the average extreme values of the spectrum of A and the mean values of their sum. The negative eigenvalue changes only by a small amount, while the larger one grows as the temperature increases. As a result, their sum moves away from 0. This behavior is also reflected in the increase of the graph energy $E(G)$ with the temperature, see Table[3]. Hence, we have another confirmation that the graph becomes less bipartite as the temperature increases.

Table 4: Mean values of the minimum λ_N and maximum λ_1 eigenvalue of the adjacency matrices A, and the mean values of their sum $\lambda_1 + \lambda_N$ for the four temperatures.

	λ_N	λ_1	$\lambda_1 + \lambda_N$
170 K	-3.54	4.58	1.04
180 K	-3.83	5.74	1.90
200 K	-3.88	5.93	2.05
240 K	-3.91	6.13	2.22

Next, we have found that the Watts-Strogatz clustering coefficients $\bar{C}$ are in a linear correlation with the transitivity indexes C (see Table[5]), with a Pearson correlation of 0.97 for all the temperatures considered. All the networks are assortative due to the fact that $|P_{2/1}| < |P_{3/2}| + C$. In particular, the quantity $|P_{2/1}|$ is larger than the average of $|P_{3/2}|$, $0 \leqslant C \leqslant 1$ hence the transitivity index C has a primary effect on the assortativity. As the temperature increases, the gap $(|P_{2/1}| - |P_{3/2}|)$ grows, therefore the transitivity index also increases. The increase in the transitivity index upon raising the temperature, related to the increase in the number of edges in the graph, corresponds, from a chemico-physical point of view, to the increase in the number of interstitial water molecules in the high density regime (see section 2).

Table 5: Mean values of the $|P_{2/1}|$, $|P_{3/2}|$, transitivity index C, Watts-Strogatz clustering coefficient $\bar{C}$, and degree assortativity coefficient r for the four temperatures.

	$\lvert P_{2/1}\rvert$	$\lvert P_{3/2}\rvert$	C	$\bar{C}$	r
170K	3.24	3.23	0.041	0.029	0.308
180K	4.21	4.16	0.132	0.102	0.296
200K	4.45	4.38	0.151	0.122	0.276
240K	4.70	4.60	0.175	0.149	0.252

Lastly, in the literature there are several works on the problem of counting the different number of cycles in the LDL and HDL phases, e.g., [30, 47]. These authors use tools derived from statistics. In contrast, in this paper we use the formulas (3.13), known from graph theory. In Figure[9], we observe the relative number of closed paths of length three, four, and five for the four temperatures. We normalize the

number of each cycle by dividing for the maximum value among all these numbers. Finally, we report the mean at each length. In Table[6], we also report the mean values of the number of cycles of length 3, 4, and 5. Our results show that the number of such cycles tends to increase with the temperature, signaling a topological difference between the two phases.

Table 6: Mean values of the number of edges and mean values of the number of cycles of length 3,4, and 5.

	edges	cycles of length 3	cycles of length 4	cycles of length 5
170 K	1487.61	66.43	77.98	394.36
180 K	1776.95	329.36	519.31	1327.49
200 K	1857.66	418.64	685.50	1695.79
240 K	1940.87	532.69	888.11	2136.90

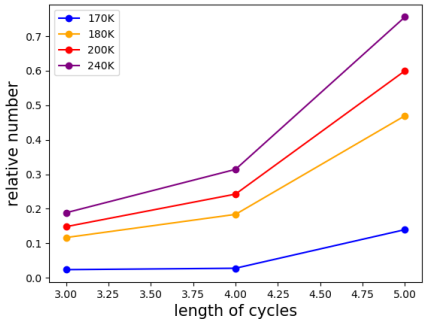


FIG. 9: Relative number of cycles of length 3, 4, 5 at 170K (blue), 180K (orange), 200K (red) and 240K (violet).

7. Conclusions and future work

In this paper, we extended the study in [28] to give a structural analysis of liquid water using molecular dynamics simulations coupled with concepts and tools from graph and network theory. We have assessed the ability of several popular centrality measures to differentiate the high-density liquid phase from the low-density liquid phase along the 1950 isobar. Since the centrality measures are computed at a molecular level, they also allow us to identify, within a single phase, the presence of high (or low) density patches. Our comparison has confirmed the usefulness of the total communicability in this context, and found that this centrality measure is also computationally very efficient.

Even if these networks are only moderately complex, network science tools prove extremely useful for obtaining fine information about the structure of liquid water and lead to a topological characterization of the two forms. Specifically, graph-theory based measures are able to show that the low-density phase is similar to a lattice, while in the high-density form this regularity is partly lost. These measures can also increase our physical understanding of the LDL and HDL forms. In fact, the changes we

observe in both centrality measures and global metrics highlight topological differences within the HDL form. We plan to deepen this aspect in future work.

Future efforts will also be aimed at improving the description of liquid water by means of a directed graph model in which the connections represent the hydrogen bonds among water molecules. Another improvement will be to consider a box with water molecules as a weighted graph, where the weight of an edge measures the length of the hydrogen bond. Finally, we would like to gain a better understanding of the clusters found by standard techniques for community detection, which do not seem to bear much relation with the low- and high-density patches identified by means of centrality measures such as TC.

REFERENCES

1. Abascal, J. & Vega, C. (2005) A general purpose model for the condensed phases of water: TIP4P/2005. *The Journal of Chemical Physics*, **123**(23), 234505.
2. Abraham, M. J., Murtola, T., Schulz, R., Páll, S., Smith, J. C., Hess, B. & Lindahl, E. (2015) GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX*, **1**, 19–25.
3. Alon, N., Yuster, R. & Zwick, U. (1997) Finding and counting given length cycles. *Algorithmica*, **17**(3), 209–223.
4. Arizmendi, O., Hidalgo, J. F. & Juarez-Romero, O. (2018) Energy of a vertex. *Linear Algebra and its Applications*, **557**, 464–495.
5. Bakó, I., Bencsura, Á., Hermannson, K., Bálint, S., Grósz, T., Chihaia, V. & Oláh, J. (2013) Hydrogen bond network topology in liquid water and methanol: a graph theory approach. *Physical Chemistry Chemical Physics*, **15**(36), 15163–15171.
6. Bakó, I., Csókás, D. & Pothoczki, S. (2021) Molecular aggregation in liquid water: Laplace spectra and spectral clustering of H-bonded network. *Journal of Molecular Liquids*, **327**, 114802.
7. Balaban, A. T. (1985) Applications of graph theory in chemistry. *Journal of Chemical Information & Computer Sciences*, **25**(3), 334–343.
8. Barabasi, A.-L. & Oltvai, Z. N. (2004) Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*, **5**(2), 101–113.
9. Benzi, M. & Boito, P. (2020) Matrix functions in network analysis. *GAMM-Mitteilungen*, **43**(3), e202000012.
10. Benzi, M. & Klymko, C. (2013) Total communicability as a centrality measure. *Journal of Complex Networks*, **1**(2), 124–149.
11. Benzi, M. & Klymko, C. (2015) On the limiting behavior of parameter-dependent network centrality measures. *SIAM Journal on Matrix Analysis and Applications*, **36**(2), 686–706.
12. Biddle, J. W., Singh, R. S., Sparano, E. M., Ricci, F., Gonzalez, M. A., Valeriani, C., Abascal, J. L., Debenedetti, P. G., Anisimov, M. A. & Caupin, F. (2017) Two-structure thermodynamics for the TIP4P/2005 model of water covering supercooled and deeply stretched regions. *The Journal of Chemical Physics*, **146**(3), 034502.
13. Biswas, A. D., Barone, V. & Daidone, I. (2021) High Water Density at Non-Ice-Binding Surfaces Contributes to the Hyperactivity of Antifreeze Proteins. *The Journal of Physical Chemistry Letters*, **12**, 8777–8783.
14. Bonacich, P. (1987) Power and centrality: A family of measures. *American Journal of Sociology*, **92**(5), 1170–1182.
15. Bucci, A. (2021) NetworkSNS, url = https://github.com/alb95/NetworkSNS.git. .
16. Bussi, G., Donadio, D. & Parrinello, M. (2007) Canonical sampling through velocity rescaling. *The Journal of Chemical Physics*, **126**(1), 014101.
17. Cuthbertson, M. J. & Poole, P. H. (2011) Mixturelike behavior near a liquid-liquid phase transition in simulations of supercooled water. *Physical Review Letters*, **106**(11), 115706.
18. Darden, T., York, D. & Pedersen, L. (1993) Particle mesh Ewald: An Nlog (N) method for Ewald sums in large systems. *The Journal of Chemical Physics*, **98**(12), 10089–10092.
19. Debenedetti, P. G., Sciortino, F. & Zerze, G. H. (2020) Second critical point in two realistic models of water.

Science, **369**(6501), 289–292.

20. Errington, J. R. & Debenedetti, P. G. (2001) Relationship between structural order and the anomalies of liquid water. *Nature*, **409**(6818), 318–321.

21. Estrada, E. (2012) *The Structure of Complex Networks: Theory and Applications*. Oxford University Press, Oxford, UK.

22. Estrada, E. (2016) When local and global clustering of networks diverge. *Linear Algebra and its Applications*, **488**, 249–263.

23. Estrada, E. & Benzi, M. (2017a) Core–satellite graphs: Clustering, assortativity and spectral properties. *Linear Algebra and its Applications*, **517**, 30–52.

24. Estrada, E. & Benzi, M. (2017b) What is the meaning of the graph energy after all?. *Discrete Applied Mathematics*, **230**, 71–77.

25. Estrada, E., Hatano, N. & Benzi, M. (2012) The physics of communicability in complex networks. *Physics Reports*, **514**(3), 89–119.

26. Estrada, E. & Knight, P. A. (2015) *A First Course in Network Theory*. Oxford University Press, Oxford, UK.

27. Estrada, E. & Rodriguez-Velazquez, J. A. (2005) Subgraph centrality in complex networks. *Physical Review E*, **71**(5), 056103.

28. Faccio, C., Benzi, M., Zanetti-Polzi, L. & Daidone, I. (2022) Low-and high-density forms of liquid water revealed by a new medium-range order descriptor. *Journal of Molecular Liquids*, **355**, 118922.

29. Fiedler, M. (1973) Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, **23**(2), 298–305.

30. Foffi, R., Russo, J. & Sciortino, F. (2021) Structural and topological changes across the liquid–liquid transition in water. *The Journal of Chemical Physics*, **154**(18), 184506.

31. Foffi, R. & Sciortino, F. (2021) Structure of high-pressure supercooled and glassy water. *Physical Review Letters*, **127**(17), 175502.

32. Formanek, M. & Martelli, F. (2020) Probing the network topology in network-forming materials: The case of water. *AIP Advances*, **10**(5), 055205.

33. Fortunato, S. (2010) Community detection in graphs. *Physics Reports*, **486**(3-5), 75–174.

34. Freeman, L. C. (1977) A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

35. Freeman, L. C. (1978) Centrality in social networks conceptual clarification. *Social Networks*, **1**(3), 215–239.

36. Gallo, P., Amann-Winkel, K., Angell, C. A., Anisimov, M. A., Caupin, F., Chakravarty, C., Lascaris, E., Loerting, T., Panagiotopoulos, A. Z., Russo, J. et al. (2016) Water: A tale of two liquids. *Chemical Reviews*, **116**(13), 7463–7500.

37. Gartner III, T. E., Zhang, L., Piaggi, P. M., Car, R., Panagiotopoulos, A. Z. & Debenedetti, P. G. (2020) Signatures of a liquid–liquid transition in an ab initio deep neural network model for water. *Proceedings of the National Academy of Sciences*, **117**(42), 26040–26046.

38. Gursoy, A., Keskin, O. & Nussinov, R. (2008) Topological properties of protein interaction networks from a structural perspective. *Biochemical Society Transactions*, **36**(6), 1398–1403.

39. Hagberg, A., Chult, D. S. & Swart, P. (2008) Exploring Network Structure, Dynamics, and Function using NetworkX. In Varoquaux, G., Vaught, T. & Millman, J., editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA.

40. Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., Gérard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C. & Oliphant, T. E. (2020) Array programming with NumPy. *Nature*, **585**(7825), 357–362.

41. Hess, B., Bekker, H., Berendsen, H. J. & Fraaije, J. G. (1997) LINCS: a linear constraint solver for molecular simulations. *Journal of Computational Chemistry*, **18**(12), 1463–1472.

42. Horn, R. A. & Johnson, C. R. (2012) *Matrix Analysis*. Cambridge University Press, Cambridge, UK.

43. Hunter, J. D. (2007) Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, **9**(3), 90–95.

44. Karathanou, K., Lazaratos, M., Bertalan, É., Siemers, M., Buzar, K., Schertler, G. F., Del Val, C. & Bondar, A.-N. (2020) A graph-based approach identifies dynamic H-bond communication networks in spike protein S

of SARS-CoV-2. *Journal of Structural Biology*, **212**(2), 107617.
45. Katz, L. (1953) A new status index derived from sociometric analysis. *Psychometrika*, **18**(1), 39–43.
46. Li, X., Shi, Y. & Gutman, I. (2012) *Graph Energy*. Springer Science & Business Media.
47. Martelli, F. (2019) Unravelling the contribution of local structures to the anomalies of water: The synergistic action of several factors. *The Journal of Chemical Physics*, **150**(9), 094506.
48. Martelli, F. (2021) Topology and complexity of the hydrogen bond network in classical models of water. *Journal of Molecular Liquids*, **329**, 115530.
49. Montes de Oca, J. M., Sciortino, F. & Appignanesi, G. A. (2020) A structural indicator for water built upon potential energy considerations. *The Journal of Chemical Physics*, **152**(24), 244503.
50. Muthachikavil, A. V., Kontogeorgis, G. M., Liang, X., Lei, Q. & Peng, B. (2022) Structural characteristics of low-density environments in liquid water. *Physical Review E*, **105**(3), 034604.
51. Newman, M. (2018) *Networks*. Oxford University Press, Oxford, UK.
52. Palmer, J. C., Poole, P. H., Sciortino, F. & Debenedetti, P. G. (2018) Advances in computational studies of the liquid–liquid transition in water and water-like models. *Chemical Reviews*, **118**(18), 9129–9151.
53. Parisutham, N. & Rethnasamy, N. (2021) Eigenvector centrality based algorithm for finding a maximal common connected vertex induced molecular substructure of two chemical graphs. *Journal of Molecular Structure*, **1244**, 130980.
54. Parrinello, M. & Rahman, A. (1980) Crystal structure and pair potentials: A molecular-dynamics study. *Physical Review Letters*, **45**(14), 1196.
55. Poole, P. H., Sciortino, F., Essmann, U. & Stanley, H. E. (1992) Phase behaviour of metastable water. *Nature*, **360**(6402), 324–328.
56. Russo, J. & Tanaka, H. (2014) Understanding waters anomalies with locally favoured structures. *Nature Communications*, **5**(1), 1–11.
57. Shiratani, E. & Sasai, M. (1996) Growth and collapse of structural patterns in the hydrogen bond network in liquid water. *The Journal of Chemical Physics*, **104**(19), 7671–7680.
58. Shiratani, E. & Sasai, M. (1998) Molecular scale precursor of the liquid–liquid phase transition of water. *The Journal of Chemical Physics*, **108**(8), 3264–3276.
59. Singh, R. S., Biddle, J. W., Debenedetti, P. G. & Anisimov, M. A. (2016) Two-state thermodynamics and the possibility of a liquid-liquid phase transition in supercooled TIP4P/2005 water. *The Journal of Chemical Physics*, **144**(14), 144504.
60. Sundar, S., Sandilya, A. A. & Priya, M. H. (2021) Unraveling the influence of osmolytes on water hydrogen-bond network: From local structure to graph theory analysis. *Journal of Chemical Information and Modeling*, **61**(8), 3927–3944.
61. Tanaka, H., Tong, H., Shi, R. & Russo, J. (2019) Revealing key structural features hidden in liquids and glasses. *Nature Reviews Physics*, **1**(5), 333–348.
62. Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P. & SciPy 1.0 Contributors (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, **17**, 261–272.
63. Wasserman, S. & Faust, K. (1994) *Social Network Analysis: Methods and Applications*, volume 8. Cambridge University Press, Cambridge, UK.
64. Yagasaki, T., Matsumoto, M. & Tanaka, H. (2019) Liquid-liquid separation of aqueous solutions: A molecular dynamics study. *The Journal of Chemical Physics*, **150**(21), 214506.
65. Yang, Y. & Fergus, M. (2020) Pair distances with PBC, url = https://yangyushi.github.io/science/2020/11/02/pbc_py.html. .
66. Yirik, M. A., Colpan, K. E., Schmidt, S., Sorokina, M. & Steinbeck, C. (2021) Review on Chemical Graph Theory and Its Application in Computer-Assisted Structure Elucidation. *Preprints*, **2021110546**, (doi: 10.20944/preprints202111.0546.v1).

67. Zahran, M., Daidone, I., Smith, J. C. & Imhof, P. (2010) Mechanism of DNA recognition by the restriction enzyme EcoRV. *Journal of Molecular Biology*, **401**(3), 415–432.

68. Zanetti-Polzi, L., Amadei, A. & Daidone, I. (2021) Segregation on the nanoscale coupled to liquid water polyamorphism in supercooled aqueous ionic-liquid solution. *The Journal of Chemical Physics*, **155**(10), 104502.

69. Zanetti-Polzi, L., Biswas, A. D., Del Galdo, S., Barone, V. & Daidone, I. (2019) Hydration Shell of Antifreeze Proteins: Unveiling the Role of Non-Ice-Binding Surfaces. *The Journal of Physical Chemistry B*, **123**, 6474–6480.