Density forms of liquid water revealed by the Total Communicability of the corresponding graph

Chiara Faccio, Scuola Normale Superiore, Pisa

SCUOLA
NORMALE
SUPERIORE

Two Days of Numerical Linear Algebra and Applications
Napoli, February 14-15, 2022

Outline

1. Motivation
2. Basic notions of graph theory
3. Centrality measures
4. Construction of the adjacency matrix
5. Results
6. Conclusions

This work is based on the paper *Low- and high-density forms of liquid water revealed by a new medium-range order descriptor*, submitted to Journal of Molecular Liquids (2021), joint work with Michele Benzi[1], Isabella Daidone[2] and Laura Zanetti-Polzi[3].

[1] Scuola Normale Superiore

[2] Department of Physical and Chemical Sciences, University of L'Aquila

[3] CNR Institute of Nanoscience

Outline

1. Motivation
2. Basic notions of graph theory
3. Centrality measures
4. Construction of the adjacency matrix
5. Results
6. Conclusions

Water

Water is a complex liquid with anomalous properties, for example:

- liquid water is denser than solid water (ice);
- very high specific heat;
- water expands instead of contracting when it cools.

One of the most popular hypotheses for explaining many of the 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).

In our work, we present a new order parameter based on graph theory, in particular on the total communicability of the corresponding graph, to identify these two density forms. Our parameter can also show that HDL forms are not homogeneous but composed of regions at different local densities.

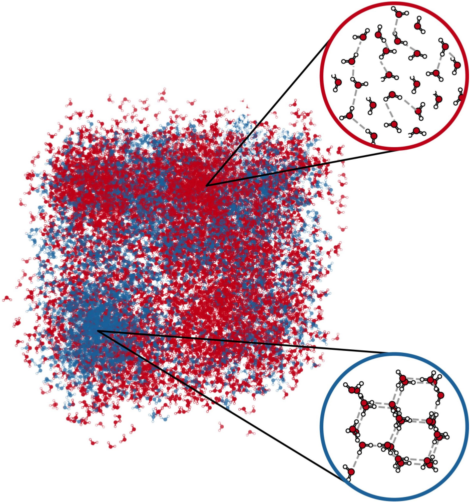
Water

Water is a complex liquid with anomalous properties, for example:

- liquid water is denser than solid water (ice);
- very high specific heat;
- water expands instead of contracting when it cools.

One of the most popular hypotheses for explaining many of the 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).

In our work, we present a new order parameter based on graph theory, in particular on the total communicability of the corresponding graph, to identify these two density forms. Our parameter can also show that HDL forms are not homogeneous but composed of regions at different local densities.



[4] HDL
Not well structurally
characterized

LDL
More ordered,
tetrahedral geometry

[4] Image of Laura Zanetti-Polzi

Outline

1. Motivation
2. **Basic notions of graph theory**
3. Centrality measures
4. Construction of the adjacency matrix
5. Results
6. Conclusions

Basic definitions

A graph $G = (V, E)$ consists of a set of nodes $V = \{v_1, ..., v_n\}$ and a set of edges $E \subseteq V \times V$. Edges of the form (v_i, v_i) are usually ignored.

G is called a *directed* graph (or a digraph) if the edges have orientations: $(v_i, v_j) \in E$ means that there exists an edge from node v_i to node v_j. G is an *undirected* graph if $(v_i, v_j) \in E \iff (v_j, v_i) \in E$.

A *weighted* graph is a graph in which a positive number (the weight) is assigned to each edge. Otherwise, it is called *unweighted*.

For an unweighted graph, the *adjacency matrix* is a matrix $A \in \mathbb{R}^{n \times n}$ such that $A_{ij} = 1$ if $(v_i, v_j) \in E$, 0 otherwise .

A graph G is *strongly connected* if any node can be reached from any other node by following edges along their directions. In this case the adjacency matrix is irreducible.

In our work, we consider undirected and unweighted graphs.

Outline

1. Motivation
2. Basic notions of graph theory
3. Centrality measures
4. Construction of the adjacency matrix
5. Results
6. Conclusions

Centrality measures

Given a graph G, the centrality of a node is a quantity which measures the importance of that node. Formally, a centrality measure is a function $f : V \to [0, \infty)$ used for ranking the nodes in the network.

There are different centrality measures, which capture different properties of the graph. For example:

- measures based on the distance from the other nodes;
- measures based on shortest paths passing through the node;
- measures based on the relative importance of the neighbours of the node.

The *degree centrality* is the simplest centrality measure and it is defined as the number of links incident upon a node: $deg(v_i) := (\mathbb{1}^T A)_i = (A\mathbb{1})_i$, where $\mathbb{1}$ is the vector of all ones. The degree centrality does not work well in predicting whether information reaches a certain node or in identifying nodes that act as a link between two clusters of nodes. It is a purely local notion.

Centrality measures

Given a graph G, the centrality of a node is a quantity which measures the importance of that node. Formally, a centrality measure is a function $f : V \to [0, \infty)$ used for ranking the nodes in the network.

There are different centrality measures, which capture different properties of the graph. For example:

- measures based on the distance from the other nodes;
- measures based on shortest paths passing through the node;
- measures based on the relative importance of the neighbours of the node.

The *degree centrality* is the simplest centrality measure and it is defined as the number of links incident upon a node: $deg(v_i) := (\mathbb{1}^T A)_i = (A\mathbb{1})_i$, where $\mathbb{1}$ is the vector of all ones. The degree centrality does not work well in predicting whether information reaches a certain node or in identifying nodes that act as a link between two clusters of nodes. It is a purely local notion.

Centrality measures

Given a graph G, the centrality of a node is a quantity which measures the importance of that node. Formally, a centrality measure is a function $f : V \to [0, \infty)$ used for ranking the nodes in the network.

There are different centrality measures, which capture different properties of the graph. For example:

- measures based on the distance from the other nodes;
- measures based on shortest paths passing through the node;
- measures based on the relative importance of the neighbours of the node.

The *degree centrality* is the simplest centrality measure and it is defined as the number of links incident upon a node:
$deg(v_i) := (\mathbb{1}^T A)_i = (A\mathbb{1})_i$, where $\mathbb{1}$ is the vector of all ones.
The degree centrality does not work well in predicting whether information reaches a certain node or in identifying nodes that act as a link between two clusters of nodes. It is a purely local notion.

Eigenvector centrality

If we assume that the graph is strongly connected, then the matrix A is irreducible. Since $A \geq 0$, by the Perron-Frobenius Theorem, there exists a unique vector $\boldsymbol{p} > \mathbf{0}$ such that $A\boldsymbol{p} = \rho(A)\boldsymbol{p}$ ($\rho(A)$ is the spectral radius of A). The *eigenvector centrality* is defined as

$$EC(v_i) = p_i$$

Walk interpretation of EC:

$$p_i = \lim_{k\to\infty} \frac{\#\{\text{walks of length k through } v_i\}}{\#\{\text{walks of length k in G}\}}$$

It takes into account how well connected a node is and how many links their connections have and so on through the network. It identifies nodes with influence over the whole network, not just those directly connected to it.

Eigenvector centrality

If we assume that the graph is strongly connected, then the matrix A is irreducible. Since $A \geq 0$, by the Perron-Frobenius Theorem, there exists a unique vector $\boldsymbol{p} > \mathbf{0}$ such that $A\boldsymbol{p} = \rho(A)\boldsymbol{p}$ ($\rho(A)$ is the spectral radius of A). The *eigenvector centrality* is defined as

$$EC(v_i) = p_i$$

Walk interpretation of EC:

$$p_i = \lim_{k\to\infty} \frac{\#\{\text{walks of length k through } v_i\}}{\#\{\text{walks of length k in G}\}}$$

It takes into account how well connected a node is and how many links their connections have and so on through the network. It identifies nodes with influence over the whole network, not just those directly connected to it.

Total communicability

Let $\beta > 0$, then the *total communicability* (Benzi & Klymko, 2013) is defined as:

$$TC(v_i) = [e^{\beta A}\mathbb{1}]_i = \sum_{k=0}^{\infty} \frac{\beta^k}{k!}[A^k\mathbb{1}]_i = 1 + \beta[A\mathbb{1}]_i + \frac{\beta^2}{2}[A^2\mathbb{1}]_i + \ldots$$

We recall that $[A^k]_{ij}$ is the number of walks of length k between nodes v_i and v_j.

The TC gives a measure of how well each node communicates with the other nodes of the network. The default value of β is 1.

Theorem: [Benzi, Klymko (2015)]

Let $G = (V, E)$ be a strongly connected and undirected graph. Then:

- for $\beta \to 0^+$, the total communicability rankings reduce to degree centrality rankings;
- for $\beta \to +\infty$, the total communicability rankings reduce to eigenvector centrality rankings;

Distinct advantages of the Total Communicability are:

- it does not require the graph to be strongly-connected;
- it can be computed efficiently using algorithms for evaluating the action of a matrix function on a vector, that is, for computing the vector $f(A)\boldsymbol{v}$ for a matrix A (usually large and sparse). In our case $f(A) = e^{\beta A}$ and $\boldsymbol{v} = \mathbb{1}$.
- the TC contains much information about the network's structure, especially if the spectral gap is small. In detail, if A is symmetric and real-valued, it can be decomposed into $A = \sum_{k=1}^{n} \lambda_k \boldsymbol{p}_k \boldsymbol{p}_k^T$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ are the eigenvalues and $\boldsymbol{p}_k$ is the eigenvector associated with λ_k. Note that the TC takes into account all the terms of the expansion:

$$TC(v_i) = e^{\beta\lambda_1}(\boldsymbol{p}_1^T \mathbb{1}) p_1(v_i) + \sum_{k=2}^{n} e^{\beta\lambda_k}(\boldsymbol{p}_k^T \mathbb{1}) p_k(v_i),$$

for all $v_i \in V$.

Distinct advantages of the Total Communicability are:

- it does not require the graph to be strongly-connected;
- it can be computed efficiently using algorithms for evaluating the action of a matrix function on a vector, that is, for computing the vector $f(A)\boldsymbol{v}$ for a matrix A (usually large and sparse). In our case $f(A) = e^{\beta A}$ and $\boldsymbol{v} = \mathbb{1}$.
- the TC contains much information about the network's structure, especially if the spectral gap is small. In detail, if A is symmetric and real-valued, it can be decomposed into $A = \sum_{k=1}^{n} \lambda_k \boldsymbol{p}_k \boldsymbol{p}_k^T$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ are the eigenvalues and $\boldsymbol{p}_k$ is the eigenvector associated with λ_k. Note that the TC takes into account all the terms of the expansion:

$$TC(v_i) = e^{\beta\lambda_1}(\boldsymbol{p}_1^T \mathbb{1})p_1(v_i) + \sum_{k=2}^{n} e^{\beta\lambda_k}(\boldsymbol{p}_k^T \mathbb{1})p_k(v_i),$$

for all $v_i \in V$.

Distinct advantages of the Total Communicability are:

- it does not require the graph to be strongly-connected;
- it can be computed efficiently using algorithms for evaluating the action of a matrix function on a vector, that is, for computing the vector $f(A)\boldsymbol{v}$ for a matrix A (usually large and sparse). In our case $f(A) = e^{\beta A}$ and $\boldsymbol{v} = \mathbb{1}$.
- the TC contains much information about the network's structure, especially if the spectral gap is small. In detail, if A is symmetric and real-valued, it can be decomposed into $A = \sum_{k=1}^{n} \lambda_k \boldsymbol{p}_k \boldsymbol{p}_k^T$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ are the eigenvalues and $\boldsymbol{p}_k$ is the eigenvector associated with λ_k. Note that the TC takes into account all the terms of the expansion:

$$TC(v_i) = e^{\beta\lambda_1}(\boldsymbol{p}_1^T \mathbb{1}) p_1(v_i) + \sum_{k=2}^{n} e^{\beta\lambda_k}(\boldsymbol{p}_k^T \mathbb{1}) p_k(v_i),$$

for all $v_i \in V$.

Distinct advantages of the Total Communicability are:

- it does not require the graph to be strongly-connected;
- it can be computed efficiently using algorithms for evaluating the action of a matrix function on a vector, that is, for computing the vector $f(A)\boldsymbol{v}$ for a matrix A (usually large and sparse). In our case $f(A) = e^{\beta A}$ and $\boldsymbol{v} = \mathbb{1}$.
- the TC contains much information about the network's structure, especially if the spectral gap is small. In detail, if A is symmetric and real-valued, it can be decomposed into $A = \sum_{k=1}^{n} \lambda_k \boldsymbol{p}_k \boldsymbol{p}_k^T$, where $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n$ are the eigenvalues and $\boldsymbol{p}_k$ is the eigenvector associated with λ_k. Note that the TC takes into account all the terms of the expansion:

$$TC(v_i) = e^{\beta\lambda_1}(\boldsymbol{p}_1^T \mathbb{1})p_1(v_i) + \sum_{k=2}^{n} e^{\beta\lambda_k}(\boldsymbol{p}_k^T \mathbb{1})p_k(v_i),$$

for all $v_i \in V$.

If G is strongly connected, $\lambda_1 > \lambda_2$, $(\boldsymbol{p}_1^T \mathbb{1}) > 0$ (by the Perron-Frobenius Theorem) and we can divide both sides by $e^{\beta\lambda_1}(\boldsymbol{p}_1^T \mathbb{1})$:

$$\frac{TC(v_i)}{e^{\beta\lambda_1}(\boldsymbol{p}_1^T \mathbb{1})} = p_1(v_i) + \sum_{k=2}^{n} e^{\beta(\lambda_k - \lambda_1)} \frac{(\boldsymbol{p}_k^T \mathbb{1}) p_k(v_i)}{(\boldsymbol{p}_1^T \mathbb{1})}.$$

Taking the limit as $\beta \to \infty$, the left-hand side $TC(v_i)$ converges to $EC(v_i)$, for all $v_i \in V$. If the spectral gap $(\lambda_2 - \lambda_1)$ is large this convergence is very fast, but if the gap is tiny the convergence will be slow. In this latter case, ignoring the contributions of the eigenvectors $\boldsymbol{p}_k$ for $k \geq 2$ implies a great loss of information, at least for β not too large.

In other words, using only the first term in the expansion $A = \sum_{k=1}^{n} \lambda_k \boldsymbol{p}_k \boldsymbol{p}_k^T$ (Eigenvector centrality) yields a poor approximation when the gap is very small.

Outline

1 Motivation

2 Basic notions of graph theory

3 Centrality measures

4 Construction of the adjacency matrix

5 Results

6 Conclusions

Construction of the adjacency matrix

In this work we aim to analyze water molecules in a box. We build a graph $G = (V, E)$, where each water molecule corresponds to a vertex v_i, and the bonds between two water molecules are the edges e_{ij} of the graph. Two molecules are bonded when the distance between oxygen atoms is ≤ 0.35 nm.

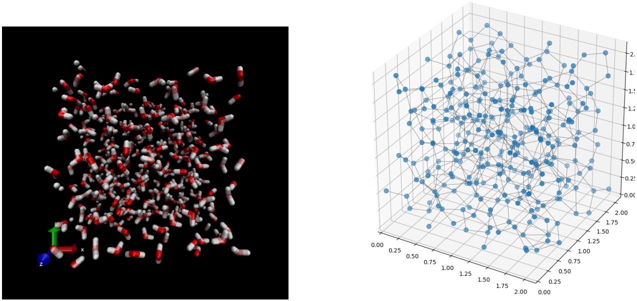


Figure: Box with 300 molecules of water

Outline

1. Motivation
2. Basic notions of graph theory
3. Centrality measures
4. Construction of the adjacency matrix
5. **Results**
6. Conclusions

Results

We analyze four temperatures along the 1950 bars isobar that crosses the liquid-liquid phase transition: 170 K (LDL phase), 180 K (HDL phase, just above the coexistence line), 200 and 240 K (HDL phase). At each temperature we analyze 10ns of the corresponding MD simulation sampled every 100 ps. For each network we then compute the centrality measures.

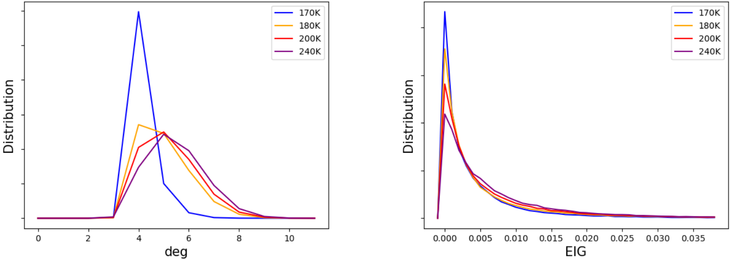


Figure: Distribution of the Degree and the Eigenvector centrality

The average spectral gap $|\lambda_2 - \lambda_1|$ is equal to 10^{-4}.

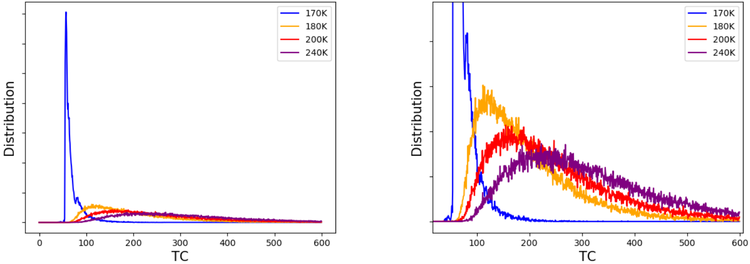


Figure: Distribution of the TC (left). The right plot highlights the TC distributions in the HDL phase.

We define two regions:

- if $TC(v_i) \leq 95$, the molecule v_i is assigned to the LDL phase
- if $TC(v_i) > 95$, the molecule v_i is assigned to the HDL phase

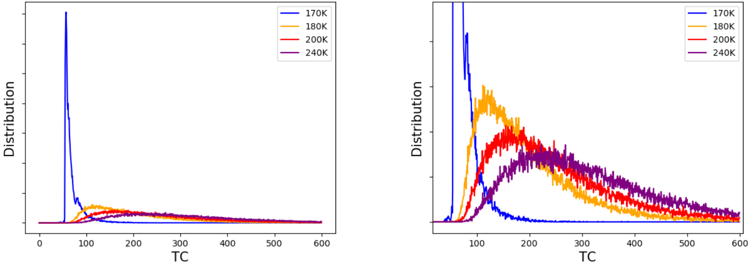


Figure: Distribution of the TC (left). The right plot highlights the TC distributions in the HDL phase.

We define two regions:

- if $TC(v_i) \leq 95$, the molecule v_i is assigned to the LDL phase
- if $TC(v_i) > 95$, the molecule v_i is assigned to the HDL phase

A number of parameters are commonly used to assign water molecules to the LDL or HDL phase along a MD simulation:

- the *local structure index* I. The set of radial oxygen-oxygen distances r_j, corresponding to the $n(i)$ neighbouring molecules that are within a cutoff distance of 0.37 nm, are ordered: $r_1 < ... < r_j < ... < r_{n(i)} < 0.37 < r_{n(i)+1}$. Then

$$I(i) = \frac{1}{n(i)} \sum_{j=1}^{n(i)} (\Delta(j;i) - \bar{\Delta}(i))^2$$

  where $\Delta(j;i) = r_{j+1} - r_j$ and $\bar{\Delta}(i)$ is the average of $\Delta(j;i)$;
- d_5 parameter is a very simple order parameter and it is based on the distance to the fifth nearest neighbor

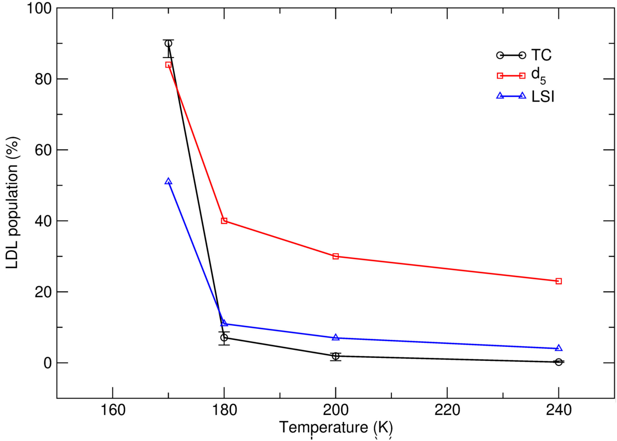


Figure: LDL population as a function of the temperature as obtained by defining the LDL phase according to the TC values (black circles), the d_5 parameter (red squares) and the local structure index (blue triangles).

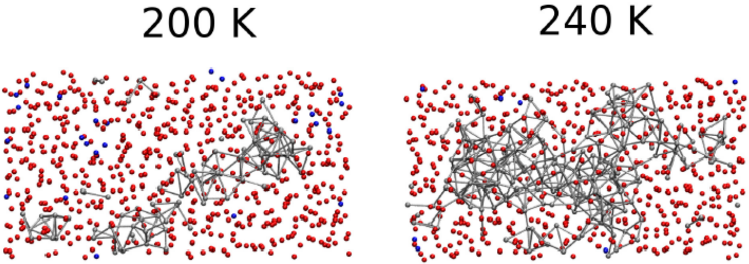


Figure: Representative snapshots of the arrangement of the patches of high-TC molecules at 200 K and 240 K. Blue nodes represent LDL molecules, red nodes HDL molecules, silver nodes molecules with high-TC value, silver edges highlight the connections among these nodes.

Outline

1. Motivation
2. Basic notions of graph theory
3. Centrality measures
4. Construction of the adjacency matrix
5. Results
6. Conclusions

Conclusions

- All the above-mentioned order parameters, including the TC, are based on the use of cutoff values to discriminate the two liquid phases. This intrinsic limitation, together with the fact that no robust benchmark is currently available for the HDL/LDL fraction in the supercooled region, makes it difficult to assess the reliability of the results obtained with one or another order parameter.

  Nonetheless, the present data suggest that the TC performs very well in distinguishing between the LDL and HDL phases.
- Since the TC is a descriptor of the structural properties at the molecular level, it is also able to identify patches at very high local density.

Conclusions

- All the above-mentioned order parameters, including the TC, are based on the use of cutoff values to discriminate the two liquid phases. This intrinsic limitation, together with the fact that no robust benchmark is currently available for the HDL/LDL fraction in the supercooled region, makes it difficult to assess the reliability of the results obtained with one or another order parameter.

  Nonetheless, the present data suggest that the TC performs very well in distinguishing between the LDL and HDL phases.
- Since the TC is a descriptor of the structural properties at the molecular level, it is also able to identify patches at very high local density.

Conclusions

- All the above-mentioned order parameters, including the TC, are based on the use of cutoff values to discriminate the two liquid phases. This intrinsic limitation, together with the fact that no robust benchmark is currently available for the HDL/LDL fraction in the supercooled region, makes it difficult to assess the reliability of the results obtained with one or another order parameter.

  Nonetheless, the present data suggest that the TC performs very well in distinguishing between the LDL and HDL phases.
- Since the TC is a descriptor of the structural properties at the molecular level, it is also able to identify patches at very high local density.

Future work

- to improve the structural descriptor of the liquid water as a directed graph in which the connections will be defined to take into account hydrogen bonds among molecules;
- to investigate additional centrality measures for the structural characterization of liquids;
- to use other global properties of the graph, for example, the Total Network Communicability, the Bipartivity measure, the Algebraic Connectivity, and some global metrics, such as the total number of triangles, the Watts-Strogatz clustering coefficient, the transitivity coefficient, etcetera.

Bibliography

- M. Benzi and C. Klymko (2013) *"Total communicability as a centrality measure"*, Journal of Complex Networks, Volume 1, Issue 2, Pages 124–149
- M. Benzi and C. Klymko (2015) *"On the limiting behavior of parameter-dependent network centrality measures"*, SIAM J. Matrix Anal. Appl., 36, pp. 686–70
- E. Estrada and P. A. Knight (2015) *"A First Course in Network Theory"*, Oxford University Press, United Kingdom, ISBN 10 978-0-19-872645-6
- C. Faccio, M. Benzi, L. Zanetti-Polzi and I. Daidone (2021) *"Low, high and very-high density forms of liquid water revealed by a medium-range order descriptor"*, arXiv preprint arXiv:2110.13747

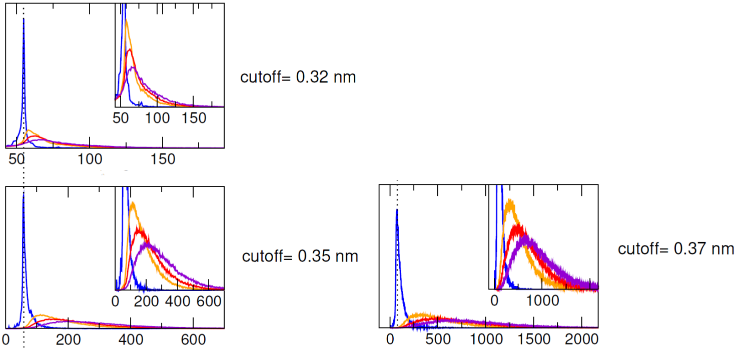


Figure: Distribution of the Total Communicability as obtained from the simulations at 1950 bars and 170 K (blue), 180 K (orange), 200 K(red) and 240 K (violet) with cutoffs 0.32 nm (top), 0.35 nm (middle) and 0.37 nm (bottom). It can be observed that the overall trend remains unaltered. Nonetheless, with the 0.32 nm cutoff the distributions at the three highest temperatures are slightly more overlapped. There is no relevant difference between the results obtained with cutoffs 0.35 and 0.37 nm.

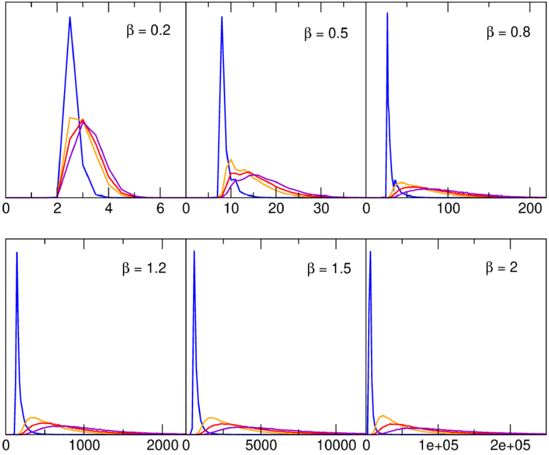


Figure: Distribution of the Total Communicability as obtained from the simulations at 1950 bars and 170 K (blue), 180 K (orange), 200 K(red) and 240 K (violet) for different values of the parameter β.

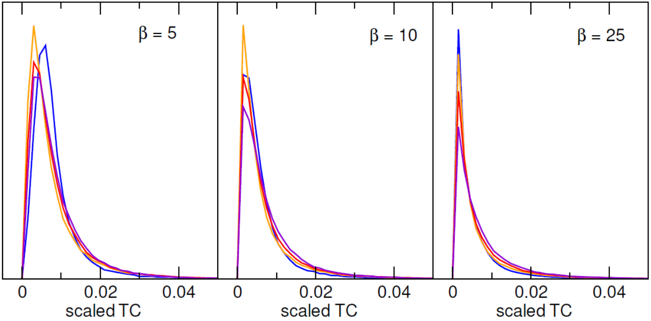


Figure: Distribution of the scaled Total Communicability as obtained from the simulations at 1950 bars and 170 K (blue), 180 K (orange), 200 K(red) and 240 K (violet) for different values of the parameter β. To compare the distributions with the corresponding distributions of the Eigenvector centrality, the TC values have been divided by $e^{\beta\lambda_1}(\boldsymbol{p}_1^T\mathbb{1})$ where λ_1 is the largest eigenvalue of the adjacency matrix A, $\boldsymbol{p}_1$ is the associated eigenvector with λ_1 ($\boldsymbol{p}_1 > 0$ by the Perron-Frobenius Therorem) and $\mathbb{1}$ is the vector of all ones.

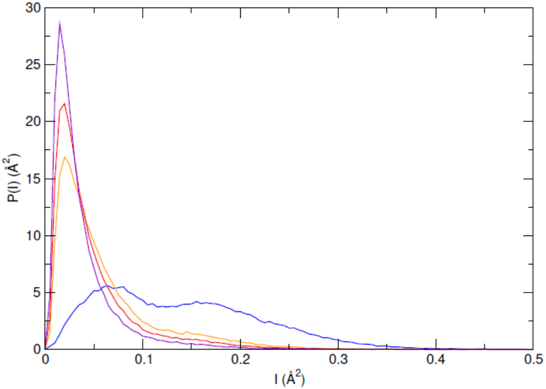


Figure: Distribution of the local structure index, LSI, as obtained from the MD simulations at 1950 bars and 170 K (blue), 180 K (orange), 200 K (red), 240 K (violet).

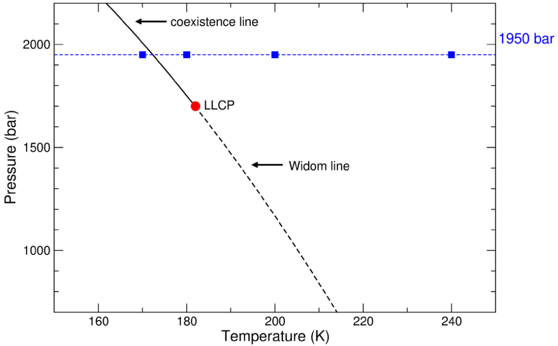


Figure: Schematic representation of the phase diagram for the TIP4P/2005 water model. The estimate of the liquid-liquid critical point (LLCP) is reported as a red circle. The temperature/pressure conditions of the present work are shown as blue squares.

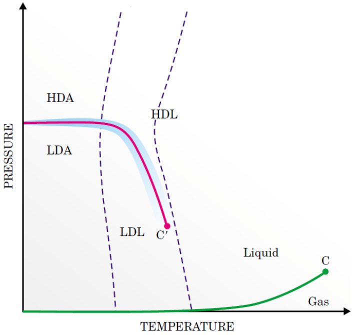


Figure: In the liquid–liquid phase transition hypothesis, a first-order phase transition (red line) occurs between two distinct forms of supercooled liquid water: low-density liquid (LDL) and high-density liquid (HDL). The first-order transition terminates at a critical point, C'. (C is the liquid–gas critical point; the liquid–gas coexistence curve is shown in green.) Because of fast crystallization, it is extremely challenging to observe experimentally. Image from "P. G. Debenedetti, *Supercooled and glassy water*, J. Condens. Matter Phys. 15 (45) (2003) R1669.".