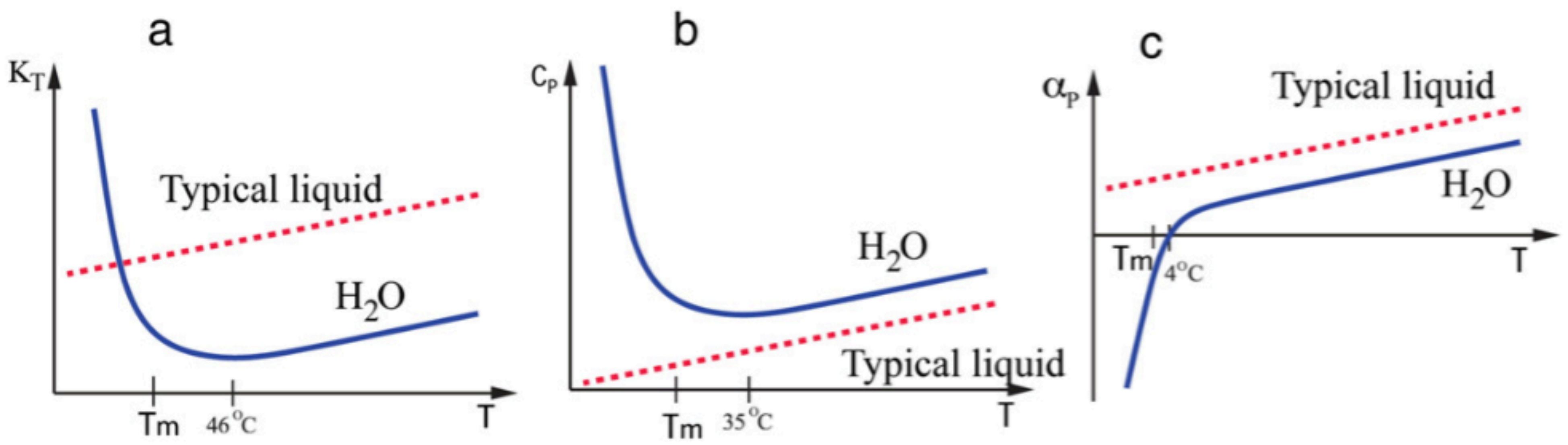
Searching for two forms of water: machine learning local molecular order

Adrián Soto[1] [2], Deyu Lu[3], Shinjae Yoo[2] [4] and M.-V. Fernández-Serra[1] [2]

1. Department of Physics and Astronomy, SUNY Stony Brook, NY, USA
2. Institute for Advanced Computational Science, SUNY Stony Brook, NY, USA
3. Center for Functional Nanomaterials, Brookhaven National Laboratory, NY, USA
4. Computational Science Initiative, Brookhaven National Laboratory, NY, USA

The anomalous behavior of liquid water

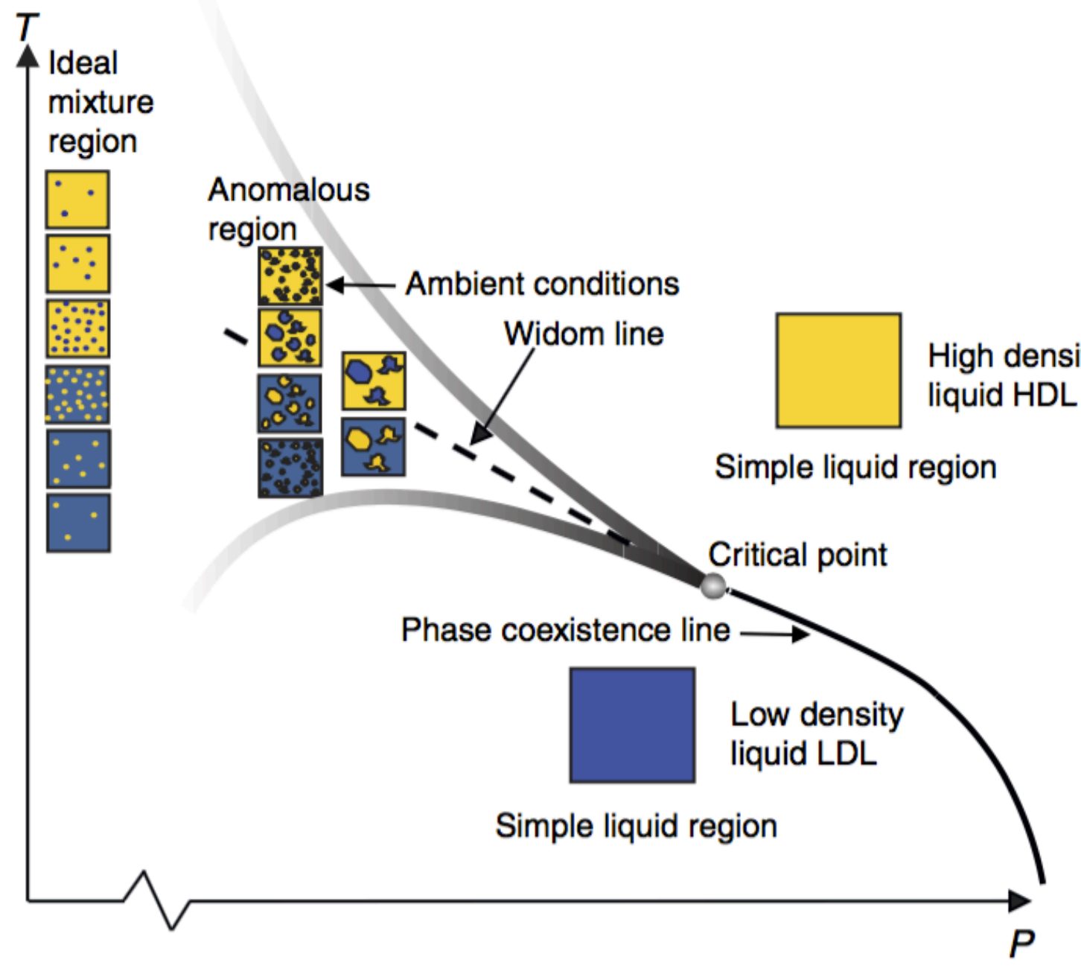


Schematic adaptation from experimental data. Source: *Pettersson et al., Journal of Non-Crystalline Solids 407 (2015) 399-417*

The isothermal compressibility (a), isobaric heat capacity (b) and thermal expansion coefficient (c) of water have minima (a,b) and vanish (c), as opposed to a typical liquid. These thermodynamic response functions are a consequence of the fluctuations in density and entropy.

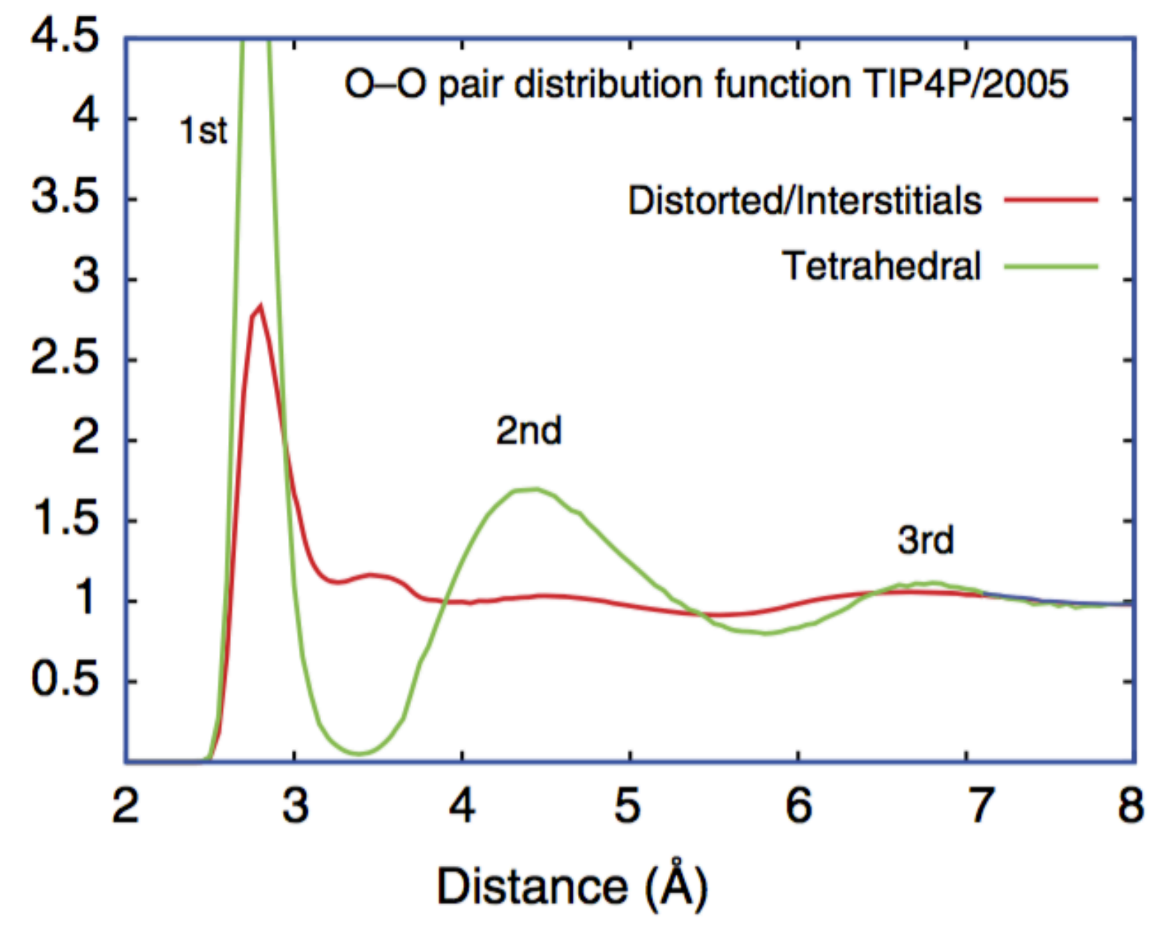
To understand the origin of these anomalies, the local geometrical properties of the water molecules need to be understood in detail at various temperatures and pressures, as well as its connection with the dynamics of the molecular motion. A great deal of effort is being put into performing experiments and more recently computer simulations address this question.

Proposed explanation: a two-component liquid

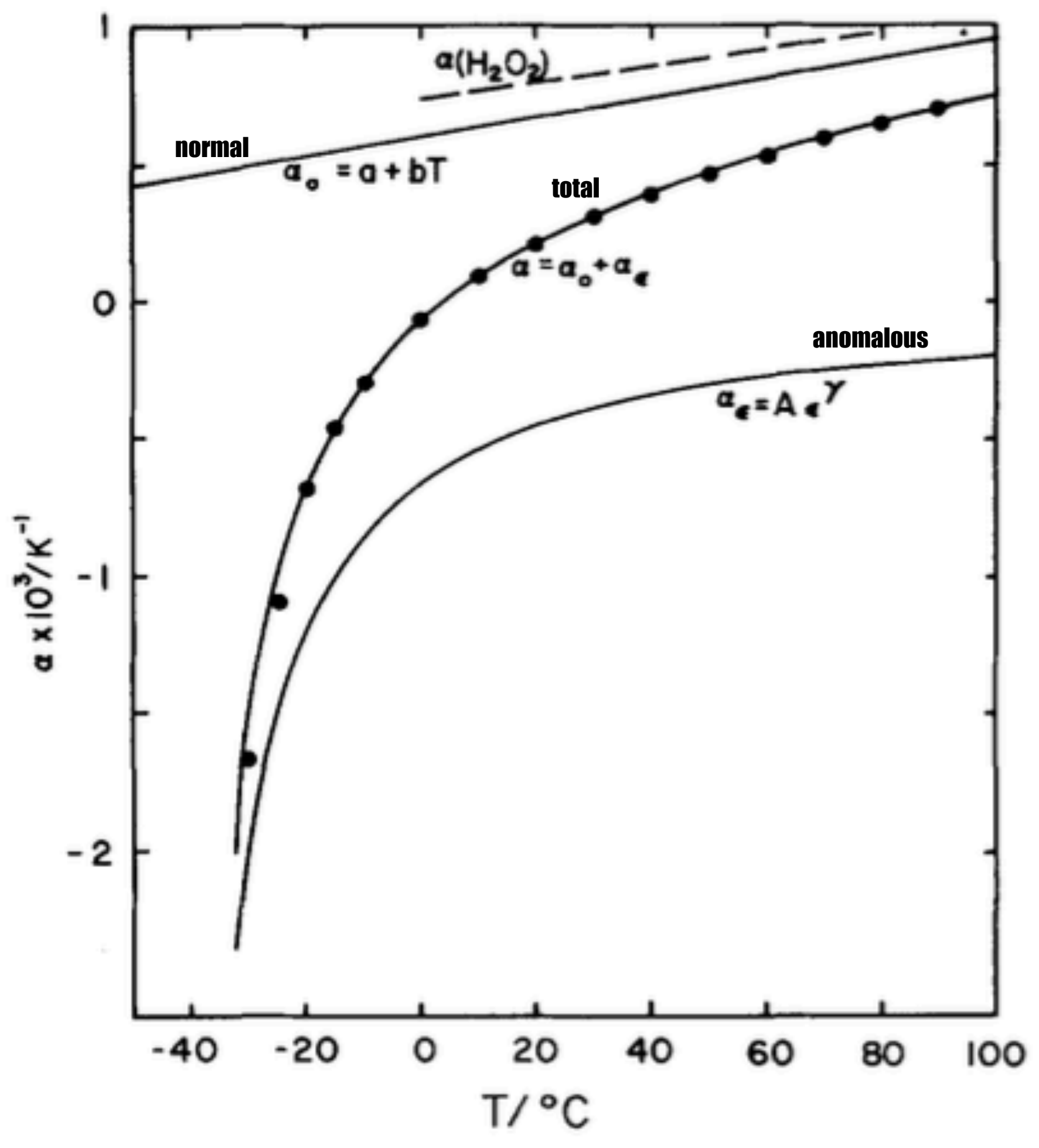


Nilsson et al., Nature Comm. 6, 8998 (2015)

A liquid-liquid phase transition has been proposed to exist at elevated pressures and supercooled temperatures. These two phases are called *high density liquid* (HDL) and a *low density liquid* (LDL). The phase separation line ends at a critical point, past which the phase transition is not well defined. However, patches of the two states form and and fluctuate, with time and length scales that depend on the pressure and the temperature.



LSI separated radial distribution functions from molecular dynamics simulation at ambient conditions.
Red: I < 0.05Å^2; Green: I > 0.05Å^2.
Nilsson et al., Nature Comm. 6, 8998 (2015)



Experimental thermal expansion coefficient of liquid water separated into a normal (α_0) and anomalous (α_ε) components.
Speedy et al., J. Chem. Phys., 65, 851 (1976)

Local order parameters

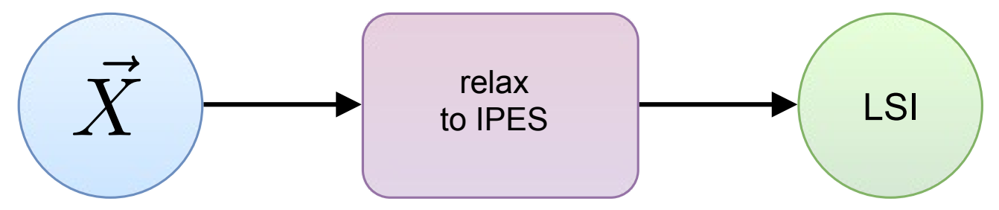
Phase transitions are characterized by an "order parameter", which changes value, as the thermodynamic variables cross the phase transition line. Near the phase transition a bimodal probability distribution is expected since the system fluctuates locally between the two phases.

In the last 3 decades a variety of parameters describing the local environment of water molecules have been proposed, among which there are: orientational tetrahedral order, q, translational tetrahedral order, Sk, local structure index (LSI), I, inverse Voronoi volume, ρ, Voronoi asphericity, η, and others. Yet none of them succeeds at finding the expected bimodal behavior.

Bimodality of the LSI at the IPES

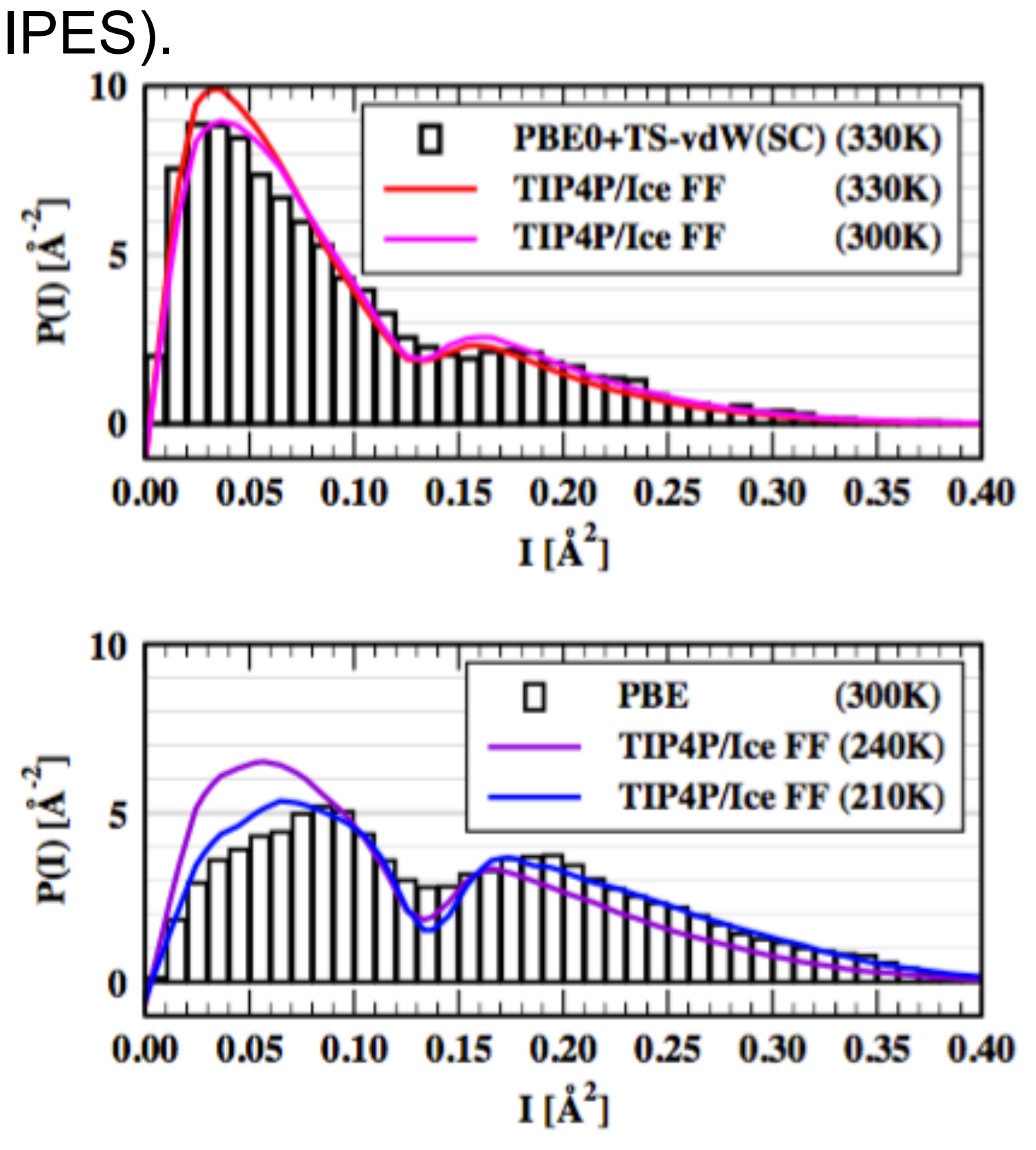
In a molecular dynamics simulation at finite temperature the atomic coordinates X can be allowed to relax to a nearby minimum of the inherent potential energy surface (IPES).

This quenching procedure effectively eliminates the thermal fluctuations. In recent studies it has been found that the local structure index shows a bimodal behavior at the IPES.



Two main drawbacks:

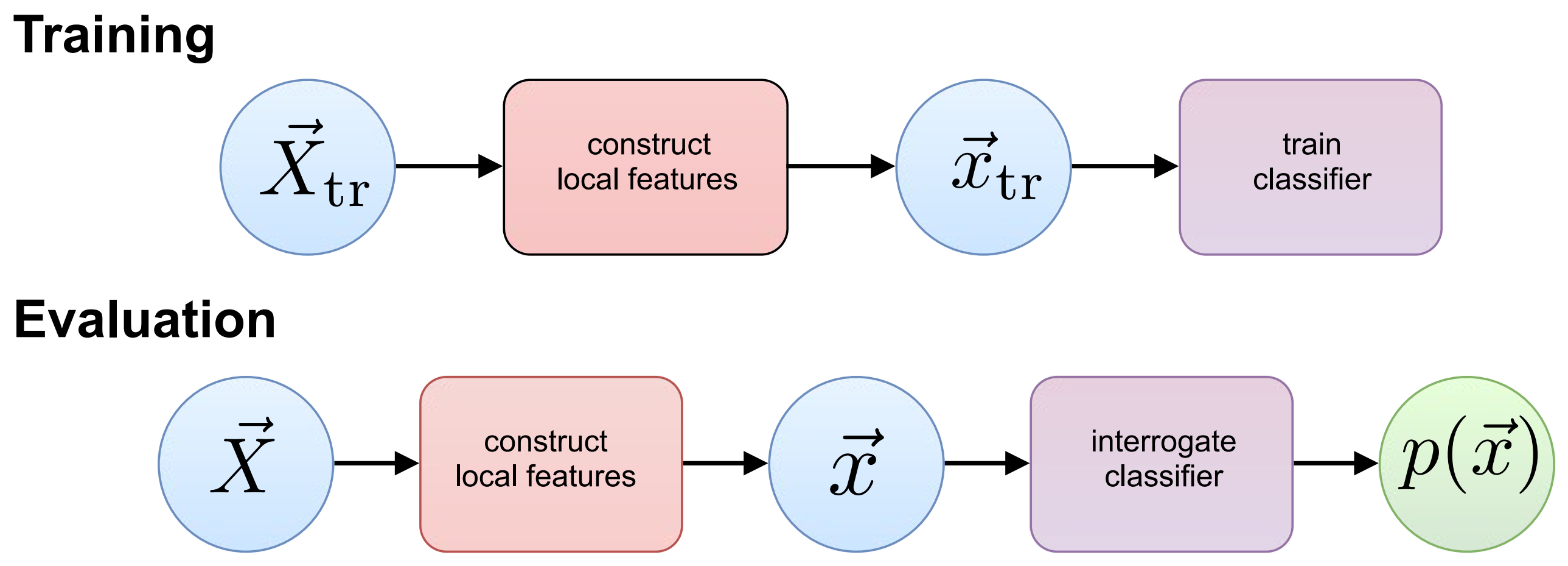
1. quenching procedure may not truthfully represent the thermodynamic state
2. requires a computationally expensive atomic coordinate relaxation; no on-the-fly evaluation



Histograms of the local structure index at the IPES
Santra et al., arXiV:1503.00020v1 (2015)

Machine Learning Local Order (MLLO)

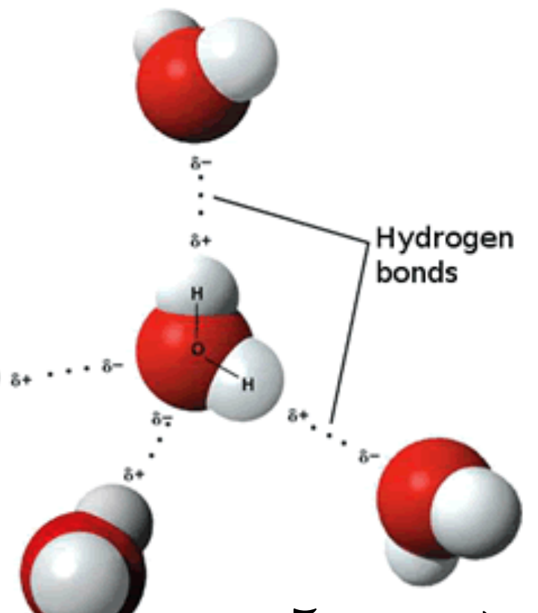
Construct feature vectors that contain the relevant information to describe the local environment of each molecule. These could be geometrical quantities (distances, angles, volumes, etc.) or physical quantities (dipole moments, local energies, etc.)



Three notable advantages of the MLLS method:

1. no quenching to IPES required
2. once the classifier is trained, it can be used in any data set
3. can be used on-the-fly

Feature vectors for H_2O



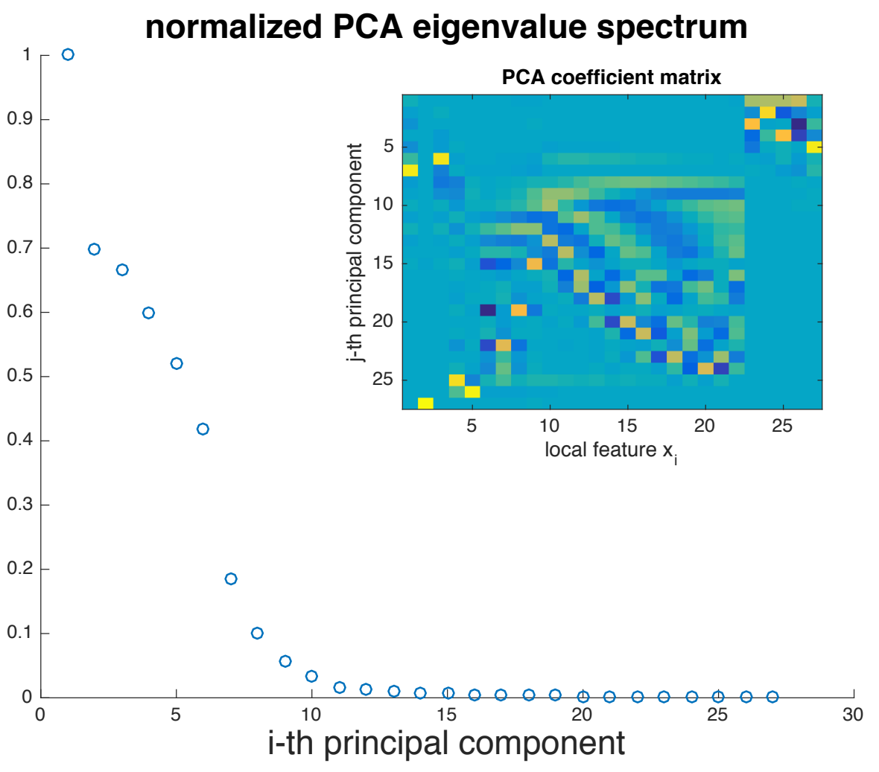
We carry out our studies on molecular dynamics simulation data, for which we have the atomic positions at many time steps. From those positions we compute local the environment feature vectors

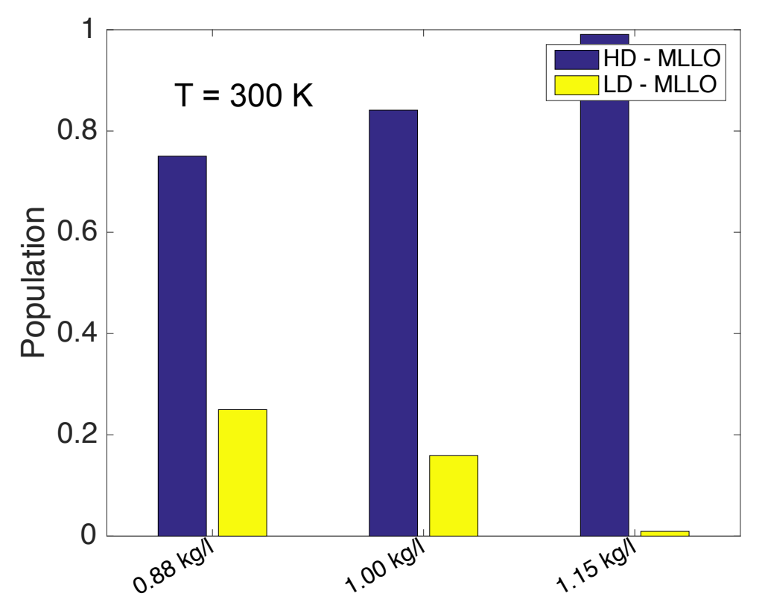
$$\vec{x} = (q, S_k, I, \rho, \eta, r^1_{OO}, \cdots, r^{17}_{OO}, \alpha^1_{OHO}, \cdots, \alpha^5_{OHO})$$

Including 5 of the previously proposed order parameters, the 17 smallest intermolecular distances (1st and 2nd H-bond coordination shells and one molecule beyond) and 5 O-H--O bonding angles (1st H-bond coordination shell and one molecule beyond).
The computational cost for feature evaluation is negligible compared to the computation of the energy at one simulation step.

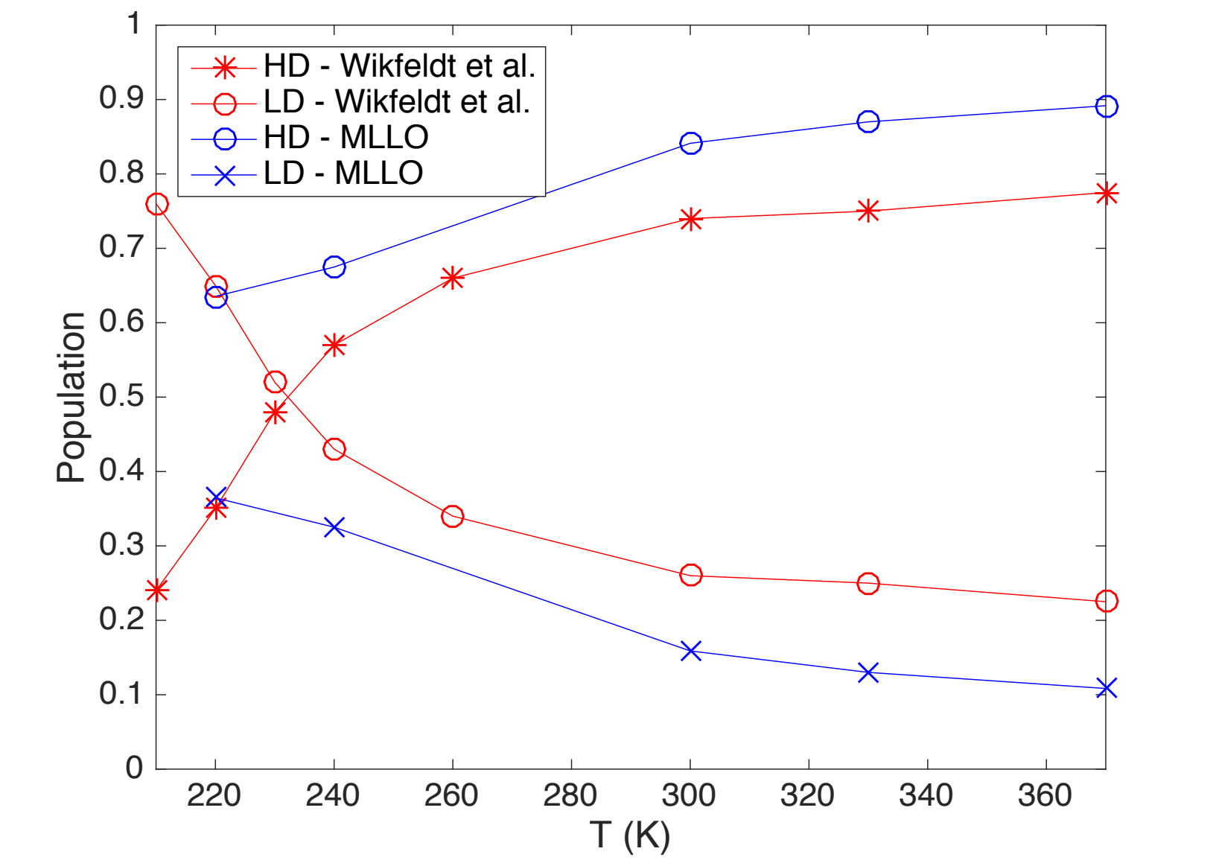
Description of the analysis and (preliminary) results

- Choose training set: random draw of N=25,000 subsamples from a MD trajectory at 300K and 1.0 kg/l with the TTM3F force field. The number of molecules in the simulation cell is 128.
- Perform a principal component analysis (PCA) of the training data
- Split the liquid in HD and LD components by an LSI value of I_{thr}=0.05Å^2 as in Nilsson et al. (2015)
- Train a support vector machine (SVM)
- Evaluate HD and LD populations





Left: A principal component analysis (PCA) shows that at least 10 principal dimensions are needed. The subspace corresponding to the 5 O-H—O angles contains most of the variance.
Right: The MLLO SVM classifier is able to correctly predict the trends of the populations of HD and LD components as the density of the liquid changes.



MLLO classification as a function of temperature. We see that the trends match previous studies based on the IPES, but the HD component is overpopulated.

Comments and outlook

- MLLO captures the temperature variations in HD and LD population found on the IPES.
- Intermolecular angles play a very important role since they characterize the collapse of the 2nd shell of the H-bond network.
- The threshold value of of I_{trh}=0.05Å^2 used to define the HD and LD components is arbitrary.
- Unsupervised learning methods could provide an agnostic way to define the HD and LD data clusters.
- Refinement is needed, since the equal population temperature is underestimated. We observe a better match with Wikfeldt's data if I_{trh} is reduced to I_{trh}=0.04Å^2
- The MLLO classification could be improved by including additional local features.
- This method is general and can be used for other interesting materials (pure liquids, alloys, solutions, …)