An Automatic, Data-Driven Definition of Atomic-Scale Structural Motifs

THÈSE N^{O} 8412 (2018)

PRÉSENTÉE LE 23 MARS 2018

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE SCIENCE COMPUTATIONNELLE ET MODÉLISATION
PROGRAMME DOCTORAL EN SCIENCE ET GÉNIE DES MATÉRIAUX

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Piero GASPAROTTO

acceptée sur proposition du jury:

Prof. V. Michaud, présidente du jury
Prof. M. Ceriotti, directeur de thèse
Dr G. Tribello, rapporteur
Dr G. Pavan, rapporteur
Prof. B. Correia, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2018

O caos é uma ordem por decifrar.
Chaos is order yet undeciphered.

— José Saramago, "Livro dos contrários"

ABSTRACT

Structure-property relationships at the atomic scale are usually understood in terms of recurrent structural motifs formed by atoms and molecules, and how they transform and interact with each other. We introduce with this thesis a novel analysis approach, capable of determining such patterns automatically. This analysis provides a unique fingerprint for metastable motifs, that is based exclusively on structural information. The rational behind the method and its functioning will be presented, followed by a discussion regarding its application to a wide range of problems in materials science and biology. We will begin by showing how it is possible to use our methodology to define adaptively the hydrogen bond in some different systems, including water, ammonia and peptides. We will then demonstrate how such definition can be used to probe the topological defects in the 3-dimensional hydrogen bond network of liquid water and will propose a method to study the non-trivial correlations among them. Furthermore, we will apply our framework to the identification of coordination environments in nanoclusters, and to the recognition of secondary-structure patterns in oligopeptides and proteins. We will prove that it is not only possible to obtain an algorithmic definition, which is unbiased and adaptive, of local motifs of matter, but also to identify and classify structures in their entirety. We will also demonstrate that a clear interpretation of the stability of the system can be obtained through the automatic analysis of atomistic simulation results, and will discuss possible applications, such as the definition of collective variables for enhanced-sampling simulation techniques or the identification of recurrent patterns in complex systems that escape an interpretation in terms of conventional structural motifs, such as intrinsically disordered proteins.

KEYWORDS Molecular dynamics, enhanced sampling, collective variables, structural fingerprints, machine learning, unsupervised learning, Bayesian classifiers, kernel density estimation, pattern recognition, clustering, dimensionality reduction, PAMM, Sketchmap, hydrogen bond, water, secondary structures, oligopeptides

SOMMARIO

Le relazioni tra struttura e proprietà su scala atomica sono solitamente contestualizzate in termini di motivi strutturali ricorrenti e di come essi si trasformano e interagiscono tra di loro. Con questa tesi viene introdotto un nuovo approccio d'analisi, capace di determinare tali strutture in modo automatico. Tale analisi fornisce un'impronta digitale univoca per ognuna di esse, basata esclusivamente sull'informazione strutturale. Saranno presentati l'idea alla base del metodo e il suo funzionamento dettagliato, seguiti da una discussione riguardo la sua applicazione a una vasta gamma di problemi in scienza dei materiali e biologia. Inizieremo mostrando come sia possibile usare la nostra metodologia per definire adattivamente il legame a idrogeno in alcuni contesti diversi, come le proteine, l'ammoniaca e l'acqua. Dimostreremo poi come tale definizione possa essere usata per identificare i difetti topologici nella rete 3-dimensionale dei legami a idrogeno dell'acqua liquida e proporremo un metodo per studiare le correlazioni non banali tra essi. Inoltre, applicheremo il nostro metodo all'identificazione degli ambienti di coordinazione nei cluster nanoscopici e al riconoscimento delle strutture secondarie in oligopeptidi e proteine. Proveremo come sia possibile non solo ottenere una definizione algoritmica obbiettiva e adattiva dei motivi locali della materia, ma anche identificare e classificare strutture nella loro globalità. Dimostreremo anche che una chiara interpretazione della stabilità del sistema può essere ottenuta attraverso l'analisi automatica dei risultati delle simulazioni atomistiche e discuteremo possibili applicazioni future, come la definizione di variabili collettive per tecniche di simulazione a campionamento accelerato o l'identificazione di strutture ricorrenti in sistemi complessi ancora poco chiari, come le proteine intrinsicamente disordinate.

PAROLE CHIAVE Dinamica molecolare, campionamento accelerato, variabili collettive, impronte digitali strutturali, apprendimento automatico, apprendimento non supervisionato, classificatore bayesiano, stima kernel di densità, riconoscimento di pattern, clustering, riduzione di dimensionalità, PAMM, Sketchmap, legame a idrogeno, acqua, strutture secondarie, oligopeptidi.

CONTENTS

ACRONYMS

AIMD *Ab Initio* Molecular Dynamics, page 8

BO Born Oppenheimer, page 8

BPSF Behler-Parrinello Symmetry Function, page 132

CV Collective Variable, page 13

DFT Density Functional Theory, page 8

EM Expectation-Maximization, page 33

FF Force Field, page 7

GGA Generalised Gradient Approximation, page 10

GMM Gaussian Mixture Model, page 33

HB Hydrogen Bond, page 57

HF Hartree-Fock, page 8

KDE Kernel Density Estimation, page 28

KS Kohn-Sham, page 9

LDA Local Density Approximation, page 10

LJ Lennard-Jones, page 113

MC Monte Carlo, page 5

MD Molecular Dynamics, page 5

MDS Multi-Dimensional Scaling, page 36

MISE Mean Integrated Squared Error, page 30

ML Machine-Learning, page 19

NN Neural Network, page 132

NQE Generalized Langevin Equation, page 12

NQE Nuclear Quantum Effects, page 11

PAMM Probabilistic Analysis of Molecular Motifs, page 3

PCA Principal Component Analysis, page 36

PCCA Perron Cluster Analysis, page 27

PDB Protein Data Bank, page 133

PDF Probability Density Function, page 26

PES Potential Energy Surface, page 113

PI Path Integral, page 11

PIGLET Path Integral Generalized Langevin Equation Thermostat, page 12

PIMD Path Integral Molecular Dynamics, page 12

PMI Probabilistic Motif Identifiers, page 51

PT Parallel Tempering, page 16

RDF Radial Distribution Function, page 76

RE Replica Exchange, page 16

RMSD Root Mean Square Displacement, page 22

SL Supervised Laarning, page 19

SS Secondary Structure, page 124

TCF Time Correlation Function, page 5

USL Unsupervised Learning, page 20

VDW Van der Waals, page 86

WHAM Weighted-Histogram Analysis Method, page 14

WT Well-Tempered, page 15

WTE Well-Tempered Ensemble, page 17

XC Exchange-Correlation, page 9

1

MOTIVATION

Materials science is an interdisciplinary field of research that tries to shed light on the fundamental laws that govern materials. Materials can be seen as very complex systems stemming from the interactions of a large number of atoms and molecules. Understanding how such building blocks arrange and transform over various time and space scales to give rise to the macroscopic properties we can observe, is of primary importance when explaining the behavior of existing materials and designing new ones.

Since the very beginning of the field, the rationalization of structure-property relationships has been driven by the interplay of two main pillars: theory and experiment. Theory identifies the different physicochemical mechanisms governing materials from the ground up, while experiments validate the theoretical hypotheses through well-thought-out measurements. The main limitation of this approach arises from the wide range of space and time scales needed to characterize fully all the possible interactions and correlations taking place in materials. The common theoretical description, in fact, consists of defining a set of equations, which apart from very few cases, cannot be solved analytically. A usual solution is the definition of approximate models of the material, such as the free electron model to describe the valence electrons in solid metals, or the use of Ising model to characterize phase transitions in ferromagnetic materials. Such approximations are inevitably inaccurate and only capable of explaining only qualitatively the experimental results qualitatively.

A step forward has been made possible with the advent of *in silico* experiments, which open a virtual window on the atomistic side of matter and thus remove the need of relying on approximate theories. Atomic-scale simulation techniques, in fact, enable the description of materials in their entirety with a great level of accuracy. Starting from relatively few assumptions on the basic interactions between atoms and molecules, they allow one to emulate and follow the complete evolution of the system and to estimate its properties. They thus establish a link between experimental data and theoretical assumptions. Given their two-fold purpose, as vehicles that enable us to test new theories and approximations as well as to suggest new possible experiments, simulations have become ubiquitous in materials science. This has been catalyzed by the tremendous increase in the available computational power in the last decades and to the increased availability of easily accessible high-performance computing facilities. In

addition to the technological advances, models have become more and more accurate, and a large number of simulation packages have appeared that simplify the task of setting up and running a simulation. As a consequence, massive amounts of data can now be produced as routine. However, the outcome of a computer simulation is ultimately nothing but an enormous quantity of unstructured numbers, that must be laboriously post-processed in order to extract meaningful information. Defining proper analysis protocols can be far from trivial, and traditional approaches, are usually not suitable for manipulating enormous data sets containing different types of multivariate data. Furthermore, simulations are often exploited for their agnostic character, which enables us to test new unknown systems or to study phenomena that are not yet clear. The drawback here is that researchers may not be able to formulate clearly the questions to be addressed.

In order to exploit the flexibility of simulations, and to extract the maximum amount of knowledge obtainable from the vast quantity of data produced new mindsets are needed, as well as new tools that are able to conduct the analysis automatically and to provide useful information to assist and guide the interpretation of results.

Machine-learning algorithms provide the essential infrastructure necessary to unravel all the potential regularities and correlations hidden in simulation data. Applications of machine-learning algorithms to computational material science have already been successfully proposed in the literature, providing, for example, insights on how different structures contribute to the stability of various systems, such as nanoclusters, small organic molecules, and peptides. Machine-learning strategies have also been used to introduce complete and accurate descriptions of local atomic environments, which not only are useful for the construction of regression models for property prediction, but also for the identification of fundamental structural motifs and their possible combinations into more complex supramolecular patterns.

This thesis focuses on the use of machine-learning techniques to guide the understanding of complex structural problems in materials science and biology, by introducing a new general protocol for the identification of repetitive global and local structural motifs sampled through atomistic simulations. Nevertheless, all the ideas proposed here are not strictly limited to atomistic simulations and could be extended to other more general problems, i.e. the interpretation of experimental data, as well as more standard pattern recognition applications, such as image and text recognition.

Identification of repetitive patterns in atomic-scale data could mean the detection of specific structures along a simulation trajectory or among those stored in a structural database. It could also refer to the

recognition of the different meta-stable motifs intrinsic to a specific system. The former is essentially a *classification* task where, for a given input structure, a proper label is assigned, while the latter is a *coarse-graining* process where, among all the possible input structures, a few are selected to represent the structural landscape of the system being studied. Both are very important and will be addressed in this work, whose initial aim was to overcome some critical limitations faced by standard analysis approaches, such as the use of an arbitrary, clear-cut geometrical criterion to define a fuzzy entities such as hydrogen bonds in liquid water and proteins.

In general, the usual paradigm for a structural analysis in atomistic simulations, involves the use of chemical intuition to guide the manual inspection of structures along the trajectory. The traditional descriptors commonly adopted to classify patterns in atomistic simulations, usually involve the introduction of ranges of parameters that are deemed to represent a specific motif. Threshold values for distances, angles and energies are typically estimated from experiments and manually adjusted to be applied to various systems. Although practical and quite effective for many simple examples, this approach brings an intrinsic level of arbitrariness that could lead to a bias in any of the following steps in the analysis. Different choices of models and parameters reflect different final results, thus one must take particular care of the fact that, simulating a system using different methods and level of approximations, could lead eventually to (slightly) different average structures. Because comparison and cross-validation of results across different methods are routine in modern material research, personal choices of parameters in structural definitions could critically propagate into the inaccurate interpretation of the final results. Moreover, given that simulations produce large volumes of very detailed, noisy, high-dimensional data, the idea of manually inspecting trajectories using the spectacles of chemical intuition to guide the discovery of possible complex motifs and meaningful correlations is naive and often unfeasible.

We propose a solution to this problem by introducing a machin-learning framework that is capable of analyzing atomistic data automatically. We dub this procedure Probabilistic Analysis of Molecular Motifs (PAMM). PAMM provides not only a clear picture of all the possible meta-stable structures explored in a simulation trajectory, but also a natural out-of-sample probabilistic definition for each one of them, which can be used afterwards to detect and quantify similar structures in new simulation outcomes. PAMM builds on the idea that meta-stable patterns are nothing more than configurations which, due to the (free) energetics of the system, result in being more probable at a certain thermodynamic condition. By exploiting a non-parametric density estimation scheme, we learn, from all the possible

configurations observed in a simulation, the underlying probability distribution function. Since each mode of the distribution must correspond to a class of configurations matching a specific meta-stable pattern, we apply an unsupervised clustering algorithm to partition the dataset between the different modes. This knowledge is used to find the best Gaussian fit to each of the modes and thus approximates the ideal probability distribution function with a Gaussian Mixture Model.

We will show that these ingredients are enough to introduce a general, flexible, probabilistic and data-driven definition for recurring patterns, which combined with non-linear dimensionality reduction schemes, greatly helps to fully characterize the configuration space of complex atomistic systems.

SUMMARY

The rest of the thesis is organized as follows. After a short introduction about atomistic simulations and machine-learning techniques in chapter II an III respectively, chapter IV contains a detailed discussion of the rationale behind PAMM and describes its functioning, and implementation. In chapter V we show how it is possible to use PAMM to introduce an unbiased, smooth and adaptive definition of hydrogen bonds, while in chapter VI we make use of this definition to study defect and correlations in the hydrogen-bond network of water. In chapter VII we propose the use of PAMM, in combination with sketchmap, to investigate the structural patterns exhibited by nanoclusters and we apply the same methodology in chapter VIII to identify recurrent motifs in biomolecules. Finally, in Chapter IX we propose some future outlooks and in chapter X we draw our conclusions.

Chapter IV is adapted from refs. [1] and [2], while chapter V is adapted from ref. [1]. Chapter VI is an adaptation of ref. [3], while chapter VII and VIII are adapted from ref. [2].

2

ATOMISTIC SIMULATIONS

Contents

Simulated experiments are becoming increasingly common in all fields of materials science, engineering and chemistry. Before the advent of simulations, the only way to predict properties of materials was by describing them with some inevitably approximate models. The problem is that real materials can be extremely complex, and crude analytic approximations lead often to qualitatively inaccurate results. This fact, together with the ever-growing computational power, and the development of more accurate numerical models has made atomistic-scale simulations very common in modern materials research.

Simulations can be used to test new theories and address questions about materials more easily than experiment can, even if experiments are still central to testing and validating the accuracy of the simulation methodologies. Through simulations, it is possible to sample all the possible configurations specific to a system and to predict its static and dynamical equilibrium properties. Two main techniques have been developed for this purpose, both of which are based on statistical mechanics: Monte Carlo (MC) and Molecular Dynamics (MD). Compared to MC methods, which allow the calculation only of equilibrium averages, MD offers the possibility of also studying time-dependent microscopic properties, such as adsorption mechanisms, reaction pathways and conformational dynamics of polymers and biomolecules. Furthermore, it is possible to compute time correlation functions (TCFs) from MD trajectories. Using linear response theory, these can be used to predict transport coefficients and spectra [4]. In this thesis, we will focus exclusively on MD techniques, which will later be described in more detail.

A more complete and formal description of atomistic simulation methods can be found in several excellent books: Statistical Mechanics: *Algorithms and Computations* by W. Krauth, *Understanding Molecular Simulation: From Algorithms to Applications* by B. Smith and D. Frenkel, and *Statistical Mechanics: Theory and Molecular Simulation* by M. Tuckerman.

2.1 MOLECULAR DYNAMICS

Molecular dynamics (MD) is a broadly-used computational tool which allows matter to be studied in its microscopic detail. By describing a material as a classical system composed of N interacting particles, its dynamical behavior and evolution over time can be predicted by solving Hamilton's equations:

$$\begin{aligned} \dot{\mathbf{r}}_i &= \frac{\partial H(\mathbf{p},\mathbf{r})}{\partial \mathbf{p}_i} = \frac{\mathbf{p}_i}{m_i} \\ \dot{\mathbf{p}}_i &= -\frac{\partial H(\mathbf{p},\mathbf{r})}{\partial \mathbf{r}_i} = -\frac{\partial V(\mathbf{r})}{\partial \mathbf{r}_i} = \mathbf{F}_i \,, \end{aligned} \tag{1}$$

where $\mathbf{r}_i$, $\mathbf{p}_i$ and $\mathbf{F}_i$ are respectively the position, momentum and force associated with the ith particle, V is the interatomic potential and H is the classical N-particle Hamiltonian.

Given some initial conditions, the state of the system is completely determined at any time by the positions $(\mathbf{r}_1, \ldots, \mathbf{r}_N)$ and the momenta $(\mathbf{p}_1, \ldots, \mathbf{p}_N)$ of each particle. Since the analytical solution of the equations of motion for a generic complex many-body problem is simply not feasible, the time evolution in MD is achieved by iteratively applying a numerical integration scheme with a discrete timestep Δt. The accuracy of such approximation depends on the choice of Δt, which is usually taken to be from 0.5 to 2 fs. The most common integration scheme is the velocity Verlet [5].

If the tractory is long enough to sample the whole phase space, Hamiltonian molecular dynamics can be used to generate equilibrium ensemble averages consistent with the microcanonical ensemble (NVE), by exploiting the so-called erogodic hypothesis:

$$\bar{A}_{\mathscr{T}} = \lim_{\mathscr{T}\to\infty} \frac{1}{\mathscr{T}} \int_0^{\mathscr{T}} dt\, A(\mathbf{p}(t), \mathbf{r}(t)) = \langle A \rangle_{NVE} \,, \tag{2}$$

where $A(\mathbf{p},\mathbf{r})$ is a function corresponding to a generic physical observable and $\mathscr{T}$ is the total length of the trajectory. However, experiments are usually performed at thermodynamic conditions that are not consistent with the microcanonical ensemble, such as constant pressure or temperature. This has led to the development of many approaches to generate alternative ensembles in MD [6–9].

2.2 INTERATOMIC POTENTIALS

In order to obtain a realistic time evolution of the system, a proper description for the interatomic interactions is compulsory. Clearly, the quality of the final results depends to a great extent on the approximations that are made at this level.

A common approach consists of introducing an empirical interatomic potential, called *force field* (FF), which gives a fairly simple, qualitative mathematical description for each of the different interactions. Several parameters are usually introduced and tuned from experiments, or more accurate simulations, to reproduce different experimental systems. A commonly adopted formulation for a FF corresponds to a series of terms that represents the different interactions that can occur among increasingly larger groups of atoms

$$V = \epsilon^{(0)} + \sum_i \epsilon_i^{(1)} + \sum_i \sum_{j<i} \epsilon_{ij}^{(2)} + \sum_i \sum_{j<i} \sum_{h<j} \epsilon_{ijh}^{(3)} + \dots , \qquad (3)$$

where, for instance $\epsilon_{ij}^{(2)}$ and $\epsilon_{ijh}^{(3)}$ represents the energy contribution coming, respectively, from the two-body and three-body interactions among the atoms i,j and h.

The first example of an empirical potential that was capable of describing a variety of molecules was probably the molecular mechanics FF (MM) [10], which was followed by improved versions (MM2, MM3 and MM4) [11–13]. Current popular FFs are UFF [14], CHARMM [15], AMBER [16], GROMOS [17], and OPLS [18] among many others. Characterized by their low computational cost, FFs allow for very large simulations (both in time and space). Impressive examples are the simulations of the tobacco mosaic virus [19], bacterial flagellar filament [20] and HIV-1 capsid [21].

The major disadvantage of most empirical FFs is the fact that the connectivity between atoms is pre-defined and constrained throughout the entire simulation. This implies the impossibility to model and predict bond forming and breaking events. Moreover, the predictive power of a FF is guaranteed only for those systems and thermodynamic conditions for which they were designed.

A way to solve some of the problems faced by force fields is to treat specific parts of the system by using more accurate methods, such as quantum mechanical models. An example is QM/MM approaches [22–26], which aim to obtain a higher predictive accuracy, while still at a low computational cost, by using cheaper approximations for those regions that are less important. QM/MM methods are very promising, but they still involve strong approximations, meaning that they can fail to reproduce exactly the physics needed to interpret experimental results, for which a more precise, accurate description is necessary [27, 28].

2.2.1 Ab Initio *Molecular Dynamics*

Ab Initio Molecular Dynamics (AIMD) offers an excellent solution to the limitations of FFs, by combining classical dynamics with electronic structure methods to explicitly describe the quantum nature of electrons. Assuming the Born-Oppenheimer (BO) approximation and neglecting non-adiabatic effects, the internuclear forces are computed at each timestep via an electronic structure calculation, starting from a set of initial nuclear positions. At this point one step in the dynamics can be done by plugging the forces, together with a set of initial momenta for each atom, into a numerical integration algorithm. New velocities and nuclear positions are obtained, which can be used to iterate the procedure and evolve the system through time.

In this context, the role of the electronic structure methods is to find an approximate solution to the non-relativistic time independent many-body Schrödinger equation:

$$\hat{H}\Psi = E\Psi, \tag{4}$$

where Ψ is the many-body wavefunction and $\hat{H}$ is the Hamiltonian of the system, which is the sum of all the interactions between nuclei and electrons, plus their kinetic energy terms. Despite its simplicity, eq. 4 is incredibly complex to solve, and an exact solution of the electronic many-body problem is intractable.

Some approximations are needed to make the electronic calculation feasible. Many methods have been suggested, from simple models such as Hückel theory [29], to semi empirical approaches [30–32], Hartree-Fock (HF) [33, 34] and correlated methods beyond HF, such as configuration interaction (CI), Møller-Plesset perturbation theory (MP, MP2, ...) [35, 36], and coupled cluster (CCD, CCSD, ...) [37–39], to Quantum Monte Carlo [40–42]. Density Functional Theory (DFT) [43–45] is probably the most widely-adopted approach in atomistic simulations of materials, mainly because of the computational cost (if N is number of basis functions, standard DFT scales as $\mathcal{O}(N^3)$, which is massively faster than other approaches – CCSD scale as $\mathcal{O}(N^6)$).

DFT is based on the principle that all the properties of a system, subject to an external potential V_{ext}, are determined by its ground state density. In particular, the energy E_0 associated with the ground state is

$$E_0 = \int d\mathbf{r}\, V_{ext}(\mathbf{r})\rho_0(\mathbf{r}) + F[\rho_0(\mathbf{r})]\,, \tag{5}$$

where ρ_0 is the ground state density and $F[\rho] = T[\rho] + E_{e-e}[\rho]$ a universal functional of the density, with $T[\rho]$ being the kinetic-energy functional and $E_{e-e}[\rho]$ the electron-electron interaction functional.

Among all the possible densities, $F[\rho_0]$ has its minimum value at the true ground state density, which can be found through a variational approach.

Usually DFT is implemented using the Kohn-Sham (KS) formulation, where a system of interacting electrons is formally mapped into a system of non-interacting electrons (N-representability), to give a set of equations that can be solved self-consistently. The effective potential for the KS Hamiltonian (H_{KS}) is

$$V_{KS}(\mathbf{r}) = V_{ext}(\mathbf{r}) + V_H(\mathbf{r}) + V_{XC}(\mathbf{r}), \tag{6}$$

where V_{ext} is the Coulomb potential due to the nuclei, V_H is the Hartree potential due to the mean field electron-electron interaction and V_{XC} is exchange-correlation potential, which describes the complex many-body electron-electron interaction. V_{XC} is the functional derivative of the exchange-correlation functional (E_{XC}) with respect to the density. Within the KS scheme, the only term that requires an approximation is E_{XC}, which affects strongly the accuracy of the results.

DFT has been applied to problems involving thousands of electrons and has been successfully applied to study a broad range of phenomena, from the virtual design of new catalysts [46] and battery materials [47], as well as the study of adsorption onto porous materials and surfaces [48, 49], prediction of the optical properties of new materials [50] and simulation of spectra [51]. DFT has also been used to study the structural and dynamical properties of a number of neat molecular liquids and solids including water [52, 53], ammonia [54], ice [55], and many others [56–60]. Of course in a practical implementation one has to deal with various different approximations, and standard DFT approaches can be affected by problems of self interaction and a poor modelling of long range dispersion interactions. A popular solution to this problem is to add a pairwise correction of the internuclear energy, by adding a pairwise C_6R^{-6} dispersive term. An example is the D3 method proposed by Grimme [61]. Other approaches are to include explicitly non-local correlations in the XC functional [62].

The biggest approximation is that of the exchange-correlation functional E_{XC}, which is not known exactly, and a huge list of approaches – spanning a very wide range of complexity – can be found in the literature. A useful classification, that aims to group the existing functionals, was proposed by Perdew [63] and is the "Jacob's ladder", which is schematically represented in fig. 1. According to this scheme, E_{XC} functionals can be grouped in classes of increasing complexity, from the Hartree approximation to exact exchange-correlation. A further classification is then possible, splitting the methods between empirical (fitted to known results properties of matter) and non-empirical

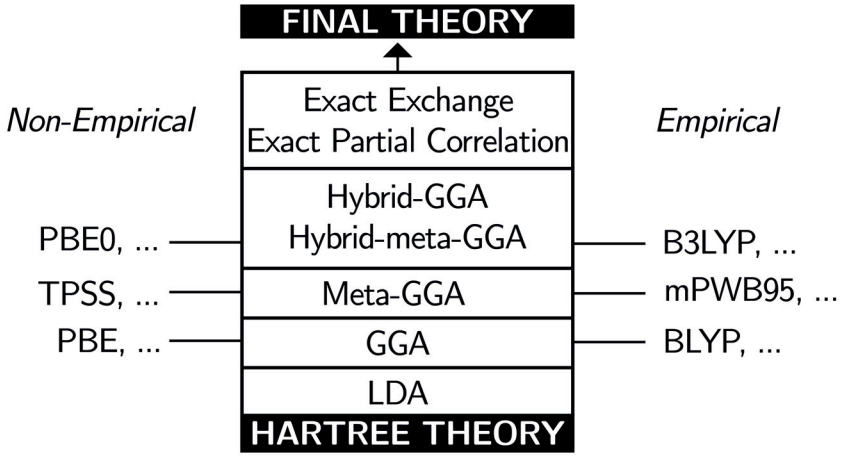


Figure 1: The schematic diagram named "Jacob's ladder" proposed by J. P. Perdew to classify exchange-correlation functionals.

(based only on physical laws). A list of the most common exchange-correlation functionals is:

- local-density approximation (LDA): E_{XC} is a functional of the local density $[n(\mathbf{r})]$ only and the corresponding XC energy is obtained from the uniform electron gas having the same density. Example functionals are Perdew-Zunger (PZ) [64], Perdew-Wang (PW) [65] and Vosko-Wilk-Nusair (VWN) [66].

- generalised gradient approximation (GGA): E_{XC} is a functional of the local density $[n(\mathbf{r})]$ and the gradient of the density $[\nabla n(\mathbf{r})]$. In practice the gradient correction to the xc energy is introduced by enhancing LDA xc energy as a function of reduced gradient $(|\nabla n(\mathbf{r})|/n(\mathbf{r})^{4/3})$. Examples are the PBE [65], BLYP [67], and BP86 [68] functionals.

- meta-GGAs: E_{XC} is a functional of the local density $[n(\mathbf{r})]$, the gradient of the density $[n(\mathbf{r})]$, and the kinetic energy density $[\nabla\psi(n(\mathbf{r}))]$. An example is the TPSS [68] functional.

- hybrid GGA: It is hybrid mixture of GGA xc functionals with some fraction of Hartree-Fock exchange. Well known examples are PBE0 [69] and B3LYP [66, 70, 71].

Another important approximation to consider, when dealing with *ab intio* methods, is that all electronic structure methods expand the unknown wave function in terms of a set of basis functions. Many basis sets can be found in the literature, and particularly popular is the use of an atom-centered localized basis. Examples are Gaussian type orbitals (GTO), such those proposed by Pople (3-21G,6-21G...) [72], and the correlation-consistent basis sets proposed by Dunning (DZVP, TZVP,...) [73, 74], but also the numeric atom centered

orbitals (NAO) [75]. Another very common choice is the use of plane wave basis sets [76].

A complete description of *ab initio* methods is not in the scope of the thesis and a more detailed description, about these and other methods, can be found in many excellent books (*Modern Quantum Chemistry* by A. Szabo and N. Ostlund, *Density Functional Theory* by E. K. U. Gross and R. Dreizler).

2.2.2 *Nuclear Quantum Effects*

In the all the previous sections, a quantum mechanical treatment has been assumed to be necessary only to achieve a better description of electrons in some particular cases. However, in many cases, it is also necessary to consider the quantum effects due to the nuclei. Nuclear quantum effects (NQE) (e.g, zero point energy and tunneling) are critical for all those systems where the motion of light nuclei (such as hydrogen or lithium) is important. Examples are water [77–79], ammonia [80], ice [55, 81], and enzymes [82] among many others.

When the internuclear potential is fitted from experiments, these quantum effects are often considered implicitly, while when *ab initio* methods are used without considering any correction for the quantum behaviour of nuclei, the resulting error on the final estimates can be as large as that due to the approximation of the exchange-correlation energy for the electrons [83, 84].

The standard procedure to include NQEs in molecular dynamics is to model them using the imaginary time path integral (PI) formalism [85, 86]. PI methods exploit the mathematical isomorphism between the quantum Boltzmann statistics of a given system and the classical Boltzmann statistic of its ring-polymer (RP) representation. This ring-polymer is composed of replicas (beads) of the physical problem, with corresponding atoms are connected by harmonic springs [87, 88].

Consider the quantum canonical partition function for a single particle in one spatial dimension. The partition function is well-defined and is given by the trace:

$$Z = \mathrm{Tr}[e^{-\beta \hat{H}}] \tag{7}$$

The path integral formalism exploits the Trotter theorem to write an extended phase space expression for the partition function

$$Z_P = \frac{1}{(2\pi\hbar)^P} \int d^P\mathbf{p} \int d^P\mathbf{r} e^{-\beta H_P(\mathbf{p},\mathbf{r})/P}, \tag{8}$$

where P is the number of beads and $H_P(\mathbf{p}, \mathbf{r})$ is the Hamiltonian of an extended ring polymer

$$\sum_{j=0}^{P-1} \left[\frac{1}{2} p_j^2 + V(r_j) + \frac{1}{2} \omega_P^2 (r_j - r_{j+1})^2 \right], \tag{9}$$

with $\omega_P = P/(\hbar\beta)$ and $r_P = r_0$ ($\beta = (k_B T)^{-1}$ is the inverse temperature, k_B the Boltzmann constant and $\hbar$ the reduced Planck constant).

The RP partition function converges to the correct quantum mechanical result when the number of beads P tends to infinity, but to get a converged result it is sufficient to use a number of beads which is a small multiple of $\beta\hbar\omega_{max}$, where ω_{max} is the largest vibrational frequency in the system. For example, if we consider a system containing an O–H bond stretch, full convergence will be achieved using 64 beads at around room temperature, therefore making the PI simulation 64 times more expensive than the correspondent classical simulation. As a consequence of this large overhead, NQEs were for many years only rarely considered in the context of *ab initio* molecular dynamics [89–91].

A promising approximate method for modeling NQEs is the use of a coloured (correlated)-noise Langevin equation (GLE) thermostat in combination with classical molecular dynamics [92, 93]. The accuracy of this methodology has been proven for many systems, showing that an inclusion of NQEs in MD is possible, at a fraction of the cost compared with standard PIMD approach. Of particular note is the so-called path integral generalized Langevin equation thermostat (PIGLET) scheme [94], which enables one to achieve a speedup of two orders of magnitude compared to standard PIMD, by combining path integral and GLE methods.

2.3 ENHANCED SAMPLING TECHNIQUES

The interesting phenomena governing technological materials and biological systems occur on a vast scale of times and lengths. Often, the behaviour of complex systems is driven by rare but important events. This is because the (free) energy landscape of a complex system consists of many metastable states divided by high kinetic barriers [95].

Long-lived (meta)stable states are a consequence of the difference between the thermal energy and the energy needed to overcome barriers. In fact, the dynamics evolves with the system fluctuating around the high-probability configurations characteristic of a (free) energy minimum for long time periods. Occasionally, larger fluctuations can take place, with the system jumping from one state to another.

An example is the autoionization of molecules in liquid water: while the average lifetime of a molecule is on the order of hours, sporadic fluctuations of the solvent can trigger a sub-picosecond event

which eventually leads to the dissociation of a water molecule into the ion pair H^+ and OH^- [96, 97]. Other well-known examples are conformational changes in proteins [98], nucleation events in phase transitions [99] and chemical reactions [100, 101].

Even with the advent of easily accessible high-performance computing (HPC) facilities and massively parallel architectures and algorithms, it is stil not possible to study physical phenomena that involve rare events directly by brute force MD. However, a large variety of methods, with the name of *enhanced sampling techniques*, have been proposed to accelerate the dynamics and address the issue of poor sampling in standard MD.

The reader is referred to the various excellent reviews present in literature for more information [102–107]. Of particular interest for the scope of this thesis are tempering-based approaches, such as parallel tempering, and those relying on collective variable biasing, such as umbrella sampling and metadynamics, for which a more detailed discussion will follow.

2.3.1 *Collective Variables*

The configurational state of a system is completely specified by the set of atomic coordinates $\{\mathbf{r}\} \in \mathbb{R}^{3N}$.

It is often convenient to reduce such great number of degrees of freedom into a few parameters, which still can describe the physics of the phenomena under study. These descriptors, depending on the field, are called order parameters, collective variables (CV) or reaction coordinates. We will mainly adopt the term CV in this thesis.

A CV can be any differentiable function of the atomic coordinates $S(\mathbf{r}_1, \ldots, \mathbf{r}_N)$ mapping a 3N-dimensional space into an M-dimensional one, with $M << 3N$.

The selection of proper CVs is probably the most difficult step in any investigation. Good CVs should be as low dimensional as possible (ideally one to three-dimensional), but at the same time they should be able to distinguish the different states of the system and the relevant intermediates.

To face the problem of choosing proper CVs, many strategies have been proposed. Of particular interest are Path CVs [108], the SPRINT graph-based CVs [109], and sketch-map CVs [110]. The last of these is particularly useful, since it tries to find the best low-dimensional description of the system, starting from a high-dimensional description and reducing the dimensionality using a machine-learning scheme.

2.3.2 *Umbrella sampling*

Umbrella Sampling is probably the most established among non–Boltzmann sampling techniques. It was first introduced by Torrie and Valleau [111] to overcome kinetic barriers and accurately estimate free energy differences.

The main idea behind umbrella sampling is to modify the original Hamiltonian by adding a bias potential (the *umbrella potential*, V_b) which forces the system to explore configurations that would not be sampled sufficiently during standard MD. It is often convenient to define the umbrella potential as a function of one or few CVs ($\mathbf{s}$), for which the equilibrium probabilty distribution reads,

$$P(\mathbf{s}) = \frac{1}{Z}\int d\mathbf{r}\, e^{-\beta V_0(\mathbf{r})}\delta(\mathbf{s} - s(\mathbf{r}))\,, \tag{10}$$

where $Z = \int d\mathbf{r}\, e^{-\beta V_0(\mathbf{r})}$ is the configurational partition function.

The probability is linked to the free energy by

$$F(\mathbf{s}) = -\frac{1}{\beta}\ln P(\mathbf{s})\,, \tag{11}$$

where $\beta = (k_B T)^{-1}$ is the inverse temperature.

Each local minimum of F corresponds to a metastable state of the system under study.

One could be interested in forcing the system to sample more (or to avoid) a specific minimum and this could be done by adding an attractive (or repulsive) bias potential $V_b(s(\mathbf{r}))$, which is a function of the CVs only. This leads to a biased distribution of the CVs ($P_b(s)$), and any unbiased property of the system can be recovered afterward, by exploiting the relationship:

$$P(\mathbf{s}) \propto P_b(\mathbf{s}) e^{\beta V_b(s(\mathbf{r}))}\,, \tag{12}$$

where $e^{\beta V_b(s(\mathbf{r}))}$ is the weight associated with the configurations having a specific value of $\mathbf{s}$.

Since usually a system of interest presents many free energy minima, it is very difficult to define *a priori* a single bias potential to enhance the sampling in each of the various minima.

A practical solution is to combine the results from different independent simulations, where different restraining potentials are applied. In this case, the unbiased statistics can be recovered by combining the data from all the trajectories using the so-called weighted-histogram analysis method (WHAM) [112].

Of course, the enhancement in sampling promoted by the application of an umbrella sampling strategy critically depends on how well the chosen CV captures the physics of the process of interest.

For more details regarding umbrella sampling and its various applications the reader is referred to Ref. [113].

2.3.3 *Metadynamics*

Metadynamics is an adaptive biasing method which solves the difficulties of building the bias potential encountered in umbrella sampling, by adding to the Hamiltonian a history-dependent bias to discourage the exploration of already-visited configurations [114]. At different times, repulsive Gaussians are deposited in the low-dimensional CV space to prevent the system from being trapped in a free energy minimum. If $S(\mathbf{r})$ is the chosen collective variable, than at the time t the bias potential is defined as:

$$V_b(\mathbf{s}, t) = \int_0^t dt' \, \omega e^{-\frac{S(\mathbf{r})-S(\mathbf{r}(t'))}{2\sigma^2}} \tag{13}$$

where σ is the standard deviation of the Gaussian, and ω is a constant that depends on the height of the Gaussian W and the deposition stride τ_G in the form

$$\omega = \frac{W}{\tau_G} \tag{14}$$

If the simulation is long enough, it is possible to recover the free energy from the relation

$$V_b(\mathbf{s}, t \to \infty) = -F(\mathbf{s}) + C\,, \tag{15}$$

where C is an additive constant.

Metadynamics accelerates sampling of rare events without any prior knowledge of the energy landscape, other than the choice of bias CV. It can also help understand new reaction pathways since the system typically escapes from a local minimum by overcoming the lowest saddle point around. Moreover, metadynamics is highly parallelizable since a faster filling of the free-energy is possible running multiple independent simulations [115]. On the other hand, standard metadynamics has the limitation that, in a single simulation, instead of converging to the free energy, it oscillates around it. Furthermore, it is not trivial to understand when to stop the simulation.

Well-Tempered (WT) metadynamics [116] solves these problems by decreasing the deposition rate with time by modulating the height of the Gaussian

$$W = \omega_0 \tau_G e^{\frac{-V_b(\mathbf{s},t)}{k_B \Delta T}}\,, \tag{16}$$

where ω_0 is the initial rate, and ΔT is a parameter controlling the extent to which the free energy is explored. Unlike standard metadynamics, the bias potential does not converge to the negative of the free energy, but to a fraction of it

$$V_b(\mathbf{s}, t \to \infty) = -\frac{\Delta T}{T + \Delta T} F(\mathbf{s}) + C\,, \tag{17}$$

where the term $\Delta T/(T+\Delta T)$ is usually referred as the bias factor. The result is an enhancement of sampling in CV space, that corresponds to an effective temperature $T+\Delta T$, where for $\Delta T \to \infty$ one recovers standard metadynamics, and ordinary MD for $\Delta T = 0$.

Finally, it is possible to reweight WT metadynamics simulations to recover the proper statistics for any observable. Many solutions have been proposed, and notable are the solutions by Bonomi et al. [117] and Tiwary et al. [118]. Fewer possibilities are present in the literature regarding the reweighting of non-WT metadynamics, with the most popular solution being that of Marinelli et al. [119].

2.3.4 *Parallel tempering*

Parallel tempering (PT) [120] is the most common among the Replica Exchange methods (RE) [121], that are built on the idea of running multiple energetically independent simulations (namely the *replicas*) which occasionally exchange configurations.

In particular, PT is an effective equilibrium Monte Carlo scheme, that satisfies detailed balance, which when applied to MD, can enhance sampling in systems having a rough free-energy landscape characterized by many local minima.

In PT, many replicas run simultaneously at different ensemble temperatures (T) with the Metropolis rule attempting exchanges of configurations between the various replicas. Each swap move does not affect the Boltzmann distribution corresponding to each ensemble, thus ensemble averages can be computed from each individual trajectory.

Usually exchanges happen between adjacent temperatures, with an acceptance probability for the "swap" between the replica i and the replica j defined using a Metropolis criterion

$$\alpha_{ij} = \min\left(1, e^{(\beta_i-\beta_j)[V(x_i)-V(x_j)]}\right), \tag{18}$$

where β is the inverse temperature and V is the potential energy. If the swap is accepted, a coordinate exchange take place, with the velocities rescaled to the new temperatures:

$$v_{new} = \sqrt{\frac{T_{new}}{T_{old}}} v_{old}\,. \tag{19}$$

The closer the replicas are in temperature, the higher the probability of swapping.

As a general rule, the highest temperature must be sufficiently high to allow the system to escape free-energy minima and explore low-probability regions of the phase space, while the low-temperature replicas should still probe efficiently the various (meta)stable states

corresponding to the free energy minima. On the other hand, the number of replicas should be large enough to ensure proper swapping among adjacent replicas. Several schemes can be adopted in order to find the optimal number and temperature of the replicas [122, 123], the most common scheme is to follow a geometric sequence between T_{min} and T_{max}. Furthermore, given that energy fluctuations scale with $\sqrt{N}$, where N is the total number of atoms in the simulation box, a large number of replicas is usually necessary to obtain a reasonable exchange rate among adjacent trajectories. An elegant way to overcome such limitations is to combine PT with the well-tempered ensemble (WTE) [124] which is the biased ensemble emerging from WT metadynamics when the energy is used as CV. Compared with the canonical ensemble, the WTE ensures similar average energy with much larger fluctuations, thus allowing for a smaller number of replicas over the same range of temperature.

3

MACHINE LEARNING

Contents

Machine-learning (ML) is a modern, highly interdisciplinary, field that is being developed with the aim of rationalizing how to get a computer to learn from data, without being explicitly programmed to perform a specific task. ML has had extraordinary progress in the last two decades, and in an ever growing number of fields, scientists now train machines with examples, rather than studying and implementing new algorithms containing an *a priori* answer for any possible future input scenario.

The gravitational center of ML is in data, and the term learning refers to the ability of a machine to improve with experience (i.e., the training data) when performing particular tasks, for example classifying inputs.

It also means extracting knowledge and inferences from data, to the extent that the primary application of ML is that of making predictions of new or missing data from what is already known. From this latter point of view, ML can be effectively seen as the successor of an older discipline, statistical model fitting, as it shares the same goal of extracting useful information from a dataset by fitting the best probabilistic model describing the data.

A vast array of ML schemes have been developed in order to deal with the great variety of data, and problems, faced across all the possible fields of application of ML.

Supervised Learning

The most common approach to ML is Supervised Learning (SL), where starting from a collection of pairs (x, y), with $x \in X$ and $y \in Y$, the aim is to produce an output y^* as a response for an input x^*. The inputs x can be vectors, images or structures computed from an MD

simulation, while the outputs y can be a label (classification tasks) or a real number (regression tasks).

The predictive ability of SL algorithms builds on a non-linear object $\mathscr{F}(x)$, that is capable of mapping X into Y. $\mathscr{F}$ is shaped from data by borrowing ideas mainly from optimization theory, and can assume various forms such as decision trees [125], random forests [126], logistic regression [127], support vector machines [128], kernels [129, 130] and neural networks [131]. A particularly promising class of models are the so-called deep-learning methods [132, 133], where multiple stacks combining simple non-linear modules (multilayered networks) are optimized using gradient-based algorithms, allowing for very complex functions to be learned.

Another paradigm of ML is reinforcement learning, where the machine produces some actions and interacts with the environment. Differently from SL, the only indication provided is a reward if the output generated is correct, or a penalty if it is wrong. This type of learning is common in control theory, game theory and neuroscience [134].

Unsupervised Learning

The other major form of ML is unsupervised learning (USL), where inferences are drawn from data without the need for labels. That is, starting from purely unstructured data, USL finds patterns, under some assumptions, about the properties characterizing the data. The most common examples of USL are dimensionality reduction and clustering.

The following sections will thus introduce the reader to some key methods, i.e. clustering, density estimation, Bayesian methods and dimensionality reduction techniques.

For the readers interested in a general overview of ML, two interesting references are *Information Theory, Inference, and Learning Algorithms* by D.J.C. MacKay and *Pattern Recognition and Machine Learning* by C. Bishop. Furthermore, excellent reviews on SL are Refs. [133, 135]. For a panorama of USL methods, the reader is referred to Ref. [136].

3.1 CLUSTERING

Clustering is probably the most common form of unsupervised machine learning. It finds applications in very diverse fields, such as image analysis, speech and text recognition, astrophysics, bioinformatics, material science, and many others.

The aim of a cluster analysis is to discover the patterns hidden in data, which is done by partitioning the elements of a dataset into groups on the basis of their (dis)similarity. Each group, or cluster, is a set of analogous patterns.

More formally, given a set of points Ω, a clustering Z consists of the partition into K sets of mutually disjoint subgroups composed of similar objects, namely the clusters $\{Z_1, Z_2, \ldots, Z_K\}$, such that $Z_i \cap$

$Z_j = \emptyset$ and $\bigcup_{i=1}^{K} Z_K = \mathcal{Q}$. If N and N_k are the number of points in $\mathcal{Q}$ and Z_k respectively, then $N = \sum_{k=1}^{K} N_k$.

Clustering is an unsupervised technique, meaning that there is no need for labels, but just for information on similarity among the data, which comes from two elements of crucial importance: the pattern representation and the similarity measure.

A pattern is usually represented by a vector of features, specific for the problem under study. For instance, when clustering atomic-scale patterns, it is often convenient to define a feature space built using functions of the atomic degrees of freedom, instead of dealing directly with the Cartesian coordinates. The chosen description should be capable of representing the local environment surrounding an atom and at the same time highlight the properties of interest.

Features representation

In the last decade, many classes of features have been proposed in the literature. Examples are the graph-based order parameters (SPRINT) by Pietrucci et al. [109], electronic-structure based descriptors, such as Kohn–Sham eigenvalue fingerprints [137], Coulomb matrices [138], symmetry functions [139], and smooth overlap of atomic position (SOAP) power spectra [109], among many others.

Defining an appropriate feature space is essential in order to obtain meaningful results, and the nature of descriptors (qualitative or quantitative, continuous or discrete) determines the choice of the other key aspect of the analysis: the similarity measure.

Measure of similarity

The (dis)similarity between two feature vectors $\mathbf{u}, \boldsymbol{\nu} \in \mathbb{R}^D$ can be built simply by taking some kind of norm their distance vector. The most common choice is the L_2-norm, which is the Euclidean distance between the two vectors :

$$d(\mathbf{u}, \boldsymbol{\nu}) = \sqrt{\sum_{i=1}^{D} (u_i - \nu_i)^2}, \tag{20}$$

The L_1-norm; namely, the Manhattan distance, is also commonly used:

$$d(\mathbf{u}, \boldsymbol{\nu}) = \sum_{i=1}^{D} |u_i - \nu_i|, \tag{21}$$

or the L_∞-norm,

$$d(\mathbf{u}, \boldsymbol{\nu}) = \max_{1 \leqslant i \leqslant D} (|u_i - \nu_i|). \tag{22}$$

All of these are special cases of the L_p-norm,

$$d(\mathbf{u}, \boldsymbol{\nu}) = \left(\sum_{i=1}^{D} |u_i - \nu_i|^p \right)^{\frac{1}{p}}, \tag{23}$$

Many other choices have been proposed in the literature. The general requirements for a proper distance function are:

1. symmetry $d(\mathbf{u}, \mathbf{v}) = d(\mathbf{v}, \mathbf{u})$
2. positivity $d(\mathbf{u}, \mathbf{v}) \geqslant 0$
3. reflexivity $d(\mathbf{u}, \mathbf{v}) = 0$, if $\mathbf{v} = \mathbf{u}$
4. triangular inequality $d(\mathbf{u}, \mathbf{v}) \leqslant d(\mathbf{v}, \mathbf{z}) + d(\mathbf{u}, \mathbf{z})$, for all $\mathbf{u}, \mathbf{v}$ and $\mathbf{z}$.

If the four conditions hold, then the distance is a *metric*. When the objects to be compared are atomic structures a straightforward choice for a metric is the root mean square displacement (RMSD) distance, which is essentially the average distance between the atomic coordinates of two superimposed structures. While very natural, RMSD is not always the best metric to compare different configurations, since it fails to capture simple symmetries, such as translation, rotation, and permutation of equivalent atoms [140]. An interesting solution is to consider the similarity among the environments, instead of their dissimilarity.

This is usually achieved by making use of kernel functions instead of distances, where a kernel $\mathcal{K}(\mathbf{u}, \mathbf{v}) : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, is a symmetric semi-definite positive function, which takes in inputs two vectors in the initial space and returns an inner product of a mapping function in feature space:

$$\mathcal{K}(\mathbf{u}, \mathbf{v}) = \langle \phi(\mathbf{u}), \phi(\mathbf{v}) \rangle_M, \tag{24}$$

where $\langle \cdot, \cdot \rangle_M$ is the inner product of $\mathbb{R}^M$, with $M > D$, and the mapping function $\phi(\mathbf{u})$ transforms $\mathbf{u}$ to $\mathbb{R}^M$ ($\phi : \mathbb{R}^D \to \mathbb{R}^M$). The use of the kernel $\mathcal{K}$ is quite convenient, because it effectively represents dot products in a higher-dimensional space $\mathbb{R}^M$, without any prior knowledge of the underlying space of functions. This is the so-called *kernel trick*.

A kernel is usually normalized, such that $0 \leqslant \mathcal{K}(\mathbf{u}, \mathbf{v}) \leqslant 1$ and $\mathcal{K}(\mathbf{u}, \mathbf{u}) = 1$.

Popular examples of kernel functions are:

- polynomial kernels: $\mathcal{K}(\mathbf{u}, \mathbf{v}) = (\gamma \langle \mathbf{u} \cdot \mathbf{v} \rangle + c)^r$ [141]
- radial basis functions: $\exp(-\gamma(\mathbf{u} - \mathbf{v})^2)$ [142]
- sigmoid kernels: $\tanh(\langle \mathbf{u} \cdot \mathbf{v} \rangle + c)$.

It is usually possible to interpret a kernel as a distance by taking

$$d(\mathbf{u}, \mathbf{v}) = \sqrt{\mathcal{K}(\mathbf{u}, \mathbf{u}) + \mathcal{K}(\mathbf{v}, \mathbf{v}) - 2\mathcal{K}(\mathbf{v}, \mathbf{u})} \tag{25}$$

Particularly interesting are two recently developed kernels explicitly designed to compare the similarity among local atomic environments: the SOAP kernel [143] and its extension, the REMatch kernel [144].

Once the pattern representation and the similarity measure have been defined, different clustering techniques are possible, depending on the algorithm used to partition the data. It is not straightforward to classify all the possible type of clustering, and of course, for any possible categorization, there would be some exception for which categories overlap. Traditionally, however, the algorithms are grouped into two classes: hierarchical methods and partitional methods.

3.1.1 *Hierarchical clustering*

Hierarchical clustering refers to a set of iterative schemes that build, for the entire dataset, a hierarchy of nested clusters (a tree). There are two strategies to achieve such a result: agglomerative clustering and divisive clustering. The former is a bottom-up approach, where each object starts as a single cluster, and is iteratively merged, in pairs, with the closest neighbor, until one large cluster covers the whole dataset. The latter is called is a top-down scheme. At the first stage all the dataset belongs to one cluster, and then it is split recursively to obtain a hierarchy of clusters, until each object is separate from the others.

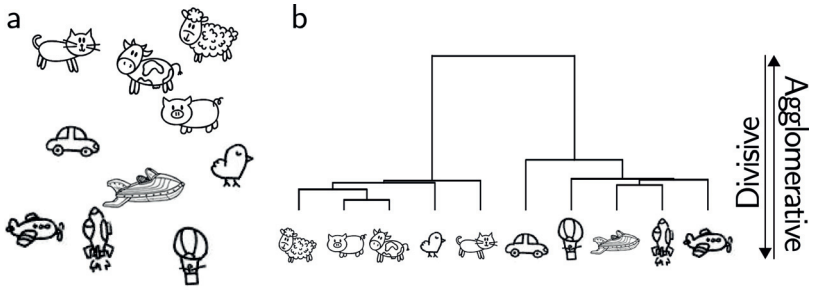


Figure 2: Dendrogram (b) representing the hierarchical clustering of an example data set containing two different classes of objects: animals and means of transportation (a).

In both cases, a connected tree is produced and visualized through a dendrogram plot. The bottom of the plot consists of the single clusters, which are sorted along the x axis in relation to the partitioning, while in the y axis the lines, representing the proximity between the pairs, are drawn, with a length that is proportional to the magnitude of the proximity measure.

It is common habit to cut the dendrogram at a given height to partition the dataset at a desired precision.

The linkage metric

The difference between different hierarchical methods stems from how the proximity (the linkage metric) between two clusters is defined. The prevalent criteria are:

- single linkage: two clusters are jointed when they have the smaller intercluster distance between two of their members,
- complete linkage: the proximity between the two clusters corresponds to the distance between the farthest pair of objects, where each cluster contribute with one object to the pair,
- average linkage: the distance between two clusters is defined as the arithmetic mean of all inter-cluster pair distances,
- centroid method: the proximity between two clusters is the distance between the geometric centroids of the two clusters,
- Ward's minimum variance method: the proximity is defined such that two clusters are merged if they produce the minimum increase in the total within-cluster variance after the merging step.

Hierarchical methods are particularly interesting, since they allow one to explore the partitioning of a dataset at different levels of granularity. However, in certain cases this aspect can also be a disadvantage, as it can be highly non-trivial to decide where to cut the tree in order to obtain the optimal partitioning. Another disadvantage is that, depending on the linkage metric used, the partitioning can be computationally expensive ($\mathcal{O}(N^3)$, where N is the number of elements being clustered).

3.1.2 *Partitional clustering*

Partitional methods break the dataset into groups rather then building a nested tree, and the main difference with hierarchical algorithms is that the partitioning into clusters is learned directly, instead of being done gradually. The algorithm does so either by relocating elements among the clusters, or by splitting the data in the regions that are most populated. The former approach is called Partitioning Relocation method, and is probably the simplest, yet the most diffuse, type of clustering algorithm.

K-means

Partitioning Relocation methods can be divided into two main groups: K-means [145] and K-medoids [146]. The former attempts to minimize the total within-cluster squared error, while the latter minimizes the total dissimilarity between the elements of each cluster. These two schemes are very similar, and consist in the iteration of two steps: an assignment step and an update step. For instance, K-means starts

from an initial set of K centroids, and proceed with the following scheme:

1. each point is assigned to the closest centroid, e.g. by partitioning the data according to Vornoi tessellation around the means,

2. compute the new centroids, i.e. the mean of each cluster, is computed.

The algorithm is repeated until a stopping criterion, such as a certain threshold in the variation, is fulfilled.

K-medoids differs from K-means in that it deals with medoids instead of centroids. A medoid is a point whose average dissimilarity to the rest of the dataset is minimal. Conceptually, medoids are similar to centroids, with the peculiarity that they are always members of the data set.

In both K-means and K-medoids, the number of clusters (K) must be specified in advance by the user. This is probably the biggest drawback of these algorithms, since the choice of K for a generic multidimensional dataset is non-trivial. To face such a problem, one usually has to perform a series of clustering calculation varying K. The optimal value of K is then found *a posteriori*, by using some method to validate clusters' quality, such as the Silhouette method [147].

Furthermore, K-means and K-medoids produce good results when the clusters are isotropic and normally distributed, which is an assumption that is often not satisfied in practice. When the clusters to be identified have very irregular shapes, a better choice is to use density-based procedures, where a cluster is seen as a dense group of connected objects, whose boundaries are defined by the density function underlying the data.

A large class of density-based methods exists, and a very popular example is model-based clustering, where after describing the data using a certain probabilistic model, the modes of the density distribution are partitioned using an Expectation-Maximization algorithm (EM) [148]. Gaussian Mixture clustering [149] is probably the most famous example. Again, the drawback of such approaches is that in general the number of clusters is an input parameter of the model.

When no prior assumption can be made about the number of clusters, one needs to make use of non-parametric methods. The term non-parametric comes from the field of statistical inference, where models are classified as parametric and non-parametric, depending on whether the model assumed to describe the data has a fixed or variable number of parameters. Clustering schemes like k-means and GMM are parametric algorithms, since the number of clusters K (and all the corresponding hyperparameters of the model) needed to represent the data, are fixed *a priori*.

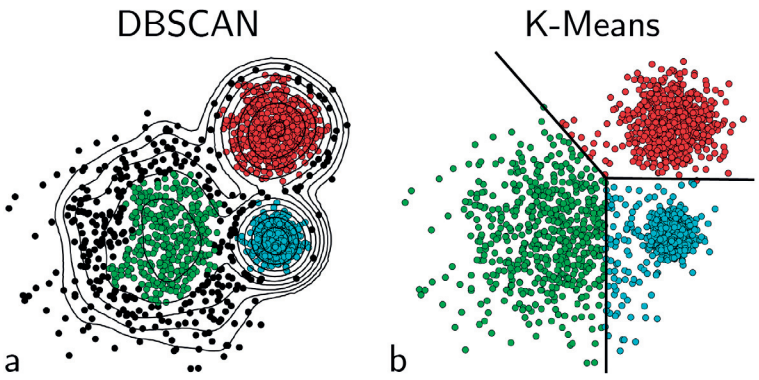


Figure 3: Clustering of points sampled from a mixture of three bivariate Gaussians: (a) DBSCAN groups the point according to the density, while (b) K-means partition the data as the Voronoi tessellation of the final centroids, which number is decided *a priori*. Notice that in (a) part of the points result as not classified (black dots).

The most well-known non-parametric density-based algorithm is Density-based spatial clustering of applications with noise (DBSCAN) [150]. DBSCAN makes use of two main hyperparameters, ϵ (neighbors cut-off distance) and n_m (the minimum number of points in a cluster) to classify the points of a dataset according to the following types:

- core: points having at least n_m neighbors within ϵ,
- border: points within ϵ from a core point but with a number of neighbors lower than n_m in a radius ϵ,
- noise: points that do not match either of the two previous types.

The clusters are formed by merging each of the core points with the points within its cut-off radius.

DBSCAN is very fast and popular, and many different variants have been proposed [151, 152]. One of the possible drawbacks is that part of the dataset could be unclassified.

Another interesting density-based non-parametric approach is the mean shift algorithm [153], which is based on a mode-seeking scheme, meaning that, given a certain estimate of the density function underlying the data ($f(\mathbf{u})$), the modes are found by seeking the points of the feature space located at $\nabla f(\mathbf{u}) = 0$, through a gradient ascent scheme.

Mean shift exploits kernel density estimation (detailed in Sec. 3.2.2) to obtain a smooth estimate of the probability density function (PDF) underlying the data ($\hat{f}(\mathbf{x})$). Each point $\mathbf{x}_i$ of the dataset is clustered by iterating two steps:

1. $\mathbf{y}_i^0 = \mathbf{x}_i$

2. $y_i^{t+1} = \frac{\sum_{i=1}^{N} K_h(\mathbf{y}_i^t - \mathbf{x}_i)\mathbf{x}_i}{\sum_{i=1}^{N} K_h(\mathbf{y}_i^t - \mathbf{x}_i)}$, where k_h is a univariate Gaussian kernel.

Starting from a point, the two steps are iterated till the closest stationary point is reached. In this way each point of the dataset is linked with the closest local mode.

Carreira-Perpiñán showed that mean-shift can be viewed as a generalized EM algorithm [154]. Notice however that the number of clusters is not specified a priori.

The main issue of mean shift is that the quality of the result strongly depend on the quality of the density estimate.

Another well-established type of partitional scheme is spectral clustering, which is a class of methods based on the solution of an eigenvalue problem built from the similarity $\mathbf{S}$ matrix of the data. The idea is to apply a standard clustering algorithm to the relevant eigenvectors of $\mathbf{S}$. An example is the Perron Cluster Analysis method (PCCA) [155–158], which is widely used in atomostic simulations for the identification of the metastable states used in Markov State Models.

3.2 DENSITY ESTIMATION

In unsupervised schemes, the learning step is usually built upon a proper probabilistic model for the data. When no prior assumptions can be expressed for the functional form of density function, one must use non-parametric techniques to estimate it.

The most common approaches for non-parametric density estimation are histograms, kernel density estimators, k-next-neighbors methods, and mixture models, such as Gaussian Mixtures, but other methods have been proposed, such as non-parametric wavelet density estimators [159], orthogonal series estimators [160] and penalized maximum likelihood estimators [161]. Good references for density estimation are [162] and [163].

3.2.1 *Histograms*

Histograms are the standard tool to visualize the distribution of values contained in a dataset, and also a simple – yet very powerful – way to estimate the probability distribution for a continuous variable [164].

The idea behind histograms as density estimators corresponds to the natural definition of the density as the number of observations

corresponding to a specific value, normalized by the total number of observations N:

$$\hat{f}(x) = \frac{1}{N} \sum_{i=1}^{N} \delta(x - x_i) \tag{26}$$

where δ is the Dirac delta function.

A frequency histogram can be viewed as a set of non-overlapping intervals, namely the bins, which serve as a counter for the number of points falling in specific portions of the space. If N_j is the number of observations falling in the j^{th} interval, the density histogram estimator for a certain x (inside the same interval) is defined as:

$$\hat{f}(x) = \frac{N_j}{Nh}. \tag{27}$$

Usually all the bins must have the same width h, to facilitate countings across different bins, which implies that a histogram is entirely defined simply by h and its boundaries.

A simple way to estimate the optimal number of bins k, given a set of N points, is Sturges' rule:

$$k = 1 + \log_2 N. \tag{28}$$

Sturges' rule is the standard rule usually introduced in statistics textbooks and is (often) the default choice in statistical packages. When the underlying distribution is skewed and far from normal, additional bins are usually required. Doane [165] proposed increasing the number of bins by the factor $\log_2\left(1 + \gamma\sqrt{N/6}\right)$, where γ is the standardized skewness coefficient of the data.

Alternative rules focus on deciding the optimal bin width h, e.g. Scott's rule [166]:

$$h = \left[\frac{24\sqrt{\pi}\sigma^3}{N}\right] \approx 3.5\sigma N^{-\frac{1}{3}}, \tag{29}$$

where σ is the sample standard deviation.

The use of histograms as density estimators is very practical and easy to implement. The drawbacks are the fact they are non-smooth estimators of the density and that the result is strongly dependent on the choice h. Another problem is the *curse of dimensionality*, since the size of a uniform grid scales exponentially with the number of dimensions, implying high computational and storage costs.

3.2.2 *Kernel density estimation*

Kernel density estimation (KDE) [1] is a non-parametric way to estimate the probability density function for a random variable. Con-

1 KDE is also known as Parzen-window density estimation, after its inventor E. Parzen, who introduced it in 1962 [167]

sider a set of N points $\{x_1, \ldots, x_N\}$, with $x_i \in \mathbb{R}^D$. The points are sampled from a distribution with an unknown density function f.

The kernel density estimator of f is defined by:

$$\hat{f}(\mathbf{x}) = \frac{1}{Nh^D} \sum_{i=1}^{N} K(\frac{\mathbf{x} - \mathbf{x}_i}{h}), \tag{30}$$

where h is the *bandwidth* and K is a symmetric, non-negative function that integrates to one, namely the kernel. Depending on the type of data, different choices can be optimal for the kernel, and common functions include Gaussian, box, triangular, Epanechnikov, and biweight functions shown in fig. 4.

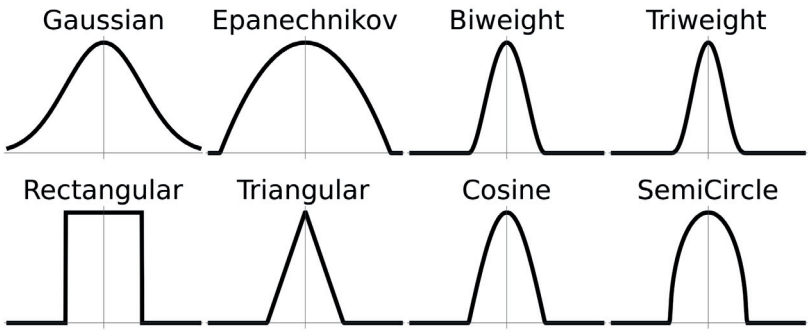


Figure 4: Some of the most commonly adopted kernels for KDE.

Cacoullos [168] and Epanechnikov [169] introduced the use of multivariate kernels:

$$\hat{f}(\mathbf{x}) = \frac{1}{N|\mathbf{H}|} \sum_{i=1}^{N} \mathbf{K}\left(\mathbf{H}^{-1}(\mathbf{x} - \mathbf{x}_i)\right), \tag{31}$$

where $\mathbf{H}$ is a positive definite $D \times D$ bandwidth matrix, which can be interpreted as the covariance matrix of $\mathbf{K}$.

An alternative, less general, method is the product kernel density estimator, which takes the form:

$$\hat{f}(\mathbf{x}) = \frac{1}{Nh_i \ldots h_D} \sum_{i=1}^{N} \left\{ \prod_{j=1}^{D} K(\frac{x_j - x_{ij}}{h_j}) \right\}, \tag{32}$$

However, in general, the choice of the functional form of the kernel is relatively unimportant and not crucial to the accuracy of the final estimates. The only parameter that eventually influencies the quality of the estimator $f(\hat{\mathbf{x}})$ is the bandwidth h [163].

The equations above have the general form of fixed-bandwidth KDE. The terminology fixed-bandwidth derives from the fact that h is kept constant for all $\mathbf{x} \in \mathbb{R}^D$, meaning that, at a specific point $\mathbf{x}$, the estimate of the density is the average of identical kernels, centered in each data point and scaled by h. The choice of the optimal

value for a fixed h is a very delicate task: a small value of h could lead to over-structuring, with spurious, false peaks due to the noise, while a large value for h usually wipe out all the details, leading to a significant over-smoothening of the peaks in the final estimate (see fig. 5).

The standard procedure to select the optimal value of h would be to perform many different KDE varying the window width and eventually analyze which one results being the best estimator. This can be done minimizing the average L_2 risk function, also known as the Mean Integrated Squared Error (MISE) test:

$$\mathrm{MISE}(h) = E\left[\int \left(\hat{f}(\mathbf{x}) - f(\mathbf{x})\right)^2 d\mathbf{x}\right]. \tag{33}$$

A faster alternative, proposed by Silverman [170], is to use asymptotic analysis and consider the h which minimize the AMISE – namely the asymptotic MISE for a number of trials $n \to \infty$ – which is:

$$h_{\mathrm{AMISE}} = \left[\frac{R(K)}{nR(f'')\left(\int K(\mathbf{x})d\mathbf{x}\right)^2}\right]^{\frac{1}{5}}, \tag{34}$$

where R is a functional corresponding to $R(\phi) = \int (\phi(\mathbf{x}))^2 d\mathbf{x}$. Even though in some cases the AMISE trick is a good approximation, Marron and Wand [171] have shown that in general it can be very poor, since it takes sample sizes in the order millions to have a good approximation, even when the KDE is done in low dimensions. Furthermore, as the ideal density function $f(\mathbf{x})$ is usually unknown and no prior assumptions can be made, in most of the real life examples, h cannot be estimated by minimizing the bias.

A variety of automatic, data-based methods have been developed for selecting the optimal bandwidth [163], with the rule-of-thumb when using univariate Gaussian kernels being that proposed by Silverman [170]:

$$h = \left(\frac{4}{(D+2)N}\right)^{\frac{1}{D+4}} \sigma, \tag{35}$$

where σ is an estimate of the standard deviation of the dataset and N the number of points.

Silverman's rule is the choice that minimizes the MISE when the data points are drawn from a univariate Normal distribution.

Another very common rule is the variation proposed by Terrell and Scott [172] and known as Scott's rule

$$h = N^{-\frac{1}{D+4}} \sigma, \tag{36}$$

which slightly over-smooths the final result.

While these rules are easy to compute, they should be used with caution as they can give very inaccurate estimates when the data is not normally distributed, as shown in fig 5. A detailed survey on bandwidth selection can be found in Ref. [173, 174].

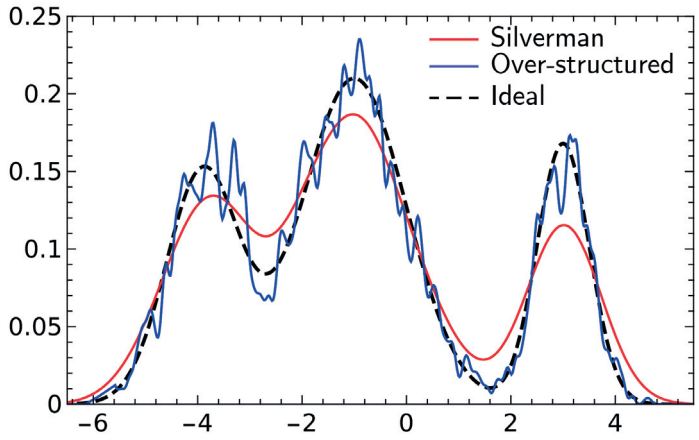


Figure 5: Ideal probability density composed by the mixture of three normal distributions (black dashed line). In red, the fixed global KDE computed from 10^3 points using Silverman's rule to select the optimal bandwidth: one can notice how the final estimate is over-smoothed. The blue line is the KDE computed using a bandwidth value that is too small. This too small bandwidth produces an undesired over-structuring in the final estimate.

Fixed-bandwidth KDE leads in general to very poor estimates of the density function, especially when data exhibits multi-modality.

A more promising approach consists of the adoption of an adaptive (or variable) bandwidth, where h is varied depending upon the location of the estimate or the samples. In principle, where $f(\mathbf{x})$ is large in magnitude – which corresponds to a region of a high density of data points – h should be small, while in the regions where f is close to zero, h should be very large, to account for the lack of statistics.

Adaptive KDE methods can be divided into two categories: balloon estimators and sample point estimators.

In the former, a different bandwidth is used for each estimation point $\mathbf{x}$:

$$\hat{f}(\mathbf{x}) = \frac{1}{N h(\mathbf{x})^D} \sum_{i=1}^{N} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h(\mathbf{x})}\right), \tag{37}$$

while in sample point estimators a different h is used for each sample point:

$$\hat{f}(\mathbf{x}) = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{h_i^D} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h_i}\right), \tag{38}$$

From a pointwise point of view, balloon estimators behave exactly as fixed kernel estimators.

A big drawback of balloon estimators is that, when considered as a global estimate, the estimator does not integrate to one.

Various strategies have been explored to automatically set the proper local value of h, which, as a starting point, is usually set to be the next-neighbour distance.

In the specific case of sample point estimation, Abramson [175] proposed a two step scheme where, starting from a pilot fixed kernel global estimator $\hat{f}(\mathbf{x})$, the local bandwidths are re-scaled following the rule $h(\mathbf{x}_i) \propto \hat{f}(\mathbf{x}_i)^{1/2}$.

Silverman [170] extended this approach and proposed a three-step global implementation of an Abramson-like estimator, where after computing a global pilot estimator, with a fixed h, one defines local bandwidth factors as:

$$\lambda_i = \left\{\hat{f}(\mathbf{x}_i)/g\right\}^{\alpha}, \tag{39}$$

where g is the geometric mean of $\hat{f}(\mathbf{x}_i)$ and $0 \leqslant \alpha \leqslant 1$ is called the sensitivity parameter. In this way the adaptive kernel is built as:

$$\hat{f}(\mathbf{x}) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{(h\lambda_i)^D} K\left(\frac{\mathbf{x}-\mathbf{x}_i}{h\lambda_i}\right). \tag{40}$$

Compared with the global procedure, an adaptive scheme will in general lead to an improved estimator. However, poorly constructed adaptive schemes can be inferior to global fixed approaches. This is especially true for sample point estimation, which, as shown by Teller and Scott [176], can suffer from a "non-locality" phenomenon, meaning that the estimate at any point can be strongly influenced by data very far away. This can be particularly delicate in high-dimensions, since, as the dimensionality increases, the probability mass of a distribution tend to move toward the tails (and thus the value of kernel with a large bandwidth will decay more slowly in high dimensions).

3.2.3 *KNN Density Estimation*

An alternative approach to the use of kernels for local density estimation, is the next-neighbors method, namely the KNN density estimation.

Once has been established the number of neighbors (K) needed to estimate the density properly, the density at a point $\mathbf{x}$ can be estimated using the formula:

$$\hat{f}(\mathbf{x}_i) = \frac{K/N}{V_i^{NN}}, \tag{41}$$

where V_i^{NN} is the volume occupied by the K neighbors and N is the total number of points.

The volume is what determine the final density – since K and N are constants – in analogy with the bandwidth parameter in KDE, in fact, KNN methods can be used to estimate the local bandwidth in adaptive KDE.

KNN methods are particularly robust when working with high-dimensional problems, even though the problem of determining the volume of the neighborhood in a generic high-dimensional manifold, can be non-trivial. Furthermore, the results of the clustering are strongly dependent on the choice of K.

A standard way of choosing a value for K is by *hyperparameter optimization* [177, 178], which is a very common practice in machine-learning, for setting the optimal parameters of a learning model, even if it is sometimes impractical when the problem becomes complex.

3.2.4 *Gaussian Mixtures*

Mixture models [179] provide flexible representations for densities that can be used to model heterogeneous data in high dimensions.

A Gaussian Mixture Model (GMM) is a probabilistic model to represent a complex probability density function as a linear combination of (K) Gaussians representing different sub-populations in the dataset:

$$\hat{P}(\mathbf{x}) = \sum_{k=1}^{K} p_k G\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right), \tag{42}$$

where the p_ks are the so-called mixing coefficients (the weight associated with each Gaussian, which are positive and $\sum_{k=1}^{K} p_k = 1$) and normalized such that $G\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)$ is a multivariate D-dimensional Gaussian distribution

$$G\left(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{\sqrt{(2\pi)^D \det|\boldsymbol{\Sigma}|}} e^{-(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})/2}, \tag{43}$$

with mean $\boldsymbol{\mu} \in \mathbb{R}^D$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{D\times D}$.

GMM is a semi-parametric scheme, since the number of Gaussians needed to model the data (K), must be decided *a priori*, while the parameters of the model $(p_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k))$ are learned from a training dataset through an expectation-maximization (EM) technique, which recalls the iterative two-step scheme introduced with K-means.

Starting from a dataset $\mathcal{X}$, made up of N feature vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, with $\mathbf{x}_i \in \mathbb{R}^D$, after defining some reasonable initial parameters for the Gaussians, one must iterate an expectation step (E-step) and a

maximization step (M-step) until the log-likelihood $\mathcal{L}$ of the model is maximized:

$$\log\mathcal{L} = \sum_{i=1}^{N} \left(\log \sum_{k=1}^{K} p_k G\left(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right). \tag{44}$$

The performance of the EM algorithm depends strongly on the choice of the initial parameters.

Another approach to optimize the paremeters is through *variational Bayesian inference* [180].

GMMs are commonly used as a parametric model to fit multimodal, asymmetric density functions and, from a machine-learning point of view, they can be exploited as unsupervised learning schemes for model-based clustering and classification tasks.

3.3 BAYESIAN CLASSIFIERS

A Bayesian classifier is a simple probabilistic function which is used to classify the data by minimizing the probability of misclassification [181]. The name comes from the fact that the classification task builds upon the application of Bayes' theorem, under the (strong) assumptions of independence between the input features – which is why they also go by the name of *naive* Bayesian classifiers.

The Bayesian framework

Bayes rule is a simple, yet very powerful, method for dealing with uncertainty in a model, and it can be expressed, in words, as:

$$\text{posterior} = \frac{\text{prior} \times \text{likelihood}}{\text{evidence}}, \tag{45}$$

which can be easily derived from basic probability theory. The prior can be seen as the probability reflecting the degree of belief about an event, before having had any evidence about it. The posterior is the updated probability for the event after having taken into account the information collected about it.

Consider a pair of random variables x and y taking values in some spaces $\mathcal{X}$ and $\mathcal{Y}$, respectively. Knowing the joint probability of the two variables $P(x, y)$ one can extract the marginal probability of x by summing (or integrating, if they are continuous variables) over all possible values of y:

$$P(x) = \sum_{y \in \mathcal{Y}} P(x, y). \tag{46}$$

Assuming that the two variables are independent one can decompose $P(x, y)$ as

$$P(x, y) = P(x)P(y|x) = P(y)P(x|y), \tag{47}$$

where $P(x)$ is the marginal probability of x and $P(x|y)$ is the conditional probability of x given a certain value of y (and the other way around for $P(y)$ and $P(y|x)$).

Combining and rearranging these two equations one can easily obtain the Bayes rule:

$$P(y|x) = \frac{P(y)P(x|y)}{P(x)}, \tag{48}$$

where in a practical implementation x might be a data point and y some model parameters.

From the Bayesian point of view, $P(y)$ represents the prior probability of the model y before knowing anything about x. $P(x|y)$ is the likelihood of the model. $P(y|x)$ is the posterior probability of y after observing x, and $P(x)$ is the normalizing constant, ensuring that the left-hand side is a valid probability distribution.

From the point of view of pattern recognition, $P(y|x)$ can be seen as a smooth function, varying from zero to one, which can be used as a fingerprint function to partition the domain of x into distinct classes (patterns or labels).

Assume now that we have a dataset $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, with $\mathbf{x}_i \in \mathbb{R}^D$ and a set of labels $\mathcal{Y} = \{1, \ldots, K\}$, a classification task is the process of associating each point of $\mathcal{X}$ with one of the elements of its class label $\mathcal{Y}$. This can be done easily through a Bayesian classifier, which combines a probabilistic model for the data with a decision rule. For instance, if the data is modeled through a Gaussian Mixture Module, with K multivariate Gaussians composing the mixture and corresponding to the "labels" used to partition the dataset, one can use the *maximum a posteriori* decision rule to associate a point $\mathbf{x}$ with the most probable Gaussian from which it was sampled as:

$$C^{Bayes}(\mathbf{x}) = \arg\max_{k \in \mathcal{Y}} P_{GMM}(k|\mathbf{x}), \tag{49}$$

where $P_{GMM}(k|\mathbf{x})$ is the Bayesian class posterior probability (referred to as responsibility) defined as:

$$P_{GMM}(k|\mathbf{x}) = \frac{p_K G(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} p_k G(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}. \tag{50}$$

3.4 DIMENSIONALITY REDUCTION

Dimensionality reduction (DR) has the goal of simplifying data so that it can be efficiently processed or visualized.

Consider a data set represented by an $n \times D$ matrix $\mathbf{X}$ where the n rows are data vectors $\mathbf{x}_i$ having dimensionality D. Assume that $\mathbf{X}$ has intrinsic dimensionality d, with $d < D$, meaning that the distribution of points $\mathbf{X}$ lies on or near a d-dimensional manifold embedded in

the D-dimensional space underlying $\mathbf{X}$. DR aims to map $\mathbf{X}$ into a new d-dimensional data set $\mathbf{Y}$, conserving the geometry of the data as much as possible.

Reducing dimensionality can be useful for characterizing variability, removing unimportant degrees of freedom, and discovering better representations of complex data. The latter is probably the most common application of DR methods.

In atomistic simulations, for instance, the degrees of freedom characterizing a system can be on the order of millions, which often makes the reduction of dimensionality necessary for obtaining an intuitive picture of the physics hidden in the data.

Several linear and nonlinear approaches have been developed to derive a meaningful low-dimensional mapping of high-dimensional spaces. The most common approach is probably principal component analysis (PCA) [182], which is a linear projection algorithm that reduces the dimensionality of the data set by projecting the data on the eigenvectors of the covariance matrix with the largest eigenvalues. PCA assumes that the data lies on a linear d-dimensional subspace embedded in the full D-dimensional original space. Other well-known reduction algorithms are the linear multi-dimensional scaling (MDS) [183], and more advanced non-linear projections techniques, such as ISOMAP [184], diffusion maps [185], kernel PCA [186], Laplacian eigenmaps [187], locally-linear embedding [188], Hessian eigenmaps [189], t-SNE [190] among the many others.

Sketch-map

In this thesis, we will use Sketch-map [191], which is a non-linear DR method specifically designed to examine high-dimensional data produced from atomistic simulations.

Sketch-map can deal with the thermal fluctuations typical of metastable configurations in a tempered simulation.

The rationale of sketchmap is very similar to that of MDS, i.e. a low dimensional projection is found by iteratively optimizing the stress function

$$\chi^2 = \sum_{i \neq j} \left[s(R_{ij}) - s(r_{ij}) \right]^2 , \tag{51}$$

which represents the mismatch between the distances in high dimensions and the distances between the corresponding low-dimensional projections. The difference with MDS consists in the transformation s, which is a non-linear switching function with a sigmoidal form:

$$s(r_{ij}) = 1 - \left(1 + \left(2^{a/b} - 1 \right) \left(r_{ij}/\sigma \right) \right)^{-b/a} , \tag{52}$$

where σ is the threshold of the switching function, which can be used to tune the length scale of the problem. The function $s(r_{ij})$ transforms to zero distances that are characteristic of fluctuations within

the same meta-stable state, and to one the distances between configurations that are completely unrelated. The exponents a and b have a relatively small effect on the projection and control the steepness of the sigmoid.

Since the transformation is applied in both the high and low dimensional spaces, this is equivalent to requiring that points that are close together stay close in the projected space, and configurations that are far apart from each other are projected in separate regions. This is a much simpler task than matching distances, and the iterative optimization (which has to be used since eq. 51 cannot easily be expressed as an eigenvalue problem) can focus on representing correctly the connectivity between nearby basins, which is arguably the most important requirement to obtain a meaningful representation of the configuration space of a compound at the atomic scale. Since the iterative minimization of eq. (51) is not trivial and very expensive, it is important to start from a good selection of reference configurations (the landmarks). Then, the projection of the other data points can be obtained based on these reference configurations using out-of-sample embedding [110].

For a comprehensive overview of dimensionality reduction methods, we refer the reader to refs. [192–194].

4

PROBABILISTIC ANALYSIS OF MOLECULAR MOTIFS

Contents

In this chapter we will introduce a general framework to analyze the results of atomistic simulations automatically and identify recurrent structural motifs in an unbiased and quantitative fashion.

The main purpose is to demonstrate that machine-learning routines can assist the interpretation of *in silico* experiments, in particular with the rationalization of recurring structural patterns.

Applying unsupervised-learning techniques and using exclusively structural information to guide the analysis, makes it possible to work with results coming from different levels of theory, or even from experiments.

This is at odds with other analysis techniques proposed in the literature, that rely on specific types of data, such as those coming from *ab initio* methods. An example is ALMO [195], which is based on an energy decomposition on top of an electronic structure calculation.

The name Probabilistic Analysis of Molecular Motifs reflects the core idea behind the method: the estimation of the stability of structural motifs is inferred by computing their probability distribution from a simulation. One could also use experimental data to train the model, for example by parsing X-Ray or NMR structures deposited in a database. This is justified by the fact that the (free) energetic stability of different molecular patterns is implicit in the frequency with which they appear and in line with the free-energetic interpretation.

This chapter is adapted from refs. [1] and [2]

Distinct clusters (patterns) are separate modes of the distribution, corresponding to the basins of attraction of local maxima.

The PAMM workflow

A PAMM analysis is based ideally on a preliminary atomistic simulation, and its general workflow – which is schematically represented in Fig. 6 – can be split into three main steps:

1. *Training set definition*: Given a set of structures, for instance, a trajectory from an MD simulation, a training dataset is built by selecting a set of representative structures, which are then mapped to some (possibly) high-dimensional feature space. This can be done by describing the groups of atoms that might be involved in recurring molecular patterns with the set of their inter-atomic distances or, more generally, with abstract descriptors based on some expansions of atomic environments. This step yields a set $\mathcal{X} = \{\mathbf{x}_i\}$ that contains N vectors of dimensionality D, that represent the molecular configurations observed in the simulation.

2. *Model definition*: A kernel density estimation is used to evaluate the PDF of $\mathcal{X}$, which is then analyzed to recognize the different modes of the distribution by applying a partitioning clustering algorithm. This splits $\mathcal{X}$ into n disjoint clusters which are used to fit a Gaussian mixture model. This procedure provides a probabilistic framework that is capable of associating regions of the D-dimensional feature space to one or more recurring patterns.

3. *Pattern classification*: the mixture model can then be used to give qualitative and quantitative insight on the system being studied. It can also possibly serve as a basis for defining more complex order parameters to describe and bias collective rearrangements of the various molecular patterns when used in combination with accelerated sampling methods, such as metadynamics and umbrella sampling.

[1] PAMM is designed to work with atomistic simulation data and to overcome the limitations of common generic methods such as K-means or DBSCAN.

A detailed discussion of each step in the algorithm will follow in the next sections.

1 A basic implementation of PAMM is available at `https://github.com/cosmo-epfl/pamm`, together with a series of Python scripts to process simulation trajectories

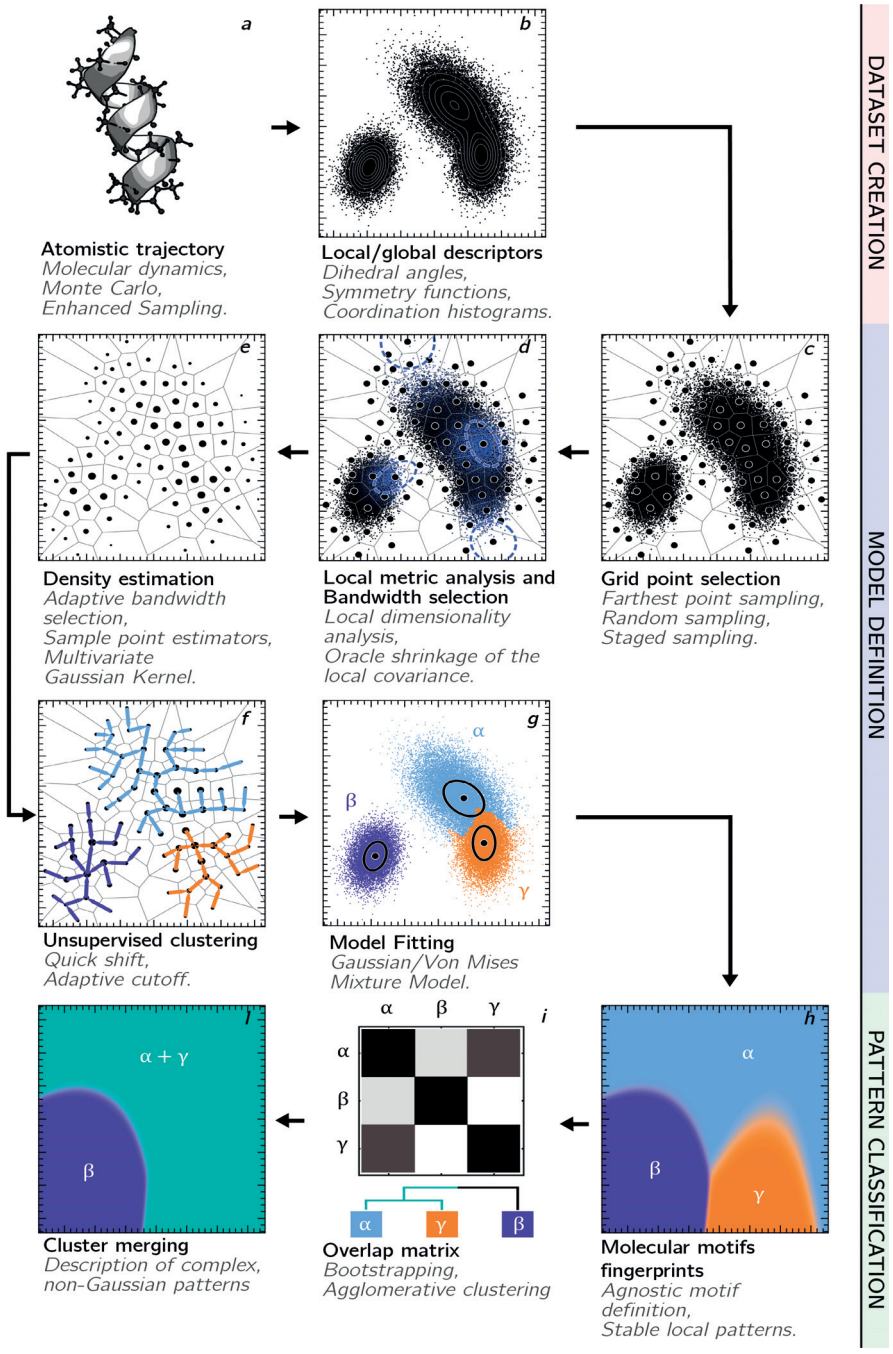


Figure 6: Schematic illustration of the PAMM workflow applied to an artificial two-dimensional dataset.

4.1 DEFINITION OF THE TRAINING SET

4.1.1 *Feature space representation*

The analysis we propose starts by introducing descriptors of the relative arrangement of groups of atoms, for which one wishes to identify recurring patterns. For most applications, it is impractical to directly use the Cartesian coordinates of the atoms to describe structures or local environments. One should preferably represent them in terms of a (possibly large) number of "order parameters" or "fingerprints", that provide an unbiased and sufficiently complete description of the geometry while fulfilling all of the important symmetries – such as being invariant to atom labelling or to rigid rotations and translations [137, 140, 196]. For example, if one wanted to recognize the existence of a bond between two atomic species, for instance, one could process the configurations from an atomistic simulation to output the list of distances between the pairs of atoms of the two species.

For more complex structural patterns, one could pick all the possible tuples of atoms of a few selected species, and describe them in terms of all the pair-wise distances among them, possibly sorting groups of distances to account for the permutation of identical atoms [197]. Seen through this lens, the atomistic simulation data is converted to a set $\mathcal{X} = \{\mathbf{x}_i\}$ containing a large number N of D-dimensional vectors, $\mathbf{x}_i \in \mathbb{R}^D$, with each vector representing either a subset or the entirety of the atoms within a structure.

It is worth stressing that the selection of a proper set of descriptors is far from being a trivial point, and it can influence the outcome of the subsequent analysis deeply. To mitigate this problem, we optimized the different ingredients in PAMM to be robust with growing dimensionality, so that many order parameters can be used simultaneously to deal, e.g., with heterogeneous systems.

One could also use more abstract descriptors that are based on more or less systematic expansions of atomic environments – for instance Behler-Parrinello symmetry functions [139], SOAP power spectra [140], SPRINT coordinates [109]. The choice of input order parameters is also important, because it determines the metric relative to which probability distributions and free energies are computed [198, 199]. A poor choice can generate spurious maxima in the probability distribution, or merge kinetically separate states into a single basin. These artifacts can be corrected, at least in part, by taking into account kinetic information in the definition of the order parameters [200, 201].

4.1.2 *Grid Selection*

To mitigate the high cost of density estimation for large datasets in high dimension, the first step in the PAMM workflow involves the selection of subset $\mathcal{Y} \subseteq \mathcal{X}$ containing $M \ll N$ points. One could think about doing the KDE on a uniform grid. This is impractical in our case, however, as PAMM should in principle work with very high-dimensional spaces. The problem is that the size of a uniform grid grows with the number of dimensions.

A better idea is thus to use sparse grids. A grid covering almost uniformly the parameter space spanned by $\mathcal{X}$ can be obtained using a greedy farthest-point sampling (FPS) procedure [202], using a *minmax* criterion [203] to select the points.

The first point $\mathbf{y}_1 \in \mathcal{X}$ is chosen randomly, and then iteratively we repeat

$$\mathbf{y}_{j+1} = \arg\max_{\mathbf{y}\in\mathcal{X}} \left[\min_{i\leqslant j} |\mathbf{y}_i - \mathbf{y}| \right]. \tag{53}$$

Each new point in the sample is chosen that it has the maximal minimum distance to the points that have already been selected. The procedure can be repeated until M points had been chosen. These M points then form a sparse grid on which the probability can be estimated (Figure 6(c)). The computational cost of the FPS selection is $\mathcal{O}(MN)$, so it can also be performed on huge datasets.

The determination of the grid by sub-sampling the full set $\mathcal{X}$ also allows one to partition the data into neighborhoods of the grid. For instance, one can construct the Voronoi polyhedra for $\mathcal{Y}$, and assign each datum x to the Voronoi set $\mathcal{V}_i$ of the closest-by grid point y_i. Different strategies for subsampling are also possible [93, 204]. It should be stressed, however, that the subsequent steps in the PAMM workflow are designed to minimize the impact of the grid size on the final outcome, and to guarantee that in the limit $M, N \to \infty$ there is no dependence at all.

4.2 MODEL DEFINITION

4.2.1 *Kernel density estimation*

Optimal bandwidth Selection

Density-based clustering algorithms depend crucially on the quality of the estimation of the underlying probability density. [205, 206] Kernel-density estimation provides a smooth, robust approach for estimating this probability density, that also leaves the flexibility to adapt to strongly anisotropic probability distributions and/or non-Euclidean geometries.

Recalling eq. 38, the KDE on a grid point $\mathbf{y}_i$ can be written as

$$P(\mathbf{y}_i) = \frac{1}{\sum_{j=1}^{N} w_j} \sum_{j=1}^{N} w_j K_{\mathbf{H}_j}(\mathbf{x}_j - \mathbf{y}_i), \tag{54}$$

where we use a very general expression in which each data point can be assigned a weight w_i (e.g. to compensate for biased sampling), and an adaptive bandwidth matrix $\mathbf{H}_j$.

The use of a grid implies that (for a fixed grid) the cost of evaluating P scales only linearly with the number of data points.

We use an anisotropic multivariate Gaussian kernel,

$$K_{\mathbf{H}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^D |\mathbf{H}|}} \exp\left[-\frac{1}{2}\mathbf{x}^T \mathbf{H}^{-1} \mathbf{x}\right], \tag{55}$$

that provides enough flexibility to adapt to strong variations of the geometric distribution of data points.

As mention before in Sec. 3.2.2, a common problem with KDE is the optimization of the bandwidth of the kernel. This is particularly severe for high-dimensional and/or sparsely populated datasets.

The shape of the kernel, encoded in the bandwidth matrix $\mathbf{H}$ in eq. (55), determines a trade-off between the statistical noise in the estimated density and a systematic error due to the smoothing of the true underlying density.

Since the true density is not known in general, the optimal bandwidth cannot be selected by minimizing the MISE. One has to resort to recipes to choose the bandwidth that are derived based on some reasonable assumptions about the underlying distribution. A particularly simple heuristic for selecting the bandwidth is the multivariate extension of Silverman's rule,

$$\mathbf{H} = \left[\frac{4}{N(D+2)}\right]^{\frac{2}{D+4}} \boldsymbol{\Sigma}, \tag{56}$$

where $\boldsymbol{\Sigma}$ is the covariance of the entire dataset, N the number of data points and D the dimensionality.

Local Kernel Metric Analysis

Silverman's rule (56) minimizes the MISE for a single multivariate normal distribution. For a general distribution, composed of many separate peaks, this generally results in an over-estimation of the bandwidth, and in loss of resolution unless an extremely large amount of data is available. As a possible solution, and to provide a mechanism to fine-tune the balance between resolution and statistical noise, we propose a simple strategy to localize the determination of the bandwidth and the dimensionality of the data. Basically, the idea is to apply Silverman's rule to subsets of the full dataset.

In order to estimate the optimal KDE bandwidth for a sample $\mathbf{x}_i$, we introduce weighting factors for each of the other sample points $\mathbf{x}_j$ around it, that are computed from a spherical Gaussian

$$u_{ij} = \exp\left[-\frac{(\mathbf{x}_j - \mathbf{x}_i)^2}{2 \cdot \sigma_i^2}\right] N w_j / \sum_j w_j, \tag{57}$$

where σ_i is a localization factor whose choice we will discuss below. The weights-adjusted sample population for the selected point is computed as $N_i = \sum_j u_{ij}$. We find it convenient to introduce two possible approaches to determine the localization parameters σ_i.

1. in cases where one expects the spatial extent of clusters to be relatively homogeneous, one can choose a fixed localization window expressed as a fraction of the overall spatial extent of the dataset, assessed as the global covariance matrix of the data, $\mathbf{\Sigma}$; this can be achieved by setting $\sigma_i = f_s\sqrt{\mathrm{Tr}\,\mathbf{\Sigma}}$. In this case, each localization region can contain a different weight-adjusted population N_i.

2. in cases where one expects clusters with very different spreads, but similar populations, it might be more convenient to use a position-dependent localization window that is adjusted so that each region contains a prescribed fraction f_p of the total number of weighted data points. Each σ_i should then be adjusted iteratively until $N_i \approx N f_p$.

Depending on the problem, one strategy can perform better than the other.

Impact of the Localization on the KDE Accuracy

To assess the accuracy of the multivariate adaptive KDE schemes, in Fig. 7, we compare the MISE analysis for bimodal distributions of two skewed Gaussians in 2,4 and 10 dimensions using both the localization approaches to estimate the bandwidth matrices, and compare it with an identical adptive KDE scheme. However, instead of a multivariate kernel, a spherical one is used (as in eq. 38). By using a spherical Gaussian we refer to a bandwidth $\mathbf{H}_i = \mathbf{1}\sigma_i$, where σ_i corresponds to the localization parameter.

We also include in the comparison the more standard, naïve approach of setting the local kernel bandwidth to the next (grid) neighbour distance.

It is clear that (at least for this test case) a multivariate Gaussian kernel shows considerably better performance compared to a spherical Gaussian particularly when the dimensionality of the problem increases. For $f_p \approx 1$ or for $f_s >> 1$ both heuristics converge to the standard version of Silverman's rule, which results in oversmoothing of this bimodal distribution.

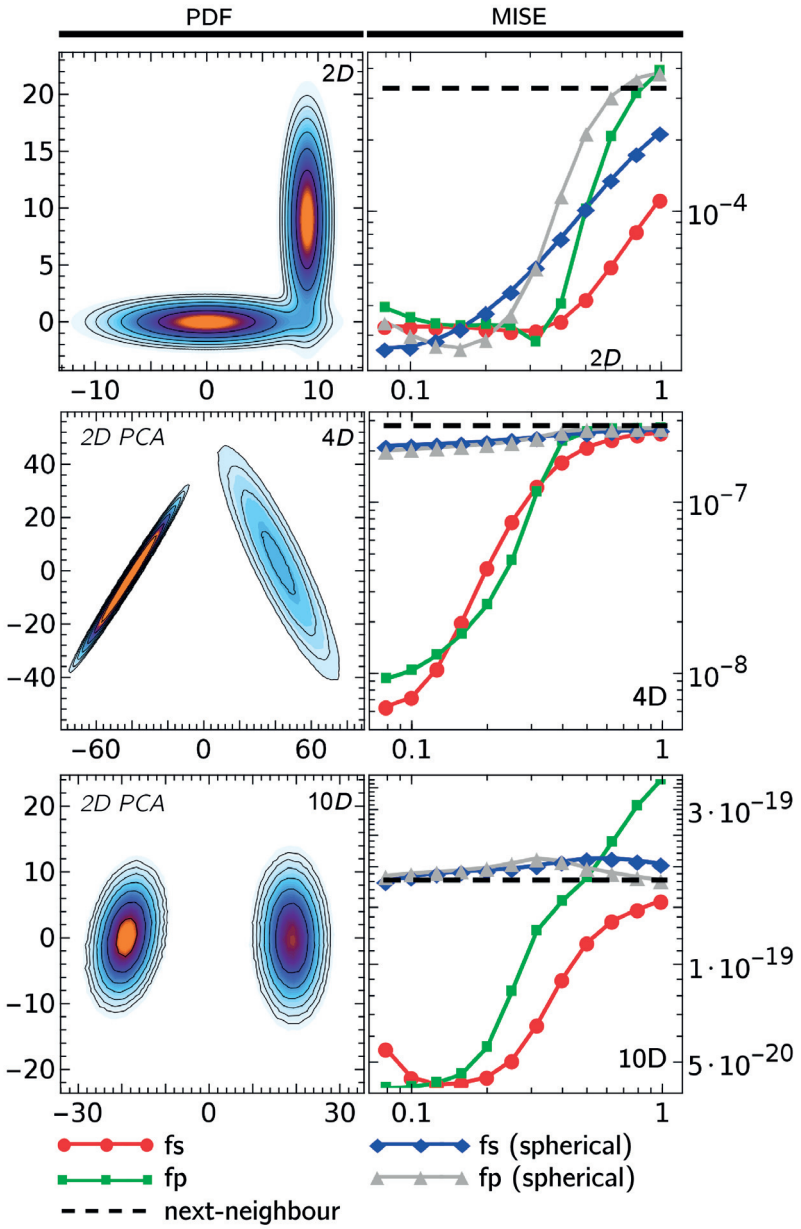


Figure 7: MISE analysis for different anisotropic Gaussian mixture distributions with increasing dimensionality. Left panels show the ideal PDFs, where, for $D > 2$, the corresponding 2D PCA projection is shown. Green and red lines represent the MISE obtained from the KDE computed using the new PAMM v2.0 scheme, scanning over various possible parameters of f_p and f_{spread} respectively. Gray and blue lines represent a modified version of the PAMM, where a 1D Gaussian kernel is used in place of the multivariate Gaussian kernel. Black dashed lines correspond to the results obtained using the next-neighbour distance as local bandwidth.

The local weights from eq. (57) can be used, for instance, to estimate the local covariance Σ_i, around each data point. Each element of the covariance is estimated using

$$[\boldsymbol{\Sigma}_i]_{kl} = \sum_{j=1}^{N} u_{ij} \cdot (x_{jk} - \bar{x}_k)(x_{jl} - \bar{x}_l)/N_i \tag{58}$$

where $\bar{\mathbf{x}} = \sum_{j=1}^{N} u_{ij}\mathbf{x}_i/N_i$.

Computing the covariance of a subset of the points can exacerbate stability problems that are also present, in general, when the sampling is insufficient or when different degrees of freedom are strongly correlated. In these circumstances, $\boldsymbol{\Sigma}_i$ can be very ill-conditioned, and can even have eigenvalues that are zero to within machine precision. This is a consequence of the fact that the usual estimator for the covariance $\boldsymbol{\Sigma}_i$ is a biased estimator of its inverse $\boldsymbol{\Sigma}_i^{-1}$. This is well-known problem and is typically addressed by introducing alternative estimators that are less strongly biased.

Here we use the so-called Oracle Approximating Shrinkage (OAS) estimator [207] that reads

$$\widetilde{\boldsymbol{\Sigma}}_i = (1 - \psi_i)\boldsymbol{\Sigma}_i + \frac{\psi_i \mathrm{Tr}(\boldsymbol{\Sigma}_i)\mathbf{I}}{D} \tag{59}$$

where

$$\psi_i = \min\left[1, \frac{\left(1 - \frac{2}{D}\right)\mathrm{Tr}(\boldsymbol{\Sigma}_i^2) + \mathrm{Tr}^2(\boldsymbol{\Sigma}_i)}{\left(N_i + 1 - \frac{2}{D}\right)\mathrm{Tr}(\boldsymbol{\Sigma}_i^2) - \frac{\mathrm{Tr}^2(\boldsymbol{\Sigma}_i)}{D}}\right]. \tag{60}$$

Furthermore, the eigenvalue spectrum of the local covariance matrix can be used to estimate an effective local dimensionality D_i based on the effective rank of $\boldsymbol{\Sigma}_i$ [208]:

$$D_i = \exp\left(-\sum_{k=1}^{D} \eta_k \log(\eta_k)\right) \tag{61}$$

where $\eta_k = \lambda_k / \sum_{k=1}^{D} |\lambda_k|$ and $\{\lambda_k\}$ is the eigenvalue spectrum of $\boldsymbol{\Sigma}_i$.

Given the local covariance matrices and an estimate of the local dimensionality, one can introduce an expression for the optimal bandwidth matrices to be used in the KDE. Assuming that each local zone resembles a normal distribution, the optimal bandwidth for $\mathbf{x}_i$ can finally be obtained as a localized version of Silverman's Rule (56):

$$\mathbf{H}_i = \left[\frac{4}{N_i(D_i + 2)}\right]^{\frac{2}{D_i+4}} \widetilde{\boldsymbol{\Sigma}}_i. \tag{62}$$

Performing this analysis for each datum would entail poor scaling of the procedure with the total number of points. One can however exploit the definition of a sparse grid to accelerate greatly the computation, without changing the spirit of the localization strategy. In

essence, one can first compute the local bandwidth $\mathbf{H}_i$ only for the grid points $\mathbf{y}_i$, and can then assign this to all the samples that belong to its Voronoi set, i.e. $\mathbf{H}_j \equiv \mathbf{H}_i$, $\widetilde{\boldsymbol{\Sigma}}_j \equiv \widetilde{\boldsymbol{\Sigma}}_i \;\; \forall \mathbf{x}_j \in \mathcal{V}_i$.

Furthermore, the evaluation of eq. (58) can be accelerated by including the contribution from grid points beyond a reasonable cutoff distance using only the position of the grid $\mathbf{y}_i$ and assigning to it a weight proportional to the total weight of points within its associated Voronoi polyhedron, rather than summing over all sample points associated to it.

Finally, we note that in cases in which sampling is particularly irregular, it can happen that the bandwidth is smaller than the distance to the nearest neighbor of a grid point. Often these outliers result from insufficient grid size and/or non-Gaussian tails of the distribution, and should not generate additional clusters. To avoid this, we increase automatically the bandwidth to match the first-neighbor distance, but issue a warning to allow for manual inspection to determine whether the outlier is of some significance.

Improving numerical stability of the KDE

One common problem when working with probability densities in high dimension is the enormous range of values they can span, which can lead to instabilities and numerical errors. To improve numerical stability in our implementation, we have used the logarithms of P throughout. Thus, if the sum or difference of probabilities should be calculated, we use the log-sum-exp (LSE) approximation [209]

$$\log \sum_i^M e^{\log P_i} = \log P_i^* + \log \sum_i^M e^{\log P_i - \log P_i^*}, \tag{63}$$

where $\log P_i^* = \max\{\log P_1, \ldots, \log P_M\}$ and $P_i \equiv P(\mathbf{y}_i)$. the KDE expression then becomes

$$\log(P_i) = \log \sum_j^N e^{\log K_{\mathbf{H}_j} + \log w_j} - \log \sum_j^N w_j \tag{64}$$

and the kernel function in the exponent can be evaluated using

$$\log K_{\mathbf{H}} = -\frac{1}{2}\left(+p \log 2\pi + \log|\mathbf{H}| + \mathbf{x}^T \mathbf{H}^{-1} \mathbf{x}\right) \tag{65}$$

This has two main advantages:

1. a numerically stable algorithm calculating logarithms of bandwidth matrix determinants, $\log|\mathbf{H}|$, is explicitly used in the kernel density estimation (cf. eq. (65)),

2. log-densities are stable even if unscaled densities exceed the natural maximum values possible for the specific data type used in the computation.

4.2.2 *Identification of Motifs*

After having estimated the density at the grid points $P(\mathbf{y}_i)$, one can proceed to use it to subdivide the distribution into several distinct clusters. As discussed above, we chose to identify clusters that represent recurring molecular patterns as maxima in the probability distribution, and to associate to each maximum all the grid points falling within its basin of attraction. In atomistic trajectories this construction has a profound physical interpretation: each identified maximum of the probability distribution can be associated with a free energy (meta-)stable minimum of the D-dimensional description of a group of atoms.

We perform a non-parametric clustering based on this idea using the Quick-Shift algorithm [210]. Starting from a random grid point which has not been assigned yet to a cluster, one connects it to the nearest grid point that has a higher probability density, i.e., $\mathbf{y}_i$ is connected to $\mathbf{y}_j$ such that

$$j = \underset{P(\mathbf{y}_j)>P(\mathbf{y}_i)}{\operatorname{argmin}} |\mathbf{y}_i - \mathbf{y}_j|. \tag{66}$$

The procedure can be interrupted based on a suitable stopping criterion, as discussed below. The final point is identified as a maximum, and tagged as the center $\mathbf{z}_k$ of a cluster. All the points in the chain are tagged as belonging to the associated set $\mathcal{Z}_k$. One can then start climbing from another unassigned point, and stop when the procedure encounters another maximum, or one of the points that have already been assigned to one of the $\mathcal{Z}_k$ clusters, with which the current chain would then be merged. After all points have been traversed, the grid $\mathcal{Y}$ will be partitioned into n disjoint sets, without the need of specifying *a priori* the number of clusters or their geometry. The stopping criterion for Quick-Shift is typically set by requiring that the Euclidean distance between the current point and the next one in the chain is below a set threshold Δ. In line with the spirit of the local metric analysis we perform to obtain the bandwidth matrix for the KDE, we select the cutoff locally. At each stage in the procedure, the cutoff is chosen based on the covariance associated with the grid point that is being considered, i.e.

$$\Delta_i = \alpha\sqrt{\operatorname{Tr}\widetilde{\boldsymbol{\Sigma}}_i}. \tag{67}$$

This choice adapts the cutoff to the local spread of the data, and is consistent with the Gaussian assumption that underlies the localization procedure. For a multivariate Gaussian, it would cluster together two points that are drawn at random from the distribution. If needed, the values of Δ_i can be further adjusted by a multiplicative factor α,

to fine-tune the resolution of the clustering procedure. In the examples we used to benchmark this method we found only small changes in the final clustering upon scaling the cutoff by $\pm 10\,\%$.

4.2.3 *Gaussian Mixture Model*

After having determined the number and position of modes of the distribution, we fit a Gaussian Mixture Model to the data. To avoid any ambiguity, we use a mixture of Gaussians with the purpose of modeling the PDF underlying $\mathcal{X}$, which is something conceptually different from GMM clustering, where the Gaussian parameters are found by optimizing the log-likelihood of the model. Our choice combines the simplicity of a Gaussian mixture model, the fuzzy, smooth nature of the posterior cluster probabilities, and the robust, deterministic partitioning of the probability density obtained by applying quick shift to an adaptive, optimally-tuned gridy-KDE.

The aim of this step is to provide a simple interpretation of the density in terms of a number of separate modes, which makes it possible to develop fingerprints for the different molecular motifs that have a transparent probabilistic interpretation. As introduced in Sec. 3.2.4, the probability distribution can be fitted to a sum of n multivariate Gaussians

$$\hat{P}(\mathbf{x}) = \sum_{k=1}^{n} p_k G(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{68}$$

where G is a multivariate Gaussian, associated with a weight p_k, with covariance matrix $\boldsymbol{\Sigma}$ and mean position $\boldsymbol{\mu}$. Rather than fitting the Gaussian parameters with an expectation-maximization algorithm, we exploit the fact that we know the number n and modes $\mathbf{z}_k$ of clusters. We set the mean of the Gaussian cluster to the mode of the cluster, and estimate the covariance with the usual expression:

$$\begin{aligned} \boldsymbol{\mu}_k = \mathbf{z}_k, \quad p_k &= \sum_{\mathbf{y}\in\mathcal{Z}_k} P(\mathbf{y}) \Big/ \sum_{\mathbf{y}\in\mathcal{Y}} P(\mathbf{y}), \\ \bar{\mathbf{y}}_k &= \sum_{\mathbf{y}\in\mathcal{Z}_k} \mathbf{y} P(\mathbf{y}) \Big/ \sum_{\mathbf{y}\in\mathcal{Z}_k} P(\mathbf{y}) \\ \boldsymbol{\Sigma}_k = \sum_{\mathbf{y}\in\mathcal{Z}_k} (\mathbf{y}-\bar{\mathbf{y}}_k)(\mathbf{y}-\bar{\mathbf{y}}_k)^T P(\mathbf{y}) &\Big/ \sum_{\mathbf{y}\in\mathcal{Z}_k} P(\mathbf{y}). \end{aligned} \tag{69}$$

The reason for introducing this additional step, after having already obtained a non-parametric clustering by quick-shift, is that a GMM lends itself quite naturally to a probabilistic interpretation. Given a configuration associated with the fingerprints $\mathbf{x}$, the expression

$$\hat{P}_k(\mathbf{x}) = p_k G(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big/ \left(\zeta + \hat{P}(\mathbf{x})\right) \tag{70}$$

corresponds to the probability that such configuration belongs to the kth cluster. Eqn. (70) can therefore be used to introduce some smooth, fuzzy Probabilistic Motif Identifiers (PMI) that constitute a data-driven definition of a molecular pattern – such as the hydrogen bond [1, 3] or the accumulation of charge around an excess proton [211].

The "background" parameter ζ – that defaults to zero and should in any case be set to a very small value – serves to provide a more physical description of outlier configurations. In practice, configurations for which all Gaussian densities are below ζ are considered to be new, unclassified states that had not been sampled properly in the initial dataset.

4.2.4 *Mixture models in periodic spaces*

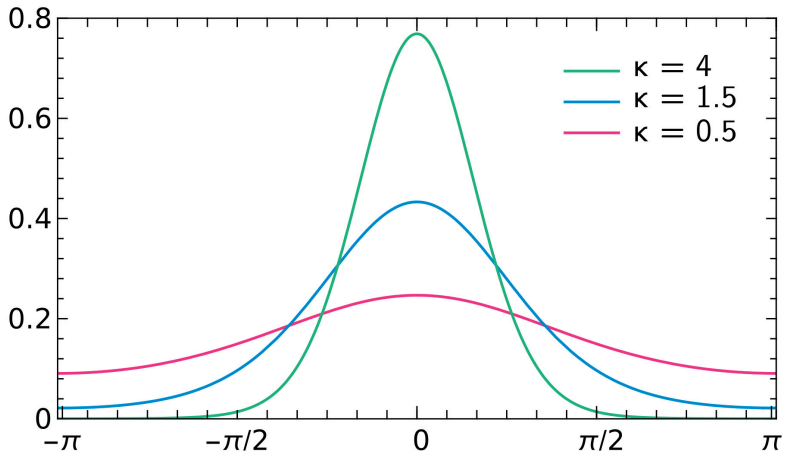


Figure 8: Example of von Mises distribution with different concentration parameters.

When working with descriptors that are periodic in nature (e.g. angles and dihedrals) the probabilistic description should be adapted to account for the non-Euclidean geometry of the space. Von Mises distributions [212, 213] are the equivalent of a Gaussian on a circle, and can be used to describe periodic data, as they are smooth across the boundaries. The multivariate extension of a von Mises distribution, however, cannot be normalized analytically, which makes it hard to use it for the KDE step in our procedure. Furthermore, when the bandwidth is negligible with respect to the periodicity, a Gaussian kernel computed while using a minimal image convention in defining distances between points is virtually indistinguishable from a von Mises distribution. For this reason, we use multivariate Gaussian kernels in the KDE step, also along periodic directions.

When determining the GMM that underlies our fingerprints, however, one cannot assume that the covariance associated with each cluster is small with respect to the periodicity of the pattern space. Given the difficulties with normalizing a multivariate von Mises distribution, we use a product of one-dimensional distributions to construct basis functions for the GMM that represent each cluster, i.e. we use

$$G(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\kappa}) = \prod_{i}^{D} \frac{e^{\kappa_i \cos(x_i-\mu_i)}}{2\pi I_0(\kappa)}, \tag{71}$$

in lieu of Eq. (43). In this expression I_0 is the modified Bessel function of order zero. The mean value and weight for each cluster are determined according to (69), whereas κ_i values are obtained using the conventional estimators for the concentration parameter of a one-dimensional von Mises distribution [214].

4.2.5 *Error Assessment*

Computing the *absolute* MISE requires knowledge of the underlying probability distribution. It is however possible to estimate the statistical error associated with a given estimate, which can be useful to fine-tune the KDE parameters and to decide on the statistical significance of the clusters that are identified in later stages of the PAMM analysis. Bootstrapping provides a very general and well-established approach to infer the statistical error in a distribution, when its analytic form is not known [215].

Bootstrapping relies on the analogy existing between the population and the sample drawn from it and consists in re-sampling with replacement a large number of sets from the given data, in order to build empirically an estimate of the probability distribution related to a certain statistical estimate. In this specific case, we exploit bootstrapping to generate N_{BS} independent samples of the KDE, $P^{(m)}(\mathbf{y}_i)$. From these, one can estimate the standard error $\delta P(\mathbf{y}_i)$ associated with the KDE at each grid point.

The bootstrapping procedure is not only useful to get an estimate of the statistical error in the KDE. By performing the (deterministic) clustering procedure we discussed above on the m-th bootstrapped estimate of $P(\mathbf{y}_i)$, one can obtain clusters $\mathcal{Z}_k^{(m)}$ that reflect the statistical fluctuations of the KDE. The comparison between the bootstrapped clusters and those obtained on the straightforward KDE can then be

used to compute indicators of how "stable" the clustering procedure can be considered. To do so, one can first compute

$$\begin{aligned} Q &= \sum_i P(\mathbf{y}_i), \quad Q_k = \sum_{\mathbf{y}\in\mathcal{Z}_k} P(\mathbf{y}), \\ Q_k^{(m)} &= \sum_{\mathbf{y}\in\mathcal{Z}_k^{(m)}} P(\mathbf{y}), \quad Q_{j|k}^{(m)} = \sum_{\mathbf{y}\in\mathcal{Z}_k\cap\mathcal{Z}_j^{(m)}} \frac{P(\mathbf{y})}{Q_k^{(m)}}, \end{aligned} \tag{72}$$

and then introduce

$$A_{ij} = \frac{1}{N_{BS}\sqrt{Q_i Q_j}} \sum_m \sum_k Q_k^{(m)} Q_{i|k}^{(m)} Q_{j|k}^{(m)}. \tag{73}$$

For each bootstrapping run, this expression determines what is the probability that – taking one of the bootstrap cluster at random, and drawing two sample points from them – one would be part of the reference cluster i, and one of the reference cluster j, renormalized over the probabilities of clusters i and j.

The diagonal elements A_{ii} report on how robust is the determination of the i-th cluster. If the i-th cluster appears identical in each bootstrapping run, A_{ii} takes a value of one, which becomes smaller if in one or more of the iterations the cluster is split over multiple clusters.

Off-diagonal terms report on how "fuzzy" the borders of the clusters are. If no bootstrapping run generates a cluster that overlaps with both $\mathcal{Z}_i$ and $\mathcal{Z}_j$, A_{ij} would take a value of zero, which increases if the clusters get merged in some of the runs, or if some of the $\mathcal{Z}_k^{(m)}$ include points from both clusters.

4.2.6 *Cluster Association*

The cluster stability matrix A_{ij} from eq. (73), can also be used to perform an additional "meta-clustering" step, that suggests ways to group together some of the clusters identified in the previous steps of PAMM, based on the notion that they were separated due to statistical error rather than because they correspond to separate free-energy basins.

We attempted two approaches, that provide satisfactory and similar results, but differ in the underlying interpretation. One possibility is to choose a threshold value for the adjacency matrix, and find the connected components of the associated graph. This approach corresponds to a sort of "flooding" scheme, in which clusters that are above a prescribed level of fuzziness are merged at once.

Alternatively, one can proceed with a hierarchical clustering procedure [216], which can also be represented in a tree-like plot which helps when it comes to interpreting the relations between different

clusters [144]. Based on the adjacency matrix one can define a distance between clusters as $d_{ij} = -\log(A_{ij}/\sqrt{A_{ii}A_{jj}})$. The pair of clusters which are closest is merged first, after which the merging is repeated iteratively until a single cluster remains. The cluster hierarchy can be represented as a binary tree, in which the vertical position of the branching point correspond to the distance between the leaves. Different strategies exist – and can be tried to improve the resolving power of the method – to define a distance between merged clusters. In this work we always use Ward's minimum-variance prescription [217], unless otherwise specified. This second strategy is more consistent with an interpretation in which fuzzy clusters correspond to clustering errors, and performs as little merging as possible to achieve the desired number of clusters, or degree of separation.

4.2.7 *Non-Gaussian Patterns.*

Besides making the clustering procedure more robust, this merging step is also useful to address the presence of strongly non-Gaussian features in fingerprint space. Even though the KDE and Quick-Shift algorithms are fully non-parametric, at many steps in our protocol we invoked the assumption that data can be (locally) described by multivariate Gaussian distributions. As a specific, and rather extreme, example of non-Gaussian behavior, let us consider the distribution depicted in Fig. 9, corresponding to three concentric rings. Figure 9 demonstrates how, in the presence of non-Gaussian clusters, the partitioning of the data by PAMM is highly unstable, leading to an adjacency matrix that shows considerable overlap between different clusters. Hierarchical clustering, illustrated using a dendrogram in Fig. 9e, shows clearly that there are three "macro-clusters" that correspond to the rings, while Gaussian features within each ring are clearly detected as being strongly connected. It is worth noting that, once different Gaussian clusters have been joined based on the adjacency matrix, it is easy to develop non-Gaussian fingerprints, by simply summing over all GMM fingerprints associated with each macrocluster $\mathcal{M}$

$$\hat{P}_{\mathcal{M}}(\boldsymbol{x}) = \sum_{k\in\mathcal{M}} p_k G(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) / \left(\zeta + \hat{P}(\boldsymbol{x})\right). \tag{74}$$

A demonstration of this non-Gaussian fingerprint is also depicted in Fig. 9d.

4.3 PATTERN CLASSIFICATION

A critical analysis of the outcome of the quick shift partitioning of the probability density, and of the adjacency among the clusters, makes

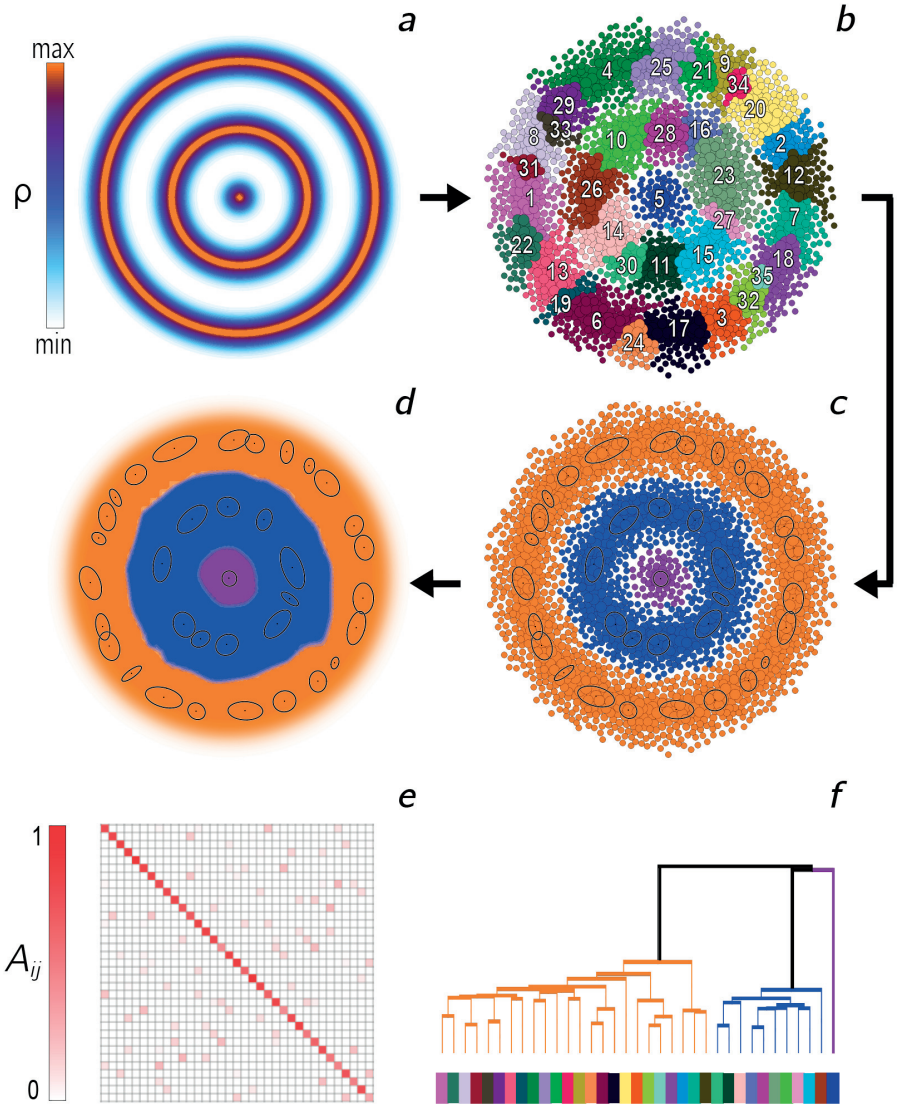


Figure 9: (a) Radially-symmetric probability density function that corresponds to concentrical circles. Clusters before (b) and after (c) merging and the PMIs corresponding to the final macroclusters (d). Panel (f) represents the dendrogram resulting from the agglomerative clustering based on the cluster stability matrix $\mathbf{A}$ (e) and using a single-linkage strategy for the merging.

it possible to associate a specific PMI with a structural pattern, described by the D dimensional feature vector.

In many cases – such as the example of the hydrogen bond that will be discussed in chapter 6 – one cluster stands out clearly as the distinct structural feature one is interested in, and one can focus further analysis on a single mode using the associated conditional probability function (68). For simplicity we will consider the selected cluster to be the one labeled by $k = 1$.

The most direct application of the definition embodied by $\hat{P}_1(\mathbf{x})$ is to use it to test whether tuples of atoms $\mathbf{R}_{ijk\ldots} = (\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k, \ldots)$ match the definition of the structural pattern by computing

$$s^{(D)}_{ijk\ldots} = \hat{P}_1(\mathbf{x}(\mathbf{R}_{ijk\ldots})). \tag{75}$$

If the components of $\mathbf{x}$ are continuous functions of the atomic coordinates, $\hat{P}_1$ is also a smooth continuous function, that takes a value close to 1 whenever a group of atoms matches the target pattern (Figure 6(h)). This makes our definition of a pattern recognition function well-suited for use as a collective variable in accelerated sampling methods,[218] possibly in conjunction with other machine learning techniques to characterize the overall connectivity induced by the selected molecular pattern [109], or similar fingerprint metrics that are guaranteed to distinguish dissimilar structures [137]. The PAMM variables corresponding to different structural descriptors can also be analyzed to yield a coarse-grained, low-dimensional map [110, 185, 219]. If necessary, one can also artificially "soften" the transition between clusters, by dividing *all* the covariance matrices $\boldsymbol{\Sigma}_k$ in the Gaussian model by a scaling factor α.

Since the $s^{(D)}_{ijk\ldots}$'s effectively count the instances of the structural pattern that are present at any given time in the trajectory, one can also combine several of these indicators together to count the number of patterns that involve a tagged atom i, or pair of atoms (i, j):

$$s^{(1)}_i = \sum_{j,k,\ldots} s^{(D)}_{ijk\ldots}, \quad s^{(2)}_{ij} = \sum_{k,\ldots} s^{(D)}_{ijk\ldots}, \quad \text{etc.} \tag{76}$$

Depending on the application being considered, $s^{(1)}_i$ can be taken to represent the total coordination of the atom i, $s^{(2)}_{ij}$ the overall bonding between atoms i and an atom j, and so on.

5

AN AGNOSTIC DEFINITION OF THE HYDROGEN BOND

Contents

5.1 THE HYDROGEN BOND

The PAMM framework we have introduced in the previous chapter is very abstract, and can be applied to any situation in which one wishes to recognize recurring motifs in an atomistic simulation. As a first application to a practical case, we chose to focus on recognizing the Hydrogen Bond (HB) in a number of different contexts.

The term "hydrogen bond" refers to a highly directional three-center interaction between two polar atoms and a hydrogen. [220] The hydrogen atom H is covalently bound to one of the polar atoms, which is designated as the donor D, and points towards the second polar atom which is designated as the HB acceptor A.

Despite the apparent simplicity of the concept, it is not easy to develop an universal definition of the HB, mostly because this entity has been used in many different contexts. The term has been associated to near-covalent interactions with an energy in excess of 30 kcal/mol, as well as to exceedingly weak ones with an energy of less than one kcal/mol. Typically, HBs are understood to have a predominantly electrostatic nature, with strongly electronegative donors and acceptors such as F, O or N. However, the observation of recurring C–H$\cdots$O units in the secondary structure of polypeptides have also been interpreted in terms of weak HBs, that have been suggested to play a significant role in stabilizing proteins. [221]

Adding to the complexity of the broad energy scale covered by HBs is the fact that in most situations of interest thermal fluctuations and the environment modulate their stability, and that they are formed and destroyed on a relatively short time scale. One sees the difficulty in giving a clear-cut definition of a chemical entity which exhibits

This chapter is an adaptation of ref. [1]

such a variability. The most generally applicable definitions rely on performing an electronic structure calculation, and on decomposing the energy of the systems in a sum of terms that can be interpreted as the binding energy of putative HBs [222–225]. Definitions that are based solely on structural information are much more practical, in that they do not require a supporting electronic structure calculation and can be applied to experimental structural data or to atomistic simulations based on empirical force fields. The downside is that these structural definitions invariably contain a degree of arbitrariness, as they are based on the heuristic introduction of ranges of structural parameters that are deemed to represent a hydrogen bond in a given context [226, 227].

Kumar *et al.* carried out a systematic comparison of many of these structural definitions in the case of liquid water [228], and recognized that the best way to assess whether a given definition makes physical sense is to compare the probability distribution of the structural parameters with the range of values associated with the hydrogen bond.

PAMM offers a very natural probabilistic way of describing this fuzzy entity, by automatically inferring from a training simulation the probability of the various recurrent metastable patterns that have been explored.

The data itself informs the definition of a range of parameters that unambiguously identify hydrogen-bonded configurations, and naturally and smoothly describes the transition between this region and configurations that are clearly not hydrogen bonded. Even though this definition is by construction system-specific, the protocol to obtain it is univocal and unbiased, as it does not rely on choosing manually threshold values for the structural parameters.

5.2 FEATURE SPACE DEFINITION

The first steps in the application of PAMM are the identification of groups of atoms that should be tested for recurring patterns and the choice of structural parameters that describe the arrangement of atoms within each group. In the case of the HB, these choices are fairly obvious. One should select an atomic species that should be considered as the putative HB donor D, one that should be considered as the acceptor A and (a subset of) the hydrogen atoms that complete the HB triplet. The geometry of each of these groups is completely determined by the three distances $d(\text{A-D})$, $d(\text{A-H})$ and $d(\text{D-H})$. To simplify comparison with other definitions, and to highlight the symmetries inherent in the problem, we decided to use combinations of these distances, namely the proton-transfer coordinate $\nu = d(\text{D-H}) - d(\text{A-H})$, the symmetric stretch coordinate $\mu =$

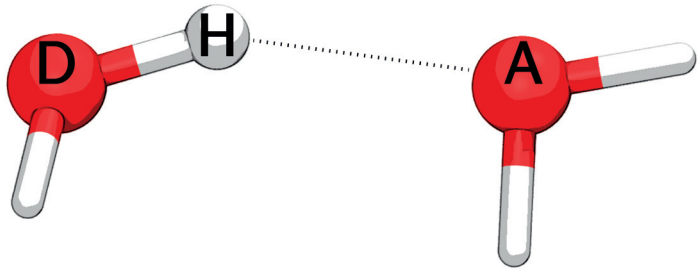


Figure 10: Simple scheme of the hydrogen bond in water. In general, a hydrogen bond can be seen as a three-center interaction between two polar atoms and a hydrogen. The *donor* atom (D) is covalently bound to the H, which is weakly bounded to the *acceptor* atom (A).

$d(\text{D-H}) + d(\text{A-H})$ and the acceptor-donor distance $r = d(\text{A-D})$ as the group descriptors. We computed these (ν, μ, r) triplets for each D-H-A group present in each snapshot extracted from the simulations, thereby obtaining the training data set $\mathcal{X}$ that we used to run PAMM. In building the probability distribution, each point was weighed by a factor $[r(\nu+\mu)(\mu-\nu)]^{-1}$, that accounts for the trivial phase space volume so that a uniform distribution of atoms would yield a constant probability density in (ν, μ, r).

5.3 ANALYSIS OF SIMULATION RESULTS

By direct inspection of configurations, one can clearly see that only one cluster (that for simplicity will be labeled by $k = 1$) corresponds to hydrogen-bonded configurations – the orange hue in Figure 11.

Having developed an effective definition for a specific HB pattern, we can proceed to probe such a pattern in the structural outcome of a simulation. During the analysis, the most direct application of the HB PMI is to use it to test whether a triplet D-H-A matches the definition of the HB pattern by computing

$$s_{DHA} = \hat{P}_1(\mathbf{x}_{DHA}). \tag{77}$$

Since s_{DHA} takes a value close to 1 whenever a group of atoms matches the *characteristic* hydrogen bonded configuration exhibited by the system, we introduced effectively an HB counting function. Combining several of these indicators together one can count the number of bonds that involve a tagged atom i. For instance, by summing over all the possible acceptor atoms O' and hydrogen atoms H, one can get a smooth order parameter to count the total number of HBs donated by a selected oxygen O:

$$s_O = \sum_{H,O'} s_{OHO'}. \tag{78}$$

Having a value of $s_O \approx 2$ means that the water molecule is donating two HBs. In the text we will use the notation s_D when referring to *donated* HBs and s_A for *accepted* HBs. If we sum over all the acceptors O' and donors O we obtain the count of the total number of HBs in which a selected hydrogen H is involved:

$$s_H = \sum_{O,O'} s_{OHO'}, \tag{79}$$

where $s_H = 0$ =*not* H-*bonded*, $s_H \approx 1$ =*standard* HB, $s_H \approx 2$ =*bifurcated* HB and so on. One could also evaluate the bonding order between pair of atoms (i, j). For instance by summing over all the possible H atoms one can probe if two tagged water molecules are hydrogen bonded or not:

$$s_{O,O'} = \sum_{H} s_{OHO'}. \tag{80}$$

We can then compute the probability distribution $P(s)$ relative to a certain counter s by normalizing the histogram $h(s)$ obtained processing the outcome of a long equilibrated simulation

$$P(s) = \langle \delta[s(\mathbf{q}) - s] \rangle. \tag{81}$$

The probability $P(s)$ can then be expressed as a *free energy*:

$$F(s) = -k_B T \ln(P(s)). \tag{82}$$

5.3.1 *Alanine dipeptide*

Let's consider the case of an empirical forcefield model of alanine dipeptide (N-acetylalanine-N'-methylamide) – one of the simplest examples of peptide bonding, displaying many of the essential features that are present in proteins.

This is an ideal test case, as it allows us to demonstrate the functioning of PAMM for different kinds of hydrogen bonds.

We will consider HBs donated by water molecules to the carbonyl of alanine O_C, HBs donated by the peptide nitrogen to the oxygens in water O_w, and investigate the significance of a more exotic, weak HB donated by the peptide C_α to the O_w atoms.

We used the CHARMM27 forcefield [229] to describe interactions within the polypeptide and a TIP3P model for the water molecules [230], with flexible bonds modelled as harmonic stretches, as implemented in LAMMPS [231]. We equilibrated a supercell containing 128 water molecules in the *NpT* ensemble, and ran subsequently 600 ns of *NVT* molecular dynamics using a Langevin thermostat with a time constant of 10 ps. [232] The configurations were saved every 1 ps.

The first kind of HB we considered is O_w–$H\cdots O_C$. To accelerate the analysis we only included configurations with $\mu < 5$Å. Figure 11

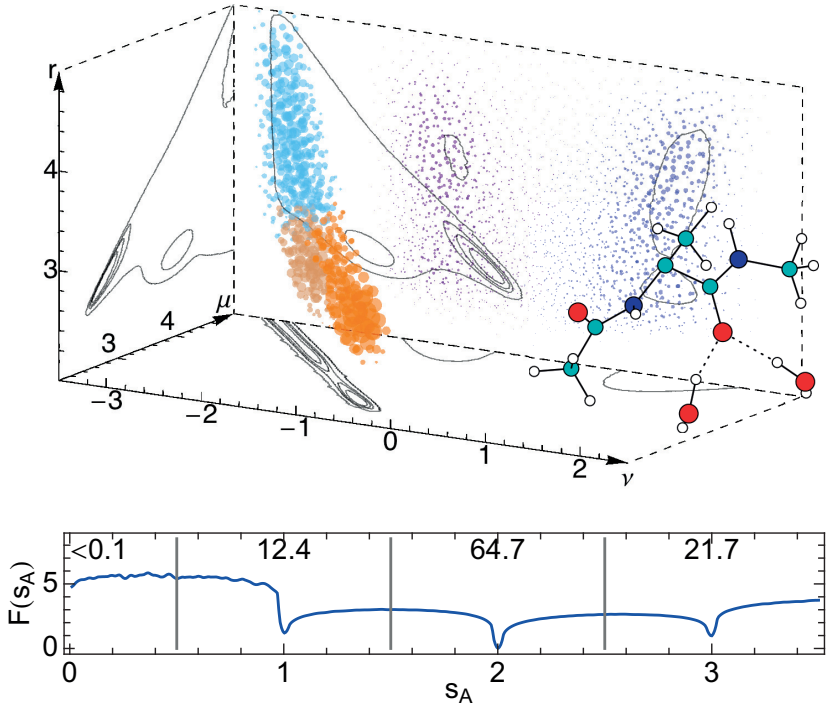


Figure 11: (Upper panel) Distribution of (ν, μ, r) configurations for O_W–H$\cdots O_C$ in a simulation of alanine dipeptide in water. Size and opacity of points correspond to the KDE of $P(\mathbf{y})$, and colors indicate the cluster each grid point has been assigned to. (Lower panel) Free energy (kcal/mol) computed from the distribution of number of accepted hydrogen bonds s_A for the oxygen atom in the carbonyl group in solvated alanine dipeptide. The histogram was smoothed with a triangular kernel of width 0.025. We also report the integrated probabilities for for having $s_A < 0.5$, $0.5 < s_A < 1.5$ and so on. The average number of accepted HBs is $\langle s_A \rangle = 2.1$ and the standard deviation is 0.6.

shows the distribution of the values of (ν, μ, r), colored according to the partitioning of the density obtained by running PAMM on the data set. Several clusters are recognized, which means that besides HBs there are other recurring patterns that can be distinguished by this analysis. One of these clusters – represented with an orange hue – can be seen by direct inspection of configurations (or by comparison with other structural definitions) to correspond clearly to hydrogen-bonded configurations. As discussed in section 4.2.3 we used the Gaussian-mixture model built based on the clustering to define the degree of confidence s_{DHA} by which we classify a certain configuration of a D donor, H hydrogen and A acceptor as a hydrogen bond, and then introduce a count of the total HBs that involve a given acceptor oxygen $s_A = \sum_{D,H} s_{DHA}$. The free energy built from the histogram of s_A is represented in the lower panel of Figure 11. The free energy is strongly peaked at integer values of s_A, because the transition between 0 and 1 is very sharp when a hydrogen bond is formed or broken. This plot shows clearly that most of the time the carbonyl is involved in receiving two hydrogen bonds, but there is also a fairly large probability of accepting one or three bonds. It would be interesting to compare these results with first-principles simulations of solvated alanine dipeptide, to verify whether the possibility of forming over and under-coordinated configurations is a consequence of the simplified modelling of the interactions between water molecules and the carbonyl.

We then moved on to look into the hydrogen bond donated by the amide group N–H$\cdots$O$_w$. Since the chemical identity of atoms is fixed in an empirical force field calculation, we specifically restricted the search to include only the amide H atom and oxygen atoms from the water molecules. We used a cut off of $\mu < 5.5$Å to disregard configurations that are clearly irrelevant to the HB search. We report the distribution of configurations and the PAMM clustering in Figure 12. Perhaps unsurprisingly, the distribution of (ν, μ, r) associated with this set of atoms differs considerably from that in Figure 11 – this is a somewhat weaker bond, which results in a less structured $P(\nu, \mu, r)$. Still, one can recognize a cluster that is clearly associated with HB configurations, that we can use to define a bond counting order parameter, that in turns can be used to compute the total number of hydrogen bonds donated by the N atom, $s_D = \sum_{A,H} s_{DHA}$. While the most likely value of s_D is one, there is a high probability of observing a N–H group donatinhg two HBs. Given the geometry of the amide group, this actually means that based on our unbiased, self-consistent definition, HBs donated by the amide group as described by the empirical force field we used have a large probability of being bifurcated, binding simultaneously to two different water molecules.

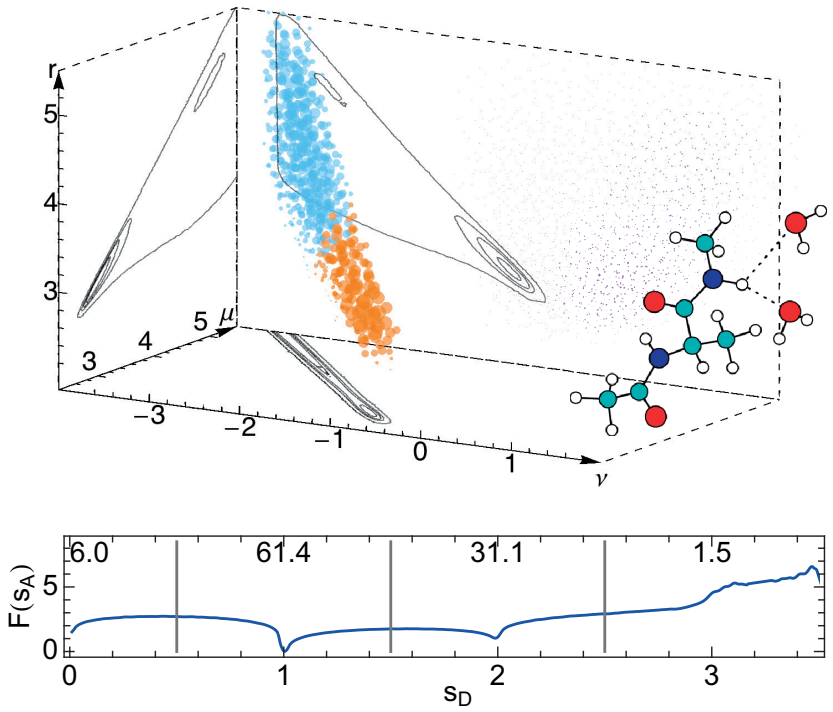


Figure 12: (Upper panel) Distribution of (ν, μ, r) configurations for N–H$\cdots$O$_w$ in a simulation of alanine dipeptide in water. Size and opacity of points correspond to the KDE of $P(\mathbf{y})$, and colors indicate the cluster each grid point has been assigned to. (Lower panel) Free energy (kcal/mol) computed from the distribution of number of donated hydrogen bonds s_D for the amide nitrogen atom in solvated alanine dipeptide. See Fig. 11 for details. The average number of accepted HBs is $\langle s_D \rangle = 1.3$ and the standard deviation is 0.5.

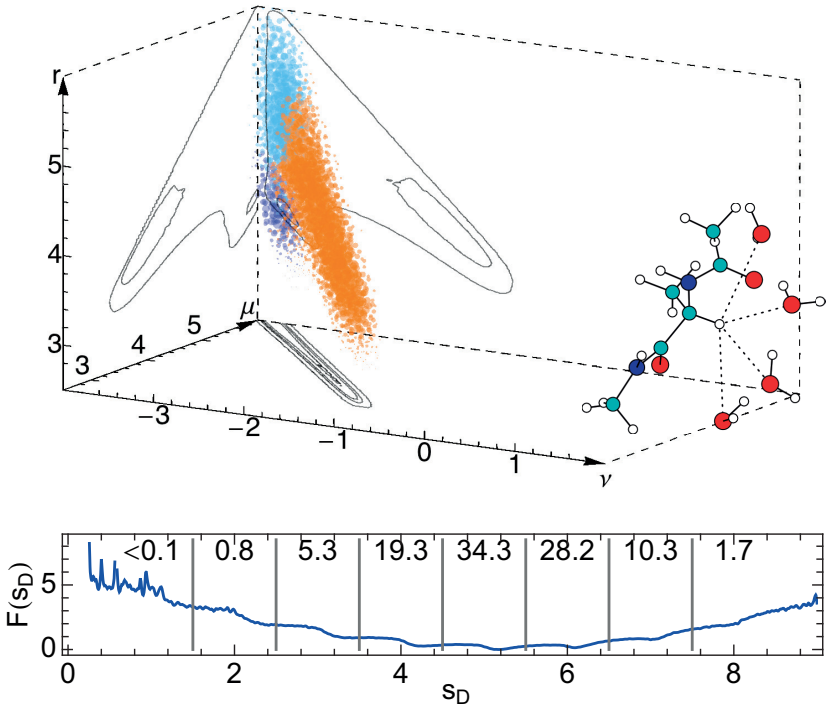


Figure 13: (Upper panel) Distribution of (ν, μ, r) configurations for C_α–H$\cdots$O$_w$ in a simulation of alanine dipeptide in water. Size and opacity of points correspond to the KDE of the $P(\mathbf{y})$, and colors indicate the cluster each grid point has been assigned to. (Lower panel) Free energy computed from the distribution of number of donated hydrogen bonds s_D for the amide carbon atom in solvated alanine dipeptide. See Fig. 11 for details. The average number of donated HBs is $\langle s_D \rangle = 5.2$ and the standard deviation is 1.1.

Finally, to verify how the PAMM algorithm behaves when applied to a selection of atoms that does not exemplify a typical hydrogen bond, we considered groups that would correspond to a C_α–H group donating a HB to water. Fig. 13 shows the partitioning of the probability density for this choice of atoms, which has some features that are reminiscent of those seen for conventional HBs, albeit with a much longer d(A–H). A more careful inspection of configurations that belong to the lobe of the probability density with the lowest μ, however, shows that these can hardly be described as HBs: in many cases the hydrogen atoms of water molecules are oriented *towards* the C_α–H group, and the distribution of s_D shows very little structure. This example demonstrates that the presence of a recurring structural motif with a signature in terms of the probability distribution in configuration space does not necessarily imply that the atoms that compose the motif are involved in some sort of chemical bonding. Here, the non-uniform structure of oxygen atoms in the vicinity of the C_α–H group is probably an indirect consequence of the hydrogen-bond interaction of water molecules with nearby carbonyl groups, and of the stiffness of the backbone of the dipeptide.

5.3.2 *Classical and quantum water*

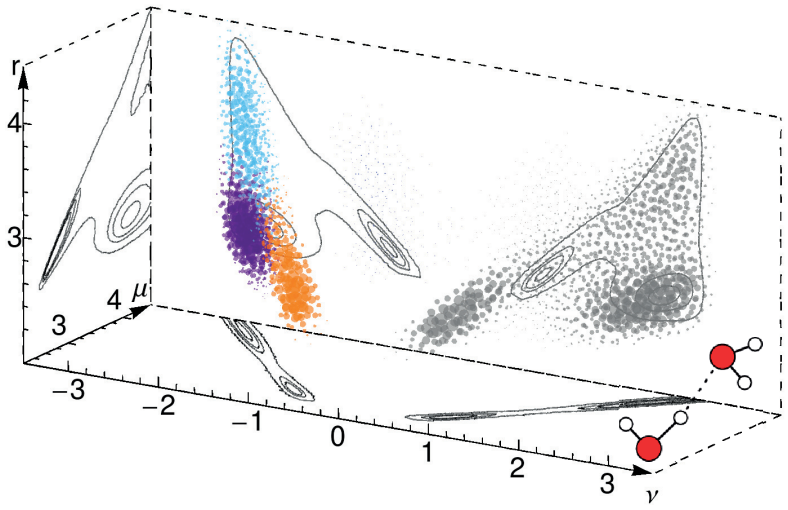


Figure 14: Distribution of (ν, μ, r) configurations for O–H$\cdots$O′ in a simulation of neat TIP4P water. Size and opacity of points correspond to the KDE of $P(\mathbf{y})$, and colors indicate the cluster each grid point has been assigned to. Clusters with $\nu > 0$ have not been colored, but have been correctly identified by PAMM.

Simulations of alanine dipeptide contain different kinds of hydrogen bonds, and allowed us to demonstrate the adaptive nature of

PAMM to derive a different, data-driven definition of the range of structural parameters that can be associated with a HB for each set of constituent atoms. In a simulation of neat water, instead, there is only one type of O–H$\cdots$O′, the slight complication being that each oxygen atom can simultaneously act as a donor and an acceptor of hydrogen bonds.

TIP4P water

We began by analyzing a simulation of a very common empirical water force field, the flexible TIP4P/2005f [233] model. A box containing 128 water molecules was first equilibrated for 2 ns at constant pressure (1 atm), constant temperature (298 K) *NpT* dynamics. A subsequent 500 ns *NVT* run was performed using a Langevin thermostat with relaxation time τ=5 ps. The configurations were saved every 1 ps. In the spirit of a fully automated analysis of the trajectory, we did not exploit knowledge of the chemical identity of water molecules, which is fixed in a simulation with a non dissociable model. The distribution of (ν, μ, r) shows clearly the dual role played by the O atoms (see Figure 14), which is apparent in the symmetry of the probability density across the $\nu = 0$ plane. Both the cluster highlighted in orange and its mirror image correspond to legitimate HB configurations, but only the former corresponds to structures in which the first oxygen is acting as the donor and the second as the acceptor.

Once s_{DHA} has been defined based on the analysis of the simulation data, it can be used to characterize in great detail how a given model of water describes hydrogen bonding. Figure 15 summarizes some of the information that can be obtained from this analysis. One can compute free energies for the number of hydrogen bonds donated (s_D) or accepted (s_A) by each oxygen atom, as well as for the total ($s_A + s_D$) and for the number of HBs that are formed by each hydrogen. A large fraction of TIP4P water molecules are tetracoordinated, with nearly 65% of oxygen atoms receiving and donating two HBs. There is a small but significant asymmetry between the distribution of s_D and that of s_A, the former being more strongly peaked at $s_D = 2$, while there is somewhat more flexibility in the count of accepted bonds. Given the rigid constraints on the covalent O–H bond, any oxygen in the simulation can donate two bonds at most, except for the case of bifurcated HBs where a single O–H moiety is involved with bonds to two different O′ atoms. The distribution of s_H shows that there is just about 2% probability of observing such bifurcated bonds. More detailed information on the topology of the HB network can be obtained by observing the joint probability distribution of s_D and s_A. While the order of magnitude of the probability of each joint configuration is determined by the product of s_D and s_A, there are significant deviations that are indicative of the correlations

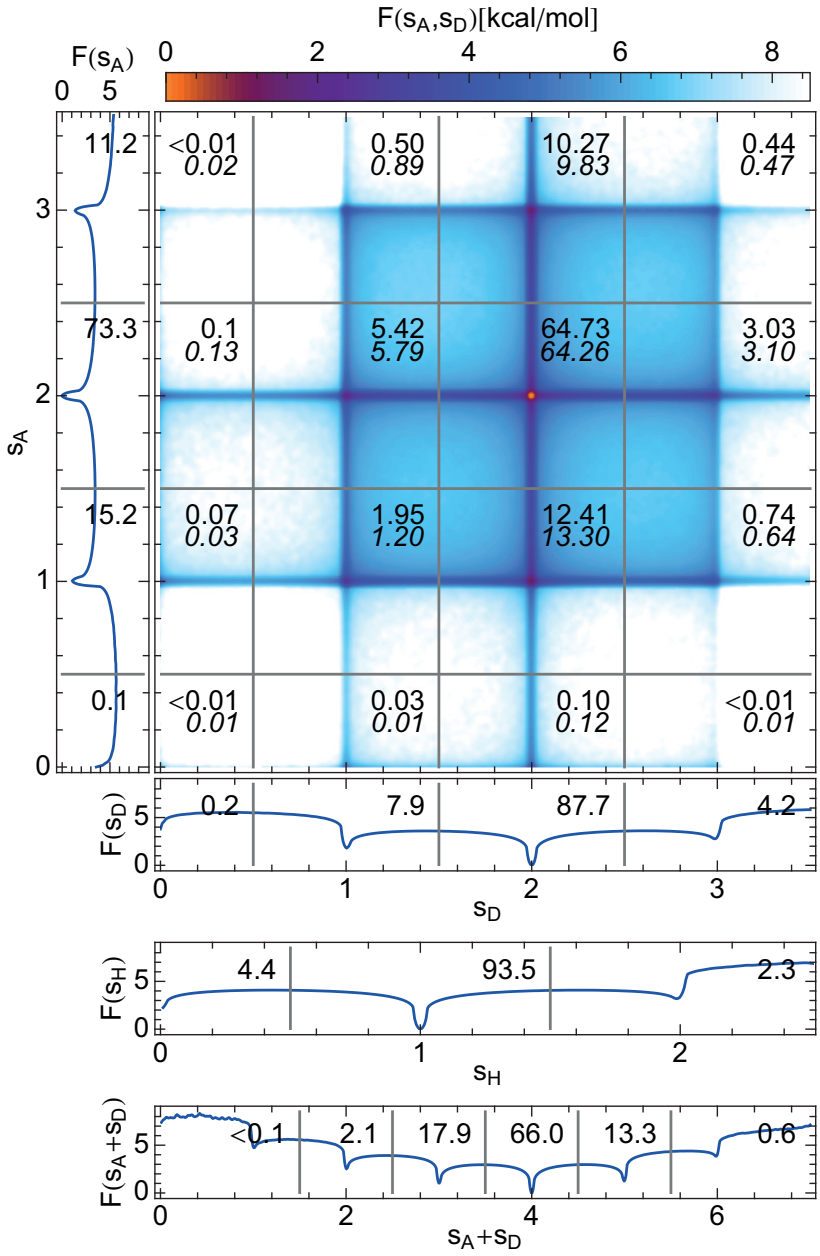


Figure 15: Hydrogen-bond counts statistics for a classical simulation of TIP4P water at room temperature. All the probability distributions have been smoothed with a triangular kernel of width 0.025, and are represented in terms of the associated free energies $F = -k_B T \ln P$, that are expressed in kcal/mol throughout. We also report integrated probabilities (in percent) to have a configuration in the vicinity of the different integer numbers of HBs. Below the values of the joint probabilities of s_A and s_D the product of the marginal probabilities are indicated, in italics.

between defects in the network. For instance, the probability of having a "linear" water that donates and accepts a single HB is almost twice the value that would be expected based on the product of the marginal distributions. Note that this analysis focuses on the connectivity of the network rather than on the geometry of the environment of each water molecule. An interesting way to extend this analysis could consider the correlation between the HB counts and the degree of tetraedrality, with the electronic structure, or with quantities that can be directly related to experimental observables.

Comparison with GMM clustering

We decided to rely on the non-parametric QS clustering procedure, because of the ambiguity introduced by the EM optimization of the model log-likelihood of GMM clustering.

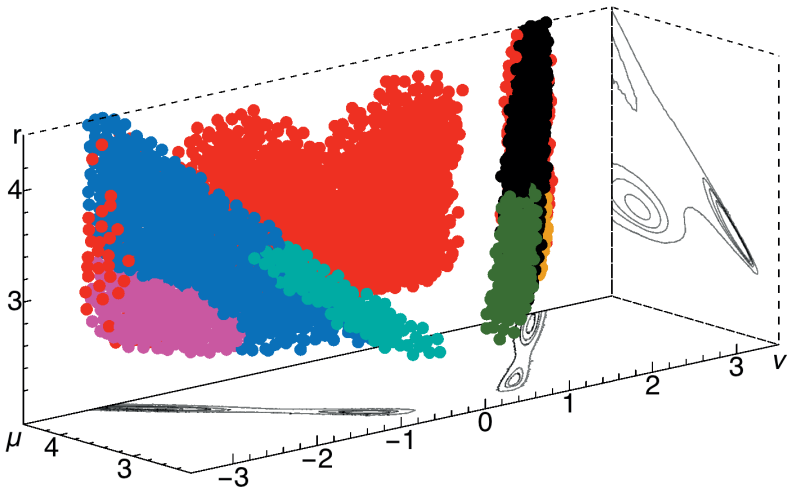
To proof the stability of our procedure, we here apply a GMM clustering, instead of QS, to the TIP4P test case.

In GMM after having chosen the number of clusters needed to fit the PDF underlying the data, one has to first fit the model and then assign each point to the Gaussian from which, it was more likely sampled from. Figure 17 shows the posterior classification obtained using PAMM, which differ considerably from those resulting from a standard GMM with a similar number of clusters (see e.g. Fig. 16a). Furthermore, varying the number of cluster (K), changes significantly the results, as shown in fig. 16b, where according to the results, the HB mode should be described by two Gaussians.

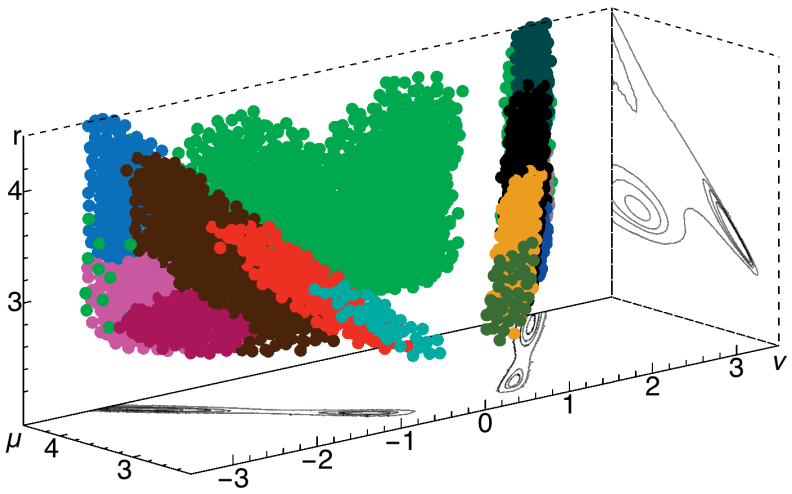
Hydrogen-bonding defects in ice Ih

The consistency of the results obtained with PAMM descriptors of the hydrogen bond and those obtained with more conventional descriptors is reassuring. It is however useful to verify the behavior of indicators such as s_A and s_D in a more ordered environment such as the tetrahedral hydrogen-bond network of ice Ih, in which one should be able to identify clearly coordination defects.

To this aim, we have performed a PAMM analysis of a simulation of ice Ih, using the same flexible TIP4P model discussed above, and a proton-disordered unit cell with 768 molecules [234]. We have then created a pair of Bjerrum coordination defects [235, 236], and separated them by the maximum distance allowed by the simulation cell, by repeatedly flipping water molecules in the lattice (Figure 18(a)). We have then equilibrated the simulation for a few tens of ps, collected some snapshots of the configurations and evaluated s_A and s_D for each oxygen. The vast majority of the O atoms have $s_A = s_D = 2$, as one would expect in a perfect tetrahedral arrangement consistent with the ice rules. We could however identify clearly a pair of oxygen



(a) $n = 7$, $\mathcal{L} = 1161405$.



(b) $n = 13$, $\mathcal{L} = 1514967$.

Figure 16: Application of GMM clustering to the (ν, μ, r) configurations for O–H$\cdots$O′ in a simulation of neat TIP4P water. The colors indicate the cluster each grid point has been assigned to (i.e. the cluster with the largest posterior probability at that point).

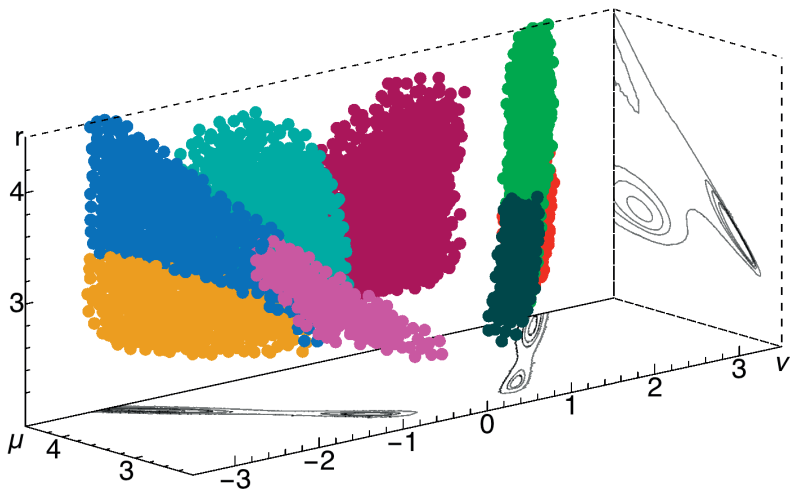


Figure 17: Classification of the FPS grid in Figure 14 based on the PMIs trained for the TIP4P water model. Points are colored according to the Gaussian with the largest value of the PMI.

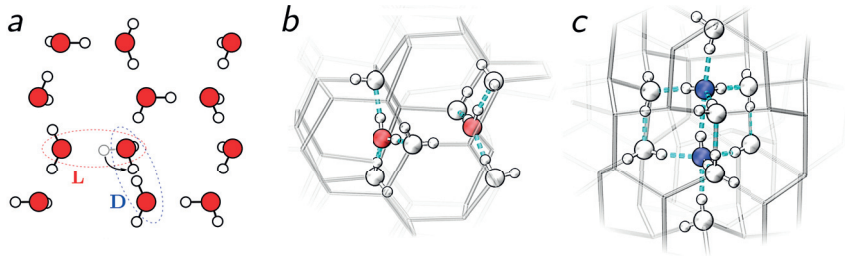


Figure 18: a) Schematic representation of how a pair of Bjerrum defects are generated by flipping the orientation of a water molecule in a lattice that satisfies the ice rules. By flipping other molecules, the L and D defects can be separated. b) and c) Configurations of equilibrated L and D defects in ice Ih. Oxygen atoms with $s_A = 1$ are colored in red, atoms with $s_A = 3$ are colored in blue, while atoms with $s_A = s_D = 2$ are in white, or hidden for clarity.

atoms with $s_A = 1$ (a L defect), and a pair with $s_A = 3$ (a relaxed D defect). Snapshots of these defective environments are represented in Figure 18.

Classical ab initio *water.*

We then moved on to perform our analysis on a first-principles simulation of liquid water. The trajectory is the classical simulation from Ref. [94], which was performed using the CP2K software package, [237, 238] with a BLYP exchange-correlation functional [239, 240] and a DZVP basis set. The simulation box contained 64 water molecules at the experimental density, and 100 ps of *NVT* dynamics were performed, with the first 5 ps discarded for equilibration. A PAMM analysis of the simulation yielded very similar clusters to those obtained from TIP4P water. The analysis of HB counts, in Figure 19, shows that BLYP water has a very regular structure, with a higher count of tetracoordinated oxygen atoms, and a very low count of defective structures. This is consistent with the well-known observation that generalized-gradient approximation models of water are overstructured compared to experiment and to empirical water models. Note that correlations in the HB network are stronger in this case than for TIP4P water, with one-donor/one-acceptor oxygen atoms being four times more likely than one would expect given the separate probabilities of $s_A \approx 1$ and $s_D \approx 1$.

Quantum ab initio *water.*

Finally, we considered a simulation that used the PIGLET technique to introduce nuclear quantum effects [94] on top of a first-principles description of the electronic structure, analogous to the one used for the classical trajectory described above. To achieve convergence of quantum properties, 6 beads were used together with a custom-tailored generalized Langevin equation thermostat, [94] as implemented in the i-PI Python interface. [241] The overall statistics of the HB network (Figure 20) are not dramatically changed by nuclear quantum effects, that only enhance marginally the probability of distorted configurations, with a decrease of ideal ($s_D \approx 2, s_A \approx 2$) configurations and an increase of bifurcated hydrogen bonds. These are not however substantial changes, and are in part due to the fact that quantum fluctuations make the PAMM definition of s_{DHA} less clear-cut than in the classical case.

Defining HBs with a method such as PAMM, that does not make any assumption on the covalent bonds present in the system, is particularly convenient in a context such as the present one. Extreme quantum fluctuations of protons along the hydrogen bond lead to transient formal autolysis events, where the hydrogen atom detaches from the

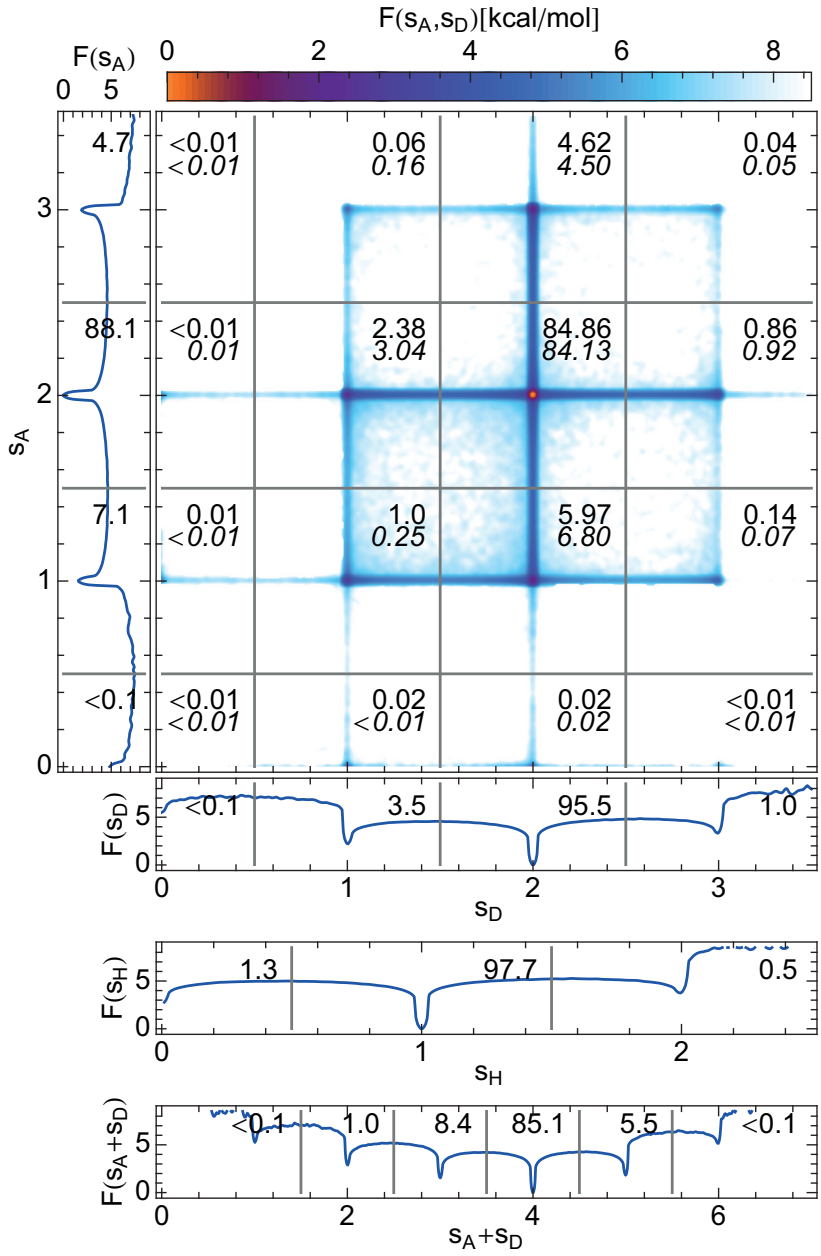


Figure 19: Hydrogen-bond counts statistics for a classical simulation of BLYP water at room temperature. See the caption of Figure 15 for a detailed explanation of the plots.

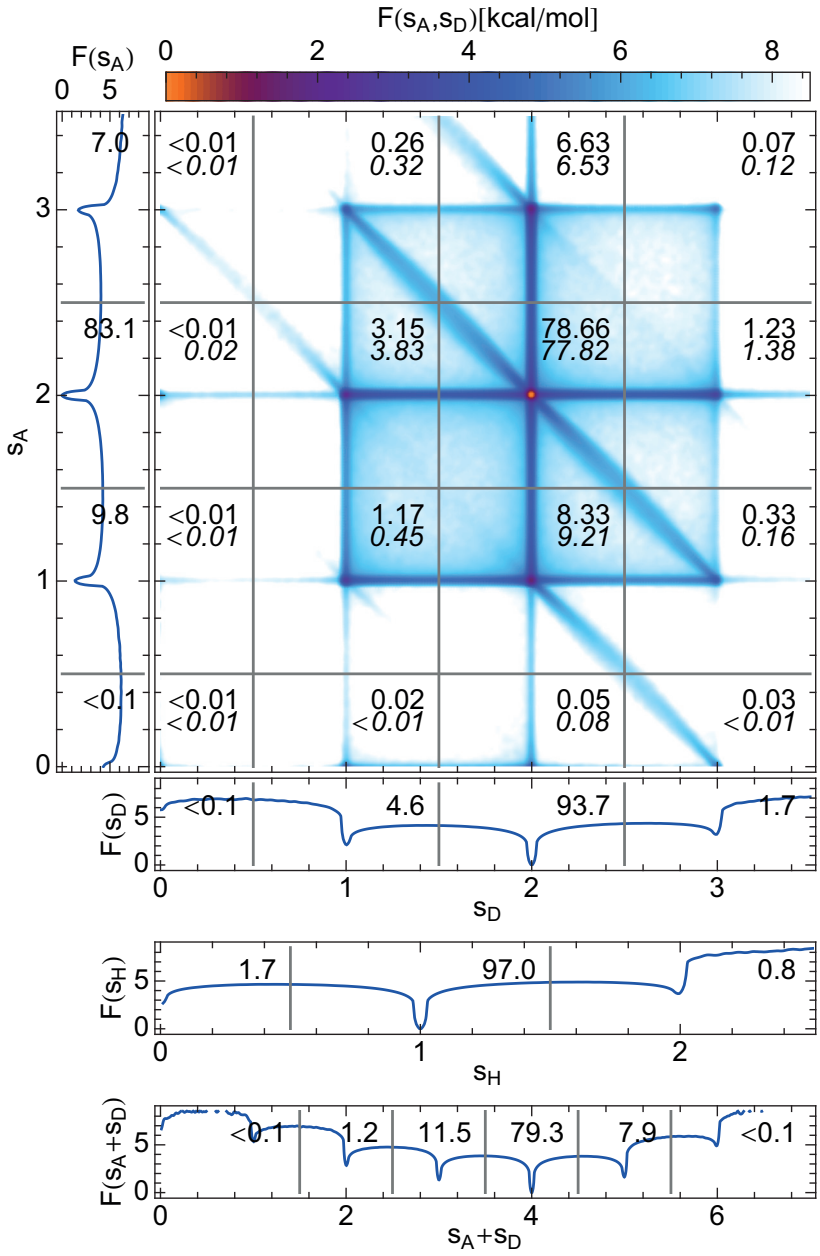


Figure 20: Hydrogen-bond counts statistics for a PIGLET simulation of BLYP water at room temperature. See the caption of Figure 15 for a detailed explanation of the plots.

donor atom and reaches out to be closer to the acceptor oxygen. [242] From a structural standpoint, PAMM does not recognize a distinct cluster corresponding to these distorted configurations, but rather a continuum of structures. Starting from a O_1–H$\cdots O_2$ bond where the hydrogen atom is covalently bound to O_1, one goes smoothly through distorted donated hydrogen bond to a configuration that is formally classified as a distorted bond *accepted* by O_1 and *donated* by O_2. These fluctuations conserve the total number of HBs between the two oxygen atoms, and only change their character from donated to accepted. These excursions are apparent in the joint probability distribution of s_D and s_A, where they show up as regions with higher probability extending diagonally between two near-integer (s_D, s_A) regions.

Hydrogen Bond dynamics in water

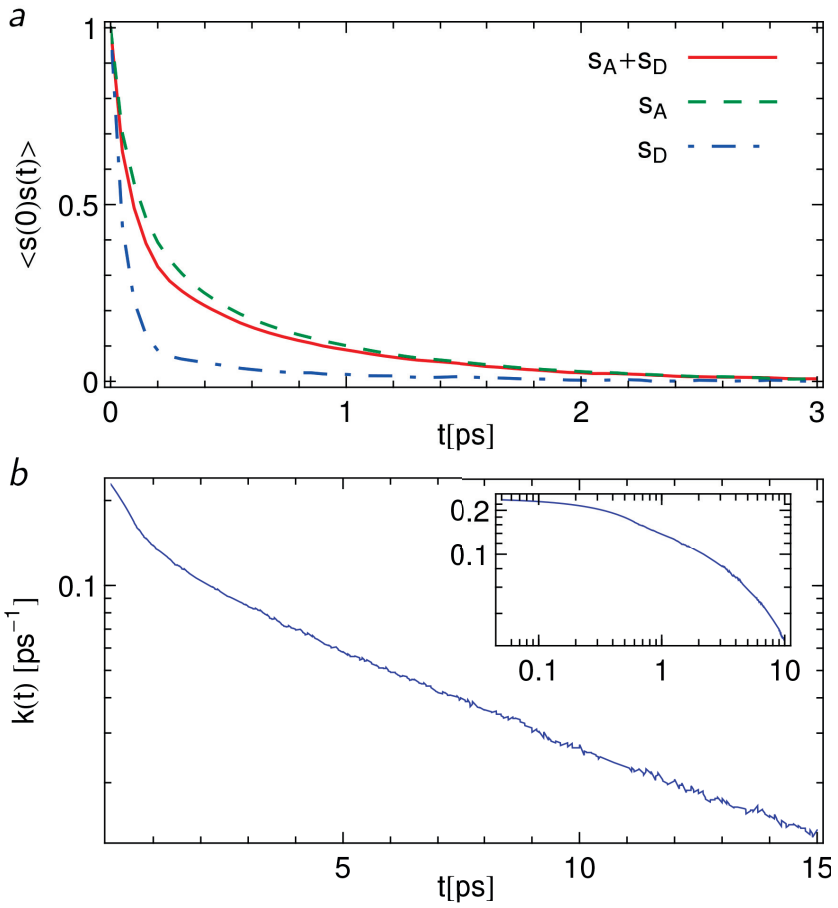


Figure 21: (a) Time correlation functions of the hydrogen-bond counts for a tagged oxygen in a simulation of TIP4P/2005f water. (b) Rate function for the hydrogen-bond formation/break-up, computed as the derivative of the correlation function $\langle s_{DA}(t)s_{DA}(0)\rangle$

The asymmetry between s_A and s_D is also apparent when one considers the dynamical behavior of the two quantities. The upper panel of Figure 21 shows the correlation functions for the counts of acceptor and donor HBs for a tagged oxygen, as well as the total. The curves were computed by analyzing several short *NVE* simulations started from independently equilibrated configurations. The correlation time for s_D is very short, because configurations where a water molecule donates less or more than two HBs are very short-lived. Configurations that are distorted from the point of view of accepted HBs are less unstable, and therefore the correlation function of s_A decays more slowly. The correlation function for the total is dominated by the slow decay of s_A, and is compatible with the results reported in Ref. [228] for traditional structural definitions of the hydrogen bond.

The hydrogen-bond count functions we have used this far do not consider the identity of individual bonds, so a quick fluctuation that momentarily breaks a HB and that is immediately re-formed is indistinguishable from a fluctuations that breaks a HB and leads immediately to the formation of a new HB with a *different* acceptor oxygen. A correlation function that is sensitive to the identity of the HB triplet, which is more easily interpreted in terms of physical observables, [243] can be readily computed by considering all the (D, A) pairs, computing for each pair $s_{AD} = \sum_H s_{DHA}$. One can then compute the rate function as the time derivative of the autocorrelation of s_{AD}, computed for each pair separately:

$$k(t) = -\frac{1}{n_A n_D}\frac{\partial}{\partial t}\sum_{A,D}\langle s_{AD}(t)s_{AD}(0)\rangle\,. \tag{83}$$

The decay takes place on a similar time scale to that observed in Ref. [243], and exhibits similar features, including the presence of multiple time scales in the decay of the rate function.

A Graph-based analysis of the HB network.

Most of the remarkable properties of liquid water stem from its ability to form dynamic, labile HB networks [244] whose connectivity changes constantly. Thus, in order to properly understand water, it is fundamentally important to be able to characterize the structure and the dynamics of the whole HB network. [245]

One can think about combining PAMM with some basic concepts of *graph theory* to describe the 3-dimensional HB network in its entirety.

A *graph* is a mathematical structure that can be used to model pairwise relations between *objects*. The objects are called *vertices* (or *nodes*), and the relations (*link*) between the nodes are called *edges*. A graph

having n vertices can be associated with an $n \times n$ matrix A which is called the *adjacency matrix* and is defined as

$$A_{i,j} = \begin{cases} 1 & \text{if } (i,j) \in E \\ 0 & \text{otherwise} \end{cases} , \tag{84}$$

where E denotes the set of edges.

In the specific context of water one can build a weighted oriented graph in which the nodes are the oxygens, while the edges are defined as in Eq. 80. A schematic example is shown in Figure 22.

One can use A to capture the dynamical behavior of the HB network by defining the function:

$$\zeta(t) = \frac{1}{N_O} ||A(t) - A(0)||_F^2 , \tag{85}$$

where the subscript F stands for Frobenius Norm.[1]

Figure 23 shows the adjacency matrix relaxation analysis applied to a *NVT* simulation of 64 water molecules at room temperature with the TIP4P-2005f empirical force field (black curve) compared with an analogous classical *NVT ab-initio* simulation using a dispersion-corrected BLYP exchange-correlation functional and a DZVP basis set (red curve). From this analysis it clearly emerges that *ab-initio* water is glassier compared to the empirical water model. From fig. 23 it appears that the timescale on which the HB network relaxes to a new configuration is of the order of tens of picosecond.

The long relaxation time of the HB network can also be probed in terms of the more familiar radial distribution function (RDF) by introducing the autocorrelation function

$$c_{rr}(t) = \frac{1}{\sigma_g^2} \langle (g_0(r^*) - \langle g(r^*) \rangle)(g_t(r^*) - \langle g(r^*) \rangle) \rangle , \tag{86}$$

1 The Frobenius norm of a $m \times n$ matrix M is defined as $||M||_F = \sqrt{\sum_{i=1}^{m} \sum_{i=1}^{n} |m_{i,j}|^2}$.

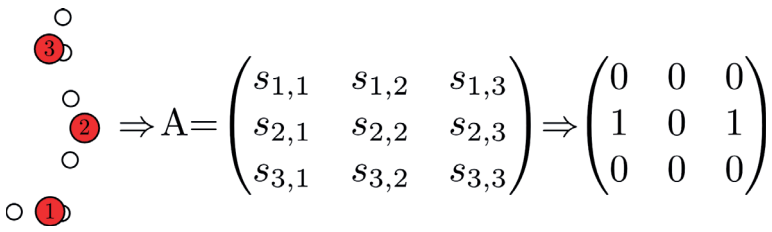


Figure 22: Schematic representation of three hydrogen bonded water molecules (left) and (right) the adjacency matrix associated to the graph built having the oxygens as nodes and $s_{O,O'}$ as edges.

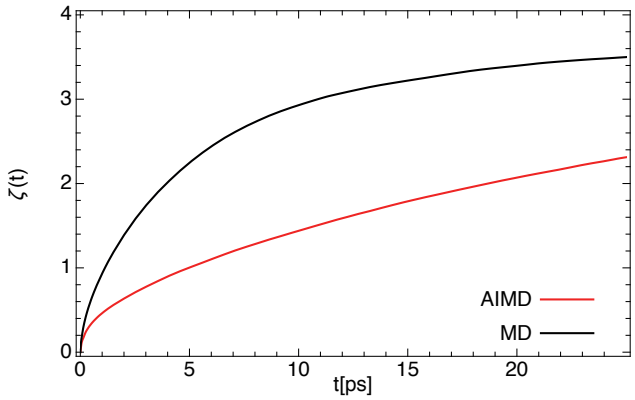


Figure 23: Adjacency matrix relaxation curve $\zeta(t)$ for a classical MD simulation (black) compared to a classical *ab-initio* MD simulation (red). The fact that $\zeta(t)$ reaches a plateau at a value close to 4 after certain time means that each water molecule has changed all four of the HBs in which it was involved.

where r^* is a fixed distance, e.g. the position of the first maximum or first minimum in the averaged RDF, $g_0(r^*)$ is the value of the RDF in r^* at the instant $t = 0$, $g_t(r^*)$ is the value of $g(r^*)$ at the time t, and $\sigma_g^2 = \langle g(r^*)^2 \rangle - \langle g(r^*) \rangle^2$.

The autocorrelation function $c_{rr}(t)$ describes how quickly the trajectory loses memory of fluctuations away from the mean, we can thus use such information to have a clear idea of the timescale in which we have a structural rearrangement of the whole system.

In Figure 24(b) we compare the analysis results – in the case of the TIP4P simulation just mentioned above – for the $c_{rr}(t)$ curves for the first maximum and first minimum RDF values (fixing r^* at the average positions shown in Figure 24(a)). Both the curves show a first initial fast decay followed by a long tail that is representative of a really slow decay. This is consistent with the PAMM network analysis result: in order to lose memory of an initial structural configuration one should wait a time on the order of tens of picoseconds.

5.3.3 *Liquid ammonia*

As a final example, we considered liquid ammonia at 180K and ambient pressure. Configurations were kindly provided by Joshua More and David Manolopoulos. [246] Simulations were performed including nuclear quantum effects by path integral molecular dynamics, using 12 beads and PIGLET [94] as implemented in i-PI. [241] Quantum Espresso [247] was used as the force back-end, with a PBE exchange-correlation functional [248] and ultra-soft pseudo-potentials. [249] The simulation box contained 32 molecules, and trajectories were per-

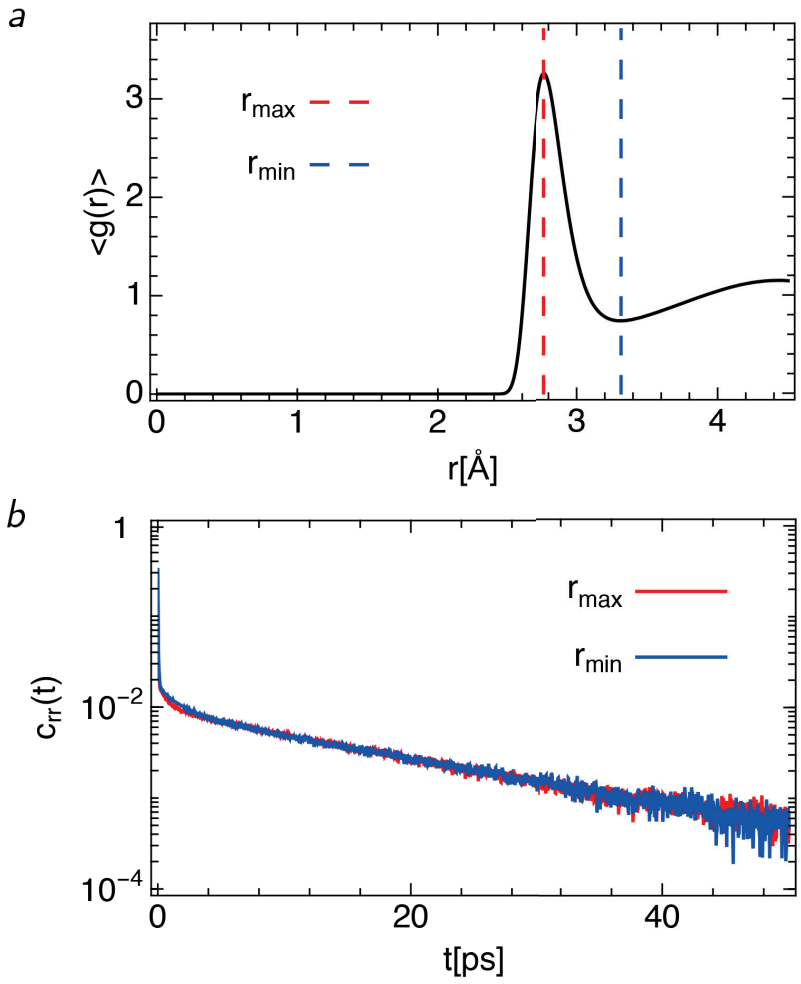


Figure 24: a) The RDF for TIP4P-2005f water at 300K: the position of the first maximum (red dashed line) and of the first minimum (blue dashed line) are marked. b) Comparison of the semi-log plots of the autocorrelation curves for the first maximum and first minimum RDF values. The fixed positions during the ACF computation are those shown in a).

formed for 10ps at constant, experimental density, with the first 2 ps discarded for equilibration.

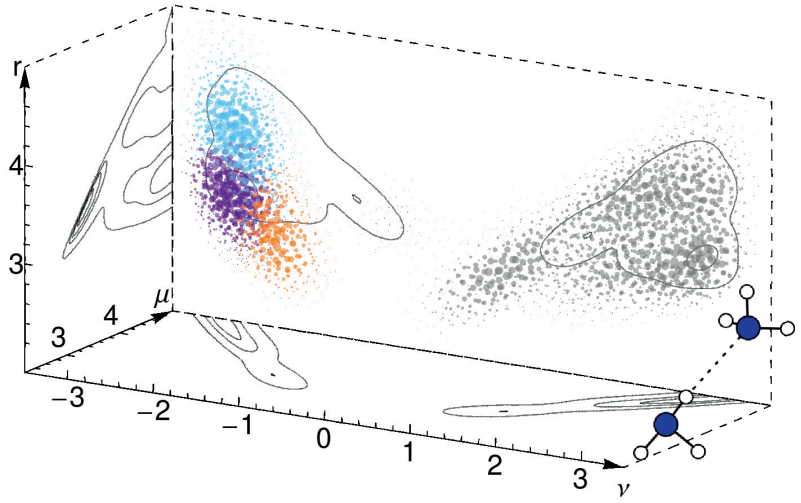


Figure 25: Distribution of (ν, μ, r) configurations for O–H$\cdots$O′ in a simulation of liquid BLYP ammonia at 180K. Size and opacity of points correspond to the KDE of $P(\mathbf{y})$, and colors indicate the cluster each grid point has been assigned to. Clusters with $\nu > 0$ have not been colored, but have been correctly identified by PAMM.

Ammonia is a less-structured liquid than water, with weaker hydrogen bonds, as it is already apparent from the probability density shown in Figure 25. Clusters are barely recognizable when using a two-dimensional (ν, r) representation, which was capable of characterizing the HB in all the other cases we considered. Clustering is much more evident using the three-dimensional (ν, μ, r) description, and PAMM clearly identifies a range of values that can be ascribed to hydrogen-bonded configurations. We expect the observation that higher-dimensional descriptors offer increased discriminating power to be general, and provide a strategy to resolve weakly structured systems. However, as the dimensionality is increased it becomes more difficult to converge the probability distribution, so longer simulations are needed and PAMM becomes more sensitive to the parameters of the procedure. As it is the case for oxygen in H_2O, nitrogen atoms act both as donors and acceptors, leading to a symmetric structure for $P(\nu, \mu, r)$.

Figure 26 shows the analysis of the distribution of s_D, s_A and s_H. Despite the low temperature, the weaker and less directional HBs result in a less clear-cut distribution of HB environments. A little more than 50% of the nitrogen atoms receive and donate 3 HBs – the ideal HB pattern that is observed in the solid phases of ammonia. Even though the simulation included nuclear quantum effects, there is no

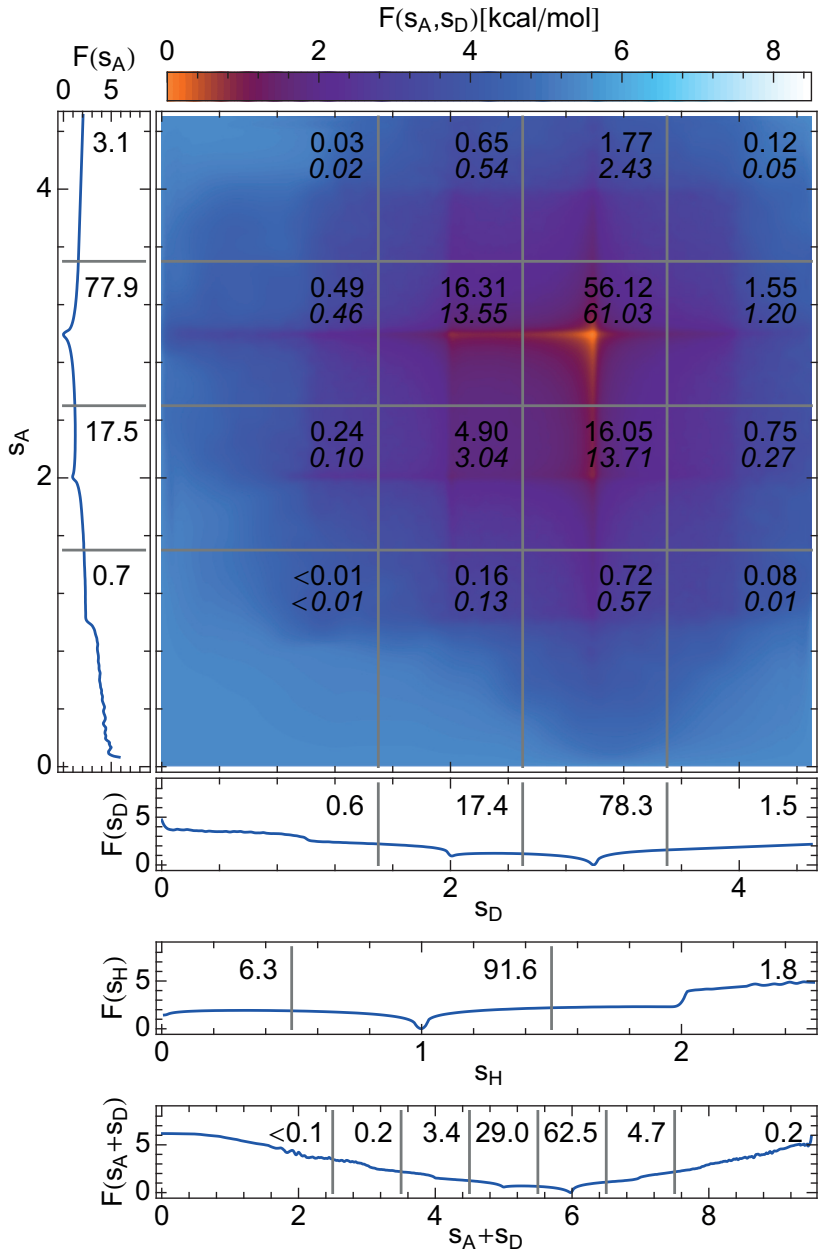


Figure 26: Hydrogen-bond counts statistics for a classical simulation of BLYP liquid ammonia at 180K. See the caption of Figure 15 for a detailed explanation of the plots.

trace of the diagonal patterns that are a manifestation of extreme proton excursions along the hydrogen bond. Molecules maintain strictly their chemical integrity, and from a structural standpoint these weak HBs appear to have a purely electrostatic character.

SUMMARY

We demonstrated the effectiveness of PAMM as a tool for recognizing a ubiquitous but hard to define entity: namely, the hydrogen bond, in a variety of different contexts. For each donor-hydrogen-acceptor triplet of atoms, PAMM automatically identifies an appropriate range of structural parameters that provides an unbiased, agnostic definition of what constitutes a hydrogen bond for a given set of atoms and a particular atomistic model. In the case of an empirical forcefield model of solvated alanine dipeptide it identifies three (very different) ranges of configurations that qualify as a distinct, recurring patterns for the carbonyl oxygen accepting a HB from water, for the ammide nitrogen donating a HB to water, and for a hypothetical weak HB involving C_α atoms and water oxygens. In the latter case, the presence of a distinct feature in the probability distribution is probably an indirect effect of the structural correlations in the water HB network, be tween the HBs between water and the electro-negative atoms in alanine dipeptide and of the rigidity of the molecular backbone.

We then assessed the behavior of PAMM when performing a more detailed analysis of hydrogen bonding in water, comparing an empirical water model and a first-principles, density functional model of water with and without a description of the quantum nature of nuclei. We introduced a compact representation of the hydrogen-bonding properties of water molecules in terms of the total number of accepted and donated HBs, that arises naturally because PAMM identifies hydrogen-bonded configurations in terms of a smoothly varying HB count function. We demonstrated that these hydrogen bond counts can be used to study the dynamics of the hydrogen bond network, giving results that are fully compatible with well-established definitions of the hydrogen bond, and to identify coordination defects in the otherwise ideal HB network of ice Ih. This analysis also highlights the presence of characteristic features that are a signature of extreme excursions of protons along the hydrogen bond observed with a quantum description of nuclei. Finally, we discussed liquid ammonia as an example of a weakly hydrogen-bonded system, that shows a much less clear-cut partitioning of the probability distribution and a more varied ensemble of hydrogen-bonding molecular environments.

In all of these cases our algorithm provides an adaptive approach to define the hydrogen bond in a unique and unbiased way, that only

uses structural information and that can be easily exploited to recognize correlation between the HB patterns involving a given molecule. Even though we have used hydrogen bonding as a representative benchmark, application of the PAMM algorithm is by no means limited to this example. It could be used to recognize complex structural patterns in a variety of materials and compounds, and can be easily applied to bias molecular simulations to accelerate the interconversion between different (meta)stable atomic configurations.

6

DEFECTS AND CORRELATIONS IN THE HB-NETWORK OF WATER

Contents

Many of the anomalous physical and chemical properties of water can be understood in terms of its highly-structured hydrogen-bond network [250, 251]. Tetrahedrally coordinated water, with two donated and two accepted H-bonds constitutes the fundamental building block of such networks. Of course, this idealized tetrahedral environment can be heavily distorted by thermal [252] and quantum [253] fluctuations. In the liquid phase, coordination defects exist and their presence, concentration, and relative arrangement contribute to the structural and dynamical properties of water [251, 254, 255]. Here we investigate the properties of such coordination defects, with a particular focus on their structural correlations, by means of first-principles molecular dynamics simulations.

Over the last three decades, considerable effort and progress has been made in the simulation of liquid water from first principles calculations. In this regard, DFT-based ab initio simulations have elucidated the importance of several factors such as the quality of the electronic structure, the treatment of nuclear quantum effects and also the role of statistical sampling [256–276] in reproducing the experimentally available oxygen-oxygen pair correlation function of water (see Figure 28). Unlike the idealized tetrahedral structure of ice, finite temperature fluctuations create defects which are a small fraction of the H-bond network and thus challenging to sample. Here we have taken exceptional precautions to ensure as extensive as possible thermodynamic sampling to collect meaningful statistics on the populations and structure of defects that in some cases contribute to less

This chapter is an adaptation of ref. [3]

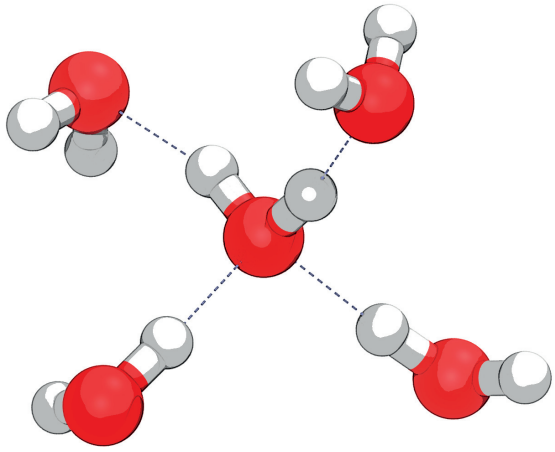


Figure 27: Ideal tetrahedrally coordinated water, with two donated and two accepted H-bonds.

than a percent of the H-bond network of liquid water. To achieve this, we used parallel tempering combined with the well-tempered ensemble [124] (PTWTE) to perform an extensive sampling of a box of 64 and 128 water molecules.

Using different models for the inter-atomic forces, or different thermodynamic conditions, may change slightly the typical geometry of a hydrogen bond.

We use PAMM to make the structural definition of a hydrogen bond and the classification of the different coordination environments independent from these effects.

Since in this study we restrict ourselves to a narrow range of thermodynamic conditions, this choice does not entail dramatic differences relative to one of the more traditional definitions. It provides, however, a robust framework that makes it straightforward to perform a similar analysis for different systems or to investigate more dramatic changes in environmental factors. We also report on the impact of different details of the electronic structure calculation, although as we will see, for a given choice of exchange-correlation functional, the main factor contributing significantly to the structure and population of defects is the use or neglect of dispersion corrections. This is consistent with previous observations which have shown that the inclusion of dispersion corrections significantly reduces the overstructuring that is seen for the most common choices of exchange-correlation density functional [268, 277, 278]

The topological constraints induced by the presence of an extended H-bond network mean that defects appear at highly correlated positions and are hence clustered together with different propensities [255, 279]. What is more, the details of the description of the inter-atomic

forces do not significantly change the population of defects, nor their relative structural correlations. Although the RDF provides useful information on the structure of the system that can be readily compared with accurate experiments [280] and is therefore regarded as the holy grail for benchmarking the quality of ab initio models, it averages over all the underlying complexity of the topology of the HB network and its directional correlations [281]. As we will show here, one should consider the RDF as arising from the combination of correlations between different ideal or defective coordination environments. We present an extremely thorough analysis of such correlations, that assesses the impact of many different computational details, and shows which of these matter, and which only cause small changes to the RDF but no profound qualitative change to the topological properties of the H-bond network. Furthermore, we use defect-resolved three dimensional distribution functions to elucidate the role of the weak interactions that are characteristic of undercoordinated environments. We find that the interactions between defects formed in the network have a remarkably directional character that could be interpreted as arising from weak rather than altogether broken hydrogen bonds – and link these angular correlations to those found in solid phases of water. While our analysis here focuses on thermodynamic, time-independent properties, we believe the structure and correlations of H-bond defects will prove crucial to understand the fluctuations and hence dynamics of liquid water in future studies.

6.1 COMPUTATIONAL METHODS

In this work we have used AIMD simulations based on DFT coupled with PTWTE. We will begin by first summarizing details of the electronic structure methods used and then describe the protocol we applied for the PTWTE simulations.

6.1.1 *Ab Initio methods*

The electronic structure calculations for computing the energies and forces were conducted using Quickstep which is part of the CP2K package [282]. The molecular dynamics and parallel tempering simulations were performed using the recently released code i-PI, that decouples the calculation of the interatomic forces from the dynamic evolution of the nuclei [241]. A convergence criterion of 5×10^{-7}a.u was used for the optimization of the wavefunction in all the simulations. Unless otherwise stated, the wavefunction was expanded in a DZVP Gaussian basis set, and an auxiliary basis set of plane waves was used to expand the electron density up to a cutoff of 300Ry. The

D3 Grimme dispersion corrections [283] for the van der Waals (VDW) interactions were used for most of the simulations. We used the BLYP generalized gradient correction [284] to the local density approximation and Goedecker-Teter-Hutter (GTH) pseudopotentials [285]. All simulations were thermostatted within the NVT ensemble using the canonical-sampling velocity-rescaling thermostat [286] with a time constant of 1fs. To maximize sampling of uncorrelated potential energy structures and accelerate replica exchanges in PT simulations, we also included a generalized Langevin equation thermostat tuned for efficient sampling [8, 93].

Extensive tests of the sensitivity of results to different computational details were performed using a box of side length 12.4138Å with 64 water molecules, corresponding to the experimental density of the system at 300K. The different PT simulations that were conducted included the following: BLYP without D3 Grimme's dispersion correction (BLYP+NOVDW), BLYP with dispersion corrections (BLYP+VDW), BLYP+VDW simulations with the TZV2P basis set (BLYP+VDW+TZV2P) and finally BLYP+VDW simulations using a plane wave cutoff of 350Ry (BLYP+VDW+350). For all the previously described PT runs, a timestep of 1fs was used. Most AIMD simulations using Born Oppenheimer molecular dynamics use a smaller timestep of 0.5 fs, which is necessary to obtain accurate real-time dynamics, but does not change significantly structural properties. In order to assess the sensitivity of our results to the choice of a larger timestep than commonly used, PT simulations were also conducted using BLYP+VDW with a timestep of 0.5fs (BLYP+VDW+0.5fs) which shows that using a larger timestep does not qualitatively change the structural properties of the system such as the diversity of different defects in the system.

The PT runs for the 64 water boxes detailed above were used to identify parameters for our production simulations with a larger box. We thus also performed parallel tempering simulations of 128 water molecules with a box size of 15.6404Å using BLYP+VDW, DZVP, 300Ry cutoff and 1fs time step. This simulation will be referred to as PTL. From our PTL simulations, we initiated four independent PIGLET simulations, using six beads and colored-noise thermostatting [94], to assess the role of nuclear quantum effects (NQE).

Besides the simulations using the BLYP functional, we also conducted simulations with the more expensive hybrid functional B3LYP implemented in CP2K [287], with D3 vdW corrections, in order to ascertain the role of electron exchange on the properties of liquid water. Due to the high cost of including Hartree-Fock exchange, we could not run replica-exchange simulations. Instead, we ran four independent simulations of about 16ps each, starting from BLYP equilibrated configurations extracted from the BLYP+VDW PT simulation consist-

ing of 64 waters, and discarding the first 2ps for equilibration with the hybrid functional.

Although the combination of hybrid functionals and dispersion corrections appear to improve the properties of ab initio water, we wanted to ascertain whether the defect correlations were not exclusive to DFT simulations. Hence, to complement our AIMD simulations, we also investigated populations and correlations of defects using force field models. In particular, we performed a classical simulation of 512 water molecules using a fixed point-charge model (TIP4P/2005f [288]) and classical and PIGLET simulations of 216 water molecules using MB-pol, a sophisticated force-field based on a many-body expansion and high-end quantum chemical reference calculations [289–291]. For non-reactive simulations of water, MB-pol is probably the most realistic theoretical model of the behaviour of molecular H_2O. Finally, we also make some comparisons of the topological defects found in liquid water to the coordination environment found in solid phases of ice.

In what follows, we will focus mostly on the results of the PTL simulations with 128 waters. Simulation details for the various systems considered for the analysis of the temperature dependence and electronic structure, are summarized in Table. 1 .

6.1.2 *Parallel tempering protocol*

Ab initio parallel tempering simulations for the 64 water boxes were performed using six replicas at the following temperatures: 290, 304, 322, 343, 365, 390K. In parallel tempering, a series of replicas are simulated at these six temperatures and thereafter exchanges between adjacent replicas are performed using a Metropolis criterion. In order to achieve efficient exchange between the replicas there must be significant overlap in the potential energy distributions of neighboring temperatures. Since a larger system exhibits smaller (relative) energy fluctuations, for PTL we used 8 replicas at the temperatures 290, 300, 310, 322, 335, 351, 369, 390K. It has recently been shown that by combining parallel tempering with the well tempered ensemble (PT+WTE), the overlap in energy between adjacent replicas is increased resulting in more frequent exchanges [124]. In the well tempered ensemble the enhancement of the fluctuations in energy is tuned by the γ factor which in our case was chosen to be 4. Exchanges between the replicas were attempted every MD step.

AIMD simulations are still prohibitively computationally expensive and hence it is rather challenging to perform a systematic benchmarking of parameters to tune the optimal number of replicas used as well as the γ factor. We decided to adopt a simplified protocol to determine the bias to generate the well-tempered ensemble. The

Models	N_{H_2O}	**vdW**	T **[K]**	δ**t [fs]**	**Time [ps]**	**Run type**
PT BLYP-GTH DZVP (300Ry)	64	No	290, 304, 322, 343, 365, 390	1	112	6 PT (NVT) runs
PT BLYP-GTH DZVP (300Ry)	64	Yes	290, 304, 322, 343, 365, 390	1	120	6 PT NVT runs
PT BLYP-GTH TZV2P (300Ry)	64	Yes	290, 304, 322, 343, 365, 390	1	39	6 PT NVT runs
PT BLYP-GTH DZVP (350Ry)	64	Yes	290, 304, 322, 343, 365, 390	1	52	6 PT NVT runs
PT BLYP-GTH DZVP (300Ry)	64	Yes	290, 304, 322, 343, 365, 390	0.5	25	6 PT NVT runs
PT BLYP-GTH DZVP (300Ry)	128	Yes	290, 300, 310, 322, 335, 351, 369, 390	1	145	8 PT NVT runs
MD BLYP-GTH DZVP (300 Ry)	64	Yes	300	0.5	30	4 independent NVT runs
MD BLYP-GTH DZVP (300 Ry)	128	Yes	300	0.5	16	4 independent NVT runs
MD B3LYP-GTH DZVP (300 Ry)	64	Yes	300	0.5	16	4 independent NVT runs
MD BLYP-GTH DZVP (300Ry) NQE	64	Yes	300	0.5	23	4 independent PIGLET NVT runs (6 beads each)
MD TIP4P/2005f	512	Yes	300	1	25000	1 NVT run
MD MBPOL	216	Yes	300	0.5	80	1 NVT run
MD MBPOL NQE	216	Yes	300	0.25	25	4 PI NVT runs (32 beads each)

Table 1: Computational details for all the simulations discussed in the main text. From left to right we report: a short title describing the model used for the potential energy surface, the presence or neglect of a correction for dispersion forces, the ensemble temperature, the integrator timestep, the length of each trajectory (e.g. in simulations with multiple independent runs, the total simulation time is the indicated length times the number of runs), and some remarks on the type of runs performed.

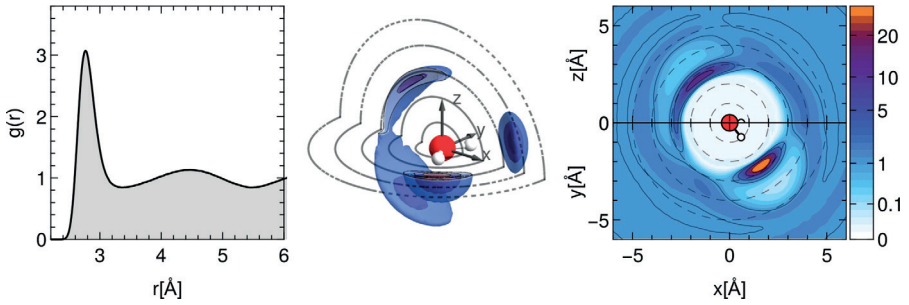


Figure 28: Oxygen-oxygen radial distribution functions computed at 300K from PTL (BLYP+VDW+PTWTE-128) runs. Left to right, the figures correspond to the usual 1D radial distribution, to the 3D distribution and finally to slices of the 3D distribution along the xy and xz molecular planes. Dashed circles are a drawn as a guide for the eye, indicating a radial grid with a 1Å spacing.

mean and fluctuations of potential energy as a function of temperature were determined based on a short (5ps) preliminary run. Then, a fixed bias was constructed and applied for the remainder of the simulation. Rather than using a (divergent) parabolic bias, we decided to give it a Gaussian shape with mean and curvature compatible with the measured fluctuations, but designed to cut off to zero for large fluctuations. Specifically, we used

$$B(V) = k_B T (\gamma - 1) \exp\left[-\frac{1}{2\gamma\delta V^2}\left(V - \bar{V}\right)^2\right], \tag{87}$$

where $\bar{V}$ and δV^2 are respectively the mean and variance of the potential energy evaluated for a given parallel tempering replica.

6.2 STRUCTURAL PATTERNS IN WATER

We will now introduce the different types of pair correlation functions that will be discussed in the rest of the chapter. The left-most panel of Fig. 28 illustrates the familiar oxygen-oxygen pair RDF obtained from the PTL simulations at 300K. The middle panel of Fig. 28 shows instead the 3D oxygen-oxygen correlation function, which contains information on angular correlations that are lost in the RDF. The contour plot in the right-most panel shows a cut of this 3D correlation function in the plane of the molecule (xy) and along the orthogonal symmetry plane (xz). The middle and right panels quantify the orientation of water molecules either accepting or donating hydrogen bonds in the first hydration shell of a particular water. The peaks at positive x correspond to the position of water molecules to which the tagged water is donating a hydrogen bond, and the broader peaks at negative x values involve the accepting side of the central molecule.

This is consistent with the notion of a ring of delocalized electron density, refered to as the 'negativity track' created by the lone-pairs which permits a larger range of local tetrahedrality [251]. Note that a peak in the density from both the donating and accepting side is a signature of highly directed nature of the hydrogen bonds. The importance of highlighting this feature will become clear when we show similar distributions for clustered defects.

One of the crucial observations we make is that there are significant correlations amongst defects in the hydrogen bond network. Since these defects tend to be rare, fleetingly lived and characterized by geometrical properties that are possibly different from idealized tetrahedral water, we wanted to ascertain how structurally intact or well-defined hydrogen bonds change under different definitions or thermodynamic conditions. In the previous chapter we have seen how, using PAMM, it is possible to give an adaptive definition of HBs, that could be made fully consistent with systems as diverse as alanine dipeptide, classical and quantum water, and liquid ammonia.

This definition has many advantages over more traditional ones. Firstly, it is probabilistic in nature and fuzzy: each O–H$\cdots$ O$'$ triplet is assigned a fingerprint that varies smoothly between 0 (no HB) and 1 (clear-cut HB). Secondly, it is adaptive: since it detects modes of the probability distribution in configuration space, the definition will change depending on temperature and water model, separating clearly the slight model dependence of HB geometry and the changes in populations of defects. Whereas a conventional geometrical definition would require a manual adjustment of its parameters [228], PAMM determines automatically, for each simulation scenario, which range of geometries should be considered to be a hydrogen bond. For the simulations conducted in this specific study – which is performed at the thermodynamic conditions to which the traditional hydrogen-bond definitions have been tuned – the choice of a PAMM definition over a conventional bond-angle definition introduces only minor differences and does not change our conclusions (see sec. 6.2.1).

6.2.1 *The role of the H-bond definition*

The rationale behind a PAMM-based HB definition is that – particularly for an inherently fuzzy chemical entity such as the hydrogen bond – the range of structural parameters that can be qualified as a bonding pattern depend on the thermodynamic conditions and inter-atomic potential, and should therefore be determined self-consistently.

This is a very different approach from what is done by most standard definitions. These approaches introduce heuristically a range of structural parameters that identifies a bonded configuration. Let

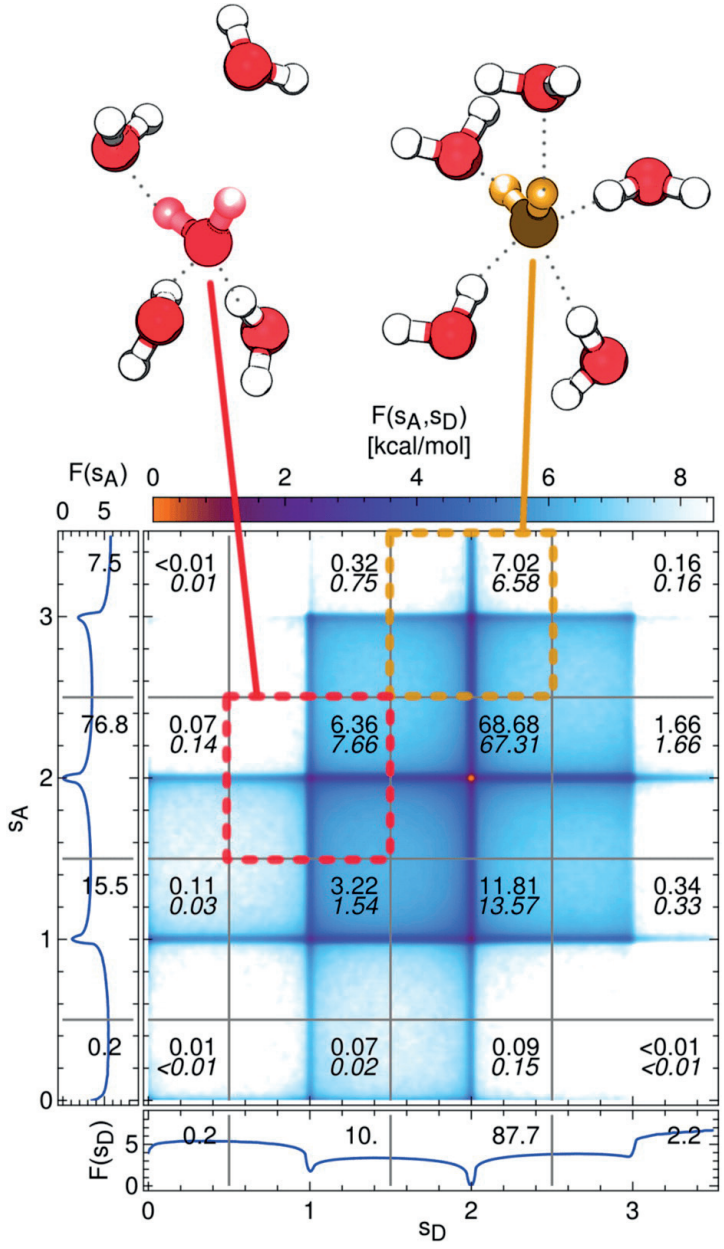


Figure 29: H-bond coordination summary for the PTL simulation data at T = 300K. s_D counts the H-bonds donated by a O atom, and s_A those accepted. Fractional values characterize fluctuations. The (s_D, s_A) range is partitioned in discrete regions that are assigned to different coordination states. For instance, the region with $0.5 \leqslant s_D < 1.5$ and $1.5 \leqslant s_A < 2.5$ is assigned to the 1_D2_A state. The numbers reported in each region correspond to the overall fraction of O atoms observed in a given state (in percent), while the number in italics is the corresponding value obtained as a product of the marginal probabilities. The larger the difference between the two values, the larger the correlations that exist between the donor and acceptor counts, for a given coordination state.

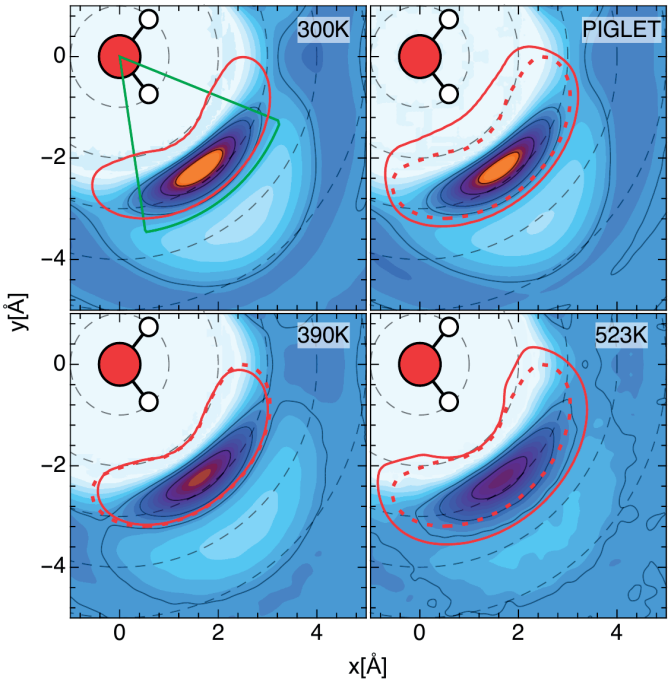


Figure 30: The four panels show a cut of the 3D O-O radial distribution function in one of the symmetry planes of a water molecule. The region for which a the O atom of a second water molecule would be identified as hydrogen-bonded is also highlighed. The top-left panel shows results for PTL runs at 300K: the green cone corresponds to a conventional definition of the H-bond [243], while the red line is the $p_1 = 0.5$ iso-contour for the self-consistent PAMM definition. The other panels, top to bottom, left to right correspond to a simulation at 300K including quantum effects, to the $T = 390$K replica of PTL, and to a classical simulation of BLYP water (with no VDW corrections) at 523K and the experimental coexistence density. In these panels, the PAMM definition obtained at 300K is reported as a dashed line, for reference, while the full line corresponds to the self-consistent PAMM definition in the various conditions.

us discuss briefly why this affects our analysis, and how it is important when extending our study to a broader range of thermodynamic conditions. To this aim, let us first compare the standard angle-distance definitions with the PAMM-based definition for a classical BLYP+VDW simulation at 300K. Such a comparison is performed in the top-left panel of Fig. 30, that shows a slice of the 3D O-O correlation function together with the region of space for which a hydrogen bond with a second O atom would be identified as formed when using the self-consistent PAMM definition (red contour) or when using the definition of Luzar and Chandler (green contour) [243].

The bond-angle definition tags two water molecules as bonded if their O-O distance is less than 3.5Å and if the O–H$\cdots$ O angle is less than 30°. In the case of the probabilistic, PAMM definition we show the contour where the fingerprint takes a value of 0.5. We consider the central molecule to have the ideal, rigid geometry typically used for empirical water models, and only vary in space the position of the putative acceptor. Both definitions single out the maximum of $g_{OO}(\mathbf{r})$ corresponding to the nearest-neighbor acceptor oxygen atom. This is not surprising, since the conventional definition has been tuned heuristically to give a reasonable description of H bonding at these thermodynamic conditions. One can see, however, that for this water model, this criterion cuts out too sharply the angular fluctuations, whereas the PAMM definition approaches more closely the saddle-point region in the distribution function.

To give an indication of how sensitive the PAMM definition is to changes in the underlying potential energy surface or thermodynamic conditions, fig. 30 also reports some other examples. As long as the general shape of the distribution function does not vary dramatically (as it is the case for instance when raising temperature at constant density) the changes in the PAMM isocontour are limited. Cases in which the distribution, or the shape of the molecule, are modified more dramatically lead to a pronounced change in the shape of the H-bond region. In particular, nuclear quantum fluctuations deform quite dramatically the geometry of a water molecule, and result in a considerably broader H-bond region. It should be noted, however, that quantum fluctuations lead to a stronger correlation between the covalent OH bond length and the H-bond definition – that cannot be fully captured by the plot in Fig. 30 that assumes a fixed molecular geometry. Increasing the temperature to 390K while keeping the density constant induces minimal changes to the H-bond definition, while simulations at 523K and the corresponding coexistence density of $798.598\mathrm{kg/m^3}$ show substantial changes to the region that is recognized as a H-bond.

To get a sense of the significance of the differences between various models, it is useful to examine how the populations of defects

change for a given water model, when one only changes the H-bond definition.

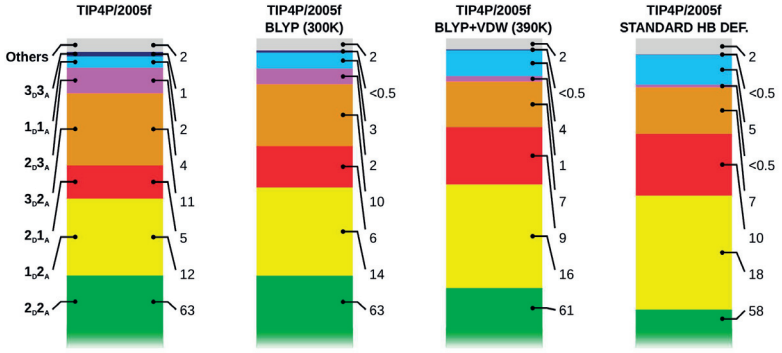


Figure 31: Bar charts showing the H-bond coordination summary for the TIP4P/2005f simulation using four different HB definitions. Left to right: PAMM-based definition trained from the TIP4P/2005f data, PAMM-based (trained from PT runs at 300K with no vdW corrections), PAMM-based (trained from the PTL runs at 390K), conventional bond/angle definition.

Fig. 31 shows bar-charts obtained for TIP4P/2005f water model using different definitions of the HB (see the caption for details). It is clear that the choice of the HB definition can introduce small changes in the relative proportion of defects, although qualitatively they all present very similar features. One should then be wary of discussing in too much detail the quantitative values of the defect populations, as that depends on the (rather arbitrary) choice of hydrogen-bond definition.

6.2.2 *Population of coordination defects*

The PAMM methodology provides a convenient approach to define three main H-bond count functions, s_D, s_A, and s_H. s_D quantifies the total number of HBs donated by a tagged O atom, s_A the number of HBs accepted by an atom and finally s_H quantifies the number of HBs that any particular hydrogen participates in [1]. These are obtained by summing the value of the PAMM HB fingerprint computed for all possible donor-H-acceptor triplets involving the tagged atom[1]. Based on these counts, one can build very informative defect stability maps as seen in Fig. 29, that summarize the relative probabilities of finding a water molecule in each of the different coordination states. It is clear that for the PTL simulations at 300K that most wa-

1 Of course, in practical implementations, a cutoff and a neighbor list can be used to maintain a linear scaling computational complexity of the analysis.

ter molecules donate and accept 2 hydrogen bonds. However, there is a sizable fraction of different types of topological defects. A nice feature of this type of analysis is that one can immediately point to asymmetries in the accepting vs donating abilities of hydrogen bonds - for example, there is a higher probability of finding water molecules that accept 2 and donate 1 hydrogen bond compared to those that accept 1 and donate 2 hydrogen bonds. We also see clearly the asymmetry in the distributions associated with a water molecule being a donor or acceptor which is consistent with previous observations by Agmon [251]. Similar maps for other simulations at a higher temperature and electronic structure approximations can be found in the SI of ref. [3].

An important advantage of PAMM relative to traditional H-bond definitions is that since the underlying fingerprint is probabilistic, and varies smoothly between zero and one, it is possible to recognize features in the transition regions between clear-cut defect states. The smooth definition of PAMM provides a qualitative description of the pathway from one defect state to another. Even though an order parameter based on a single site cannot capture quantitatively the complex collective modes that underlie the rearrangement of the H-bond network [292], the height of the barrier between two defect states gives an indication of the relative propensity towards a transition. For instance, one can see that in the overwhelming majority of cases the state of an O atom evolves by increasing or decreasing by one the donated or accepted hydrogen-bond count, while concerted transitions do not contribute significantly. By looking at cuts in the free energy surface at constant s_A or s_D, one can also get an idea of how the free energy barriers along the pathways to make or break hydrogen bonds change for molecules that are initially undercoordinated or overcoordinated (see fig. 32).

For instance, the barrier for the $1_D2_A \rightarrow 1_D1_A$ transition is lower by $\sim 33\%$ than the barrier for the $2_D2_A \rightarrow 2_D1_A$ transition. Similarly, the barrier for the $2_D1_A \rightarrow 1_D1_A$ transition is lower than the $2_D3_A \rightarrow 1_D3_A$ transition roughly by a factor of two. Interestingly, these qualitative trends appear to be quite robust to the choice of different approximations made in treating the underlying electronic potential. However, it should be noted that small differences in these barriers will lead to much larger changes in dynamical properties, thus making it even more challenging to achieve statistical convergence.

Although we will not exploit this aspect here, it is worth stressing that strictly speaking, PAMM identifies coordination of O atoms, rather than water molecules, which means that this classification could be used transparently also in the presence of charged defects and charge fluctuations, since one is not relying on the definition of molecular entities. For instance, in the presence of an excess proton, one

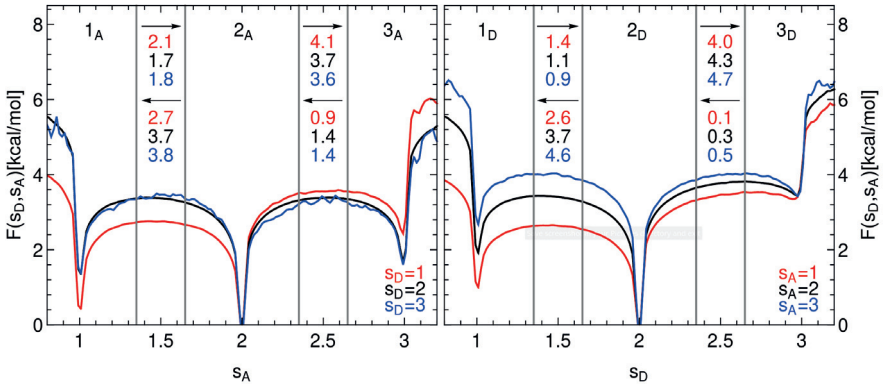


Figure 32: Cuts in the free energy surface at constant s_D (left) and s_A (right) for the PTL simulation. The three curves are colored according to the specified H-bonding state, as indicated. The numbers above the curves correspond to the free-energy barriers (in kcal/mol) in the two directions, that are computed by integrating over the free-energy basins around integer H-bond counts.

would expect to see an increase in density in the region with $s_D \approx 3$ and $s_A < 1$, or detect the quantum fluctuations of a proton along a hydrogen bond by the appearance of diagonal features (see Ref. [1]) that correspond to one hydrogen bond momentarily changing its character from acceptor to donor.

6.2.3 *Defect correlations and the RDF*

The hydrogen bond maps of s_A and s_D shown in fig. 29 provide a convenient way to classify water environments based on their coordination state [293, 294]. Given that the probability maxima are very clear-cut and with a rather obvious structure as seen in Fig. 29, we subdivided the map manually labelling e.g. $n_D m_A$ an oxygen atom that has $s_A \in [m_A - 0.5, m_A + 0.5)$ and $s_D \in [n_D - 0.5, n_D + 0.5)$.

Armed with this classification, we can proceed to investigate whether different $n_D m_A$ states are distributed randomly in the network, or whether significant correlations exist between them. Two-body spatial correlation functions provide powerful tools to recognize such correlations, and allow us to disentangle the underlying factors that control the shape of the overall O-O RDF shown in Fig. 28.

The simplest analysis involves coordination-resolved RDFs $n_D m_A - n'_D m'_A$ – that report on the probability of finding an O atom in a coordination state $n_D m_A$ and another in $n'_D m'_A$ at a distance r from one another. In addition, one can also look at the 3D distributions, as shown in Fig. 28, which give deeper insight into the angular position of waters within the first hydration shell. We will label as

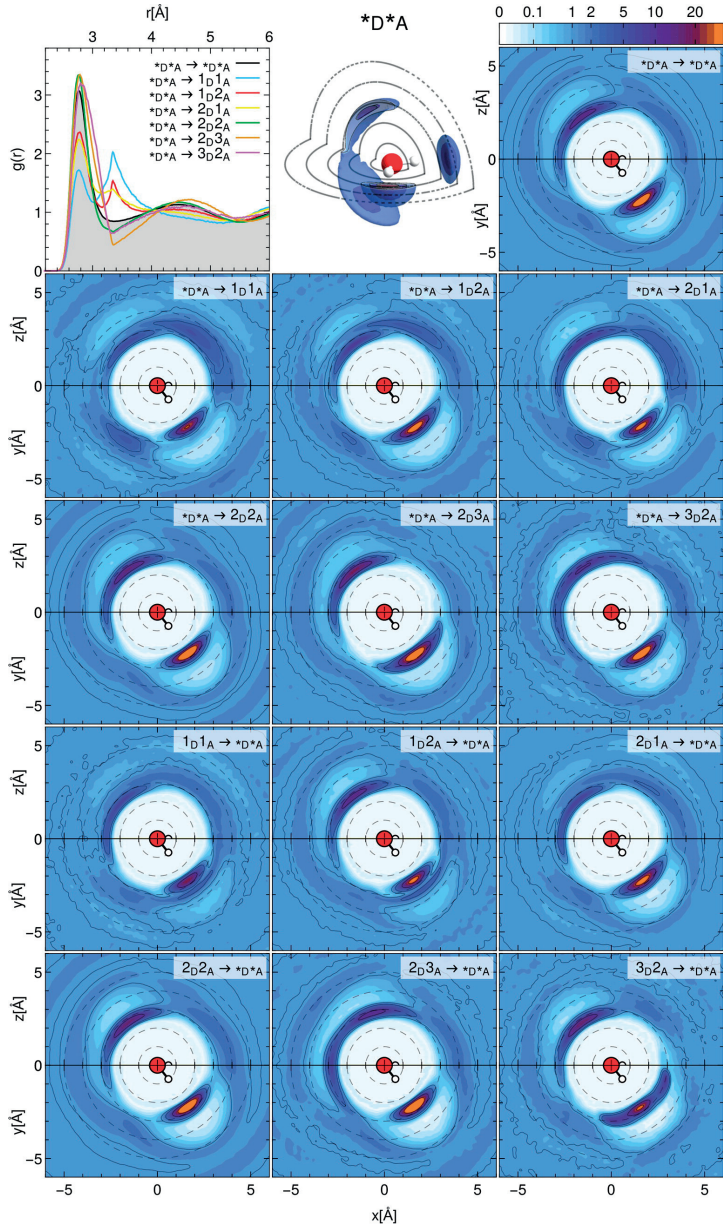


Figure 33: The figure reports concisely the O-O correlation functions involving a defective H-bond O environment, and a second oxygen atom without specification of its H-bonding state. The first row reports the baseline O-O correlations, as in Fig.28. The 1D radial distribution function reports all the distributions of oxygen atoms around the most important defect environments, with the baseline shaded in grey. The following six panels report slices along high-symmetry molecular planes of the 3D distribution of the various defects around a non-specified O atom, while the last six report the mirror distribution of an arbitrary O atom around the specified defective species.

$n_D m_A \rightarrow n'_D m'_A$ the distribution of $n'_D m'_A$ evaluated in the reference frame of a $n_D m_A$ molecule. It is important to recognize that for each pair of species, the distributions associated with $n_D m_A$ and $n'_D m'_A$ are *not* symmetric. If there is enhanced probability of finding $n'_D m'_A$ waters in the donor region of a $n_D m_A$ molecule, one can expect that viewed from the point of view of $n'_D m'_A$ this same correlation will amount to an enhancement in the *acceptor* region. Figure 33 provides an example of this kind of analysis, where we consider correlations between an un-specified $\star_D \star_A$ O atom and the main coordination states. The complete series of radial and 3D correlation functions, for all temperatures and electronic structure methods we considered in the present study can be found in the SI of ref [3]. Here we will only comment on the most significant correlations, that can shed some light on the complex topological features of the H-bond network of liquid water.

Let us start by commenting on the O-O radial distribution functions. It is tempting to analyze the changes of the overall O-O RDF with temperature and simulation details in terms of the components resolved into the different coordination states. Contrary to the inherent structure analysis, that identifies "low-density/high-tetrahedrality" and "high-density/low-tetrahedrality" by quenching instantaneous liquid configurations [295], the analysis we perform here includes snapshots that are fully consistent with the given thermodynamic state point. We have seen in Section 6.2.2 that the majority of water molecules corresponds to thermal fluctuations around a tetrahedral $2_D 2_A$ environment. All defective environments conspire to modify the shape of the short-range region of the RDF: over-coordinated defects do so by broadening the first peak, but would themselves increase the depth of the first minimum. Undercoordinated environments, instead, enhance the density in the interstitial region. This analysis thus provides an alternative interpretation of the structure of the RDF in terms of correlations between coordination defects on top of a dominant tetrahedral network, without explicitly invoking the existence of two thermodynamically distinct low and high-density water networks. Changing the simulation temperature, or the details of the electronic structure, modulate both the populations and the shape of the RDF of defect resolved individual components, although many of the qualitative features associated with the relative population of defects as well as the strutural correlations between them, are still conserved.

It is perhaps worth stressing that any decomposition of this kind that dissects a structural observable based on prior structural analysis, risks being tautological to some extent. For instance, the lowering of the first peak, and appearance of a second peak in the interstitial region for undercoordinated defects could be regarded just as hydrogen

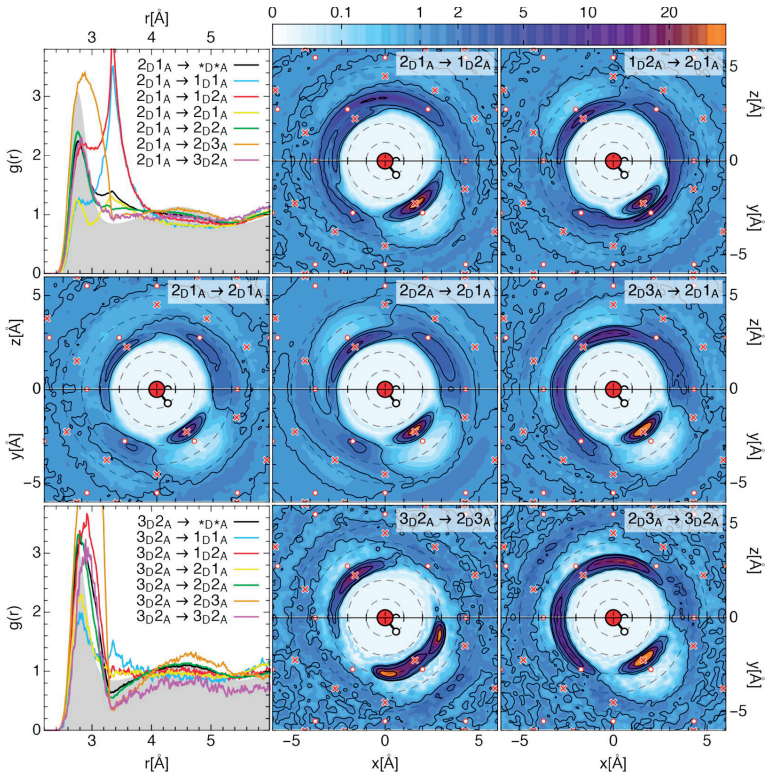


Figure 34: Defect-resolved structural correlations in the hydrogen-bond network of vdW-corrected BLYP water. We focus on a few particularly significant pairs of under-coordinated and over-coordinated defects. The structural correlations resemble some features seen in solid phases of ice. The crosses and circles correspond, respectively, to the positions of the nearest oxygen atoms in ice Ih, and ice VIII expanded to match the density of room temperature water.

bonds caught in the act of breaking up, rather than as a significant feature in the water network. An analysis of the 3D DF can identify more clearly the nature of the correlations between defects, telling apart artifacts of the analysis from genuine features of the H-bond network.

Focusing first on the under-coordinated water environments, and looking in the direction of the donated H-bond in the $1_D2_A \rightarrow \star_D\star_A$ and $1_D1_A \rightarrow \star_D\star_A$ correlation functions, one can see a sharp second peak that could indeed be interpreted as arising from the tail of the distribution of a "normal" H-bond, that is identified as broken by the PAMM fingerprint function (or the conventional structural definition). On the other hand, inspection of the mirror distributions $\star_D\star_A \rightarrow 1_D2_A$ and $\star_D\star_A \rightarrow 1_D1_A$ does not show a similar sharp peak just next to the "normal" acceptor peak. Rather, it reveals the presence of strong angular correlations in anomalous directions, that could be regarded as the manifestation of a weaker type of H-bond rather than truly unbound configurations or broken hydrogen bonds. Even though all simulations in the present work were thermostatted, making it not possible to extract rigorous dynamical information, it is clear from inspection of the trajectories that these weak hydrogen bonds, while forming a smaller part of the population in the hydrogen bond network, are not just fleetingly formed transition-state structures, but rather involve meta-stable states.

Fig. 33 quantifies the structural correlations for water molecules in the vicinity of different topological defects in the H-bond network, by fixing the coordination state of only one of the two oxygen atoms involved. Of course, one could proceed to look into pair correlation functions for which both species are in a prescribed coordination state. This more detailed analysis reveals that in many cases there appear to be strong correlations between the position of defective coordination states. In other terms, when fluctuations in the generally tetrahedral network generate topological defects, such low-probability environments appear to be clustered close to each other. Henchman and co-workers have performed a similar analysis looking at the RDFs exclusively for water molecules that are different acceptor types [255]. In particular, they observed for example, that water molecules that were single acceptors and triple acceptors tend to be close to each other. Here we considered more than 50 possible pairs of environments: all the results, for different models and simulation details, are reported in the SI of ref. [3], while here we focus on the most striking features that we could identify (fig. 34).

Undercoordinated water molecules do indeed tend to be strongly clustered together. The defect-resolved RDF between the 2_D1_A (red line) and 1_D2_A environments shows a very pronounced peak at about 3.5Å. Inspection of the directionally-resolved RDF reveals that this

sharp peak at least partly originates from the PAMM analysis that singles out a hydrogen bond in the act of breaking into a 2_D1_A–1_D2_A pair. However, the very broad angular spread of the peak at 3.5Å clearly paints a more complicated picture in which the weakening of the hydrogen bond is associated with greater conformational flexibility on both the donor and the acceptor side with respect to a tetrahedral 2_D2_A environment (see Fig. 33). In other terms, one can see this feature of hydrogen bonding as related to a form of entropic stabilization. In addition, 2_D1_A defects are associated with unusual angular correlations, with two very distinct peaks that can be seen at about 3.5Å distance, well separated from typical H-bond directions. These are seen in the $2_D1_A \rightarrow 1_D2_A$, $2_D2_A \rightarrow 1_D2_A$, and also in the $2_D3_A \rightarrow 2_D1_A$ correlations discussed in Ref. [255]. More generally, strong, anomalous angular correlations are observed for all undercoordinated species (see also the whole series of defect-resolved 3D DFs in the SI of ref. [3]), reinforcing the notion of the presence of a "weak" H-bonding mode that does not match the structural parameters range of a full-fledged hydrogen bond, but that contributes to the stability of coordination defects in liquid water.

In the liquid phase, the disorder and larger angular flexibility of water molecules makes it difficult to pinpoint the specific geometric features that lead to the peculiar angular correlations. We thus turned to high density phases of water at lower temperature to understand the origins of these correlations. In Figure 34 we overlaid markers that indicate the position of the neighboring O atoms in ice Ih (crosses) and for a model of ice VIII *expanded to match the density of room temperature water* (circles) on the 3D O-O distribution functions. The standard hydrogen-bonded peaks correspond perfectly to the position of nearest neighbors in hexagonal ice, but the additional angular correlations found around undercoordinated waters are closely related to the coordination environments in ice VIII. A relationship between "interstitial" waters and high-density phases of ice was suggested in Ref. [296] as a method to asses the accuracy of different electronic structure methods in describing defective environments in water. Our analysis confirms this intuition, and suggests that an even more representative benchmark could be obtained by expanding the unit cell to match the density of water at ambient conditions.

Fig. 34 also shows that 2_D3_A and 3_D2_A environments are very strongly correlated. In particular the $3_D2_A \rightarrow 2_D3_A$ distribution function shows two distinct peaks at the brim of the donated H-bond region, suggesting that in most cases these configurations are associated with bifurcated H-bonds. The inverse distribution $2_D3_A \rightarrow 3_D2_A$ shows the characteristic trigonal distribution of accepted H-bonds, however with broader angular fluctuations that once again point at the increased flexibility associated with correlated defective environments.

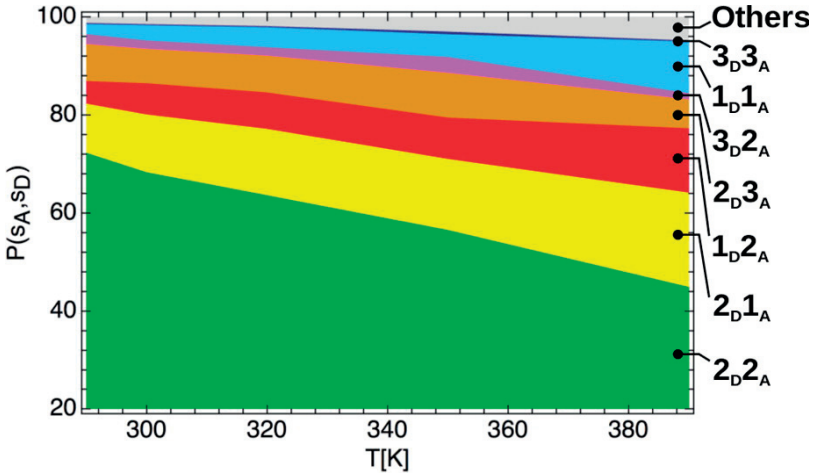


Figure 35: Temperature dependence (at constant room-temperature density) of the fraction of the main H-bonding defect states in the PTL simulations (BLYP+VDW, 128 water molecules).

6.3 COMPARISON OF WATER MODELS

In the preceding discussions we have focused on the structural correlations of defects that we have found by examining our production runs of the 128 water system at 300K (PTL). In the following, we will now describe how defect populations and correlations change as a function of finite temperature, box size, the quality of the electronic structure and nuclear quantum effects. We find that topological defects appear to be present with similar concentrations across all the simulations. Furthermore, the structural correlations between them also appear to be qualitatively and sometimes even quantitatively conserved. However, the relative concentration of different defects changes in a subtle manner.

We begin by showing how the proportion of different types of structural defects change in water as a function of temperature at constant, room-temperature density. As one moves from 290K to 390K (Fig. 35), there is a clear decrease, by about 25%, in the proportion of water molecules that accept and donate two HBs. This is in turn accompanied by an increase in the number of structural defects. In particular, in going from 290K to 390K there is a significant increase in the number of water molecules accepting 2 and donating 1 HBs, accepting 1 and donating 2 hydrogen bonds and a smaller increase in the the number of defects accepting and donating 1 hydrogen bond. The concentration of overcoordinated defects such as the 2_D3_A and 3_D2_A stays more or less constant, at least in the constant density conditions we are currently simulating. Despite the increase in the concentration of defects with growing temperature, the 3D correlation plots reveal

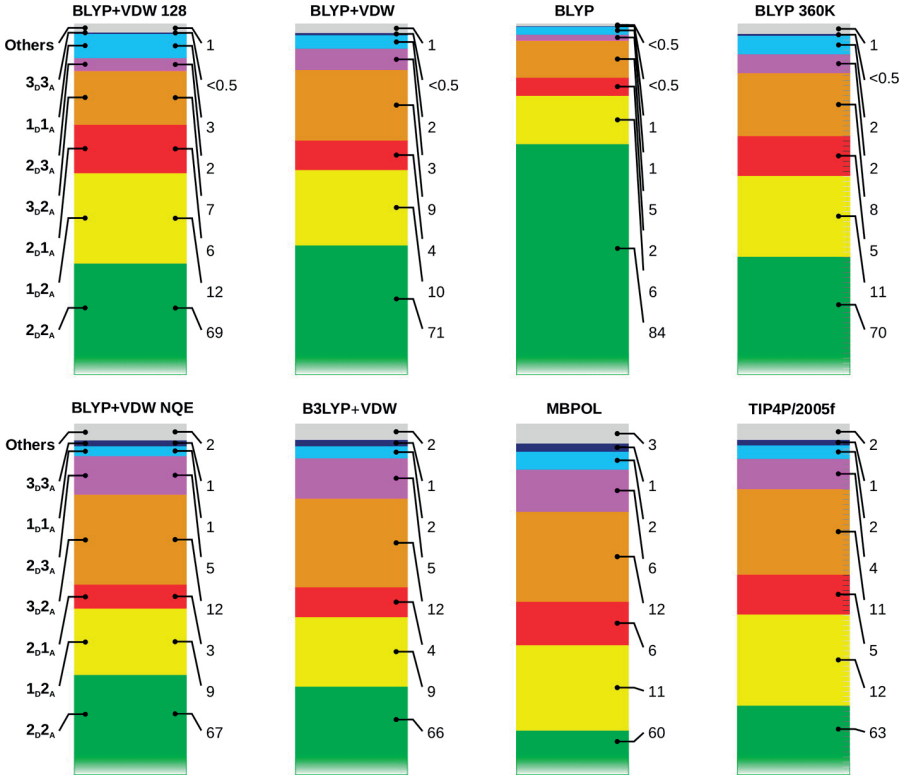


Figure 36: The bar-charts show the percentage of the main defect states considered in the text for different models explored in this work. The label on top of each chart indicates the simulation protocol (functional, system size, temperature, the inclusion of van-der-Waals corrections). On the extreme left, in bold, the defect type corresponding to each segment of the stack plots is indicated. The numbers on the right of each bar-chart indicate the percentage of each defect for the specified batch of simulations. Note that the segment corresponding to the majority, tetrahedral 2_D2_A environments has been truncated for clarity.

the presence of well formed directed hydrogen bonds up to 390K (see Fig. 39–41).

Fig. 36 shows the percentage of the main defects in the water network that we obtained with various simulation protocols. We compare the effect of finite box size (moving from 128 to 64 water molecules, that we used for most comparisons), the role of nuclear quantum effects and the use of model potentials such as TIP4P/2005 and the more recently developed MB-pol water model. For the most part, we see that the relative proportion of coordination defects is not very sensitive to the simulation protocol. In the case of MB-pol, the number of 2_D2_A waters is lower than that of BLYP+VDW PTWTE 64 by about 10%, which is compensated by a slight increase in the under-

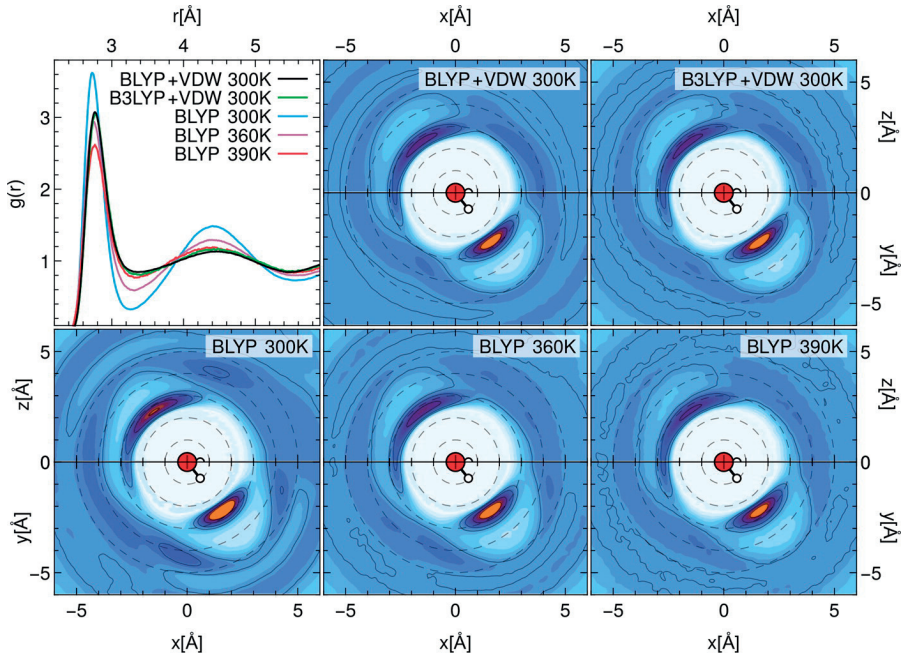


Figure 37: Radial and 3D O-O distribution functions, for selected simulation conditions. The bottom row compares BLYP-based simulations without vdW corrections at three different temperatures.

coordinated 1_D2_A and 2_D1_A defects in the network. It is thus rather comforting to see that all these various protocols for simulating liquid water at ambient conditions produce qualitatively consistent results at least with respect to the types of coordination defects in the hydrogen bond network. On the other hand, the predictions for physical properties of water such as the RDF and the diffusion constant, *do* change significantly between these models. Seeing how such changes depend on small differences in the structure and stability of topological defects gives some rationale for the difficulty in obtaining a quantitatively accurate description of the structural and dynamical properties of water.

The clear outlier between the defect population plots in Fig. 36 is the case of the BLYP functional without vdW corrections that leads to a dramatic increase of the fully-coordinated tetrahedral environments. Indeed, it is well appreciated that standard generalized gradient approximation functionals used in AIMD simulations lead to the overstructuring of the hydrogen bonds in liquid water. A common trick that has been suggested in early CPMD simulations, and which has been used in many AIMD simulations thereafter, is to increase the temperature in the simulation. This trick was applied to account for the supercooled nature of DFT water at 300K [297]. Typical values that have been used are 330K but there have also been sugges-

tions that temperatures above 400K are needed [277]. When this is done, it is empirically observed that the RDF gets less structured and agrees more closely with experiments. Indeed, we see that by raising the temperature to 360K (the right-most stack plot in the first row of fig. 36) one can reproduce quite accurately the defect populations obtained with vdW corrections.

Earlier, we alluded to the fact that one can think of the ensemble-averaged RDF as coming from individual contributions involving structural correlations between tetrahedral waters together with those from the clustering of different types of defects. To appreciate a bit better the challenge in converging this property, in Fig. 37 we compare the RDFs obtained with vdW corrections to those with the bare BLYP functional, at 300K, 360K and 390K, to illustrate how the choice of simulation temperature and the inclusion of dispersion interactions conspire to affect different parts of the distribution. It is clear that van-der-Waals interactions cannot be fully mimicked by an increase in simulation temperature. Without dispersion, the simulations at 360K reproduce the height of the first peak of the distribution, but the long-range oscillations in the RDF are still considerably stronger than the vdW-corrected reference. One has to increase the temperature up to 390K to approach the corrected long-range behavior, at the expense however of lowering too much the height of the first peak. Angular correlations, that are seen in the 3D distribution functions, clarify the source of this discrepancy. The bare GGA simulation shows pronounced peaks in the second coordination shell, that correspond to the angular positions seen in ice Ih (see also Fig 34). Raising the temperature broadens this peak, and shifts it towards the interstitial region that is characteristic of under-coordinated defects. At the same time, the increased temperature enhances the fluctuations and lowers the height of the first-neighbor peak. There is no temperature at which thermal fluctuations match the effect of vdW corrections on these two components simultaneously – performing simulations at an artificially increased temperature is a poor substitute for a model that describes properly dispersion interactions. vdW interactions thus play an important role in tuning both the short (first shell) and long-range (second shell and beyond parts of the RDF although the effect is more pronounced on the latter. Similar conclusions have also been made by Weeks and co-workers examining classical models of water such as SPC/E with molecular field theories [298].

There is currently an ongoing lively debate regarding the level of electronic structure theory required to reproduce structural and dynamical properties of liquid water. In particular, several studies have advocated for the need of including a certain amount of Hartree-Fock exchange in the exchange-correlation functional [276, 287]. For this

reason, we performed AIMD simulations with the B3LYP hybrid functional. Although we could not afford replica-exchange simulations, we ran four independent simulations at 300K, for a total of more than 60ps. The distributions with BLYP and B3LYP (both including dispersion interactions) show that the relative proportion of different types of coordination states are almost quantitatively the same. Also, radial and angular-resolved distribution functions are remarkably similar. Thus, while it has been previously observed that the use of hybrid functionals alone gives much better RDFs than bare GGAs [276], it appears that dispersion corrections can similarly remedy the most blatant deficiencies of BLYP, and that combining the two does not have a major effect compared to applying the exchange or vdW corrections separately.

The delicate balance between an ice-Ih-like and a defective hydrogen-bond network cannot be ascribed to a single approximation in the electronic structure framework. What is more, one should not focus too much on the RDF as the only benchmark to assess the quality of a water model: as we have shown, angular correlations and defect-resolved distribution functions contain much more detailed information, and other physical properties (such as thermodynamic and dynamical properties [299], or isotope fractionation ratios [300]) should also be included to avoid the risk of obtaining an RDF that matches experiment for the wrong reasons. Empirical vdW corrections seem to be enough to reproduce the experimental RDF, but there is strong evidence that this is largely due to a cancellation of errors between three and four-body terms [301], and it has been shown that the description of the water monomer energy is very poor in the absence of an exact exchange correction [302].

6.3.1 *Impact of simulation details on defect correlations*

The main effect of changing the model of inter-atomic forces is the modulation of the population of H-bond defects. Defect-defect correlations are robust to details in the inter-molecular potential, and can be traced to topological constraints in the network. In what follows we report, for selected defects, coordination-state-resolved 3D oxygen distribution functions computed from all of the different models we have considered. All show remarkably similar features, and even the BLYP simulations with no vdW corrections – that stand out as the outlier in all of our analyses – display the same angular correlations between under-coordinated defects that are observed for the more accurate models. The main structural difference induced by neglect of dispersion corrections is the over-structuring of the fully tetrahedral H-bond network, that can be understood easily as it underlies long-range order in hexagonal ice. For completeness, the full listing of

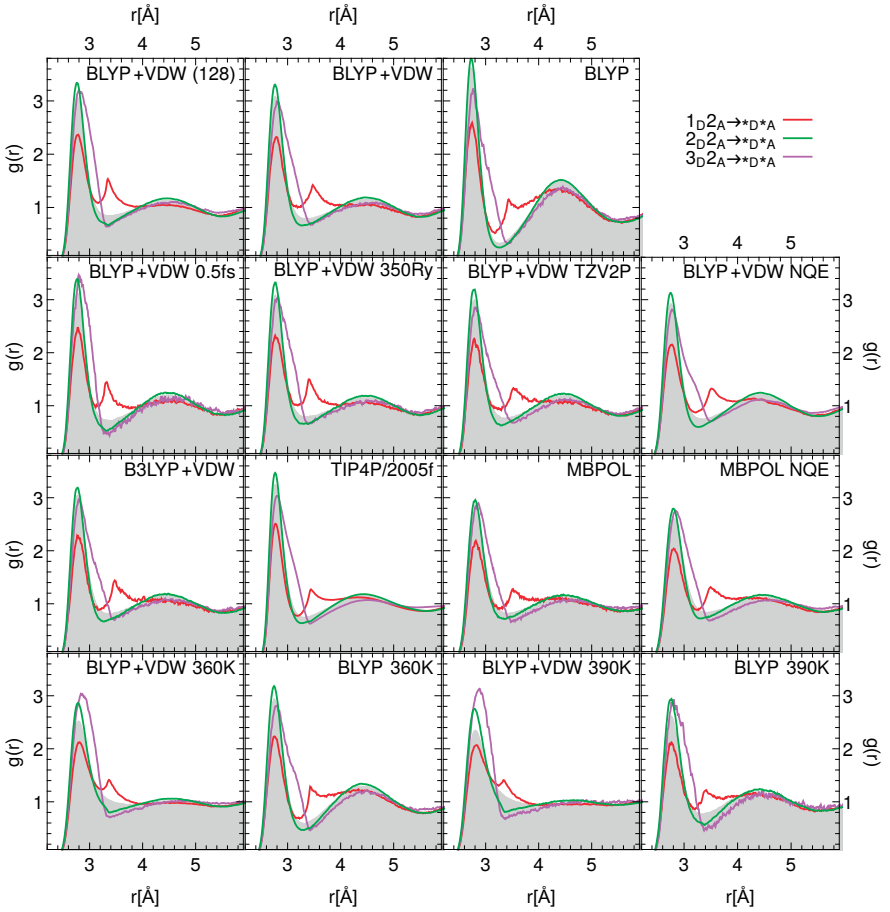


Figure 38: O-O radial distribution function for selected H-bond defects, computed for all the simulation protocols discussed in the main text. The baseline shaded in gray corresponds to the total O-O RDF.

correlation plots, for all water models and defect pairs, is provided as an electronic archive in the SI of ref. [3].

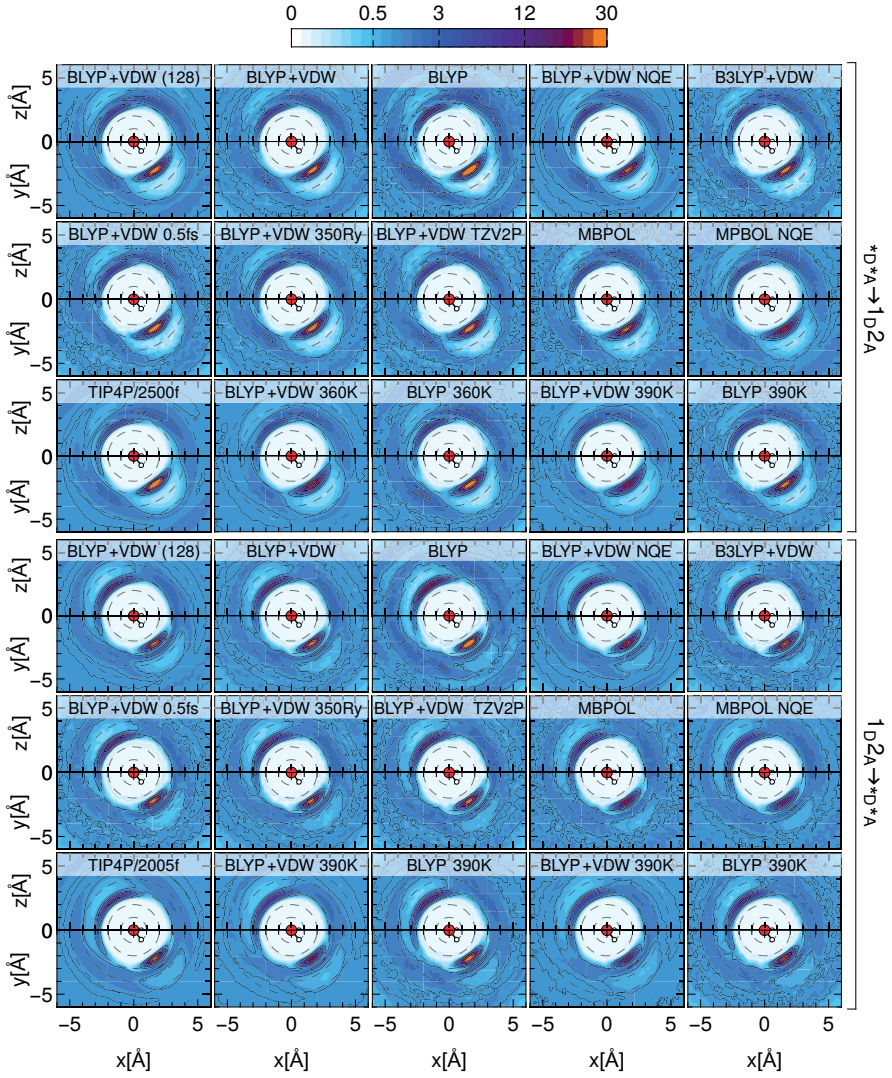


Figure 39: 3D O-O distribution function for 1_D2_A defects, computed for all the simulation protocols discussed in the main text. The first 15 figures are obtained considering the distribution of the tagged O around an arbitrary water molecule, and the second 15 figures to the mirror distribution, with the tagged O atom fixed at the origin.

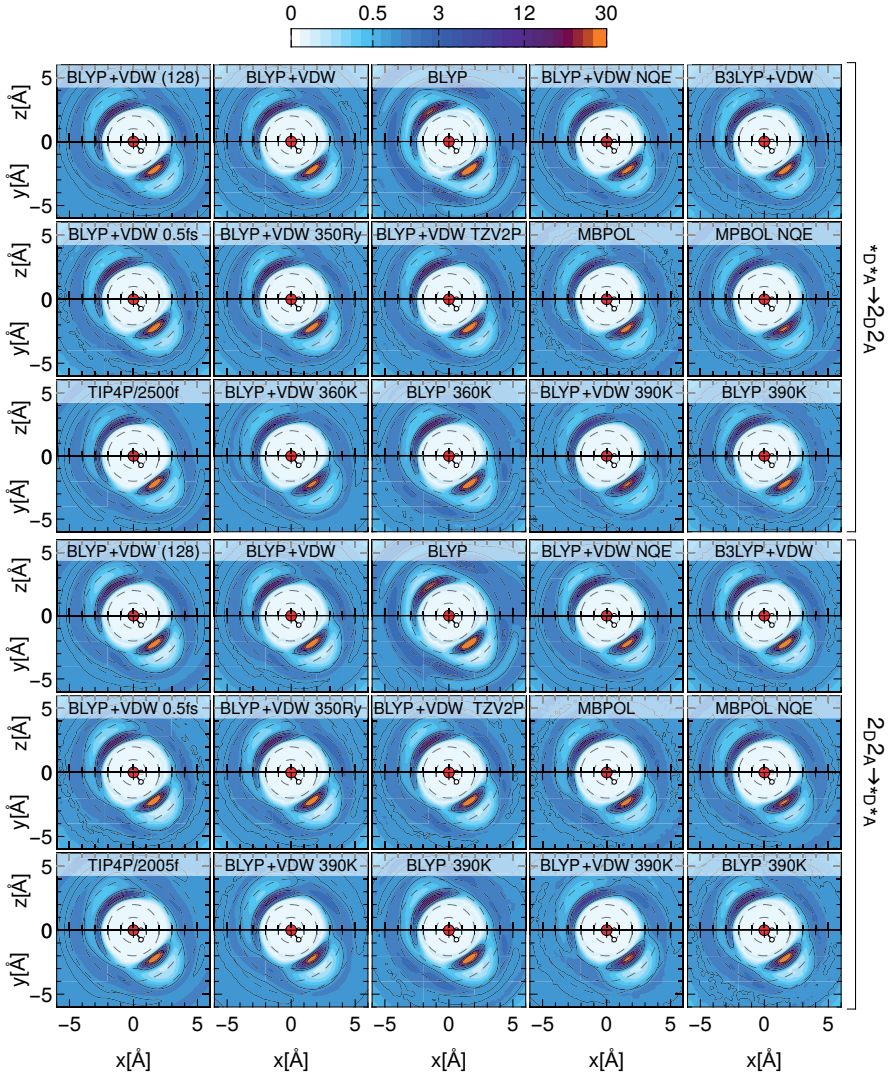


Figure 40: 3D O-O distribution function for 2_D2_A defects, computed for all the simulation protocols discussed in the main text. The first 15 figures are obtained considering the distribution of the tagged O around an arbitrary water molecule, and the second 15 figures to the mirror distribution, with the tagged O atom fixed at the origin.

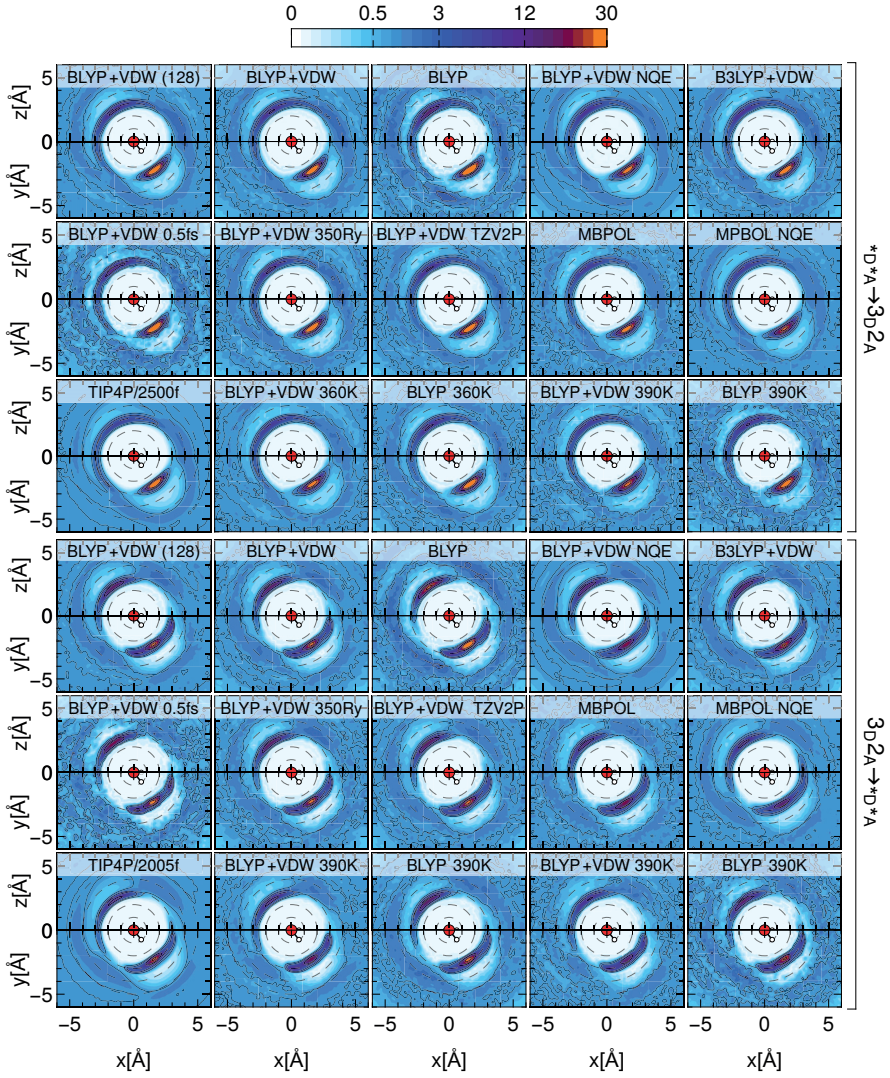


Figure 41: 3D O-O distribution function for 3_D2_A defects, computed for all the simulation protocols discussed in the main text. The first 15 figures are obtained considering the distribution of the tagged O around an arbitrary water molecule, and the second 15 figures to the mirror distribution, with the tagged O atom fixed at the origin.

SUMMARY

In this chapeter we have used PAMM, combined with extensive AIMD simulations under a variety of different modelling conditions to understand the complex hydrogen bond landscape of liquid water. We have shown that due to topological constraints in the hydrogen bond network, water molecules that deviate from idealized tetrahedral structures cluster with each other with different propensities. In particular, we find that under-coordinated environments display consistently strong angular correlations that can be traced to a weaker, but still directional, mode of the hydrogen bond. This alternative H-bonding mode, that is not recognized as such by definitions that are trained to identify the majority tetrahedral environment, is related to the correlations found in a (dilated) ice VIII lattice. Although the focus of this work has been on structural correlations, these features have important bearing on understanding dynamical processes in liquid water. In particular, the clustering of these defects suggests that the breakage and formation of hydrogen bonds are correlated over an extended part of the network, and provides a mechanism to lower the barriers associated with hydrogen bond dynamics. In this regard, the type of analysis we have presented here forms a framework to rationalize, in a microscopic way, the balance between entropic and enthalpic forces that drive fluctuations in the hydrogen bond network.

Here, we took extra care to ensure that the AIMD simulations we used for analysis of the HB network were sampled extensively using ab initio replica exchange molecular dynamics. Furthermore, we examined the sensitivity of the topological properties of the hydrogen network to temperature, the use of dispersion corrections, the inclusion of exact exchange and nuclear quantum effects. Within the framework of ab initio methods, the inclusion of dispersion corrections appears to have the most significant impact on the RDF compared to the inclusion of exact exchange or nuclear quantum effects. It is rather comforting to see, however, that regardless of the details of the choice of the water potential, the qualitative predictions for the defect populations and structures are very similar. While there has been a lot of effort in trying to come up with simulation recipes to reproduce the experimental RDF, we find that the differences one observes from using different simulation protocols are quite subtle, and that for instance it is not possible to fully mimic the effect of dispersion corrections by altering the temperature of the simulation. Examining the sensitivity of defect distributions to the use of different simulation protocols, shows that there are noticeable differences in the relative populations of the various defects when varying minor computational details. It should however be stressed that it is very hard to assess the statistical convergence of the minority popula-

tions, and that therefore we do not believe one can draw meaningful conclusions on the significance of these differences.

It is clear that the radial distribution function masks a lot of complexity in the underlying hydrogen bond network and that the details of its structure will be modulated by the balance in the relative proportions of different defects and how they cluster with each other – two aspects that can be disentangled by looking at three dimensional distribution functions resolved in different defect pairs. We believe that the analysis framework we introduce here, based on a self-consistent, data-driven definition of the hydrogen bond and the study of non-trivial correlations between topological defects, will prove to be particularly effective when investigating different portions of the phase diagram of bulk water, the role of charged defects as well as the behaviour of water in confinement and at interfaces.

7

STRUCTURAL PATTERNS IN NANOCLUSTERS

Contents

Particles containing a relatively small number of atoms or molecules (from few to thousands) are commonly referred to as "clusters". Examples are fullerenes, boranes, and carboranes, among many others.

Given the high proportion of surface, clusters manifest tunable and unique properties compared to their bulk counterparts. This fact explains their essential role in many technological fields, such as catalysis and nanotechnology [303]. A great example is that of Quantum dots, which have fascinating optical, magnetic and electrical properties [304, 305], and represent a classical textbook example of a promising class of nanomaterials.

Depending on their size and composition, clusters can be characterized by very complex potential energy surface (PES), which reflects a vast range of behaviors. Indeed, they provide a powerful theoretical test case to clarify the relationship between structural properties and PES and gain insights regarding more complex systems, that behave similarly.

In this chapter, we choose to focus only on the study of Lennard-Jones (LJ) clusters, since their structural and thermodynamical properties have been extensively studied, providing a perfect test case for PAMM.

A LJ pair potential governs the interaction between atoms:

$$U(\mathbf{r}_{ij}) = 4\epsilon \left[\left(\frac{\sigma}{\mathbf{r}_{ij}} \right)^{12} - \left(\frac{\sigma}{\mathbf{r}_{ij}} \right)^{6} \right], \tag{88}$$

where ϵ is the well depth and $2^{\frac{1}{6}}\sigma$ is the equilibrium separation for a diatomic molecule. This potential describes dipole fluctuation attractive interactions that decay as r^{-6}, and a somewhat arbitrary r^{-12} repulsive wall at short inter-atomic separations, that models the Pauli repulsion between electron clouds. The Lennard-Jones potential is a

This chapter is adapted from ref. [2]

good model for the interaction between noble gases atoms, as well as an inexpensive model of an isotropic pair-wise interaction between atoms.

We will use PAMM to learn automatically the structural patters of a LJ cluster of 38 atoms. The LJ_{38} cluster has been used very often as a benchmark of minimization algorithms, free-energy techniques, and structure-recognition methods, because its potential energy landscape contains a very deep, narrow enthalpic minimum corresponding to a truncated *fcc* lattice, and a broad basin containing a multitude of defective structures corresponding to icosahedral symmetry [93, 306–308]. Fig. 42 shows the structures corresponding to the lowest energy minima of LJ_{38} [309].

Here we used data from the $T = 0.18T^{\star}$ replica of a long parallel tempering trajectory [93], that contains structures that are representative of both the solid phases and liquid-like, highly defective configurations.

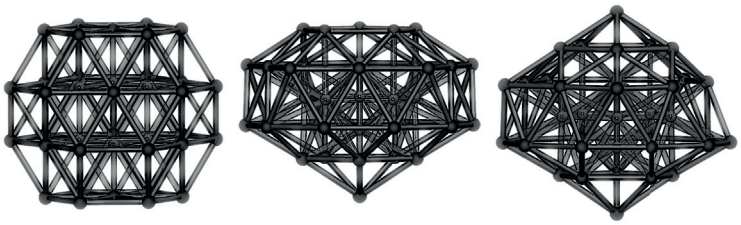


Figure 42: From left to right: the LJ_{38} global minimum, which is an fcc truncated octahedron, and the second and the third lowest energy minima, which are both incomplete Mackay icosahedra.

7.1 LOCAL MOTIFS IN LJ_{38}

Several of the possible environments (e.g. corner, edge, facet and core atoms) can be roughly identified by the number of nearest neighbors, that can be characterized based on purely radial information. Here we use the coordination number $c(i)$ for particle i, that we define as

$$c(i) = \sum_{j \neq i}^{N} \frac{1}{\exp\left(\frac{r_{ij} - r_c}{\gamma}\right) + 1} \tag{89}$$

where $\mathbf{r}_{ij}$ indicates the vector between particle i and j and $r_{ij} = |\mathbf{r}_{ij}|$ is the Euclidean distance between those two atoms. In order for the Fermi function to count the number of first neighbors, we set $r_c = 1.45\sigma$ and $\gamma = 0.2\sigma$. While neighbor counts can be very effective in identifying motifs in the minimum-energy structures, finite-temperature simulations can be considerably fuzzier. Furthermore,

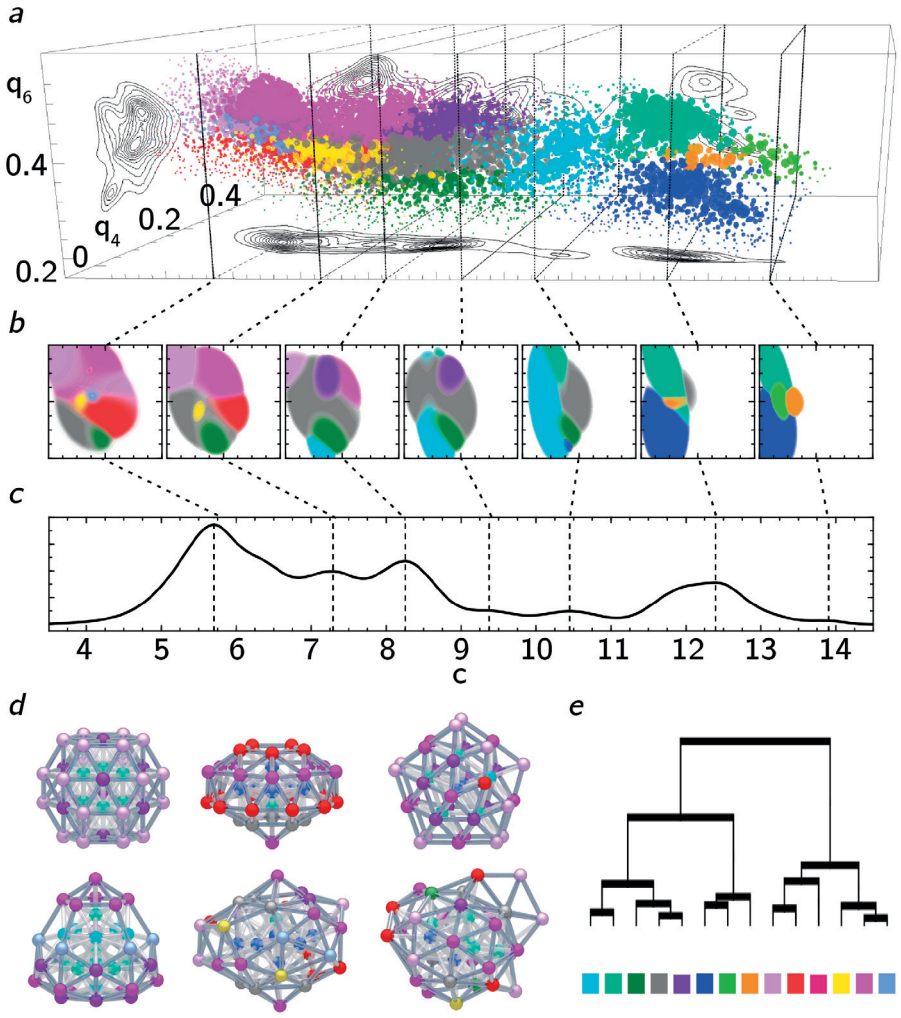


Figure 43: A PAMM analysis of local environments in a LJ_{38} cluster simulated at $T^{\star} = 0.18$. (a) The probability distribution in fingerprint space, that include a smooth coordination number c, and the Steinhardt order parameters q_4 and q_6. Grid points have an area proportional to the KDE of the probability density, and are colored according to the quick-shift clustering. Marginal probabilities are also drawn as contour plots. (b) Slices of the probabilistic motif indicators that result from the GMM built on the PAMM clusters. Each cluster is indicated with the corresponding solid color when its PMI is equal to 1, while the opacity linearly decreases to 0 when the value of the PMI is 0. (c) The probability distribution for c only. (d) Representative configurations of the LJ_{38} clusters, with atoms colored according to the dominant PMI. (e) Binary tree representation of the hierarchical clustering based on the adjacency matrix of the PAMM clusters.

the coordination number cannot distinguish between bulk environments from a *fcc* lattice and the core of an icosahedral motif. To address these problems, and to demonstrate the behavior of PAMM in a non-trivial example, we supplemented $c(i)$ with angular information. We used the local bond order parameters (or Steinhardt order parameters) [310] $q_4(i)$ and $q_6(i)$, which are known to be able to distinguish body center cubic (*bcc*), face centered cubic (*fcc*) or hexagonal close packed (*hcp*) crystal lattices [311–313].

We use the definition

$$q_l(i) = \sqrt{\frac{4\pi}{2l+1} \sum_{m=-l}^{l} |q_{lm}(i)|^2} \tag{90}$$

in which the complex vector $q_{lm}(i)$ is defined as

$$q_{lm}(i) = \frac{1}{c(i)} \sum_{j \neq i}^{N} \frac{Y_{lm}(\mathbf{r}_{ij})}{\exp\left(\frac{r_{ij}-r_c}{\gamma}\right)+1} \tag{91}$$

The functions $Y_{lm}(\mathbf{r}_{ij})$ are the spherical harmonics and the loop runs over all particles, since the Fermi function singles out contributions from just the first coordination shell. Since $c(i)$, $q_4(i)$ and $q_6(i)$ have very different ranges, when combining the different order parameters to give a 3D descriptor of the environments, we centered and scaled them so that each component has unit variance.

Figure 43 demonstrates the outcome of a PAMM analysis for the finite-temperature trajectory. A point based approach with a smoothing parameter $f_{\text{points}} = 0.02$ was used to compute the KDE, while for the clustering step the scaling factor α was set to 1. PAMM can recognize motifs based on both coordination number and angular correlations based on the three-dimensional joint probability density (Fig. 43a) – identifying several more clusters than it would be possible based on $c(i)$ alone (Fig. 43c). The PMIs for the different motifs (Fig. 43b) can be used to recognize the motifs in each snapshot of the simulation (Fig. 43d), and could be used, e.g., to bias a molecular dynamics simulation which triggers transitions between different structures. In this relatively simple case clusters are clear-cut, robust to sub-sampling and to variations of the PAMM parameters, and with an approximately Gaussian shape. Even though an agglomerative meta-clustering step is therefore not necessary, we did compute the cluster stability matrix, and generated the associated hierarchical clustering tree (Fig. 43e). Inspection of the binary tree is insightful, showing that individual clusters can be grouped in two main branches, corresponding to surface and bulk atoms. As we will also see in other cases, it appears that the hierarchical merging of the PAMM clusters can be interpreted much in the same way as a disconnectivity graph [314, 315], reporting on the relations between different basins in pattern space.

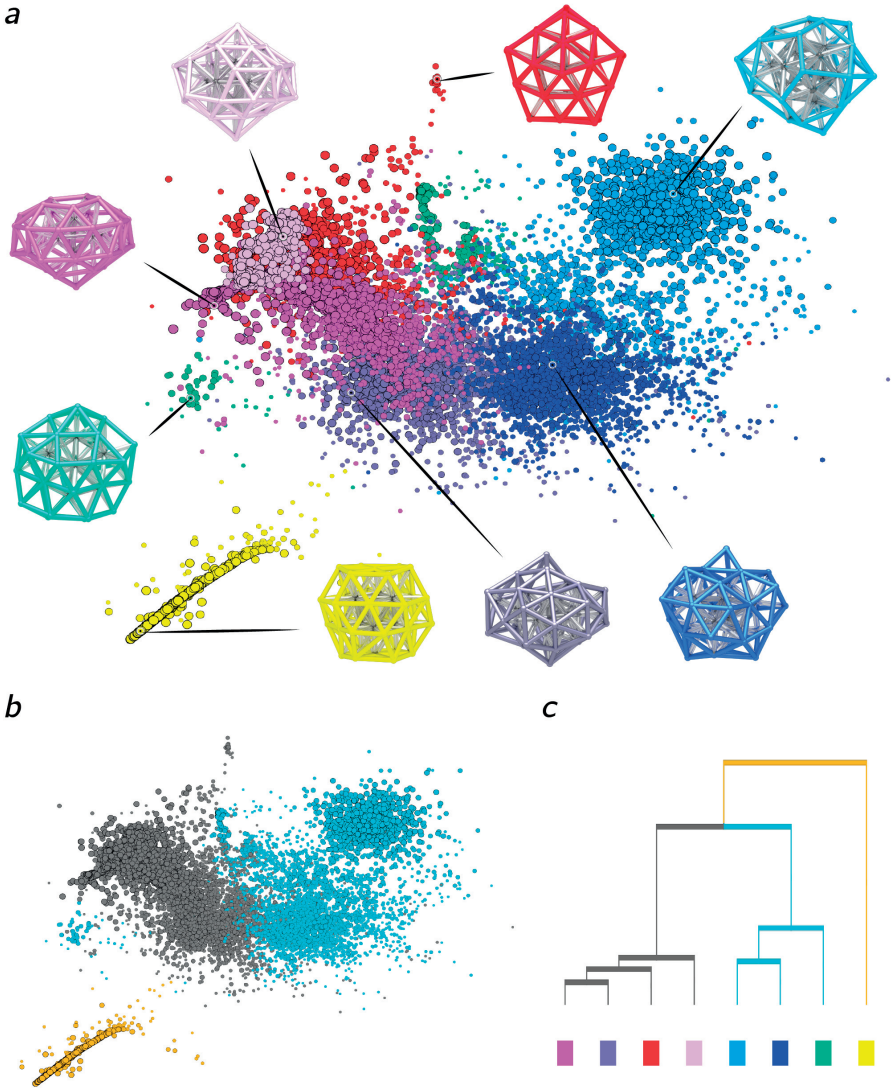


Figure 44: PAMM classification of configurations extracted from a simulation of LJ_{38} at $T^\star = 0.18$. (a) The clusters generated from the first stage of PAMM using the parameters $f_p = 0.05$ for the (point-based) KDE smoothing and $\alpha = 1$ for the scaling of the Quick-Shift cutoff. Configurations from the trajectory are represented using a two-dimensional sketch-map representation. The size of points reflects the probability density. Points are colored according to the cluster to which they belong. (b) The initial clusters are merged according to the hierarchical clustering procedure. (c) Sketch-map representation colored according to the macro-clusters, together with a snapshot representative of the highest-probability region for each motif.

7.2 GLOBAL CLASSIFICATION FOR LJ_{38}

LJ_{38} is one of the classical examples of simple clusters exhibiting a major structural transition, between a truncated-octahedral structure and a defective icosahedron. Since we consider a temperature close to the melting point, liquid-like configurations should be present in the trajectory. In order to identify different structures, we use the same 15-dimensional descriptors based on a smooth histogram of coordination numbers that were introduced in Ref. [316], and use them both as the basis for a PAMM analysis and as the input for the construction of a sketch-map representation, for which we used the same parameters and reference map as in Ref. [316]. We decided not to tune the fingerprints that had already been used in the literature so as to test the stability of PAMM-based clustering when dealing with sub-optimal inputs. For instance, some of the histogram bins have near-zero variance, which would yield singular bandwidth matrices if we did not stabilize the estimator with the OAS. Furthermore, the sharp cutoff function used in defining the coordination number histogram leads to the presence of a large number of basins, corresponding to minute differences in the structures. However, by using a point-based localization with $f_p = 0.05$, we implicitly consider a probabilistic model in which each cluster contains at least 5% of the data. As a result, the smearing parameter and the Quick-Shift threshold are large enough that many of these minute clusters coalesce, leaving just 8 motifs detected (Figure 44a).

The number and type of clusters, however, depend rather sensitively on the PAMM parameters. If one inspects the structural features that are associated with the initial PAMM clusters (see Figure 44a) it becomes clear that PAMM clustering does identify portions of configuration space that are clearly distinct, corresponding e.g. to different defective "gemstone" structures, to the truncated octahedron, and to liquid-like structures with different kinds of surface geometry. A dendrogram representation of the hierarchical clustering based on the adjacency matrix (Figure 44b) shows that some of the clusters are very sensitive to statistical noise, corresponding in high adjacency. One can identify three very stable meta-clusters, that are represented in (Figure 44c). PAMM recognizes that the most prominent features of the free-energy landscape comprise liquid-like states, more or less disordered icosahedral fragments and – even more clearly separated –the truncated octahedron. Once again it is remarkable how the adjacency matrix, despite being constructed just as a heuristic indicator of cluster stability, reflects the connectivity of the free-energy landscape. Although this far we were not able to reveal a formal connection, this aspect will be the subject of further investigation. One final consideration concerns the relationship between the

clustering and the dimensionality reduction approaches to describe complex atomistic systems. A sketch-map representation does offer a more intuitive, bird's eye view of the landscape. However, in order to achieve such a high level of coarse graining, the dimensionality reduction algorithm introduces considerable distortion, including also discontinuities in the projection [110, 316, 317]. For this reason, it is important to perform the clustering step in the high-dimensional space to avoid having a classification which is biased by sketch-map artifacts. For instance, Figure 44 shows clearly that one of the clusters identified by PAMM is split in two by the projection, without a continuous path connecting the fragments on the map. Thus, combining clustering and dimensionality-reduction approaches might provide a strategy to compensate for the shortcomings of the two methods, and obtain deeper understanding of the structural features of the system.

7.2.1 *Comparison of clustering schemes*

Many clustering approaches have been proposed during the past years, and could be used within PAMM to identify the modes of the pattern probability distribution.

To proof the efficiency of PAMM, fig. 45 shows a comparison of some of the most known clustering algorithms used instead of quick-shift for clustering the 15D-dimensional dataset describing the globlal structure of LJ_{38} discussed above. The data are standard-scaled before performing the clustering and the low-dimensional representation is generated using sketch-map.

From the results shown in Fig. 45 emerges the issues related to the use of a standard clustering schemes, which is the strong dependence of the partitioning on the choice of the initial parameters. Furthermore, PAMM provides an estimate of the quality of the clusters and the connectivity among them, which are non-trivial to extract from more standard techniques, such as K-Means or DBSCAN.

7.2.2 *Comparison of dimensionality reduction schemes*

Several different approaches have been developed to obtain low dimensional representations of complex databases. Figure 46 shows a comparison of the 2D representations of the Lennard-Jones cluster dataset as obtained with six different dimensionality-reduction techniques: Sketchmap [110], t-SNE [190], Diffusion maps [318], ISOMAP [184], Principal Component Analysis [319], Multidimensional scaling [182].

Even though all the algorithms can qualitatively understand the shape of the high-dimensional manifold, sketch-map appears capable of distinguishing more efficiently different basins.

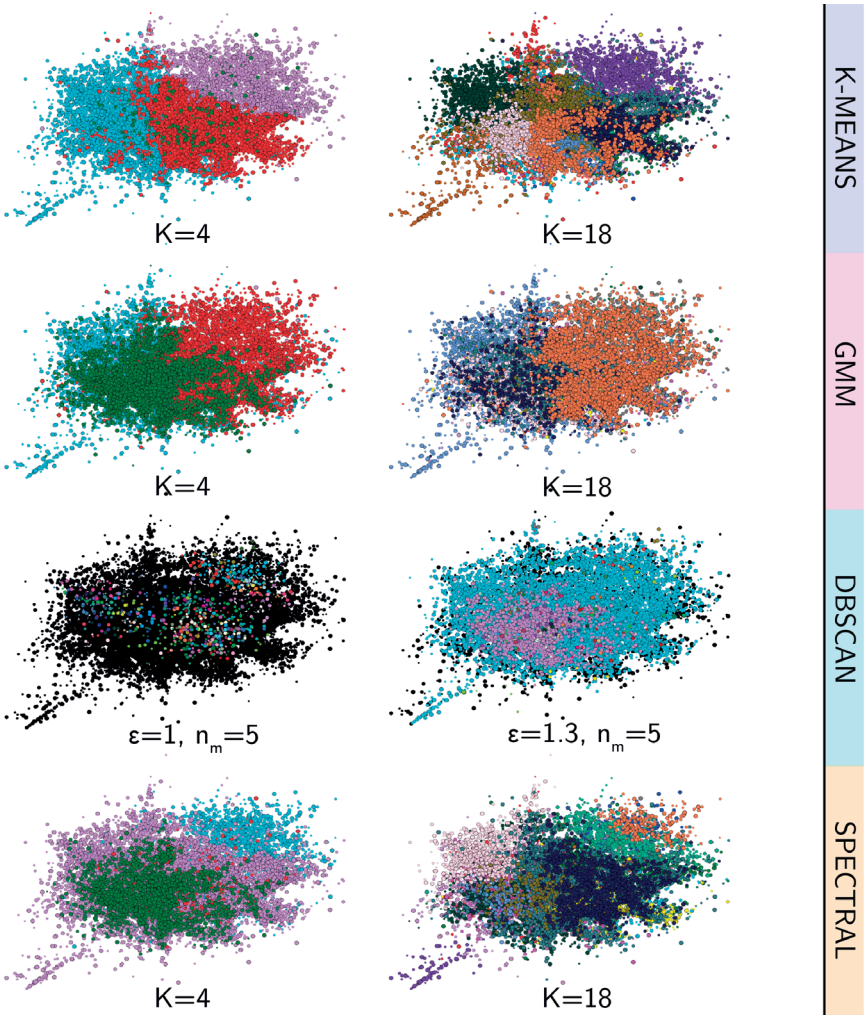


Figure 45: Global clustering for LJ_{38} from some of the most common algorithms. For each technique, various parameters have been tested.

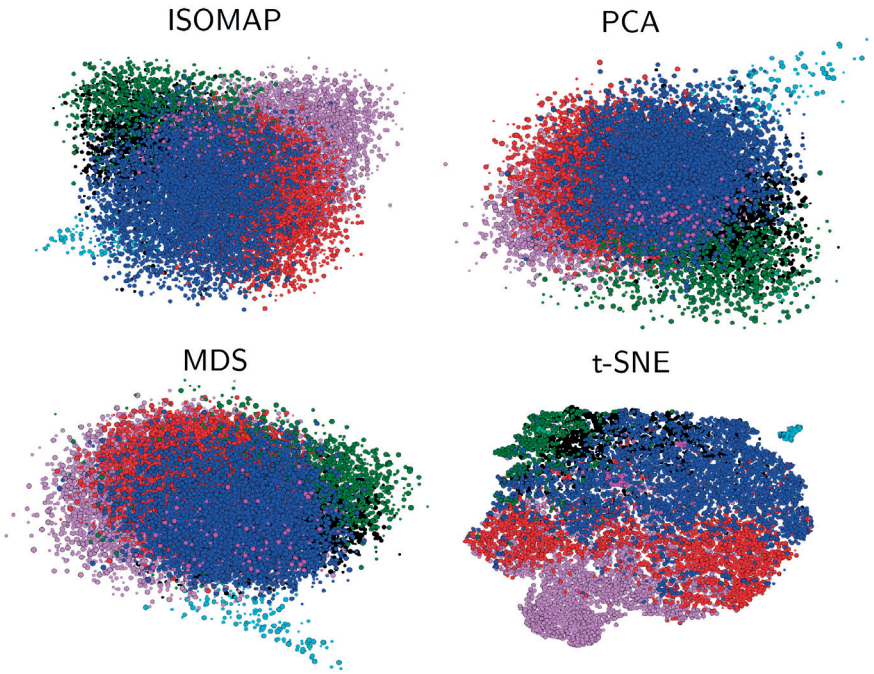


Figure 46: 2D representation of the LJ_{38} dataset, as obtained using different dimensionality reduction algorithms. Points are colored according to PAMM clustering.

8

STRUCTURAL PATTERNS IN PEPTIDES AND PROTEINS

Contents

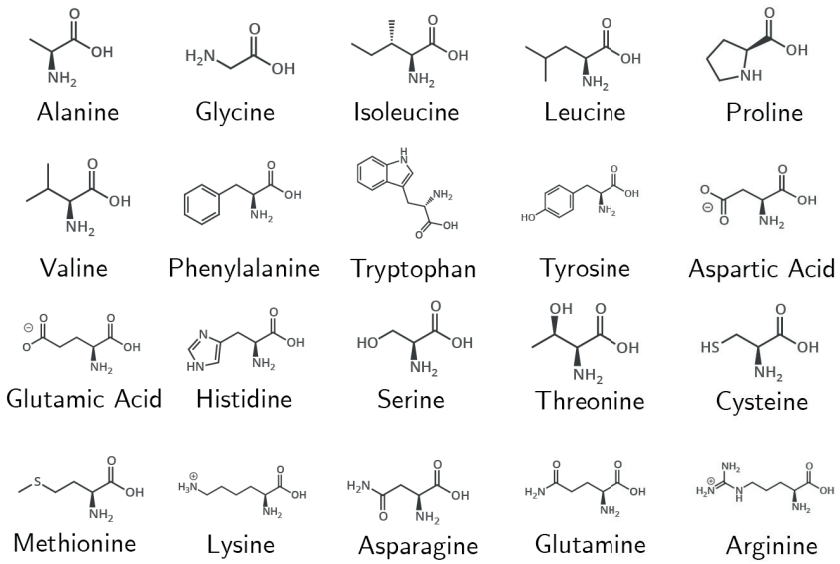


Figure 47: Structures of the 20 commonly occurring natural amino acids.

Polypeptides and proteins are macromolecules corresponding to a polymer chain composed of 20 elementary building blocks, namely the naturally occurring amino acids, schematically depicted in Fig. 47

In each amino acid, a sp^3 carbon atom (labeled α), is covalently connected to an amino group, a carboxylic acid group, a hydrogen atom, and a side-chain (usually labeled as R). What distinguishes different amino acids is the identity of R, whose nature can be hydrophobic, charged, or polar. Apart from glycine, for which the side-chain is just a hydrogen, the C_α is a chiral center that, depending on where the side-chain is attached, splits the amino acids into L and D opti-

This chapter is adapted from ref. [2]

cal stereoisomers. Naturally occurring proteins are composed only of amino acids of type L.

Proteins play a central role in living systems and, depending on their function, they can have very different structures. The investigation of the structure-function paradigm in proteins is one of the most active fields of research in structural biology and the complexity of biomolecular structures is usually understood by looking at their recurrent patterns, such as helices and sheets. Among the most com-

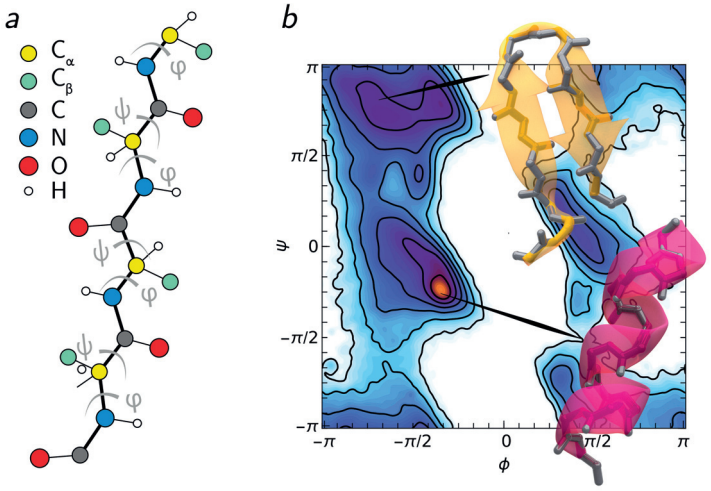


Figure 48: (a) Schematic representation of a generic polypeptide. For each residue, the ϕ and ψ dihedral angles are also shown. (b) Example of a Ramachandran plot.

monly used tools to visualize and understand structures in biology is the Ramachandran plot [320]. An example is shown in fig. 48(b). It is a 2D-plot in which one visualizes, for each residue in the peptide chain, the dihedral angle ϕ against ψ, which can be regarded as the most natural local structural parameter to describe backbone conformations, as schematically described in fig.48(a).

The Ramachandran plot computed for naturally occurring polypeptides provides a clear indication of the protein's propensity to form few well-known secondary structures (SSs), such as helices, sheets, and coils. Many algorithms exist to identify secondary structure patterns, that are based on the identification of hydrogen bonds [321, 322]. More recently, it has been shown that secondary structure conformations can be classified solely on the basis of backbone dihedrals of a protein [323, 324]. However, SSs are somewhat arbitrarily defined, since helices and sheets are often not in their ideal shapes in protein structures. For this reason, polypeptides offer a perfect test case to demonstrate the flexibility of PAMM.

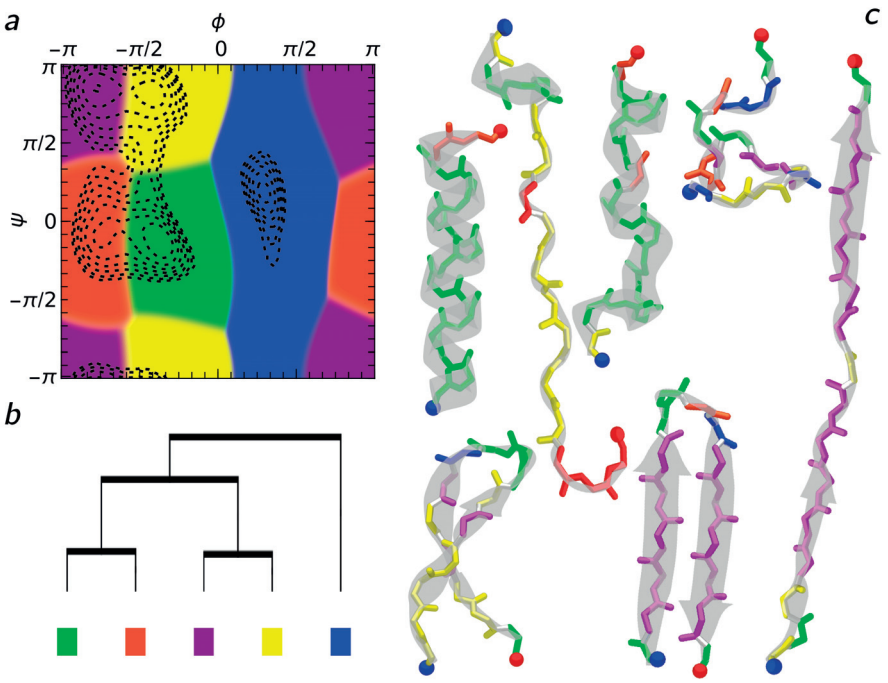


Figure 49: (a) PMIs for the backbone dihedrals of a β-hairpin are shown as a function of the Ramachandran angles. Different regions are colored based on the cluster that is associated with the dominant PMI. The underlying probability distribution is shown using black contours, that are equally spaced on a logarithmic scale. (b) Dendrogram representation of the hierarchical clustering of the adjacency matrix of the PAMM clusters. (c) A few representative structures of the molecules are shown with the residues colored according to the dominant PMI. The color code corresponds to that used in panel (a).

8.1 LOCAL MOTIFS OF A β-HAIRPINE PEPTIDE

Secondary structure in proteins is a textbook example of how molecular motifs can behave as building blocks of complex supramolecular structures, and thus a perfect benchmark for our pattern recognition algorithm. As a demonstration of our classification approach to realistic periodic data, we have applied our method to analyze the data of the backbone dihedral pairs (ϕ, ψ) from a replica exchange simulation of a 16-residue C-terminal fragment of the immunoglobulin binding domain B1 of the Streptococcus protein G in explicit solvent (GB1, amino acids sequence Ace-GEWTYDDATKTFTVTE-Nme) (for further details of the simulation see Ardevol et al. [325]). For each residue and each frame of the trajectory we computed the backbone dihedral angles ϕ and ψ, and performed a PAMM analysis. The underlying KDE was estimated by using a point-based KDE smoothing $f_p = 0.15$, and step-scaling $\alpha = 1.0$ for the subsequent Quick-Shift clustering. The resulting PMIs result in a partitioning of the Ramachandran $\phi - \psi$ plot [326] into 5 regions, that correspond roughly to β sheets, α helices, turns, etc. (see Fig. 49), and that are clearly associated with local probability maxima in the KDE (shown as black contours in Fig. 49a).

Even though the clusters are identified as von Mises modes, with a single basis function assigned to each of them, it is clear that the PMI correspond very accurately to the partitioning of the probability density in basins of attraction, with the transition zones between two PMIs following closely the dividing surface between basins. In Fig. 49b we also show a few reference structures selected from the trajectory. The aminoacids in the backbone have been colored based on the dominant PMI to which they are associated.

It is easy to recognize the PMIs associated with well known secondary structure elements by comparing reference structures to the PAMM partitioning of the Ramachandran plot. One can clearly identify, e.g. α-helices or the antiparallel β-sheets that are abundant in the simulation data for the GB1 fragment. Chains of dihedrals that correspond to a β-sheet conformation are also seen in extended structures. We also find several instance of turn-type T1 motifs [324], that in our case concentrate in the unstructured portions of the polypeptide.

Meanwhile, it is clear that there is not a 1:1 correspondence between PMIs and "traditional" secondary structure motifs. Some of the PMIs correspond to portions of the Ramachandran plot that are traditionally assigned to polyproline II (PPII) helices and left-handed helices, yet we could not identify a significant presence of stable structures associated with these motifs. Rather, extended structures and distorted β strands are associated with the PPII region, while left-handed helical patterns appear, together with many other PMIs,

within disordered, "random-coil" configurations. This observation highlights the potential of a data-driven approach for the definition of molecular patterns. Well-separated clusters in fingerprint space are recognized even though they do not appear clearly when the trajectory is inspected and well-established structural motifs are searched for. This agnostic behavior could be very useful, for instance, in rationalizing the behavior of intrinsically-disordered proteins [327, 328], or to study polypeptides in unusual environments, such as at inorganic interfaces or in combination with synthetic polymers. Furthermore, an automated probability analysis allows one to extend the definition of pattern space by combining e.g. several backbone dihedrals, or by combining dihedrals and H-bonding indicators. This procedure could give rise to more precise identification on secondary-structure patterns, and will be the subject of future research. Finally, the hierarchical clustering of the five PAMM motifs (Fig. 49b) shows another example of how the adjacency matrix built by a bootstrapping analysis reflects the structure in the free-energy landscape of patterns, with the linkage distance corresponding roughly to the free-energy barrier between basins.

8.2 STRUCTURAL CLASSIFICATION FOR A β-HAIRPINE PEPTIDE

The structural landscape for the GB1 oligopeptide provides another suitable benchmark for the application of PAMM to the clustering for the high-dimensional data. This β-hairpin fragment has been studied extensively by metadynamics [329], and has been used to demonstrate the comparison between wild type and mutant proteins using sketch-map [325]. Such analysis revealed a rugged free-energy landscape, containing many metastable states including a helical configuration, several mis-folded hairpin configurations as well as the native fold.

As in the case of the the LJ_{38} cluster, we used a simple-minded choice of high-dimensional descriptors – namely, the 30 backbone dihedrals – as the input representation. This choice minimizes the bias on the clustering procedure, but implies that configurations that differ by minor details (e.g. the configuration of the terminal aminoacids) are considered as distinct structures. Indeed, a global classification of the GB1 peptide produces a multitude of clusters (Figure 50(a)), and varies considerably depending on the clustering details and bootstrapping resampling. Hierarchical cluster merging, however, simplifies considerably this picture (Figure 50(b)), making it possible to group states that are associated with a helical configuration, the native fold, and a few misfolded hairpin configurations (Figure 50(c)). Clustering also highlights the limitations of the sketch-map projection, that gives a contiguous representation of the high-probability

clusters (helix and native hairpin) but scatters higher-free-energy states at the periphery of the map.

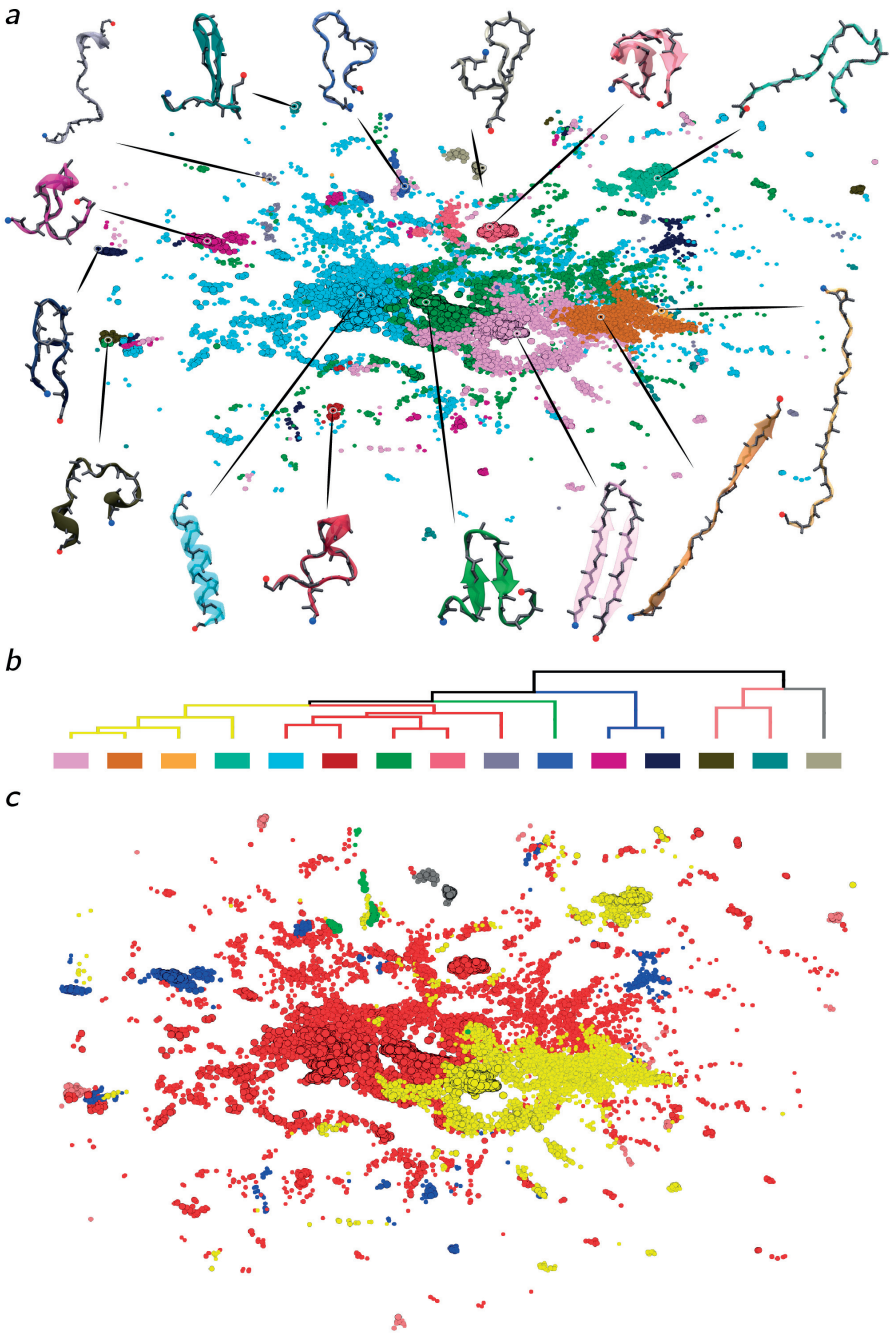


Figure 50: PAMM classification of configurations extracted from a simulation of the GB1 hairpin fragment. (a) Clusters generated in the first stage of PAMM using the parameters 0.1 for the (point-based) KDE smoothing and $\alpha = 1$ for the scaling of the Quick-Shift cutoff. Configurations from the trajectory are represented using a two-dimensional sketch-map representation. The size of points reflects the probability density. Points are colored according to the cluster to which they belong. A snapshot representative of the highest-probability region for each cluster is also shown. (b) The initial clusters are merged according to the hierarchical clustering procedure. (c) Sketch-map representation colored according to the macro-clusters.

9

FUTURE OUTLOOKS

Contents

We have shown how it is possible to build an automatic, unbiased framework to learn patterns in the complex feature spaces generated from atomistic simulations. However, the feature space is determined by the chosen order parameters, which play a critical role in the following analysis. It would be very interesting to introduce new general descriptors that are capable of efficiently describing local environments automatically. A possible idea is to use the symmetry functions introduced by Behler and Parrinello as discussed in sec. 9.1.

Another promising extension of this thesis' work is the use of PAMM to classify and clarify the structural properties of proteins and peptides. PAMM can provide new flexible definitions of complex patterns, which can be used to identify recurrent motifs in intrinsically disordered proteins and in proteins in contact with non-physiological environments, such as inorganic interfaces. It can also be used to improve the standard classifiers used in scoring functions for docking. This idea will be discussed in sec. 9.2.

Finally, a very interesting extension would be the use of PAMM to design elaborate order parameters to enhance sampling when combined with biasing schemes. A simple example, will be discussed in sec. 9.3.

9.1 AUTOMATIC DEFINITION OF THE FEATURE SPACE

The distinction of local atomic environments is not only crucial for rationalizing complex structure-property relationships in simulations, but also for building effective CVs capable of adequately sampling interesting regions of phase space using enhanced sampling schemes, such as metadynamics.

We introduced PAMM as an agnostic, unsupervised method, assuming however that the choice of the descriptors generating the feature manifold was the optimal one. Unfortunately, in many realistic scenarios, the choice of suitable descriptors is far from trivial.

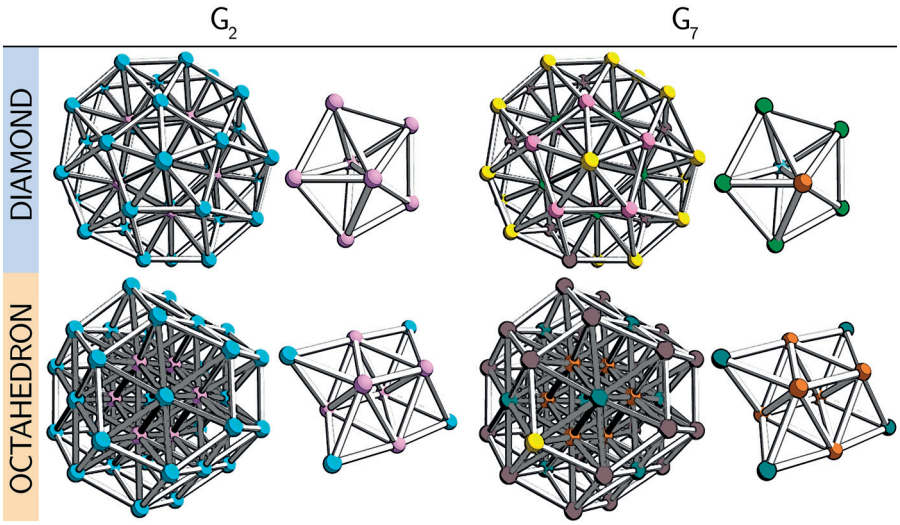


Figure 51: Classification of the local environments for two different structures of the LJ_{38}. At the right side of each structure, a zoom of the core atoms is shown. Two different SFs where used, G_2 with parameters $r_c = 12.3, r_s = 2$ and $\eta = 0.05$, and G_7 with $r_c = 20.3, \eta = 0.01, \alpha = 0.3$. The SFs were applied to the simulation trajectory to transform the Cartesian coordinates into a 1D dataset, on which a PAMM analysis was run to find distinct stable patterns. While with this choice of parameters G_2 can distinguish only the surface from the core, G_7 can classify a core atom in a diamond-like structure as different from those of a truncated octahedron, as well as distinguish surface atoms with different symmetries and neighborhood.

We here propose the use of the same symmetry functions (BPSFs) used in neural networks (NN) for energies and forces calculations. Such symmetry functions, introduced by Behler and Parrinello [196], are used to decompose the system into smaller units, corresponding to the environments surrounding each atom.

BPSFs can be of two types: radial (two-body) and angular (three-body). The former provides information regarding the radial distribution of surrounding atoms, while the latter also includes information regarding angular distributions between triplets of atoms. Both obey the following requirements:

- rotational and translational invariance
- permutational invariance for the same element types

Many SFs have been proposed [330], and few examples are:

- $G_2^i = \sum_j e^{-\eta(\mathbf{r}_{ij} - r_S)^2} f_c(\mathbf{r}_{ij})$

- $G_3^i = 2^{1-\xi} \sum_{j,k\neq j} (1+\gamma cos\theta_{ijk})^{\xi} e^{-\eta(\mathbf{r}_{ij}+\mathbf{r}_{ik}+\mathbf{r}_{jk})^2} f_c(\mathbf{r}_{ij}) f_c(\mathbf{r}_{ik}) f_c(\mathbf{r}_{jk})$
- $G_7^i = \frac{1}{2} \sum_{j,k\neq j} sin\left[\eta(\theta_{ijk}-\alpha)\right] f_c(\mathbf{r}_{ij}) f_c(\mathbf{r}_{ik})$,

where j and k are the indexes of the atoms surrounding i, θ_{ijk} is the angle between the three atoms, and $r_S, \gamma, \xi, \eta, \alpha$ are the parameters that have to be adjusted to probe different angles and distances. The function $f_c(\mathbf{r})$ is used to ensure that the symmetry function goes smoothly to zero at a fixed cutoff value r_c.

Each BPSF corresponds to a different fingerprint, and carries different pieces of spatial information. To test the ability of symmetry functions to distinguish local environments, we applied them to the LJ_{38} dataset introduced in chapter 7. Fig. 51 shows how choosing different BPSFs and different parameters it is possible to control how sensitive the SF is to the different environments.

One could think about combining many BPSFs into a feature vector, thus obtaining a manifold in which all the significant patterns stand out as clear, distinct modes in the underlying PDF.

The critical issue in this case would be to find a general method able to decide, for very different scenarios, the optimal number of BPSFs needed for an efficient description of the system, and to guess reasonable parameters for each BPSF.

9.2 STRUCTURAL PATTERNS IN PROTEINS*

Secondary structure in proteins is usually understood in terms of hydrogen bonds patterns in the peptide backbone or as regular patterns in the values of the Ramachandran angles ϕ and ψ. In both the approaches, the classification is clear-cut and based on the arbitrary assumption of some geometrical or energetic parameters.

In chapter 6 we have demonstrated how it is possible to use PAMM to introduce a flexible and adaptive definition of the hydrogen-bond. An exciting extension of this study, would be to use PAMM to learn the different HB patterns formed along the backbone chain. This analysis could help when it comes to understanding whether hydrogen bonds with different donor and acceptor species, or in different parts of the protein (main or side chain) have different preferred geometries.

Fig. 52, shows the PMIs for different types of HB occurring in a protein backbone. We performed the analysis on the X-Ray experimental structures deposited in the Protein Data Bank (PDB) and resolved to 1.2Å or better. This high resolution was necessary to ensure an accurate account of all the atomic positions in the protein, including the hydrogens. However, also in high-resolution structures, it could also

*This work is done in collaboration with B. A. Helfrecht.

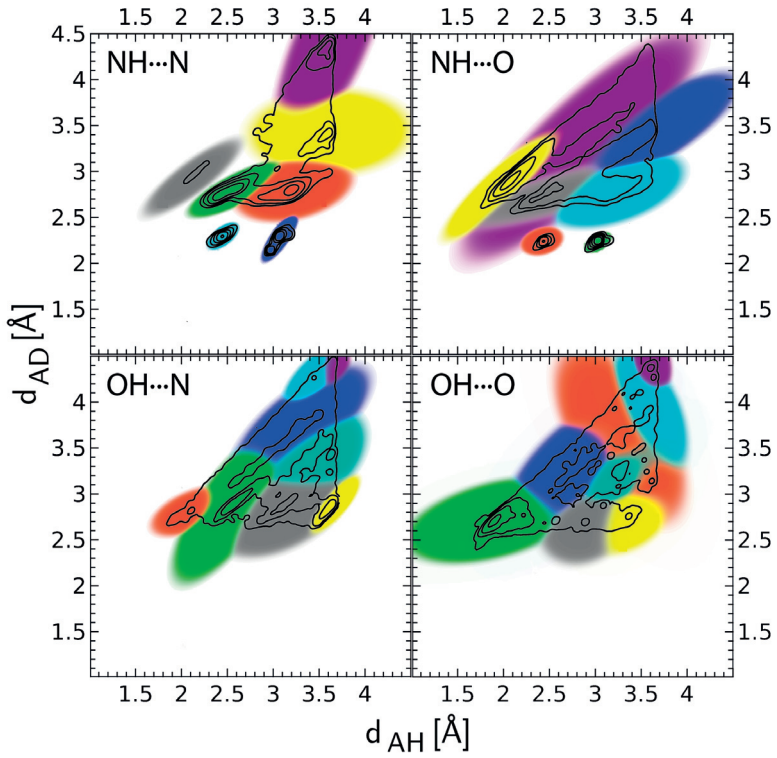


Figure 52: PAMM PMIs for the different types of hydrogen bonds in a protein backbone. The training was done from the high-resolution (better than 1.2Å) X-ray structures deposited in the PDB. The HBs correspond to the PMIs covering the region with $d_{AH} \approx 1.9$Å and $d_{AD} \approx 2.8$Å. One can notice how the ranges of parameters in which PAMM identifies a structure as a HB vary substantially in the different scenarios.

be that hydrogen positions are set to a predefined distance from the donor atom. To avoid any bias in the clustering analysis, we decided to use only the donor-acceptor and acceptor-hydrogen distances to represent D–H$\cdots$A triplet of atoms.

From the preliminary results shown in fig. 52, it is evident that the backbone HBs involving different species have very different characteristics, which implies that the substitution of a generic, clear-cut, geometrical definition with a flexible PAMM-based description could be very reasonable.

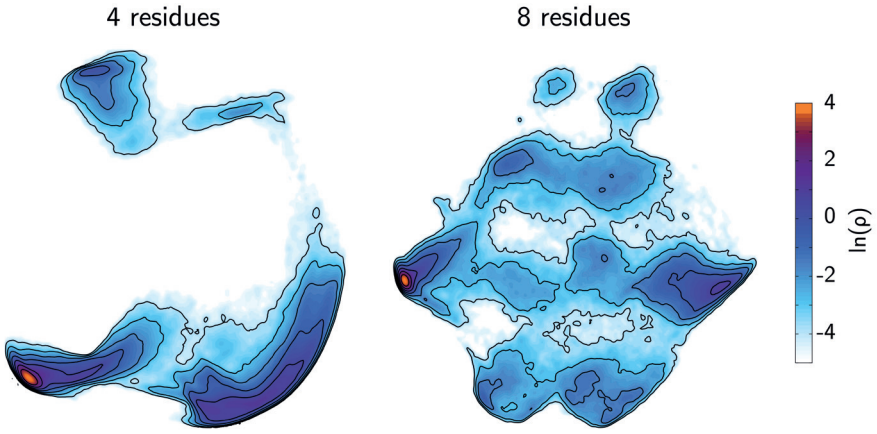


Figure 53: 2D projections of the distribution of points in two different feature spaces built from the backbone dihedrals of high-quality X-Ray structures deposited in the PDB: (left) each point corresponds to four Ramachandran angles coming from two consecutive amino acids, (right) each point corresponds to eight Ramachandran angles coming from four consecutive amino acids. Since dihedral angles are periodic variables, the PCA was done following the approach proposed in ref. [331], where each angle ϕ_n is represented by its equivalent vector $(\cos(\phi_n), \sin(\phi_n))$ on the unit circle. The contour plots are shown in logarithmic scale, with the isolines separated by a unit in the logarithm of the density.

Another natural application of PAMM, would be to use it for the automatic definition of secondary structures. We proved, in chapter 8, that an automatic classification of SSs based on the partitioning of the Ramachandran plot is possible. However, the description of secondary structures based only on the 2D Ramachandran angles is very simplistic. One can think about a more accurate description using PAMM that learns recurrent patterns in higher dimensional feature spaces, where the information relative to a local motif comes from more than one amino acid. Increasing the dimensionality of the

learning space should enable the classification of patterns at a higher resolution.

To support this idea, fig. 53 shows the logarithm of the histogram of the distribution of points in two different high-dimensional feature spaces, where one feature point corresponds to the Ramachandran angles computed from two and four consecutive amino acids respectively. The analysis is done on the high-resolution X-Ray structures deposited in the PDB (resolved to 1.8Å). We used PCA to project the points in two dimensions, following the approach proposed in ref. [331] to deal with periodicity.

One can see how, increasing the dimensionality of the feature space, increases the number of modes that appear in the PDF.

9.3 ENHANCED SAMPLING

A very promising application of PAMM is the construction of complex order parameters to be used, in combination with biased-sampling schemes, to modify the statistic of specific patterns.

Here we report some preliminary results on the use of PAMM to modify the statistics of a particular defective state in the HB network of liquid water, with the final aim of accelerating the dynamics of water molecules.

Speeding up the motion of liquid water could have many advantages, from a better understanding of the dynamics of neat water, thanks to the use of more expensive and accurate models, to the possibility of more efficiently sampling the interactions among water molecules and solutes, such as a protein.

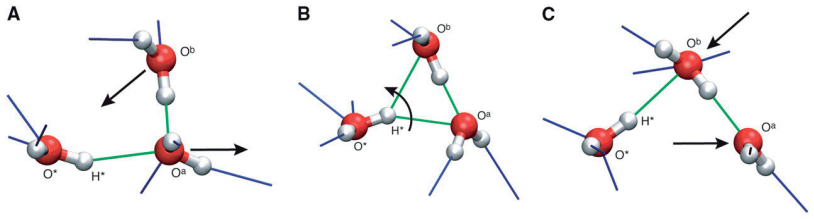


Figure 54: Schematic representation of the H-bond exchange mechanism in liquid water. Three consecutive snapshots of an actual H-bond switching event are shown, where the green lines indicate the water molecules involved in the HB exchange, while the black arrows shows the key movements of the molecules between the first and the second coordination shells. This figure is adapted from [332].

Laage *et al.* [332] proposed a model (*Extended Jump Model*, EJM) to describe the reorientation mechanism of water molecules, in which a molecule forms a new HB through a large-amplitude angular jump,

rather than by a sequence of small diffusive steps, which is the commonly accepted picture (*Debye small-step* diffusion model).

A clear picture of the mechanism is shown in Figure 54, where a rotating water molecule breaks a HB with an overcoordinated neighbor in the first coordination shell, and form a HB with an undercoordinated water coming from the second coordination shell. The HB cleavage and the molecular reorientation occur concertedly and not successively, as usually considered. In this picture the transition state in the water reorientation mechanism is some state that involve a bifurcated HB. PAMM can easily recognize this particular pattern

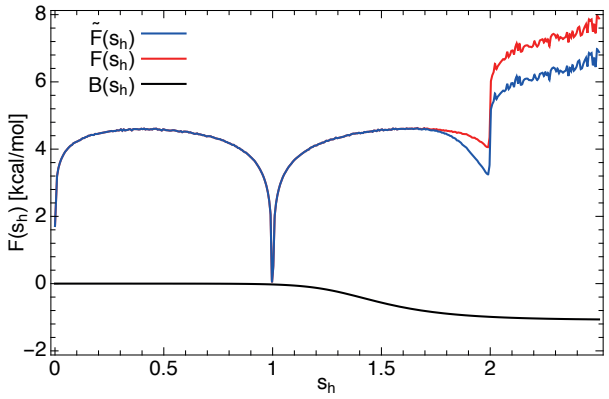


Figure 55: Effect on the free-energy, in a TIP4P-2005f simulation at 300K, due to the application of the bias potential introduced in eq. 95. The parameter used in this example are $h = 1$, $\mu = 1.9$, $\sigma = 0.05$ and $\alpha = 0.85$.

using the s_H counter introduced in sec. 4.3. If a water molecule is involved in standard standard HB patterns, each of of its hydrogens have $s_H \approx 1$. In the case a bifurcated HB is formed $s_H \approx 2$.

A possible way to speedup the dynamics of water molecules, would be to increase the number of the extended jump events. To this aim, a PAMM-based collective variable can be introduced to encourage the sampling of configurations involving a bifurcated HB.

The general idea is to use umbrella sampling to add a bias term $B(s(\mathbf{q}))$ to the potential $V(\mathbf{q})$ to sample configurations according to a modified probability distribution $\tilde{P}(s_H)$ in which bifurcated HBs have a higher weight compared to the original distribution $P(s_H)$

$$\tilde{V}(\mathbf{q}) = V(\mathbf{q}) + B(s(\mathbf{q})) \ . \tag{92}$$

The change in the free energy due to the biasing procedure is

$$\tilde{F}(s) = -\frac{1}{\beta} \ln \int d\mathbf{q} e^{-\beta[V(\mathbf{q})+B(s(\mathbf{q}))]} \delta(s - s(\mathbf{q})) = F(s) + B(s) \ , \tag{93}$$

where we exploited the fact that $e^{-\beta B(s)}$ can be brought outside the integral, as δ selects only those configurations with $s(\mathbf{q}) = s$. The

unbiased statistics of any configuration-dependent property $A(\mathbf{q})$ relative to F can be obtained by reweighting

$$\langle A \rangle_{\text{unbiased}} = \langle A e^{\beta B(s((q)))} \rangle_{\text{biased}} \tag{94}$$

Since the bias has to favour just the free energy minimum corresponding to bifurcated HB configurations, a convenient choice is to define $B(s_H)$ as a Fermi-like function:

$$B(s_H) = h \left(\left(\frac{1}{1 + e^{\frac{s_H - \mu}{\sigma}}} \right)^{\alpha} - 1 \right) \quad . \tag{95}$$

where h is the height of the step and α is a smoothing factor, while μ ans σ set the position and the width of the step respectively.

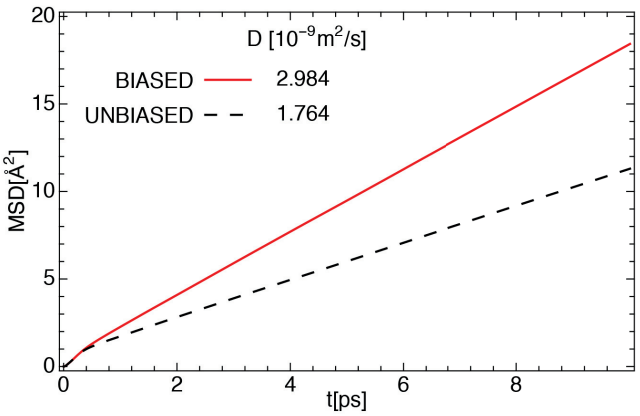


Figure 56: MSD curve of an unbiased *NVT* TIP4P-2005f at 300K (red) compared with a biased one (dashed black), using the bias potential shown in Fig. 55. The corresponding diffusion coefficients D are also reported.

To test this idea, we run two *NVT* simulations, one of neat TIP4P-2005f water at 300K and another one adding a bias potential as in Figure 55. To check how the introduction of the bias influences the dynamics of water molecules, we computed the mean square displacement (MSD), from which is possible to extract the diffusion coefficient D using the relation:

$$\langle (r(t) - r(0))^2 \rangle = 6Dt \tag{96}$$

Figure 55 shows that the introduction of the bias effectively enhances the diffusion of water molecules. Unfortunately, the RDF computed from the biased simulation (fig. 57(a)), clearly shows that, under the action of the bias, the structure of the liquid results is modified substantially. Figure 57(b) shows that obtaining the original RDF by reweighting is not possible.

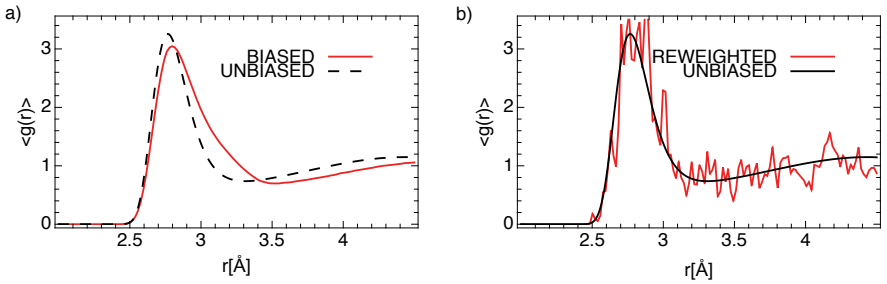


Figure 57: a) Comparison between the RDFs of an unbiased *NVT* TIP4P-2005f simulation at 300K (red) and a biased simulation (black) that was done by adding the umbrella potential described in the main text and shown in fig. 55. b) Comparison between the RDF computed from the unbiased simulation and that obtained by reweighting from the biased simulation.

From these results, it appears that our PAMM-based collective variable is not selective in capturing only the transition state of water reorientation events.

To understand these results, we processed the unbiased trajectory to collect all the EJM events as described in Ref. [332] and computed the s_H, s_D and s_A values for the molecules involved in the HB exchange. By construction, in all the EJM transition states, the hydrogen involved in the molecular jump has $s_H \approx 2$, but this appears to be just a small fraction among all the possible events in which $s_H \approx 2$.

Fig. 58 shows the hydrogen bond counting maps for the water molecules involved in an EJM event.

Many complex scenarios are compatible with the recipe proposed by Laage, and further research should be done to find a more sophisticated descriptor that is able to bias the EJM events in a more efficient fashion. A suggestion of the possible defective states of the HB network that could be involved in an HB exchange event is shown in fig. 58, where the HB counting maps for the molecules involved in EJM event are reported.

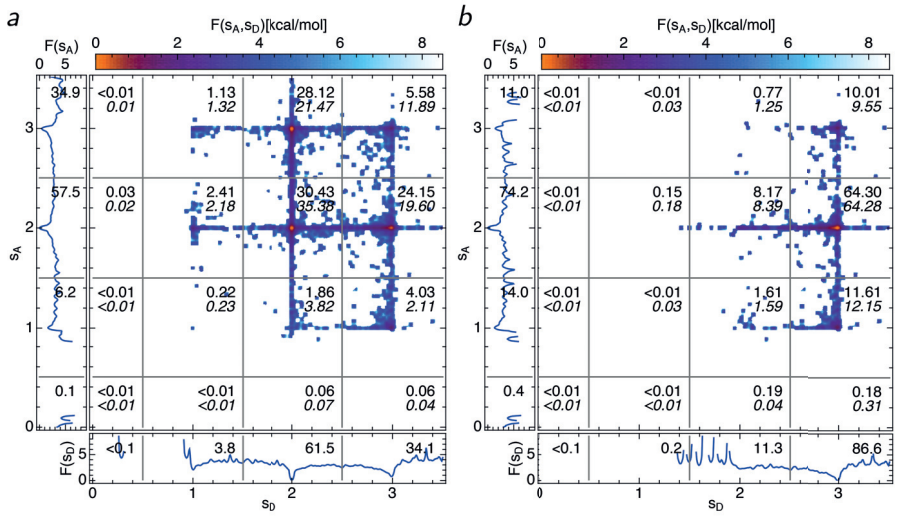


Figure 58: Hydrogen-bond counts statistics for a classical simulation of TIP4P water at room temperature selecting only water molecules involved in an EJM event: while in (a) all the molecules involved are considered, in (b) only the central water molecule is taken into account. The probability distributions are represented in terms of the associated free energies $F = -k_B T \ln P$, that are expressed in kcal/mol throughout. We also report integrated probabilities (in percent) to have a configuration in the vicinity of the different integer numbers of HBs. Below the values of the joint probabilities of s_A and s_D the product of the marginal probabilities are indicated, in italics.

10

CONCLUSION

Machine-learning algorithms constitute a promising approach to guide the rationalization and analysis of complex, large-scale atomistic simulations. In this thesis, we have presented a general machine-learning framework, the Probabilistic Analysis of Molecular Motifs (PAMM), to analyze the outcome of atomistic simulations in order to search for the essential atomic-scale patterns that characterize the behavior of materials and molecules. PAMM can be exploited in atomistic simulations to reproduce in an automatic, data-driven manner the process of recognizing recurring structural patterns that is behind our intuitive understanding of chemical bonding and structuring in materials and biomolecules. A PAMM analysis starts by learning a probabilistic model underlying structural data through non-parametric density estimation methods, followed by the partitioning of the PDF through a density-based clustering approach. We eventually model the PDF as a mixture of Gaussians, by fitting each local mode with a multivariate Gaussian. A natural application of such a model is the construction of Bayesian classifiers to classify patterns in atomistic simulations, which enable the fuzzy definition of chemical entities in terms of smoothly varying posterior probabilities.

PAMM is a general method that is capable of dealing with periodic, high-dimensional and/or sparsely sampled data. It combines several state-of-the-art techniques to guarantee robust, reliable and efficient clustering. We also introduced a new original way to estimate the quality of the clusters and to infer the connectivity among them. Indeed, this method performs repeated clustering attempts on top of re-sampled distributions constructed by bootstrapping, and uses these to construct an adjacency matrix that characterizes the stability and the overlaps between clusters. Combined with a hierarchical "meta-clustering" and a binary tree representation this further analysis serves two purposes. First, it provides a way to improve the quality of PAMM clusters – by merging non clear-cut motifs and allowing the representation of strongly non-Gaussian modes in the probability. Second, it makes it possible to recognize the relations between the fine-grained details of the free-energy landscape and the main basins on a coarser scale. In all the examples we considered, the hierarchical clustering reflects qualitatively the structure of the free energy and resembles a disconnectivity graph. It might be possible to construct the adjacency matrix such that this analogy is made quantitative, which will be the subject of future reaserch.

We demonstrated that PAMM is able to identify fuzzy entities such as hydrogen bond in a variety of different contexts. Based on a PAMM analysis, we then introduced a compact representation of the hydrogen-bonding properties of liquid water as a function of the total number of accepted and donated HBs per water molecule, which arises naturally from the use of a smoothly varying PAMM counting function. We demonstrated that these hydrogen bond counts can be used to clarify the structure and dynamics of the hydrogen bond network, which we studied extensively to probe the non-trivial correlations between topological defects and the influence of different simulation protocols.

To prove the generality of PAMM, we applied it to classify coordination environments in a LJ_{38} cluster and secondary-structure motifs in a 16-residues β-hairpin peptide. Although our method is geared towards recognizing *local* molecular patterns, we also showed it can cluster overall structures for both LJ_{38} and the oligopeptide, which allows for the further development of the method as an approach to identify *global* conformational states. This could be applied simply for classification purposes, as well as for Markov state models to cluster simulation data into microstates and discretize the state space of Markov chains.

The combination of PAMM clustering, binary-tree merging of the motifs and the sketch-map representation of high-dimensional free energy landscapes provides multiple insights into the behavior of complex atomistic systems, and overcomes some of the limitations of the methods, such as the presence of discontinuities in sketch-map projections. The probabilistic motif identifiers associated with each cluster can be used as fuzzy, smoothly varying collective variables in biased molecular dynamics schemes to accelerate sampling and reconstruct the underlying free-energy landscape. The Gaussian Mixture Model associated with the PMI could also be used as an ansatz for the target probability distribution in a variational-sampling scheme [333], with the populations of the clusters as the variational parameters.

To summarize, the research performed in this thesis tested the capability of machine-learning schemes within the domain of structural identification in atomic-scale structures. The development of general analysis tools that are capable of identifying and classifying metastable patterns of matter at the atomic scale is a remarkable challenge, which must cope with the growing complexity and high variability of the systems that can be simulated through atomistic simulations.

Data-driven approaches to recognize the building blocks of complex materials constitute a necessary ingredient to assist the analysis and interpretation of large-scale atomistic trajectories, and provide a

natural representation of free-energy landscapes in terms of the most stable configurations of the system. Such a flexible description of the structural space can be used with a twofold aim: it can be exploited to accelerate configurational sampling in complex computer experiments, as well as to classify structural patterns along simulation trajectories, or in structural databases, in an unbiased and adaptive way.

BIBLIOGRAPHY

[1] Piero Gasparotto and Michele Ceriotti. "Recognizing molecular patterns by machine learning: An agnostic structural definition of the hydrogen bond." In: *Journal of Chemical Physics* 141 (2014), p. 174110 (cit. on pp. 4, 39, 51, 57, 94, 96).

[2] Piero Gasparotto, Robert Meissner, and Michele Ceriotti. "Recognizing Local and Global Structural Motifs At the Atomic Scale." In: *Journal of Chemical Theory and Computation* 14.2 (2018), pp. 486–498 (cit. on pp. 4, 39, 113, 123).

[3] Piero Gasparotto, Ali A. Hassanali, and Michele Ceriotti. "Probing Defects and Correlations in the Hydrogen-Bond Network of ab Initio Water." In: *Journal of Chemical Theory and Computation* 12 (2016), pp. 1953–1964 (cit. on pp. 4, 51, 83, 95, 98, 100, 101, 107).

[4] Ryogo Kubo, Morikazu Toda, and Natsuki Hashitsume. *Statistical physics II: nonequilibrium statistical mechanics*. Vol. 31. Springer Science & Business Media, 2012 (cit. on p. 5).

[5] William C Swope et al. "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters." In: *The Journal of Chemical Physics* 76.1 (1982), pp. 637–649 (cit. on p. 6).

[6] Hans C Andersen. "Molecular dynamics simulations at constant pressure and/or temperature." In: *Journal of Chemical Physics* 72 (1980), pp. 2384–2393 (cit. on p. 6).

[7] Shuichi Nosé. "A unified formulation of the constant temperature molecular dynamics methods." In: *Journal of Chemical Physics* 81 (1984), p. 511 (cit. on p. 6).

[8] Michele Ceriotti, Giovanni Bussi, and Michele Parrinello. "Langevin Equation with Colored Noise for Constant-Temperature Molecular Dynamics Simulations." In: *Physical Review Letters* 102 (2009), p. 20601 (cit. on pp. 6, 86).

[9] Giovanni Bussi, Tatyana Zykova-Timan, and Michele Parrinello. "Isothermal-isobaric molecular dynamics using stochastic velocity rescaling." In: 130 (2009), p. 74101 (cit. on p. 6).

[10] U Burkert and NL Allinger. "Molecular Mechanics ACS." In: *Washington, DC* (1982), p. 52 (cit. on p. 7).

[11] Norman L Allinger. "Conformational analysis. 130. MM2. A hydrocarbon force field utilizing V1 and V2 torsional terms." In: *Journal of the American Chemical Society* 99.25 (1977), pp. 8127–8134 (cit. on p. 7).

[12] NL Allinger, YH Yuh, and JH Lii. "J Am Chem Soc 111: 8551;(b) Lii JH." In: *Allinger NL (1989) J Am Chem Soc* 111 (1989), p. 8566 (cit. on p. 7).

[13] Norman L Allinger, Kuohsiang Chen, and Jenn-Huei Lii. "An improved force field (MM4) for saturated hydrocarbons." In: *Journal of Computational Chemistry* 17.5-6 (1996), pp. 642–668 (cit. on p. 7).

[14] Anthony K Rappé et al. "UFF, a full periodic table force field for molecular mechanics and molecular dynamics simulations." In: *Journal of the American chemical society* 114.25 (1992), pp. 10024–10035 (cit. on p. 7).

[15] Alex D MacKerell Jr et al. "All-atom empirical potential for molecular modeling and dynamics studies of proteins." In: *The journal of physical chemistry B* 102.18 (1998), pp. 3586–3616 (cit. on p. 7).

[16] Wendy D Cornell et al. "A second generation force field for the simulation of proteins, nucleic acids, and organic molecules." In: *Journal of the American Chemical Society* 117.19 (1995), pp. 5179–5197 (cit. on p. 7).

[17] Walter RP Scott et al. "The GROMOS biomolecular simulation program package." In: *The Journal of Physical Chemistry A* 103.19 (1999), pp. 3596–3607 (cit. on p. 7).

[18] William L Jorgensen and Julian Tirado-Rives. "The OPLS [optimized potentials for liquid simulations] potential functions for proteins, energy minimizations for crystals of cyclic peptides and crambin." In: *Journal of the American Chemical Society* 110.6 (1988), pp. 1657–1666 (cit. on p. 7).

[19] Peter L Freddolino et al. "Molecular dynamics simulations of the complete satellite tobacco mosaic virus." In: *Structure* 14.3 (2006), pp. 437–449 (cit. on p. 7).

[20] Akio Kitao et al. "Switch interactions control energy frustration and multiple flagellar filament structures." In: *Proceedings of the National Academy of Sciences* 103.13 (2006), pp. 4894–4899 (cit. on p. 7).

[21] Gongpu Zhao et al. "Mature HIV-1 capsid structure by cryo-electron microscopy and all-atom molecular dynamics." In: *Nature* 497.7451 (2013), pp. 643–646 (cit. on p. 7).

[22] A Warshel and M Karplus. "Calculation of ground and excited state potential surfaces of conjugated molecules. I. Formulation and parametrization." In: *Journal of the American Chemical Society* 94.16 (1972), pp. 5612–5625 (cit. on p. 7).

[23] ISY Wang and M Karplus. "Dyanmics of organic reactions." In: *Journal of the American Chemical Society* 95.24 (1973), pp. 8160–8164 (cit. on p. 7).

[24] Arieh Warshel and Michael Levitt. "Theoretical studies of enzymic reactions: dielectric, electrostatic and steric stabilization of the carbonium ion in the reaction of lysozyme." In: *Journal of molecular biology* 103.2 (1976), pp. 227–249 (cit. on p. 7).

[25] Martin J Field, Paul A Bash, and Martin Karplus. "A combined quantum mechanical and molecular mechanical potential for molecular dynamics simulations." In: *Journal of Computational Chemistry* 11.6 (1990), pp. 700–733 (cit. on p. 7).

[26] Ute F Röhrig et al. "QM/MM Car-Parrinello Molecular Dynamics Study of the Solvent Effects on the Ground State and on the First Excited Singlet State of Acetone in Water." In: *ChemPhysChem* 4.11 (2003), pp. 1177–1182 (cit. on p. 7).

[27] Hans Martin Senn and Walter Thiel. "QM/MM methods for biomolecular systems." In: *Angewandte Chemie International Edition* 48.7 (2009), pp. 1198–1229 (cit. on p. 7).

[28] Richard A Friesner and Victor Guallar. "Ab initio quantum chemical and mixed quantum mechanics/molecular mechanics (QM/MM) methods for studying enzymatic catalysis." In: *Annual Review of Physical Chemistry* 56 (2005), pp. 389–427 (cit. on p. 7).

[29] Roald Hoffmann. "An extended Hückel theory. I. hydrocarbons." In: *The Journal of Chemical Physics* 39.6 (1963), pp. 1397–1412 (cit. on p. 8).

[30] Michael James Steuart Dewar. "Molecular orbital theory of organic chemistry." In: (1969) (cit. on p. 8).

[31] John A Pople and David L Beveridge. "Molecular orbital theory." In: *Co., NY* (1970) (cit. on p. 8).

[32] John Norman Murrell and Alan John Harget. "Semi-empirical self-consistent-field molecular orbital theory of molecules." In: (1972) (cit. on p. 8).

[33] John C Slater. "A simplification of the Hartree-Fock method." In: *Physical Review* 81.3 (1951), p. 385 (cit. on p. 8).

[34] Charlotte Froese Fischer. "Hartree–Fock method for atoms. A numerical approach." In: (1977) (cit. on p. 8).

[35] PJ Knowles and NC Handy. "A new determinant-based full configuration interaction method." In: *Chemical physics letters* 111.4-5 (1984), pp. 315–321 (cit. on p. 8).

[36] Chr Møller and Milton S Plesset. "Note on an approximation treatment for many-electron systems." In: *Physical Review* 46.7 (1934), p. 618 (cit. on p. 8).

[37] Hendrik J Monkhorst. "Calculation of properties with the coupled-cluster method." In: *International Journal of Quantum Chemistry* 12.S11 (1977), pp. 421–432 (cit. on p. 8).

[38] Bogumil Jeziorski and Hendrik J Monkhorst. "Coupled-cluster method for multideterminantal reference states." In: *Physical Review A* 24.4 (1981), p. 1668 (cit. on p. 8).

[39] John F Stanton and Rodney J Bartlett. "The equation of motion coupled-cluster method. A systematic biorthogonal approach to molecular excitation energies, transition probabilities, and excited state properties." In: *The Journal of chemical physics* 98.9 (1993), pp. 7029–7039 (cit. on p. 8).

[40] Kurt Binder. "Introduction: Theory and "technical" aspects of Monte Carlo simulations." In: *Monte Carlo Methods in Statistical Physics*. Springer, 1986, pp. 1–45 (cit. on p. 8).

[41] WMC Foulkes et al. "Quantum Monte Carlo simulations of solids." In: *Reviews of Modern Physics* 73.1 (2001), p. 33 (cit. on p. 8).

[42] David Ceperley and Berni Alder. "Quantum monte carlo." In: *Science* 231.4738 (1986), pp. 555–560 (cit. on p. 8).

[43] Walter Kohn and Lu Jeu Sham. "Self-consistent equations including exchange and correlation effects." In: *Physical review* 140.4A (1965), A1133 (cit. on p. 8).

[44] Robert G Parr. "Density functional theory of atoms and molecules." In: *Horizons of Quantum Chemistry*. Springer, 1980, pp. 5–15 (cit. on p. 8).

[45] F Matthias Bickelhaupt and Evert Jan Baerends. "Kohn-Sham Density Functional Theory: Predicting and Understanding Chemistry." In: *Reviews in Computational Chemistry, Volume 15* (2007), pp. 1–86 (cit. on p. 8).

[46] Alvaro Valdes et al. "Solar hydrogen production with semiconductor metal oxides: new directions in experiment and theory." In: *Phys. Chem. Chem. Phys.* 14.1 (2012), pp. 49–70 (cit. on p. 9).

[47] Yifei Mo, Shyue Ping Ong, and Gerbrand Ceder. "First-principles study of the oxygen evolution reaction of lithium peroxide in the lithium-air battery." In: *Physical Review B* 84.20 (2011), p. 205446 (cit. on p. 9).

[48] Javier Carrasco et al. "Insight into the description of van der Waals forces for benzene adsorption on transition metal (111) surfaces." In: *The Journal of Chemical Physics* 140.8 (2014), p. 084704 (cit. on p. 9).

[49] John Landers, Gennady Yu Gor, and Alexander V Neimark. "Density functional theory methods for characterization of porous materials." In: *Colloids and Surfaces A: Physicochemical and Engineering Aspects* 437 (2013), pp. 3–32 (cit. on p. 9).

[50] Andrew Zangwill. "A half century of density functional theory." In: *Physics today* 68.7 (2015), pp. 34–39 (cit. on p. 9).

[51] Jens K Nørskov et al. "Density functional theory in surface chemistry and catalysis." In: *Proceedings of the National Academy of Sciences* 108.3 (2011), pp. 937–943 (cit. on p. 9).

[52] Kari Laasonen et al. ""Ab initio"liquid water." In: *The Journal of chemical physics* 99.11 (1993), pp. 9080–9089 (cit. on p. 9).

[53] Michiel Sprik, Jürg Hutter, and Michele Parrinello. "Ab initio molecular dynamics simulation of liquid water: Comparison of three gradient-corrected density functionals." In: *The Journal of chemical physics* 105.3 (1996), pp. 1142–1152 (cit. on p. 9).

[54] M Diraison, GJ Martyna, and ME Tuckerman. "Simulation studies of liquid ammonia by classical ab initio, classical, and path-integral molecular dynamics." In: *The Journal of chemical physics* 111.3 (1999), pp. 1096–1103 (cit. on p. 9).

[55] Magali Benoit, Dominik Marx, and Michele Parrinello. "Tunnelling and zero-point motion in high-pressure ice." In: *Nature* 392.6673 (1998), pp. 258–261 (cit. on pp. 9, 11).

[56] Jeff Greeley, Jens K Nørskov, and Manos Mavrikakis. "Electronic structure and catalysis on metal surfaces." In: *Annual Review of Physical Chemistry* 53.1 (2002), pp. 319–348 (cit. on p. 9).

[57] Mauro Boero, Michele Parrinello, and Kiyoyuki Terakura. "First Principles Molecular Dynamics Study of Ziegler- Natta Heterogeneous Catalysis." In: *Journal of the American Chemical Society* 120.12 (1998), pp. 2746–2752 (cit. on p. 9).

[58] Dominik Marx et al. "The nature of the hydrated excess proton in water." In: *Nature* 397.6720 (1999), pp. 601–604 (cit. on p. 9).

[59] Ali A Hassanali et al. "The fuzzy quantum proton in the hydrogen chloride hydrates." In: *Journal of the American Chemical Society* 134.20 (2012), pp. 8557–8569 (cit. on p. 9).

[60] Michele Ceriotti et al. "Nuclear quantum effects in ab initio dynamics: Theory and experiments for lithium imide." In: *Physical Review B* 82.17 (2010), p. 174306 (cit. on p. 9).

[61] Stefan Grimme. "Density functional theory with London dispersion corrections." In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.2 (2011), pp. 211–228 (cit. on p. 9).

[62] Max Dion et al. "Van der Waals density functional for general geometries." In: *Physical review letters* 92.24 (2004), p. 246401 (cit. on p. 9).

[63] John P Perdew and Karla Schmidt. "Jacob's ladder of density functional approximations for the exchange-correlation energy." In: *AIP Conference Proceedings*. Vol. 577. 1. AIP. 2001, pp. 1–20 (cit. on p. 9).

[64] John P Perdew and Alex Zunger. "Self-interaction correction to density-functional approximations for many-electron systems." In: *Physical Review B* 23.10 (1981), p. 5048 (cit. on p. 10).

[65] John P Perdew, Kieron Burke, and Matthias Ernzerhof. "Generalized gradient approximation made simple." In: *Physical review letters* 77.18 (1996), p. 3865 (cit. on p. 10).

[66] Seymour H Vosko, Leslie Wilk, and Marwan Nusair. "Accurate spin-dependent electron liquid correlation energies for local spin density calculations: a critical analysis." In: *Canadian Journal of physics* 58.8 (1980), pp. 1200–1211 (cit. on p. 10).

[67] Chengteh Lee, Weitao Yang, and Robert G Parr. "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density." In: *Physical review B* 37.2 (1988), p. 785 (cit. on p. 10).

[68] John P Perdew. "Density-functional approximation for the correlation energy of the inhomogeneous electron gas." In: *Physical Review B* 33.12 (1986), p. 8822 (cit. on p. 10).

[69] Carlo Adamo and Vincenzo Barone. "Toward reliable density functional methods without adjustable parameters: The PBE0 model." In: *The Journal of chemical physics* 110.13 (1999), pp. 6158–6170 (cit. on p. 10).

[70] Axel D Becke. "Phys. Re V. A 1988, 38, 3098.(b) Becke." In: *J. Chem. Phys* 98 (1993), p. 5648 (cit. on p. 10).

[71] PJ Stephens et al. "Ab initio calculation of vibrational absorption and circular dichroism spectra using density functional force fields." In: *The Journal of Physical Chemistry* 98.45 (1994), pp. 11623–11627 (cit. on p. 10).

[72] RHWJ Ditchfield, W J_ Hehre, and John A Pople. "Self-consistent molecular-orbital methods. IX. An extended Gaussian-type basis for molecular-orbital studies of organic molecules." In: *The Journal of Chemical Physics* 54.2 (1971), pp. 724–728 (cit. on p. 10).

[73] David E Woon and Thom H Dunning Jr. "Gaussian basis sets for use in correlated molecular calculations. V. Core-valence basis sets for boron through neon." In: *The Journal of chemical physics* 103.11 (1995), pp. 4572–4585 (cit. on p. 10).

[74] David E Woon and Thom H Dunning Jr. "Benchmark calculations with correlated molecular wave functions. I. Multireference configuration interaction calculations for the second row diatomic hydrides." In: *The Journal of chemical physics* 99.3 (1993), pp. 1914–1929 (cit. on p. 10).

[75] Volker Blum et al. "Ab initio molecular simulations with numeric atom-centered orbitals." In: *Computer Physics Communications* 180.11 (2009), pp. 2175–2196 (cit. on p. 11).

[76] Georg Kresse and Jürgen Furthmüller. "Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set." In: *Physical review B* 54.16 (1996), p. 11169 (cit. on p. 11).

[77] Joseph A Morrone and Roberto Car. "Nuclear quantum effects in water." In: *Physical review letters* 101.1 (2008), p. 017801 (cit. on p. 11).

[78] Michele Ceriotti et al. "Nuclear quantum effects in water and aqueous systems: Experiment, theory, and current challenges." In: *Chemical reviews* 116.13 (2016), pp. 7529–7550 (cit. on p. 11).

[79] Michele Ceriotti et al. "Nuclear quantum effects and hydrogen bond fluctuations in water." In: *Proceedings of the National Academy of Sciences* 110.39 (2013), pp. 15591–15596 (cit. on p. 11).

[80] Miguel A Morales et al. "Nuclear quantum effects and nonlocal exchange-correlation functionals applied to liquid hydrogen at high pressure." In: *Physical review letters* 110.6 (2013), p. 065702 (cit. on p. 11).

[81] Bet Pamuk et al. "Anomalous nuclear quantum effects in ice." In: *Physical review letters* 108.19 (2012), p. 193003 (cit. on p. 11).

[82] Salomon R Billeter et al. "Hybrid approach for including electronic and nuclear quantum effects in molecular dynamics simulations of hydrogen transfer reactions in enzymes." In: *The Journal of Chemical Physics* 114.15 (2001), pp. 6925–6936 (cit. on p. 11).

[83] Joseph A. Morrone and Roberto Car. "Nuclear Quantum Effects in Water." In: *Physical Review Letters* 101.1 (2008), p. 017801 (cit. on p. 11).

[84] Mariana Rossi, Piero Gasparotto, and Michele Ceriotti. "Anharmonic and Quantum Fluctuations in Molecular Crystals: A First-Principles Study of the Stability of Paracetamol." In: *Physical Review Letters* 117 (2016), p. 115702 (cit. on p. 11).

[85] Richard P Feynman, Albert R Hibbs, and Daniel F Styer. *Quantum mechanics and path integrals*. Courier Corporation, 2010 (cit. on p. 11).

[86] RP Feynman. *Statistical Mechanics, A Set of Lectures, California, Institute of Technology*. 1972 (cit. on p. 11).

[87] David Chandler and Peter G Wolynes. "Exploiting the isomorphism between quantum theory and classical statistical mechanics of polyatomic fluids." In: *Journal of Chemical Physics* 74 (1981), pp. 4078–4095 (cit. on p. 11).

[88] M Parrinello and A Rahman. "Study of an F center in molten KCl." In: *Journal of Chemical Physics* 80 (1984), p. 860 (cit. on p. 11).

[89] Mark E Tuckerman et al. "Efficient and general algorithms for path integral Car-Parrinello molecular dynamics." In: *Journal of Chemical Physics* 104 (1996), pp. 5579–5588 (cit. on p. 12).

[90] Dominik Marx and Michele Parrinello. "Ab initio path integral molecular dynamics: Basic ideas." In: *Journal of Chemical Physics* 104 (1996), p. 4077 (cit. on p. 12).

[91] D Marx et al. "The nature of the hydrated excess proton in water." In: *Nature* 397 (1999), pp. 601–604 (cit. on p. 12).

[92] Michele Ceriotti, Giovanni Bussi, and Michele Parrinello. "Colored-Noise Thermostats á la Carte." In: *Journal of Chemical Theory and Computation* 6.4 (2010), pp. 1170–1180 (cit. on p. 12).

[93] Michele Ceriotti, Giovanni Bussi, and Michele Parrinello. "Colored-Noise Thermostats à la Carte." In: *Journal of Chemical Theory and Computation* 6 (2010), pp. 1170–1180 (cit. on pp. 12, 43, 86, 114).

[94] Michele Ceriotti and David E Manolopoulos. "Efficient First-Principles Calculation of the Quantum Kinetic Energy and Momentum Distribution of Nuclei." In: *Physical Review Letters* 109 (2012), p. 100604 (cit. on pp. 12, 71, 77, 86).

[95] Peter Hänggi, Peter Talkner, and Michal Borkovec. "Reaction-rate theory: fifty years after Kramers." In: *Reviews of modern physics* 62.2 (1990), p. 251 (cit. on p. 12).

[96] M Eigen. "Proton Transfer, Acid-Base Catalysis, and Enzymatic Hydrolysis. Part I: ELEMENTARY PROCESSES." In: *Angewandte Chemie International Edition* 3.1 (1964), pp. 1–19 (cit. on p. 13).

[97] Phillip L Geissler et al. "Autoionization in liquid water." In: *Science* 291.5511 (2001), pp. 2121–2124 (cit. on p. 13).

[98] Stefano Piana, Kresten Lindorff-Larsen, and David E Shaw. "Protein folding kinetics and thermodynamics from atomistic simulation." In: *Proceedings of the National Academy of Sciences* 109.44 (2012), pp. 17845–17850 (cit. on p. 13).

[99] R Martoňák, Alessandro Laio, and Michele Parrinello. "Predicting crystal structures: the Parrinello-Rahman method revisited." In: *Physical Review Letters* 90.7 (2003), p. 075503 (cit. on p. 13).

[100] Bernd Ensing et al. "Metadynamics as a tool for exploring free energy landscapes of chemical reactions." In: *Accounts of chemical research* 39.2 (2006), pp. 73–81 (cit. on p. 13).

[101] Peter Hänggi, Peter Talkner, and Michal Borkovec. "Reaction-rate theory: fifty years after Kramers." In: *Reviews of modern physics* 62.2 (1990), p. 251 (cit. on p. 13).

[102] Christoph Dellago and Peter G. Bolhuis. "Transition Path Sampling and Other Advanced Simulation Techniques for Rare Events." In: *Advanced Computer Simulation Approaches for Soft Matter Sciences III*. Ed. by Christian Holm and Kurt Kremer. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 167–233. ISBN: 978-3-540-87706-6 (cit. on p. 13).

[103] Cameron Abrams and Giovanni Bussi. "Enhanced sampling in molecular dynamics using metadynamics, replica-exchange, and temperature-acceleration." In: *Entropy* 16.1 (2013), pp. 163–199 (cit. on p. 13).

[104] Alessandro Barducci, Massimiliano Bonomi, and Michele Parrinello. "Metadynamics." In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.5 (2011), pp. 826–843 (cit. on p. 13).

[105] Pratyush Tiwary and Axel van de Walle. "A Review of Enhanced Sampling Approaches for Accelerated Molecular Dynamics." In: *Multiscale Materials Modeling for Nanomechanics*. Springer, 2016, pp. 195–221 (cit. on p. 13).

[106] A Laio and F L Gervasio. "Metadynamics: a method to simulate rare events and reconstruct the free energy in biophysics, chemistry and material science." In: *Reports on Progress in Physics* 71 (2008), p. 126601 (cit. on p. 13).

[107] David J Earl and Michael W Deem. "Parallel tempering: Theory, applications, and new perspectives." In: *Physical Chemistry Chemical Physics* 7.23 (2005), pp. 3910–3916 (cit. on p. 13).

[108] Davide Branduardi, Francesco Luigi Gervasio, and Michele Parrinello. "From A to B in free energy space." In: *Journal of Chemical Physics* 126 (2007), p. 054103 (cit. on p. 13).

[109] Fabio Pietrucci and Wanda Andreoni. "Graph Theory Meets Ab Initio Molecular Dynamics: Atomic Structures and Transformations at the Nanoscale." In: *Physical Review Letters* 107 (2011), p. 085504 (cit. on pp. 13, 21, 42, 56).

[110] Gareth A Tribello, Michele Ceriotti, and Michele Parrinello. "Using sketch-map coordinates to analyze and bias molecular dynamics simulations." In: *Proceedings of the National Academy of Sciences* 109 (2012), pp. 5196–201 (cit. on pp. 13, 37, 56, 119).

[111] Glenn M Torrie and John P Valleau. "Nonphysical sampling distributions in Monte Carlo free-energy estimation: Umbrella sampling." In: *Journal of Computational Physics* 23.2 (1977), pp. 187–199 (cit. on p. 14).

[112] S Kumar et al. "The weighted histogram analysis method for free-energy calculations on biomolecules. I. The method." In: *J. Comp. Chem.* 13 (1992), pp. 1011–1021 (cit. on p. 14).

[113] Johannes Kästner. "Umbrella sampling." In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 1.6 (2011), pp. 932–942 (cit. on p. 14).

[114] Alessandro Laio and Michele Parinello. "Escaping free-energy minima." In: *Proceedings of the National Academy of Sciences* 99 (2002), pp. 12562–12566 (cit. on p. 15).

[115] Paolo Raiteri et al. "Efficient reconstruction of complex free energy landscapes by multiple walkers metadynamics." In: *The Journal of Chemical Physics A* 110.8 (2006), pp. 3533–3539 (cit. on p. 15).

[116] Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. "Well-tempered metadynamics: A smoothly converging and tunable free-energy method." In: *Physical review letters* 100.2 (2008), p. 020603 (cit. on p. 15).

[117] Massimiliano Bonomi, Alessandro Barducci, and Michele Parrinello. "Reconstructing the equilibrium Boltzmann distribution from well-tempered metadynamics." In: *Journal of computational chemistry* 30.11 (2009), pp. 1615–1621 (cit. on p. 16).

[118] Pratyush Tiwary and Michele Parrinello. "A time-independent free energy estimator for metadynamics." In: *The Journal of Physical Chemistry B* 119.3 (2014), pp. 736–742 (cit. on p. 16).

[119] Fabrizio Marinelli et al. "A kinetic model of trp-cage folding from multiple biased molecular dynamics simulations." In: *PLoS computational biology* 5.8 (2009), e1000452 (cit. on p. 16).

[120] Ulrich HE Hansmann. "Parallel tempering algorithm for conformational studies of biological molecules." In: *Chemical Physics Letters* 281.1 (1997), pp. 140–150 (cit. on p. 16).

[121] Robert H Swendsen and Jian-Sheng Wang. "Replica Monte Carlo simulation of spin-glasses." In: *Physical Review Letters* 57.21 (1986), p. 2607 (cit. on p. 16).

[122] Simone Marsili et al. "ORAC: A molecular dynamics simulation program to explore free energy surfaces in biomolecular systems at the atomistic level." In: *Journal of computational chemistry* 31.5 (2010), pp. 1106–1116 (cit. on p. 17).

[123] Aminata Kone and David A Kofke. "Selection of temperature intervals for parallel-tempering simulations." In: *The Journal of chemical physics* 122.20 (2005), p. 206101 (cit. on p. 17).

[124] M Bonomi and M Parrinello. "Enhanced sampling in the well-tempered ensemble." In: *Physical Review Letters* 104 (2010), p. 190601 (cit. on pp. 17, 84, 87).

[125] J. Ross Quinlan. "Induction of decision trees." In: *Machine learning* 1.1 (1986), pp. 81–106 (cit. on p. 20).

[126] Leo Breiman. "Random forests." In: *Machine learning* 45.1 (2001), pp. 5–32 (cit. on p. 20).

[127] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*. Vol. 398. John Wiley & Sons, 2013 (cit. on p. 20).

[128] Marti A. Hearst et al. "Support vector machines." In: *IEEE Intelligent Systems and their applications* 13.4 (1998), pp. 18–28 (cit. on p. 20).

[129] Bernhard Scholkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001 (cit. on p. 20).

[130] Andrea Grisafi et al. "Symmetry-Adapted Machine-Learning for Tensorial Properties of Atomistic Systems." In: *Physical Review Letters* 120.3 (2018), p. 036002 (cit. on p. 20).

[131] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. "Multilayer feedforward networks are universal approximators." In: *Neural networks* 2.5 (1989), pp. 359–366 (cit. on p. 20).

[132] Alberto Testolin and Marco Zorzi. "Probabilistic models and generative neural networks: towards an unified framework for modeling normal and impaired neurocognitive functions." In: *Frontiers in computational neuroscience* 10 (2016) (cit. on p. 20).

[133] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." In: *Nature* 521.7553 (2015), pp. 436–444 (cit. on p. 20).

[134] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. Vol. 1. 1. MIT press Cambridge, 1998 (cit. on p. 20).

[135] Michael I Jordan and Tom M Mitchell. "Machine learning: Trends, perspectives, and prospects." In: *Science* 349.6245 (2015), pp. 255–260 (cit. on p. 20).

[136] Zoubin Ghahramani. "Unsupervised learning." In: *Advanced lectures on machine learning*. Springer, 2004, pp. 72–112 (cit. on p. 20).

[137] Ali Sadeghi et al. "Metrics for measuring distances in configuration spaces." In: *Journal of Chemical Physics* 139 (2013), p. 184118 (cit. on pp. 21, 42, 56).

[138] Matthias Rupp et al. "Fast and accurate modeling of molecular atomization energies with machine learning." In: *Physical Review Letters* 108.5 (2012), p. 058301 (cit. on p. 21).

[139] Jörg Behler. "Atom-centered symmetry functions for constructing high-dimensional neural network potentials." In: *Journal of Chemical Physics* 134 (2011) (cit. on pp. 21, 42).

[140] Albert P. Bartók et al. "Machine-learning approach for one- and two-body corrections to density functional theory: Applications to molecular and condensed water." In: *Physical Review B* 88 (2013), p. 054104 (cit. on pp. 22, 42).

[141] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. "A training algorithm for optimal margin classifiers." In: *Proceedings of the fifth annual workshop on Computational learning theory*. ACM. 1992, pp. 144–152 (cit. on p. 22).

[142] Martin D Buhmann. *Radial basis functions: theory and implementations*. Vol. 12. Cambridge university press, 2003 (cit. on p. 22).

[143] Sandip De et al. "Comparing molecules and solids across structural and alchemical space." In: *Physical Chemistry Chemical Physics* 18 (2016), pp. 13754–13769 (cit. on p. 23).

[144] Sandip De et al. "Mapping and classifying molecules from a high-throughput structural database." In: *Journal of Cheminformatics* 9.1 (2017), p. 6 (cit. on pp. 23, 54).

[145] James MacQueen et al. "Some methods for classification and analysis of multivariate observations." In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. 14. Oakland, CA, USA. 1967, pp. 281–297 (cit. on p. 24).

[146] Leonard Kaufman and Peter J Rousseeuw. "Partitioning around medoids (program pam)." In: *Finding groups in data: an introduction to cluster analysis* (1990), pp. 68–125 (cit. on p. 24).

[147] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65 (cit. on p. 25).

[148] Todd K Moon. "The expectation-maximization algorithm." In: *IEEE Signal processing magazine* 13.6 (1996), pp. 47–60 (cit. on p. 25).

[149] William H Press. *Numerical recipes 3rd edition: The art of scientific computing*. Cambridge university press, 2007 (cit. on p. 25).

[150] Martin Ester et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *Kdd*. Vol. 96. 34. 1996, pp. 226–231 (cit. on p. 26).

[151] Leland McInnes, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering." In: *The Journal of Open Source Software* 2.11 (2017), p. 205 (cit. on p. 26).

[152] Kamran Khan et al. "DBSCAN: Past, present and future." In: *Applications of Digital Information and Web Technologies (ICADIWT), 2014 Fifth International Conference on the*. IEEE. 2014, pp. 232–238 (cit. on p. 26).

[153] Keinosuke Fukunaga and Larry Hostetler. "The estimation of the gradient of a density function, with applications in pattern recognition." In: *IEEE Transactions on information theory* 21.1 (1975), pp. 32–40 (cit. on p. 26).

[154] Miguel A Carreira-Perpinan. "Gaussian mean-shift is an EM algorithm." In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.5 (2007), pp. 767–776 (cit. on p. 27).

[155] Marcus Weber, Wasinee Rungsarityotin, and Alexander Schliep. *Perron cluster analysis and its connection to graph partitioning for noisy data*. Konrad-Zuse-Zentrum für Informationstechnik Berlin, 2004 (cit. on p. 27).

[156] Susanna Röblitz and Marcus Weber. "Fuzzy spectral clustering by PCCA+: application to Markov state models and data classification." In: *Advances in Data Analysis and Classification* 7.2 (2013), pp. 147–179 (cit. on p. 27).

[157] Peter Deuflhard and Marcus Weber. "Robust Perron cluster analysis in conformation dynamics." In: *Linear algebra and its applications* 398 (2005), pp. 161–184 (cit. on p. 27).

[158] John D Chodera and Frank Noé. "Markov state models of biomolecular conformational dynamics." In: *Current opinion in structural biology* 25 (2014), pp. 135–144 (cit. on p. 27).

[159] Paul Doukhan and J Leon. "Déviation quadratique d'estimateurs de densité par projections orthogonales." In: *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics* 310 (1990), pp. 425–430 (cit. on p. 27).

[160] Grace Wahba. "Data-based optimal smoothing of orthogonal series density estimates." In: *The annals of statistics* (1981), pp. 146–156 (cit. on p. 27).

[161] IJ Goodd and Ray A Gaskins. "Nonparametric roughness penalties for probability densities." In: *Biometrika* 58.2 (1971), pp. 255–277 (cit. on p. 27).

[162] L Devroye and Nonparametric Density Estimation Györfi. *The L1 View*. 1985 (cit. on p. 27).

[163] David W Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015 (cit. on pp. 27, 29, 30).

[164] Karl Pearson. "X. Contributions to the mathematical theory of evolution.—II. Skew variation in homogeneous material." In: *Phil. Trans. R. Soc. Lond. A* 186 (1895), pp. 343–414 (cit. on p. 27).

[165] David P Doane. "Aesthetic frequency classifications." In: *The American Statistician* 30.4 (1976), pp. 181–183 (cit. on p. 28).

[166] David W Scott. "On optimal and data-based histograms." In: *Biometrika* 66.3 (1979), pp. 605–610 (cit. on p. 28).

[167] Emanuel Parzen. "On estimation of a probability density function and mode." In: *The annals of mathematical statistics* 33.3 (1962), pp. 1065–1076 (cit. on p. 28).

[168] Theophilos Cacoullos. "Estimation of a multivariate density." In: *Annals of the Institute of Statistical Mathematics* 18.1 (1966), pp. 179–189 (cit. on p. 29).

[169] Vassiliy A Epanechnikov. "Non-parametric estimation of a multivariate probability density." In: *Theory of Probability & Its Applications* 14.1 (1969), pp. 153–158 (cit. on p. 29).

[170] Bernard W Silverman. *Density estimation for statistics and data analysis*. Vol. 26. CRC press, 1986 (cit. on pp. 30, 32).

[171] J Steve Marron and Matt P Wand. "Exact mean integrated squared error." In: *The Annals of Statistics* (1992), pp. 712–736 (cit. on p. 30).

[172] George R Terrell and David W Scott. "Oversmoothed nonparametric density estimates." In: *Journal of the American Statistical Association* 80.389 (1985), pp. 209–214 (cit. on p. 30).

[173] M Chris Jones, James S Marron, and Simon J Sheather. "A brief survey of bandwidth selection for density estimation." In: *Journal of the American Statistical Association* 91.433 (1996), pp. 401–407 (cit. on p. 31).

[174] Berwin A Turlach et al. *Bandwidth selection in kernel density estimation: A review*. Université catholique de Louvain Louvain-la-Neuve, 1993 (cit. on p. 31).

[175] Ian S Abramson. "On bandwidth variation in kernel estimates-a square root law." In: *The annals of Statistics* (1982), pp. 1217–1223 (cit. on p. 32).

[176] George R Terrell and David W Scott. "Variable kernel density estimation." In: *The Annals of Statistics* (1992), pp. 1236–1265 (cit. on p. 32).

[177] Chris Thornton et al. "Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms." In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2013, pp. 847–855 (cit. on p. 33).

[178] Matthias Feurer, Jost Tobias Springenberg, and Frank Hutter. "Initializing Bayesian Hyperparameter Optimization via Meta-Learning." In: *AAAI*. 2015, pp. 1128–1135 (cit. on p. 33).

[179] William H Press. "Numerical Recipes: The art of scientific computing." In: *Cambridge Uni* (2007) (cit. on p. 33).

[180] Tobias Reitmaier and Bernhard Sick. "The responsibility weighted Mahalanobis kernel for semi-supervised training of support vector machines for classification." In: *Information Sciences* 323 (2015), pp. 179–198 (cit. on p. 34).

[181] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Vol. 31. Springer Science & Business Media, 2013 (cit. on p. 34).

[182] Joseph B Kruskal. "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis." In: *Psychometrika* 29.1 (1964), pp. 1–27 (cit. on pp. 36, 119).

[183] Trevor F Cox and Michael A A Cox. *Multidimensional scaling*. CRC Press, 2010 (cit. on p. 36).

[184] J B Tenenbaum, V de Silva, and J C Langford. "A global geometric framework for nonlinear dimensionality reduction." In: *Science* 290 (2000), pp. 2319–2323 (cit. on pp. 36, 119).

[185] Andrew L Ferguson et al. "Systematic determination of order parameters for chain dynamics using diffusion maps." In: *Proceedings of the National Academy of Sciences* 107 (2010), pp. 13597–602 (cit. on pp. 36, 56).

[186] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. "Nonlinear Component Analysis as a Kernel Eigenvalue Problem." In: *Neural Computation* 10 (1998), pp. 1299–1319 (cit. on p. 36).

[187] Mikhail Belkin and Partha Niyogi. "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation." In: *Neural Computation* 15 (2003), pp. 1373–1396 (cit. on p. 36).

[188] S T Roweis and L K Saul. "Nonlinear dimensionality reduction by locally linear embedding." In: *Science* 290 (2000), pp. 2323–6 (cit. on p. 36).

[189] David L Donoho and Carrie Grimes. "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data." In: *Proceedings of the National Academy of Sciences* 100 (2003), pp. 5591–5596 (cit. on p. 36).

[190] Laurens van der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE." In: *Journal of Machine Learning Research* 9.Nov (2008), pp. 2579–2605 (cit. on pp. 36, 119).

[191] Michele Ceriotti, Gareth A Tribello, and Michele Parrinello. "From the Cover: Simplifying the representation of complex free-energy landscapes using sketch-map." In: *Proceedings of the National Academy of Sciences* 108 (2011), pp. 13023–8 (cit. on p. 36).

[192] Lawrence Cayton. "Algorithms for manifold learning." In: *Univ. of California at San Diego Tech. Rep* 12 (2005), pp. 1–17 (cit. on p. 37).

[193] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. "Dimensionality reduction: a comparative." In: *J Mach Learn Res* 10 (2009), pp. 66–71 (cit. on p. 37).

[194] Christopher JC Burges et al. "Dimension reduction: A guided tour." In: *Foundations and Trends® in Machine Learning* 2.4 (2010), pp. 275–365 (cit. on p. 37).

[195] Rustam Z Khaliullin et al. "Unravelling the origin of intermolecular interactions using absolutely localized molecular orbitals." In: *The Journal of Physical Chemistry A* 111.36 (2007), pp. 8753–8765 (cit. on p. 39).

[196] Jörg Behler and Michele Parrinello. "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces." In: *Physical Review Letters* 98 (2007), p. 146401 (cit. on pp. 42, 132).

[197] GA Gallet and Fabio Pietrucci. "Structural cluster analysis of chemical reactions in solution." In: *Journal of Chemical Physics* 139 (2013), p. 074101 (cit. on p. 42).

[198] Akio Kitao and Nobuhiro Go. "Investigating protein dynamics in collective coordinate space." In: *Current Opinion in Structural Biology* 9.2 (1999), pp. 164–169. ISSN: 0959-440X (cit. on p. 42).

[199] Davide Branduardi, Francesco Luigi Gervasio, and Michele Parrinello. "From A to B in free energy space." In: *The Journal of Chemical Physics* 126.5 (2007), p. 054103. eprint: `https://doi.org/10.1063/1.2432340` (cit. on p. 42).

[200] Feliks Nüske et al. "Variational approach to molecular kinetics." In: *Journal of chemical theory and computation* 10.4 (2014), pp. 1739–1752 (cit. on p. 42).

[201] Pratyush Tiwary and BJ Berne. "Spectral gap optimization of order parameters for sampling complex molecular systems." In: *Proceedings of the National Academy of Sciences* 113.11 (2016), pp. 2839–2844 (cit. on p. 42).

[202] Daniel J. Rosenkrantz, Richard E. Stearns, and II Philip M. Lewis. "An Analysis of Several Heuristics for the Traveling Salesman Problem." In: *SIAM Journal on Computing* 6.3 (1977), pp. 563–581 (cit. on p. 43).

[203] Y Eldar et al. "The farthest point strategy for progressive image sampling." In: *IEEE transactions on image processing : a publication of the IEEE Signal Processing Society* 6 (1997), pp. 1305–15 (cit. on p. 43).

[204] Sandhya Prabhakaran et al. "Automatic Model Selection in Archetype Analysis." In: *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2012, pp. 458–467 (cit. on p. 43).

[205] M Ester et al. *A density-based algorithm for discovering clusters in large spatial databases with noise*. Tech. rep. AAAI Press, Menlo Park, CA (United States), 1996 (cit. on p. 43).

[206] Alex Rodriguez and Alessandro Laio. "Machine learning. Clustering by fast search and find of density peaks." In: *Science* 344 (2014), pp. 1492–1496 (cit. on p. 43).

[207] Y. Chen et al. "Shrinkage Algorithms for MMSE Covariance Estimation." In: *IEEE Transactions on Signal Processing* 58.10 (2010), pp. 5016–5029 (cit. on p. 47).

[208] O. Roy and M. Vetterli. "The effective rank: A measure of effective dimensionality." In: *15th European Signal Processing Conference*. 2007, pp. 606–610 (cit. on p. 47).

[209] William H. Press. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, 2007. ISBN: 0521880688 (cit. on p. 48).

[210] Andrea Vedaldi and Stefano Soatto. "Quick Shift and Kernel Methods for Mode Seeking." In: *Computer Vision – ECCV 2008: 10th European Conference on Computer Vision, Marseille, France, October 12-18, 2008, Proceedings, Part IV*. Ed. by David Forsyth, Philip Torr, and Andrew Zisserman. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008, pp. 705–718 (cit. on p. 49).

[211] Mariana Rossi, Michele Ceriotti, and David E Manolopoulos. "Nuclear Quantum Effects in H+ and OH–Diffusion along Confined Water Wires." In: *The Journal of Physical Chemistry Letters* 7.15 (2016), pp. 3001–3007 (cit. on p. 51).

[212] Kanti V Mardia and Peter E Jupp. "Distributions on spheres." In: *Directional Statistics* (2000), pp. 159–192 (cit. on p. 51).

[213] Kanti V. Mardia et al. "A Multivariate Von Mises Distribution with Applications to Bioinformatics." In: *The Canadian Journal of Statistics / La Revue Canadienne de Statistique* 36.1 (2008), pp. 99–109 (cit. on p. 51).

[214] Suvrit Sra. "A short note on parameter approximation for von Mises-Fisher distributions: and a fast implementation of $I_s(x)$." In: *Computational Statistics* 27.1 (2012), pp. 177–190 (cit. on p. 52).

[215] B. Efron. "Bootstrap Methods: Another Look at the Jackknife." In: *The Annals of Statistics* 7.1 (1979), pp. 1–26 (cit. on p. 52).

[216] Fionn Murtagh and Pedro Contreras. "Algorithms for hierarchical clustering: an overview." In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 2.1 (2012), pp. 86–97 (cit. on p. 53).

[217] Joe H Ward Jr. "Hierarchical grouping to optimize an objective function." In: *Journal of the American statistical association* 58.301 (1963), pp. 236–244 (cit. on p. 54).

[218] A Laio and M Parrinello. "Escaping free-energy minima." In: *Proceedings of the National Academy of Sciences* 99 (2002), pp. 12562–12566 (cit. on p. 56).

[219] Mary a Rohrdanz, Wenwei Zheng, and Cecilia Clementi. "Discovering mountain passes via torchlight: methods for the definition of reaction coordinates and pathways in complex macromolecular reactions." In: *Annual review of physical chemistry* 64 (2013), pp. 295–316 (cit. on p. 56).

[220] Elangannan Arunan et al. "Definition of the hydrogen bond (IUPAC Recommendations 2011)." In: *Pure and applied chemistry* 83.8 (2011), pp. 1637–1641 (cit. on p. 57).

[221] Zygmunt S Derewenda, Linda Lee, and Urszula Derewenda. "The Occurence of C–H··· O Hydrogen Bonds in Proteins." In: *Journal of Molecular Biology* 252.2 (1995), pp. 248–262 (cit. on p. 57).

[222] Alan E Reed, Larry A Curtiss, and Frank Weinhold. "Intermolecular interactions from a natural bond orbital, donor-acceptor viewpoint." In: *Chemical Reviews* 88.6 (1988), pp. 899–926 (cit. on p. 58).

[223] A Martın Pendás, MA Blanco, and E Francisco. "The nature of the hydrogen bond: A synthesis from the interacting quantum atoms picture." In: *Journal of Chemical Physics* 125.18 (2006), p. 184112 (cit. on p. 58).

[224] R. Julian Azar and Martin Head-Gordon. "An energy decomposition analysis for intermolecular interactions from an absolutely localized molecular orbital reference at the coupled-cluster singles and doubles level." en. In: *The Journal of Chemical Physics* 136.2 (2012), p. 024103. ISSN: 00219606 (cit. on p. 58).

[225] Thomas D Kühne and Rustam Z Khaliullin. "Nature of the asymmetry in the hydrogen-bond networks of hexagonal ice and liquid water." In: *J. Chem. Am. Soc.* 136 (2014), pp. 3395–9 (cit. on p. 58).

[226] Robin Taylor and Olga Kennard. "Hydrogen-bond geometry in organic crystals." In: *Accounts of chemical research* 17.9 (1984), pp. 320–326 (cit. on p. 58).

[227] Masakazu Matsumoto. "Relevance of hydrogen bond definitions in liquid water." In: *Journal of Chemical Physics* 126.5 (2007), p. 054503 (cit. on p. 58).

[228] R Kumar, J R Schmidt, and J L Skinner. "Hydrogen bonding definitions and dynamics in liquid water." In: *Journal of Chemical Physics* 126 (2007), p. 204107 (cit. on pp. 58, 75, 90).

[229] Alexander D. MacKerell et al. "All-atom empirical potential for molecular modeling and dynamics studies of proteins." In: *The Journal of Physical Chemistry B* 102.18 (1998), pp. 3586–3616 (cit. on p. 60).

[230] William L Jorgensen et al. "Comparison of simple potential functions for simulating liquid water." In: *Journal of Chemical Physics* 79.2 (1983), pp. 926–935 (cit. on p. 60).

[231] Steve Plimpton. "Fast Parallel Algorithms for Short-Range Molecular Dynamics." In: *J. Comp. Phys.* 117 (1995), pp. 1–19 (cit. on p. 60).

[232] T Schneider and E Stoll. "Molecular-dynamics study of a three-dimensional one-component model for distortive phase transitions." In: *Physical Review B* 17 (1978), pp. 1302–1322 (cit. on p. 60).

[233] Miguel A. González and José L. F. Abascal. "A flexible model for water based on TIP4P/2005." en. In: *The Journal of Chemical Physics* 135.22 (2011), p. 224516. ISSN: 00219606 (cit. on p. 66).

[234] J A Hayward and J R Reimers. "Unit cells for the simulation of hexagonal ice." In: *Journal of Chemical Physics* 106 (1997), pp. 1518–1529 (cit. on p. 68).

[235] Niels Bjerrum. "Structure and properties of ice." In: *Science* 115.2989 (1952), pp. 385–390 (cit. on p. 68).

[236] Maurice de Koning et al. "Orientational defects in ice Ih: An interpretation of electrical conductivity measurements." In: *Physical Review Letters* 96.7 (2006), p. 075501 (cit. on p. 68).

[237] Joost VandeVondele et al. "Quickstep: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach." In: *Computer Physics Communications* 167 (2005), pp. 103–128 (cit. on p. 71).

[238] S Goedecker, M Teter, and J Hutter. "Separable dual-space Gaussian pseudopotentials." In: *Physical Review B* 54 (1996), pp. 1703–1710 (cit. on p. 71).

[239] A D Becke. "Density-functional exchange-energy approximation with correct asymptotic behavior." In: *Physical Review A* 38 (1988), p. 3098 (cit. on p. 71).

[240] Chengteh Lee, W Yang, and R G Parr. "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density." In: *Physical Review B* 37 (1988), p. 785 (cit. on p. 71).

[241] Michele Ceriotti, Joshua More, and David E. Manolopoulos. "i-PI: A Python interface for ab initio path integral molecular dynamics simulations." In: *Computer Physics Communications* 185 (2014), pp. 1019–1026 (cit. on pp. 71, 77, 85).

[242] Michele Ceriotti et al. "Nuclear quantum effects and hydrogen bond fluctuations in water." In: *Proceedings of the National Academy of Sciences* 110 (2013), pp. 15591–6 (cit. on p. 74).

[243] Alenka Luzar and David Chandler. "Hydrogen-bond kinetics in liquid water." In: *Nature* 379 (1996), pp. 55–57 (cit. on pp. 75, 92, 93).

[244] D. Eisenberg and W. Kauzmann. *The Structure and Properties of Water*. Oxford (UK): Oxford University Press, 1968 (cit. on p. 75).

[245] Ali Hassanali et al. "Proton transfer through the water gossamer." In: *Proceedings of the National Academy of Sciences* 110 (2013), pp. 13723–13728 (cit. on p. 75).

[246] Joshua More and David Manolopoulos. private communication. 2014 (cit. on p. 77).

[247] S Baroni et al. *PWSCF* (cit. on p. 77).

[248] Jp P Perdew, K Burke, and M Ernzerhof. "Generalized Gradient Approximation made simple." In: *Physical Review Letters*. Physical Review Lett. (USA) 77 (1996), p. 3865 (cit. on p. 77).

[249] D Vanderbilt. "Soft self-consistent pseudopotentials in a generalized eigenvalue formalism." In: *Physical Review B* 41 (1990), pp. 7892–7895 (cit. on p. 77).

[250] D L Bergman. "Topological properties of the hydrogen-bond network in liquid water." In: *Chemical Physics* 253.2–3 (2000), pp. 267–282 (cit. on p. 83).

[251] Noam Agmon. "Liquid Water: From Symmetry Distortions to Diffusive Motion." In: *Accounts of Chemical Research* 45.1 (2012), pp. 63–73 (cit. on pp. 83, 90, 95).

[252] Thomas D. Kühne and Rustam Z. Khaliullin. "Electronic signature of the instantaneous asymmetry in the first coordination shell of liquid water." In: *Nature Communications* 4 (Feb. 2013), p. 1450 (cit. on p. 83).

[253] Michele Ceriotti et al. "Nuclear quantum effects and hydrogen bond fluctuations in water." In: *Proceedings of the National Academy of Sciences* 110.39 (2013), pp. 15591–15596 (cit. on p. 83).

[254] F. Sciortino and S. L. Fornili. "Hydrogen bond cooperativity in simulated water: Time dependence analysis of pair interactions." In: *The Journal of Chemical Physics* 90.5 (1989), pp. 2786–2792 (cit. on p. 83).

[255] Richard H. Henchman and Sheeba Jem Irudayam. "Topological hydrogen-bond definition to characterize the structure and dynamics of liquid water." In: *Journal of Physical Chemistry B* 114 (2010), pp. 16792–16810 (cit. on pp. 83, 84, 100, 101).

[256] Barbara Kirchner, Philipp J. di Dio, and Jürg Hutter. *Real-World Predictions from Ab Initio Molecular Dynamics Simulations*. Vol. 307. Topics in Current Chemistry. Springer-Verlag Berlin, 2012, pp. 109–153 (cit. on p. 83).

[257] Michiel Sprik, Jürg Hutter, and Michele Parrinello. "Ab initio molecular dynamics simulation of liquid water: Comparison of three gradient-corrected density functionals." In: *The Journal of Chemical Physics* 105.3 (1996), pp. 1142–1152 (cit. on p. 83).

[258] Teodora Todorova et al. "Molecular Dynamics Simulation of Liquid Water: Hybrid Density Functionals." In: *Journal of Physical Chemistry B* 110.8 (2006), pp. 3685–3691 (cit. on p. 83).

[259] Manuel Guidon et al. "Ab–initio molecular dynamics using hybrid density functionals." In: *The Journal of Chemical Physics* 128.21, 214104 (2008), p. 214104 (cit. on p. 83).

[260] Pier Luigi Silvestrelli and Michele Parrinello. "Water Molecule Dipole in the Gas and in the Liquid Phase." In: *Physical Review Letters* 82 (16 Apr. 1999), pp. 3308–3311 (cit. on p. 83).

[261] Jeffrey C. Grossman et al. "Towards an assessment of the accuracy of density functional theory for first principles simulations of water." In: *The Journal of Chemical Physics* 120.1 (2004), pp. 300–311 (cit. on p. 83).

[262] M. V. Fernandez-Serra and Emilio Artacho. "Network equilibration and first-principles liquid water." In: *The Journal of Chemical Physics* 121.22 (2004), pp. 11136–11144 (cit. on p. 83).

[263] P. H.-L. Sit and Nicola Marzari. "Static and dynamical properties of heavy water at ambient conditions from first-principles molecular dynamics." In: *The Journal of Chemical Physics* 122.20, 204510 (2005), p. 204510 (cit. on p. 83).

[264] Joost VandeVondele et al. "The influence of temperature and density functional models in ab initio molecular dynamics simulation of liquid water." In: *The Journal of Chemical Physics* 122.1 (2005), p. 014515 (cit. on p. 83).

[265] Hee-Seung Lee and Mark E. Tuckerman. "Structure of liquid water at ambient temperature from ab initio molecular dynamics performed in the complete basis set limit." In: *The Journal of Chemical Physics* 125.15 (2006), p. 154507 (cit. on p. 83).

[266] I-Feng W. Kuo et al. "Liquid Water from First Principles: Investigation of Different Sampling Approaches." In: *Journal of Physical Chemistry B* 108.34 (2004), pp. 12990–12998 (cit. on p. 83).

[267] Cui Zhang et al. "First Principles Simulations of the Infrared Spectrum of Liquid Water Using Hybrid Density Functionals." In: *Journal of Chemical Theory and Computation* 7.5 (Mar. 2011), pp. 1443–1449. ISSN: 1549-9618 (cit. on p. 83).

[268] I-Chun Lin et al. "Importance of van der Waals Interactions in Liquid Water." In: *The Journal of Physical Chemistry B* 113.4 (2009), pp. 1127–1131 (cit. on pp. 83, 84).

[269] Biswajit Santra, Angelos Michaelides, and Matthias Scheffler. "Coupled cluster benchmarks of water monomers and dimers extracted from density-functional theory liquid water: The importance of monomer deformations." In: *The Journal of Chemical Physics* 131.12 (2009) (cit. on p. 83).

[270] Romain Jonchiere et al. "Van der Waals effects in ab initio water at ambient and supercritical conditions." In: *The Journal of Chemical Physics* 135.15, 154503 (2011), p. 154503 (cit. on p. 83).

[271] Cui Zhang et al. "Structural and Vibrational Properties of Liquid Water from van der Waals Density Functionals." In: *Journal of Chemical Theory and Computation* 7.10 (2011), pp. 3054–3061 (cit. on p. 83).

[272] Andreas Møgelhøj et al. "Ab Initio van der Waals Interactions in Simulations of Water Alter Structure from Mainly Tetrahedral to High-Density-Like." In: *The Journal of Physical Chemistry B* 115.48 (2011), pp. 14149–14160 (cit. on p. 83).

[273] Bin Chen et al. "Hydrogen Bonding in Water." In: *Physical Review Letters* 91.21 (Nov. 2003), p. 215503 (cit. on p. 83).

[274] Joseph A. Morrone and Roberto Car. "Nuclear Quantum Effects in Water." In: *Physical Review Letters* 101 (1 2008), p. 017801 (cit. on p. 83).

[275] Mauro Del Ben et al. "Bulk liquid water at ambient temperature and pressure from mp2 theory." In: *J. Phys. Chem. Letters* 4 (2013), pp. 3753–3759 (cit. on p. 83).

[276] Robert A. DiStasio et al. "The individual and collective effects of exact exchange and dispersion interactions on the ab initio structure of liquid water." In: *The Journal of Chemical Physics* 141.8, 084502 (2014), pp. - (cit. on pp. 83, 105, 106).

[277] Soohaeng Yoo and Sotiris S. Xantheas. "Communication: The effect of dispersion corrections on the melting temperature of liquid water." In: *The Journal of Chemical Physics* 134.12, 121105 (2011), pp. - (cit. on pp. 84, 105).

[278] Omololu Akin-Ojo and Feng Wang. "Effects of the dispersion interaction in liquid water." In: *Chemical Physics Letters* 513.1–3 (2011), pp. 59–62 (cit. on p. 84).

[279] H. Eugene Stanley and J. Teixeira. "Interpretation of the unusual behavior of H2O and D2O at low temperatures: Tests of a percolation model." In: *The Journal of Chemical Physics* 73.7 (1980), pp. 3404–3422 (cit. on p. 84).

[280] A. K. Soper and M. G. Phillips. "A new determination of the structure of water at 25C." In: *Chemical Physics* 107.1 (Aug. 1986), pp. 47–60. ISSN: 0301-0104 (cit. on p. 85).

[281] Ali Hassanali et al. "Proton transfer through the water gossamer." In: *Proceedings of the National Academy of Sciences* (2013), p. 13723 (cit. on p. 85).

[282] Joost VandeVondele et al. "QUICKSTEP: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach." In: *Computer Physics Communications* 167 (2005), pp. 103–128 (cit. on p. 85).

[283] Stefan Grimme et al. "A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu." In: *Journal of Chemical Physics* 132 (2010), p. 154104 (cit. on p. 86).

[284] C. Lee, W. Yang, and R. G. Parr. "Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density." In: *Physical Review* B37.2 (1988), pp. 785–789 (cit. on p. 86).

[285] S. Goedecker, M. Teter, and J. Hutter. "Separable dual-space Gaussian pseudopotentials." In: *Physical Review* B54.3 (1996), pp. 1703–1710 (cit. on p. 86).

[286] G Bussi, D Donadio, and M Parrinello. "Canonical sampling through velocity rescaling." In: *Journal of Chemical Physics* 126 (2007), p. 14101 (cit. on p. 86).

[287] Manuel Guidon, Jürg Hutter, and Joost VandeVondele. "Auxiliary Density Matrix Methods for HartreeFock Exchange Calculations." In: *Journal of Chemical Theory and Computation* 6 (2010), pp. 2348–2364 (cit. on pp. 86, 105).

[288] Miguel a González and José L F Abascal. "A flexible model for water based on TIP4P/2005." In: *Journal of Chemical Physics* 135 (2011), p. 224516 (cit. on p. 87).

[289] Volodymyr Babin, Claude Leforestier, and Francesco Paesani. "Development of a "First Principles" Water Potential with Flexible Monomers: Dimer Potential Energy Surface, VRT Spectrum, and Second Virial Coefficient." In: *Journal of Chemical Theory and Computation* 9.12 (2013), pp. 5395–5403 (cit. on p. 87).

[290] Volodymyr Babin, Gregory R. Medders, and Francesco Paesani. "Development of a "First Principles" Water Potential with Flexible Monomers. II: Trimer Potential Energy Surface, Third Virial Coefficient, and Small Clusters." In: *Journal of Chemical Theory and Computation* 10.4 (2014), pp. 1599–1607 (cit. on p. 87).

[291] Gregory R. Medders, Volodymyr Babin, and Francesco Paesani. "Development of a "First-Principles" Water Potential with Flexible Monomers. III. Liquid Phase Properties." In: *Journal of Chemical Theory and Computation* 10.8 (2014), pp. 2906–2910 (cit. on p. 87).

[292] Damien Laage and James T. Hynes. "On the Molecular Mechanism of Water Reorientation." In: *The Journal of Physical Chemistry B* 112.45 (2008), pp. 14230–14242 (cit. on p. 95).

[293] C.P P Lawrence and J.L L Skinner. "Ultrafast infrared spectroscopy probes hydrogen-bonding dynamics in liquid water." In: *Chemical Physics Letters* 369 (2003), pp. 472–477 (cit. on p. 96).

[294] B Auer et al. "Hydrogen bonding and Raman, IR, and 2D-IR spectroscopy of dilute HOD in liquid D2O." In: *Proceedings of the National Academy of Sciences* 104 (2007), pp. 14215–14220 (cit. on p. 96).

[295] K. T. Wikfeldt, A. Nilsson, and L. G. M. Pettersson. "Spatially inhomogeneous bimodal inherent structure of simulated liquid water." In: *Physical Chemistry Chemical Physics* 13 (44 2011), pp. 19918–19924 (cit. on p. 98).

[296] M. J. Gillan et al. "First-principles energetics of water clusters and ice: A many-body analysis." In: *Journal of Chemical Physics* 139 (2013), p. 244504 (cit. on p. 101).

[297] Wei Chen et al. "Role of dipolar correlations in the infrared spectra of water and ice." In: *Physical Review B* 77 (24 June 2008), p. 245114 (cit. on p. 104).

[298] Richard C. Remsing, Jocelyn M. Rodgers, and John D. Weeks. "Deconstructing Classical Water Models at Interfaces and in Bulk." In: *Journal of Statistical Physics* 145.2 (2011), pp. 313–334. ISSN: 1572-9613 (cit. on p. 105).

[299] Thomas D. Kühne, Matthias Krack, and Michele Parrinello. "Static and dynamical properties of liquid water from first principles by a novel car-parrinello-like approach." In: *Journal of Chemical Theory and Computation* 5 (2009), pp. 235–241 (cit. on p. 106).

[300] Lu Wang, Michele Ceriotti, and Thomas E. Markland. "Quantum fluctuations and isotope effects in ab initio descriptions of water." In: *Journal of Chemical Physics* 141 (2014), p. 104502 (cit. on p. 106).

[301] Gregory R. Medders et al. "On the representation of many-body interactions in water." In: *Journal of Chemical Physics* 143 (2015), p. 104102 (cit. on p. 106).

[302] Miguel a. Morales et al. "Quantum Monte Carlo benchmark of exchange-correlation functionals for bulk water." In: *Journal of Chemical Theory and Computation* 10 (2014), pp. 2355–2362 (cit. on p. 106).

[303] Hellmut Haberland. *Clusters of atoms and molecules: theory, experiment, and clusters of atoms*. Vol. 52. Springer Science & Business Media, 2013 (cit. on p. 113).

[304] JJ Shiang et al. "Cooperative phenomena in artificial solids made from silver quantum dots: the importance of classical coupling." In: *The Journal of Physical Chemistry B* 102.18 (1998), pp. 3425–3430 (cit. on p. 113).

[305] Constantine Yannouleas and Uzi Landman. "Spontaneous symmetry breaking in single and molecular quantum dots." In: *Physical review letters* 82.26 (1999), p. 5325 (cit. on p. 113).

[306] David J. Wales and Jonathan P. K. Doye. "Global Optimization by Basin-Hopping and the Lowest Energy Structures of Lennard-Jones Clusters Containing up to 110 Atoms." In: *The Journal of Physical Chemistry A* 101.28 (1997), pp. 5111–5116. eprint: `http://dx.doi.org/10.1021/jp970984n` (cit. on p. 114).

[307] David J. Wales, Mark A. Miller, and Tiffany R. Walsh. "Archetypal energy landscapes." English. In: *Nature* 394.6695 (Aug. 1998). Copyright - Copyright Macmillan Journals Ltd. Aug 20, 1998; Last updated - 2012-11-14; CODEN - NATUAS, pp. 758–760 (cit. on p. 114).

[308] Lívia B. Pártay, Albert P. Bartók, and Gábor Csányi. "Efficient Sampling of Atomic Configurational Spaces." In: *The Journal of Physical Chemistry B* 114.32 (2010), pp. 10502–10512 (cit. on p. 114).

[309] Jonathan PK Doye, Mark A Miller, and David J Wales. "The double-funnel energy landscape of the 38-atom Lennard-Jones cluster." In: *The Journal of Chemical Physics* 110.14 (1999), pp. 6896–6906 (cit. on p. 114).

[310] Paul J Steinhardt, David R Nelson, and Marco Ronchetti. "Bond-orientational order in liquids and glasses." In: *Physical Review B* 28 (1983), pp. 784–805 (cit. on p. 116).

[311] Daniele Moroni, Pieter Rein ten Wolde, and Peter G. Bolhuis. "Interplay between Structure and Size in a Critical Crystal Nucleus." In: *Physical Review Letters* 94 (23 2005), p. 235703 (cit. on p. 116).

[312] Benoit Coasne et al. "Freezing of argon in ordered and disordered porous carbon." In: *Physical Review B* 76 (8 2007), p. 085416 (cit. on p. 116).

[313] Shuji Ogata. "Monte Carlo simulation study of crystallization in rapidly supercooled one-component plasmas." In: *Physical Review A* 45 (2 1992), pp. 1122–1134 (cit. on p. 116).

[314] Oren M. Becker and Martin Karplus. "The topology of multidimensional potential energy surfaces: Theory and application to peptide structure and kinetics." In: *Journal of Chemical Physics* 106 (1997), p. 1495 (cit. on p. 116).

[315] P N Mortenson, D A Evans, and D J Wales. "Energy landscapes of model polyalanines." In: *Journal of Chemical Physics* 117 (2002), p. 1363 (cit. on p. 116).

[316] Michele Ceriotti, Gareth A. Tribello, and Michele Parrinello. "Demonstrating the Transferability and the Descriptive Power of Sketch-Map." In: *Journal of Chemical Theory and Computation* 9 (2013), pp. 1521–1532 (cit. on pp. 118, 119).

[317] Federico Comitani et al. "Mapping the conformational free energy of aspartic acid in the gas phase and in aqueous solution." In: *Journal of Chemical Physics* 146 (2017), p. 145102 (cit. on p. 119).

[318] Ronald R Coifman et al. "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps." In: *Proceedings of the National Academy of Sciences of the United States of America* 102.21 (2005), pp. 7426–7431 (cit. on p. 119).

[319] Ian T Jolliffe. "Principal Component Analysis and Factor Analysis." In: *Principal component analysis*. Springer, 1986, pp. 115–128 (cit. on p. 119).

[320] C Ramakrishan and GN Ramachandran. "Stereochemical criteria for polypeptide and protein chain conformation." In: *Biophys. J* 5 (1965), pp. 909–933 (cit. on p. 124).

[321] Dmitrij Frishman and Patrick Argos. "Knowledge-based protein secondary structure assignment." In: *Proteins: Structure, Function, and Bioinformatics* 23.4 (1995), pp. 566–579 (cit. on p. 124).

[322] Wolfgang Kabsch and Christian Sander. "Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features." In: *Biopolymers* 22.12 (1983), pp. 2577–2637 (cit. on p. 124).

[323] Scott A. Hollingsworth et al. "$(\phi, \psi)_2$ Motifs: A Purely Conformation-Based Fine-Grained Enumeration of Protein Parts at the Two-Residue Level." In: *Journal of Molecular Biology* 416.1 (2012), pp. 78–93 (cit. on p. 124).

[324] Gabor Nagy and Chris Oostenbrink. "Dihedral-Based Segment Identification and Classification of Biopolymers I: Proteins." In: *Journal of Chemical Information and Modeling* 54.1 (2014), pp. 266–277 (cit. on pp. 124, 126).

[325] Albert Ardevol et al. "Probing the Unfolded Configurations of a β-Hairpin Using Sketch-Map." In: *Journal of Chemical Theory and Computation* 11 (2015), pp. 1086–1093 (cit. on pp. 126, 127).

[326] Gopalasamudram Narayana Ramachandran, Chandrasekharan Ramakrishnan, and V Sasisekharan. "Stereochemistry of polypeptide chain configurations." In: *Journal of Molecular Biology* 7.1 (1963), pp. 95–99 (cit. on p. 126).

[327] Peter Tompa et al. "A million peptide motifs for the molecular biologist." In: *Molecular cell* 55.2 (2014), pp. 161–169 (cit. on p. 127).

[328] Véronique Receveur-Bréchot and Dominique Durand. "How random are intrinsically disordered proteins? A small angle scattering perspective." In: *Current Protein and Peptide Science* 13.1 (2012), pp. 55–75 (cit. on p. 127).

[329] Massimiliano Bonomi et al. "The Unfolded Ensemble and Folding Mechanism of the C-Terminal GB1 β-Hairpin." In: *J. Chem. Am. Soc.* 130 (2008), pp. 13938–13944 (cit. on p. 127).

[330] Jörg Behler. "Atom-centered symmetry functions for constructing high-dimensional neural network potentials." In: *The Journal of chemical physics* 134.7 (2011), p. 074106 (cit. on p. 132).

[331] Alexandros Altis et al. "Dihedral angle principal component analysis of molecular dynamics simulations." In: *The Journal of chemical physics* 126.24 (2007), p. 244111 (cit. on pp. 135, 136).

[332] Damien Laage and James T Hynes. "A molecular jump mechanism of water reorientation." In: *Science* 311 (2006), pp. 832–5 (cit. on pp. 136, 139).

[333] Omar Valsson and Michele Parrinello. "Variational Approach to Enhanced Sampling and Free Energy Calculations." In: *Physical Review Letters* 113 (2014), p. 090601 (cit. on p. 142).

CURRICULUM VITAE

Piero Gasparotto was born in Marostica, Italy, on the 11th of December 1987. After having received his diploma at ITIS G. Chilesotti (2006), he enrolled in Materials Science at the University of Padua. Here he obtained his B. Sc. in 2009, with a thesis on "Interactions among nanoparticles and biological systems", and his M. Sc. in 2012, with a thesis concerning "Effect of terminal oligo(ethylene glycol) on the structure of alkylthiol SAMs: a Molecular Dynamics investigation", both under the supervision of Prof. Alberta Ferrarini. Since October 2013 he enrolled in a doctorate at the department of Materials Science at the École Polytechnique Fédérale de Lausanne, supervised by Prof. Michele Ceriotti. His present research deals primarily with the use of machine-learning methods to assist the interpretation of atomic-scale simulation results.

LIST OF PUBLICATIONS

1. "Recognizing molecular patterns by machine learning: an agnostic structural definition of the hydrogen bond.", P. Gasparotto and M. Ceriotti, J. Chem. Phys. **141**, 174110 (2014)
2. "Probing Defects and Correlations in the Hydrogen-Bond Network of *ab Initio* Water.", P. Gasparotto, Ali A. Hassanali and M. Ceriotti, J. Chem. Theory Comput. **12**, 1953 (2016)
3. "Anharmonic and quantum fluctuations in molecular crystals: a first-principles study of the stability of paracetamol", M.Rossi, P. Gasparotto and M. Ceriotti, Phys. Rev. Lett. **117**, 115702 (2016)
4. "Recognizing Local and Global Structural Motifs At the Atomic Scale", P. Gasparotto, R.H. Meißner and M. Ceriotti, J. Chem. Theory Comput. **14**(2), 486 (2018)

ACKNOWLEDGEMENTS

There are many people I would like to acknowledge, and few lines are not enough to express my gratitude to them all.

First and foremost, I would like to thank my supervisor, Michele Ceriotti, for the opportunities he gave me and for the countless things he taught me. Besides being a first-class scientific mentor, he is an outstanding example of professionality. Having seen him building a group from scratch and accomplishing many challenges brilliantly, he will always be an inspiring example. I also want to say thanks to Alberta Ferrarini, because of the support she gave me since the time of my bachelor: it is in part thanks to her if I ended up writing a Ph.D. thesis in computational materials science. Many thanks to my family that, even if detached from my professional life, have always been close. Special thanks to my mother for loving me and being patient even when I am at my worst. To my brother and my sister, for supporting and inspiring me since the beginning. To my father, who gave the tools that made me stronger in life, as well as the world in which my first memories and feelings still live. Thanks to Evel for the inspiring discussions and the support he gave during the last ten years: I believe it is because of him that I decided to have a look to the world of machine learning and to never give up with research. Thanks to Santa, because even if we have been far away for a long time, I have always felt his friendship and anytime I find in him a great confidant. Thanks to Michelone: we started this long path together, far from being model students, and yet, look at us! We are among the few who did it! Thanks to the friends in Vicenza, especially Sorzatone and Zanella, to remember me how sweet is home and where my heart belongs. Thanks to Carol for the beautiful friendship we have, which started here, during our swiss years, around the clubs of Lausanne. Thanks to Baldi for being always ready to support me and to solve all my problems. To Anelli for having been a great flatmate, as well as a confident and a top-class roman friend. Daniele can't miss from this list: thanks for having been patient with me and having shared his time and his office with me for four long years. Thanks to Robert for having been an excellent collaborator and a friend. Thanks to all those who have helped me with my thesis and my work, especially David, for being kind and for proofreading all my texts: he did not read this little piece, so forgive me if it looks like an illiterate dog wrote it. Many thanks to all the people I met during these years for making my Swiss age sweeter. To Erica, for the unique, never-boring love stories she continually gives me.

I am not really good at this kind of things and, of course, many others should be present on this list. Anyway, even if their names do not appear here, be sure that I am not forgetting anyone of them and I will bring them with me in the next chapter of my life.

Thanks again.

P. G.