Politecnico
di Torino

ScuDo
Scuola di Dottorato ~ Doctoral School
WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation
Doctoral Program in Materials Science and Technology (35thcycle)

Machine learning of molecular motifs in soft supramolecular systems

By

Andrea Gardin

******

Supervisor(s):
Prof. G.M. Pavan, Supervisor

Doctoral Examination Committee:
Prof. Laio Alessandro, Referee, Scuola Internazionale Superiore di Studi Avanzati, Trieste, IT
Prof. Magdau Ioan-Bogdan, Referee, University of Cambridge, Cambridge, UK
Prof. Csányi Gábor, University of Cambridge, Cambridge, UK
Prof. Chiavazzo Eliodoro, Politecnico di Torino, Turin, IT
Prof. Salvalaglio Matteo, Univercity College London, London, UK

Politecnico di Torino
2023

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Andrea Gardin
2023

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

I would like to dedicate this thesis to the memory of my loving grandparents.

Acknowledgements

I would like to acknowledge the few people that truly supported me through these years, finding some time to hang out, laugh, and share experiences. Thank you!

Abstract

Defects are ubiquitous in our world and they forcefully populate every aspect of our life. Although, defects are often random and sporadic, making them challenging to detect and quantify, they can also provide valuable insight into properties and behavior of materials. Within the field of materials science, defects play a central role, driving a huge plethora of phenomena. However, to this day, we are still missing a comprehensive and general theory on how to systematically detect and learn how to control defects, especially in soft-matter systems. In modern days, following the improvements in several technological areas, we are acquiring new tools that can help scientists to better understand and rationalize elusive phenomena, such as the complex structural dynamic behavior of some soft materials. Recently, Machine Learning (ML) techniques are gaining a lot of popularity, as they can decompose virtually anything in mathematical entities, which can later be used as a baseline to formulate new theories, explain trends and/or draw data-driven conclusions. The work presented herein focuses on a variety of soft self-assembled systems that posses an intrinsic structural and dynamical complexity (*e.g.*, supramolecular fibres, micellar aggregates, bilayers and nanoparticles) and on how to compare, classify and control them. In this sense, a universal workflow is proposed and discussed, suitable for the complete structural description and comparison of various soft materials, focusing on the concept of structural defects as common ground. The work I have conducted builds on the concept that defects, which actively populate such assemblies, determine and influence their structural and dynamical features. However, detecting such entities in a unambiguous and unbiased way is not an easy task. Using machine learning based tools we developed a data-driven “defectometer”, an analysis tool, that can be employed to compare and characterize virtually all kind of supramolecular structures, based on their intrinsic “defectivity”. We believe that this represents a first step towards the rational design and engineering of new types of materials with controllable structural features.

Contents

List of Figures

Nomenclature

Roman Symbols

$\bar{\mathbf{p}}$	SOAP frame-average
$\langle\bar{\mathbf{p}}\rangle$	SOAP simulation-average
$\mathbf{C}(\tau)$	Time-lagged autocorrelation function
$\mathbf{p}_i$	SOAP power-spectra referred to atom *i*
$\mathbf{r}_i$	Cartesian coordinates of atom *i*
$\mathscr{U}$	Potential energy function
A_t	Instantaneous configuration of N atoms at time *t* of an MD simulation

Greek Symbols

α_i	Atomic label or index of atom *i*
$\boldsymbol{\gamma}_i$	Generic descriptor feature vector of atom *i*
$\boldsymbol{\Sigma}$	Covariance matrix
Γ_t	Instantaneous configuration of N atoms at time *t* of an MD simulation in the descriptor feature space
$\kappa(\cdot,\cdot)$	Kernel function

Acronyms / Abbreviations

AA	All-Atoms
AFM	Atomic Force Microscopy

BTA	1,3,5-benzenetricarboxamide
BTA	benzenetricarboxamides
BTT	benzotrithiophene
CD	Circular Dichroism
CG	Coarse-Grained
COG	Center of geometry
CV	Collective Variable
FES	Free Energy Surface
IC	Independent Component
LJ	Lennard-Jones potential
LSP	Living Supramolecular polymerisation
MC	Monte Carlo
MD	Molecular Dynamics
MetaD	Metadynamics
ML	Machine learning
MSM	Markov State Model
NDI	naphthalenediimide
NP	Nanoparticle
NPT	Isothermal-Isobaric ensamble
NVE	Micro-Canonical ensamble
NVT	Canonical ensamble
PAMM	Probabilistic Analysis of Molecular Motifs
PBC	Periodic Boundary Conditions

PC	Principal Component
PCA	Principal Component Analysis
PES	Potential Energy Surface
RDF	radial distribution functions
SA	Self-Assembly
SCA	Self-Closing Assembly
$SOAP$	Smooth Overlap of Atomic Positions
TEM	Transmission Electron Microscopy
$tICA$	Time-structure independent components analysis
$WT-MetaD$	Well-Tempered Metadynamics

Chapter 1

Introduction

1.1 Soft matter

Soft matter refers to a branch of condensed matter physics, which, as opposed to hard condensed matter, deals with molecular systems without a well-defined short- and long-range order (*e.g*, structural network or lattice), which in turn makes them easily deformed by thermal fluctuations and external forces. This definition covers a wide variety of molecular systems: from polymers to colloids, from liquid crystals to surfactants, from soap bubbles to macromolecules solutions. This enormous variety of materials, while representing a dominant part in the scientific research of physicists, chemists, and engineers, is also gaining increasing industrial importance. Many examples of the natural occurring soft matter molecular architectures originate from the intricate process of *self-assembly*: an automated process used by nature to "forge" complex and hierarchical aggregates exploiting the non-covalent interactions that exist between specific molecular sub-units (*i.e.*, monomers). All the successful strategies that nature adopted through years of evolution, are becoming the underlying inspiration to build functional soft nanostructures, with a so called "bottom-up" approach.[1]

In the following chapters and sections of this thesis work, we will focus on this kind of soft-materials, namely constituted by non-covalent spontaneous aggregation of monomeric species. All these soft systems share distinctive physical features, which infuse in them very general and appealing features, like innate dynamic character, stimuli responsiveness, dynamical adaptivity, and others. In these systems, the bonding energies between the monomers are of a similar scale of room temperature thermal fluctuations[2, 3], making the assembled structure prone to shape shifting and rearrangements. Moreover, the lack of a fixed structural lattice, and the fact that monomers grow into mesoscopic[1] size objects via self-assembly, make the physical behavior of soft matter very hard to predict and rationalize into universal theories.

In the reminder of this chapter we will go more into the details of the self-assembly mechanism underpinning the formation of soft materials, the hierarchical structures that can originate from it, and how the structural and topological features of the monomers impact, on a higher scale, the overall physical and chemical properties of such self-assembled materials.

1.2 Self-assembly, thermodynamics, and defects

The main phenomenon driving the creation of soft materials is the so-called *self-assembly* (SA) of freely dispersed molecular units, into more complex and hierarchical structures. The SA is considered both a physical and a chemical process, where a disordered isotropic distribution of building blocks self-organize in higher order structures, showing distinctive and specific patterns (Fig. 1.1).

The concept is rather general and it encloses a huge variety of different systems. Self-assembly can be divided in two major branches: (i) *static* or *equilibrium* SA, and (ii) *dynamic* or *out-of-equilibrium* SA.[4] Regardless of the specific underlying processes, certain patterns are always dictated by the intimate chemical features of the building blocks, and from the physical conditions in which the aggregation phenomenon occurs.[5–9] Most importantly, all the specific features that constitute the interacting monomers, participate in determining the resulting properties of the aggragate (*e.g.*, stimuli responsiveness, dynamical adaptivity), which could be appealing from both a scientific and technological point of view.[10–12] Ideal aggregation theory is well established for certain classes of self-assembled systems in specific model cases, where usually both the monomer shape and the interactions interactions considered have to obey crucial restrictions.[13, 14] However, a universal and generalised explanation for the observed self-assembling equilibrium behavior of soft matter systems is yet to be formalised.

[1] Size ranges between nanometers, for small clusters of molecules, up to materials measuring micrometres.

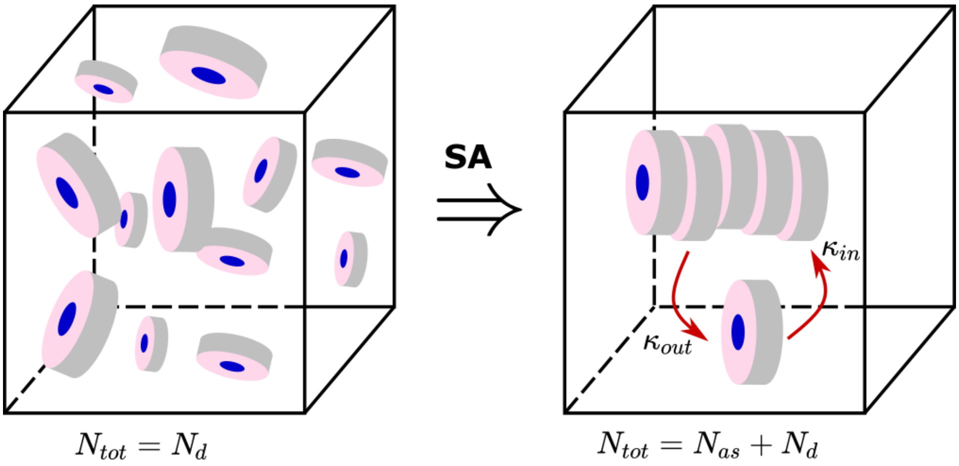


Fig. 1.1 Schematic visualization of a self assembly process. In the example depicted above, starting with an isotropic distribution of N_{fd} freely diffusing monomers (left side), the self assembly (SA) spontaneously occurs giving rise to a longitudinal 1-Dimensional ordered assembly (of N_{as} monomers) in thermodynamic equilibrium with the remaining free monomers (right side).

Self-assembly, as a natural phenomenon, has to obey the laws of thermodynamics. The first law states:

$$\Delta U = W + Q, \tag{1.1}$$

meaning that the internal energy transferred to a system ΔU, has to match the work done W on and the heat Q transferred to that system. The second law introduces the state function Entropy S, as a macroscopic property of the system, relating the transfer of energy as heat in an isolated system, as originally stated by Clausius. Entropy increases due to all spontaneous (irreversible) processes and it is maximised when the system reaches equilibrium.[15] The concept of entropy is often paired with a measure of (dis)order in a system: the higher the disorder in a system, and the higher is its entropy value. However, concepts like order and disorder are not accurate notions to quantify the changes of entropy, and often turn out to be misleading. From a microscopic point of view, Boltzmann introduced the concept of entropy as being correlated to the number of micro-states ω, in the phase space, that are accessible to the system in a specific thermodynamic state,[15]

$$S = k_B \log \omega + \text{const.}, \tag{1.2}$$

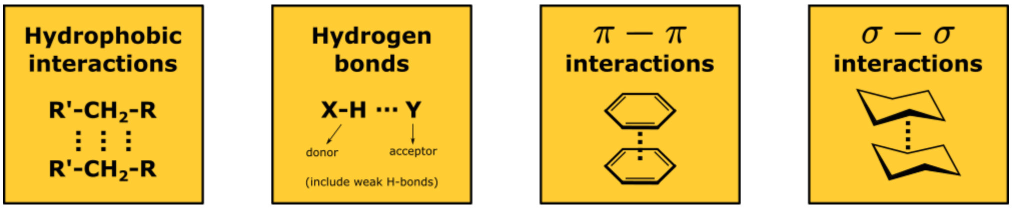


Fig. 1.2 Main intermolecular interactions classes which drive the self assembly process. Hydrophobic interactions exist between hydrocarbon chains or similar nonpolar regions of a large molecule. Hydrogen bonds form between atoms having a large difference in electronegativity. Most of the times, the donor is usually a Hydrogen atom and the acceptor is one amongst N, O, and halogens; a wide range of bond strengths are possible depending on solvent and the specific functional group combinations. $\pi-\pi$ and $\sigma-\sigma$ are weaker interactions compared to the previous ones, and originate from the interaction of the electronic orbitals of the $\pi-$bonds and $\sigma-$bonds respectively.

where k_B is the Boltzmann constant. On top of that, entropy can be further partitioned into different components, associated to different degrees of freedom, for example,

$$S = s_{trans} + s_{rot} + s_{vib}, \tag{1.3}$$

if one takes into account the translational, rotational, and vibrational contributions. These contributions might have different and often a crucial impact in the overall energy balance that drives the self-association process, as it will be explained briefly.

Following thermodynamics, the spontaneous self-assembly of monomers into soft-aggregate materials is achieved *via* the minimization of the free-energy state function, *i.e.*, the Helmoltz free energy: [2]

$$\Delta G = \Delta H - T\Delta S, \tag{1.4}$$

acquired either by decreasing the internal energy or increasing the entropy contributions.

During the SA, molecules are expected to go from a disordered isotropic (with no preferred spatial orientation) state of freely dispersed components to a more ordered one (Fig. 1.1). Intuitively, this process should be unfavorable from an entropic point of view, as it "forces" a certain degree of spatial ordering in the molecular arrangement. Consequently, the system needs a sufficiently strong gain in internal

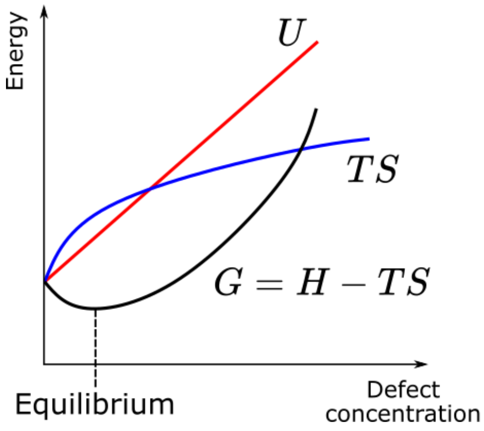
[2] Depending on the operating state conditions, these statements also holds for the Helmoltz free energy ($\Delta F = \Delta U - T\Delta S$).

energy, that has to compensate for the loss in entropy, and to make the whole process occur spontaneously. This contribution can come from the specific chemistry, as well as the physics, of the entire system, like *e.g*, weak and non-covalent intermolecular interactions (Fig. 1.2).

A straightforward example of a system that undergo spontaneous changes towards a more ordered phase is when a fluid phase crystallises into an ordered crystal lattice. This happens since the molecule in the crystal phase are in thermal contact with its environment: while some molecules are freezing, the heat that is released in the surroundings "pays" for the decrease in entropy of the crystallization. However, the process is not always energy-driven, as many times a phase transitions towards molecular order is in fact entropy-driven. In these cases, the state function unintuitively increases with the increase in molecular order, making the concept of entropy central to gain insights on the microscopic mechanisms that define self assembly phenomena.[16–18] The first historical example comes from the work of Onsager on fluids system of hard thin rods.[19] In his theory, Onsager describes the transition from an isotropic (orientationally disordered) fluid phase to a nematic[3] (orientationally ordered) phase. The transition is demonstrated to be completely entropy driven, with the ordered phase having effectively higher entropy values. In fact, under certain physical conditions,[19, 20] orientationally disordered rods can increase translational entropy by becoming more aligned, at the expense of some orientational entropy, making the overall balance in Eq. 1.3 favorable for the phase transition to happen. Similar arguments can be made for fluids made out of simple hard spheres or other similar colloidal objects, which can self assemble into a wide range of crystals, liquid crystals and quasicrystals phases, depending on the physical properties of the colloidal building blocks used.[5, 17, 18, 21, 22]

These examples are simple cases for which the SA phenomenon is either entropy or enthalpy driven. In general, soft-assembled materials are mostly constituted by flexible building blocks, which interact through complex interactions: hydrophobic effects, electrostatic interactions, and other non-covalent contributions (*e.g.* hydrogen bonds, $\pi - \pi$ interactions, and $\sigma - \sigma$ interactions, Fig. 1.2). Notably, these interactions determine the energetic (enthalpic) character of SA, as the minimisation of repulsive and maximisation of attractive molecular interactions overcome the entropy price.[3, 4] For example, the hydrophobic effect is the principle behind

[3] In a nematic phase, the rods are, on average, aligned parallel to each other, but with a spread in their orientations around the average alignment direction.



$$\Delta F = F_{def} - F_{perf}$$
$$\Delta H = H_{def} - H_{perf} > 0$$
$$\Delta S = S_{def} - S_{perf} > 0$$

Fig. 1.3 Example of energetic balance for a generic condensed lattice system. The complete absence of defects is possible only for the ideal system at the absolute zero. Once the temperature rises above zero, a certain concentration of defects is always expected from the energy balance.

the tendency of hydrocarbons (*e.g.* oils) not to mix with water molecules.[23] As a consequence, water molecules in the proximity of an hydrophobic surface form entropically unfavourable ordered structures (clathrates). Whereas, apolar[4] carbons chains, dispersed in a polar solvent (*e.g.*, water), tend to aggregate into spherical droplets limiting as much as possible their surface contact with the unfavorable solvent. Amphiphilic molecules, *i.e.* building blocks that posses both polar and apolar parts, in aqueous solutions tend to aggregate into ordered structures (micelles, bilayers, vescicles) as well, minimizing the contact with water of their apolar section.

In the paragraphs above we briefly rationalised how the formation of these self-assembled structures obeys the general laws of Thermodynamics, and how the resulting aggregates can originate from subtle energy balances. In most cases, a free-energy minimum is reached as a result of the balancing of intermolecular forces, although specific structural patterns emerge as a trade-off of thermodynamics and kinetics factors during the assembly process. The presence of defects in a molecular structure is a concept that is well known, as defects statistically appear in every system with a temperature greater than the absolute zero.[24] Even for hard condensed matter, punctual defects in a crystal lattice are expected from simple thermodynamic considerations (Fig. 1.3). The minimum in the free energy corresponds to a given concentration of defects, which gradually increases with the temperature.[24] More in general, the appearance of disorder/defects in a material, can emerge in two main

[4] Hydrophobic and apolar are used as synonyms in this context.

ways: from *intrinsic* and *extrinsic* sources.[24] Intrinsic disorder is influenced by the entropy of a system and the overall energetic balance of defects formation (Fig. 1.3). Extrinsic disorder arises from kinetically driven phenomena, (*e.g.*, mass transport mechanisms), chemical contamination, size irregularities, substrate effects, and others, related to the specific environment in which the self assembly takes place. The two kinds of phenomena affect both hard- and soft-condensed matter, and they often can be exploited to engineer in the material interesting properties.

In the field of soft-assembled materials, the concept of (structural) defects has not yet been thoroughly explored, mainly due to the dynamic nature of the supramolecular interactions (Fig. 1.2), which keep monomers "glued" in the assembly, making the structural characterization a considerably difficult task.[25] Soft-assembled structures, such as supramolecular fibers, tubules, micelles, and many others, are often idealised as perfect, average static structures, although they possess a very relevant dynamic and disordered character. Defects highly populate such hierarchical supramolecular structures, having a complex and crucial impact on their properties and behavior. Using advanced molecular simulations it is possible to effectively reproduce, and follow in time, the molecular behavior of many soft-matter aggregates, at very high resolution (submoleculer, < 5 Å), highlighting the presence and key role of structural defects.[26–29] Moreover, combining molecular simulations with Machine Learning (ML) approaches it is now possible to fully characterise the internal structure and dynamics of complex supramolecular polymers, such as those formed by the 1,3,5-benzenetricarboxamide (BTA) building blocks.[30, 31] Exploiting an agnostic description of atomic environments, dimensionality reduction, and clustering algorithms (these topics will be discussed in detail through Chapter 2, Sec. 2.2 and 2.3) Ref. [30] reports a complete description of how defects originate and evolve along the fiber "backbone", also with their qualitative mechanisms and pathways, paving the way for a defects based characterization of soft materials.

1.3 Self-assembly of soft supramolecular systems

The majority of the studied self-assembled structures is characterised by a SA process where the aggregation of monomers reaches an equilibrium state of a finite-size superstructure, *i.e.*, with a well defined spatial growth in one or more dimensions.[32] The finite-size originates when the entropic penalty of a monomer that joins the

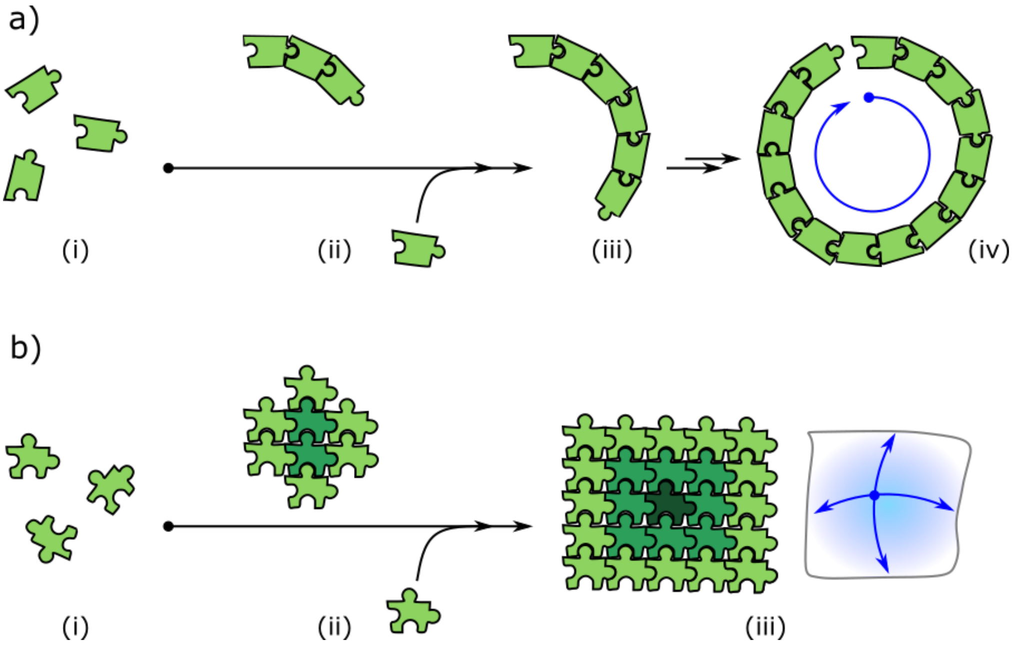


Fig. 1.4 Illustration of two examples of SA growth pathways. They both follow the same evolution toward the formation of the assembly: (i) starting from free monomers dispersed in a solvent, the process of SA (ii) begins and the building blocks (depicted as jigsaw puzzle pieces) aggregate following their specific interaction field. The aggregation continues (iii, iv) till the equilibrium structure is reached. a) "Self-closing" assembly aggregation, where the peculiar shape of the subunits allow the formation of at least one periodic direction of assembly. b) "Open-boundary" assembly aggregation, where subunits aggregate toward a pseudo unlimited size (still regulated by the balance of energies that participate in the growth phase).

assembly becomes unfavorable with respect to the internal energy gain given by the cohesive forces in play. Thus, the supramolecular aggregate grows continuously drawing units from the "reservoir" of freely dissociated ones, until it reaches its equilibrium mesoscopic/macroscopic size.

Many examples of these superstructures can be found in biological systems, where they perform a lot of delicate and vital tasks: selective encapsulation and transport,[33–35] optical response,[36, 37] and modular stiffness and strength[38, 39].

The ability of monomers to reversibly associate in one specific and thermodynamically defined state, is very appealing towards the synthesis of materials, because it does not require any supervision, since all the information is already "encoded" in the monomer features, and the aggregation happens spontaneously. In contrast, most common methodologies for synthetic materials (*e.g.*, hard-condensed materials) result in unlimited growth mechanisms, giving rise to crystalline or liquid-crystalline macroscopic phases, where the size is not controlled by the underlying thermodynamic process, but only by the extent of the reservoir (*i.e.*, availability of atomic/monomer).

In the previous section, we introduced the concepts of self assembly, going over the general thermodynamics principles behind the phenomenon. In the following, we will analyse more in details the requirements for certain structural, or topologies, to emerge as outcome of the SA process, and briefly review examples of possible natural occurring supramolecular architectures.

Monomers aggregation can be split into two general additional macro-classes (Fig. 1.4), self-closing assemblies (SCA) and open-boundary assemblies (OBA).[32] Self-closing assembly (Fig. 1.4a) refers to an aggregation that achieves a finite target size due to anisotropic binding of neighbouring building blocks, characterised by a tapered or wedge shape (*e.g.* banana-shaped or "sharp" cone-shaped) and/or by a specific non isotropic intermolecular interaction potential. The cohesive bonding leads to specific shape constraints, determining the periodic directions in which the blocks assemble. A simple example is reported in Fig. 1.4a, where a slightly banana-shaped building block assembles in a 1-D circular fashion, that can eventually close itself (*i.e.*, 1-D periodicity). All the self-assembled architectures that close upon themselves in any or all spatial directions, achieving a finite numbers of sub-units, belong to the SCA category.[32] Examples can be found in several natural occurring biological structures, and correspond to architectures that exhibit a spherical closed cover, like shells, capsules and micellar aggregates, or that have one direction of unlimited growth and another direction of well-defined periodicity, like tubules and other amphiphiles aggregates. In Ref. [32] the authors thoroughly go over all these examples, discussing thermodynamic trends and energetic balances that drive the formation and the growth of diverse architectures (these details are outside the scope of this thesis and are not discussed herein).

The second main class of self assembled materials is the self-limited with open-boundaries assemblies (OBA). These are characterised and differ from the SCA by having an open boundary, a surface that separates the inner part of the aggregate by the rest of the system (these arguments are also treated more in detail in Ref. [32]). The equilibrium OBA structure (*i.e.*, the self limiting equilibrium size of the assem-

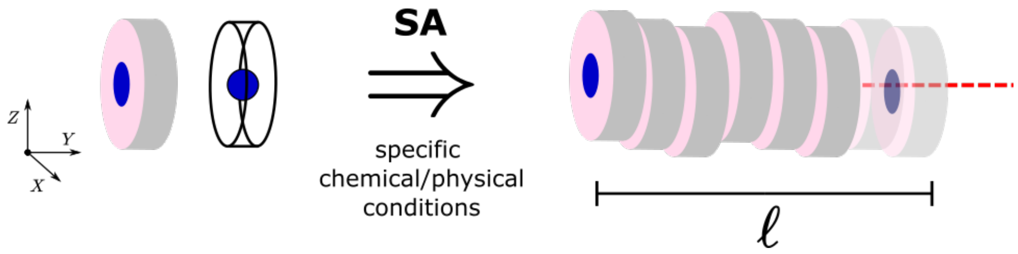


Fig. 1.5 Simplistic visualization of a molecular subunit that polymerizes along a straight line, creating an aggregate of length ℓ, proportional to the number of connected monomers.

bly) will be defined by energy balances, which depend on the geometric properties of the aggregate, *e.g.* dimensionality of the unlimited growth *versus* the limited one, and the chemical properties of monomers and solvents. A few examples of OBA are protein complexes[40], charged nanoparticles[41] and colloidal particles in low dielectric solvents[42], where the final structural assembly is driven by a balance between electrostatic long-range and short-ranged dispersion interactions of the system components. Additional factors that can drive the equilibrium shape of OBA aggregates are a global geometric frustration, or frustration gradients along the materials, caused by chiral elements in their structures. These materials are denoted as geometrically frustrated assemblies and include self-twisting protein bundles[43], chiral smectics and membranes[44–47].

In the next paragraphs we will present some real examples of N-Dimensional supramolecular assemblies, discussing few key features, along with biological and possible real or synthetic uses. We will mainly focus on supramolecular polymers, which are the types that later will be also discussed in detail the results chapters (Chapter 3 to 5).

1.3.1 1-Dimensional supramolecular assemblies

The 1-D assemblies category includes all architectures that are based on a unidirectional stacking of building blocks (Fig. 1.5); these are often referred to as *supramolecular* polymers (or fibers).[48, 49] The resulting equilibrium structure can be quite heterogeneous since it is strongly affected from the intimate features of the monomers: topology, chemical identity (*e.g.*, specific functional groups, polarity, and electrostatics), solvent, and other species present in the system (Fig. 1.6). This is one of the most studied classes of soft-assembled materials due to the simplicity of the unidirectional self-assembly, although thanks to the highly editable nature of both building blocks and aggregation pathways, the technological application are seemingly infinite.[49] Supramolecular polymers fundamentally differ from conventional polymers as the main binding forces originate from reversible and highly directional secondary interactions (Fig. 1.2). The directionality and strength of the interactions can be precisely tuned so that the resulting fiber maintains its polymeric properties in solution, *i.e.*, it still behave in a way that can be described by the ordinary theories of polymer physics. However, as already mentioned in the introductory part of this section, the high reversibility of the non-covalent nature of the bonds make sure that supramolecular polymers are typically formed under conditions of thermodynamic equilibrium, where concentration of the sub-units, temperature, and absolute strength of the bonding interactions control length and shape of the chains. Figure 1.6 reports an interesting example of two supramolecular polymers, based on the 1,3,5-benzenetricarboxamide (BTA) monomer scaffold, that exhibit drastically different architectures when the side arms of the monomer and the solvent are changed, hence, changing the strength and overall balance of the non covalent interactions between the self-assembling monomers.[26, 30, 50] The polar soluble variant (Fig. 1.6b) is found to possess a higher degree of structural complexity and a consequently higher dynamic nature, compared to its counterpart soluble in organic solvent, which is rather inert in those conditions (this relationship will be discussed in the following sections).[28, 30, 50]

The synthetic design of a polymer that possess a determined range of desired properties is of great relevance.[51] The first step towards the rational design of supramolecular synthetic materials can be traced back 20-25 years ago, where a monomer containing two 2-ureido-4[1H]-pyrimidinone (UPy) bases connected by hydrogen bonds, was polymerised into an interlocked network structure.[52] The UPy base could be further decorated with a variety of functional organic moieties, thus changing drastically the bulk properties of the resulting polymer. Following an increase in interest a lot of different monomer scaffolds and synthetic strategies to assemble supramolecular polymers have been developed during the years.[48] One of the most powerful strategies of structural engineering comes from the creation of supramolecular *co-polymers*, *i.e.*, polymers that present some sort of alternating pattern of multiple different monomers in their structure. Given two monomer types, indexed A and B, the different ways they can be mixed together are: (i) *alternating* sequence co-polymers $-[A-B]_q-$, where different monomer species form com-

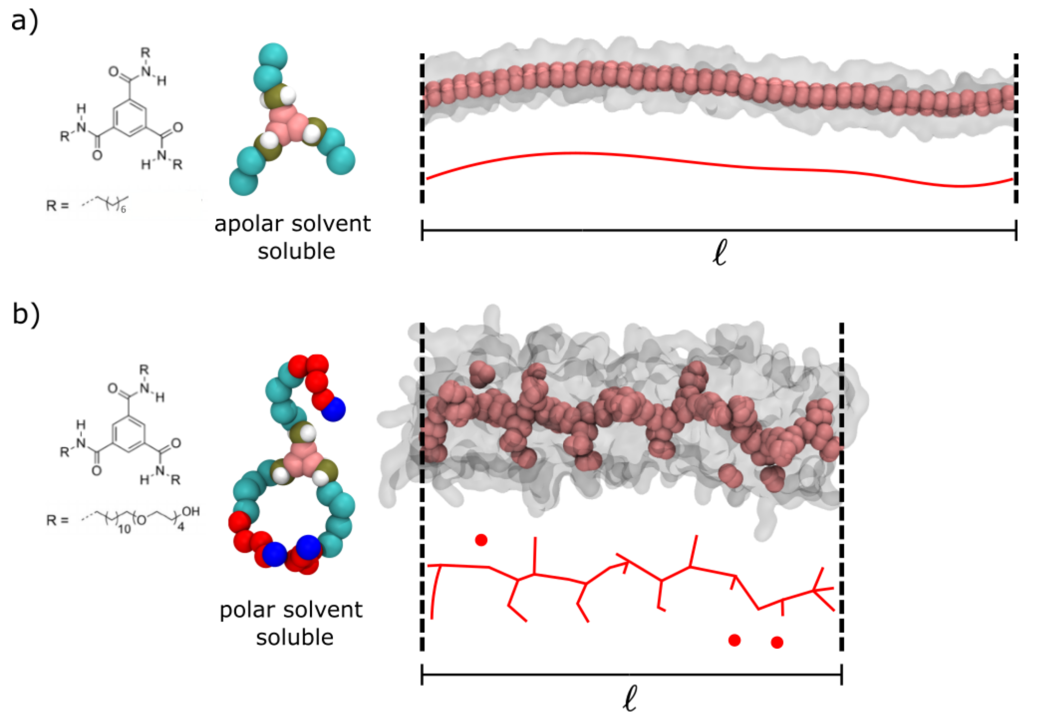


Fig. 1.6 Real examples of a supramolecular polymer based on the 1,3,5-benzenetricarboxamide (BTA) monomer scaffold.[50] a) Example of a organic solvent soluble BTA monomer that is able to stack in a seemingly perfect supramolecular fiber. b) When considering a mutated version of the same BTA base, water soluble, the obtained equilibrium aggregate visually appears twisted and contorted (the assembly length is reduced). To clearly visualize the assembled structures only the central pink beads are showed in the image and all the others are transparent grey. The red lines below, trace the "backbone" of the aggregate, and the dashed black lines represent the periodic boundaries of the simulation box, thus considering supramolecular fibers as infinite in the simulation.

plementary bonds only between each other and the pattern is repeated q times. (ii) *Block* co-polymers $-[(A)_n-(B)_m]_q-$, where the triplet of numbers, n,m,q, is fixed and repeating. (iii) *Random* sequence co-polymers $-[(A)_n-(B)_m]_q-$, where the numbers, n,m,q, vary in such a way that A and B occur randomly. The two latter examples occur if the sub-units are self-complementary, *i.e.*, they bind to their own kind, but are also able to bind to other kinds of sub-units. The random case is the most common, whereas a block structure arises if the $A-B$ interaction is considerably weaker than the $A-A$ and $B-B$. (iv) *Blend* polymers, if different monomers cannot physically mix, the polymer forms as a blend of chains each of one containing a single monomer species. These are often referred to as "narcisistically self-sorted" polymers, due to their peculiar mixing nature.[53] In principle, these mixing strate-

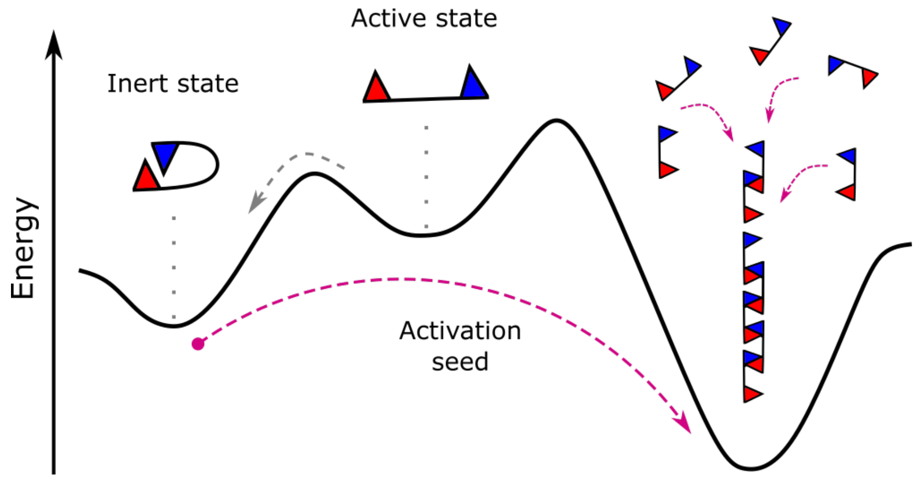


Fig. 1.7 Illustration of a general LSP pathway complexity. The monomer has two main conformational states, inert and active. In normal conditions, the former would be the the more stable configuration. The addition of an activation seed serves as catalyst, stabilizing the active state and driving the polymerization. This overall mechanism is often referred as off-pathway polymerization, as it requires a “detour” from the most stable thermodynamical pathway.

gies allow for a fine tuning of the temperature dependent properties of the polymer: by selecting the appropriate monomers one can tune the stress relaxation along the chains (*i.e.*, rigidity of the chain), the melting temperature, and also other stimuli responsive features (*e.g.*, self healing, specific light wavelenghts sensitivity) that are crucial from a technical point of view.[49]

Natural occurring supramolecular polymers includes examples where the subunits are represented by proteins that self-assemble into fibroidal structures, such as, *e.g.*, the G-actin and Tubulin, in actin fibres and microtubules.[54, 55] However, these fiber-like structures are typically rather short to be classified as polymers and the actual polymerization mechanism takes place out of (thermodynamic) equilibrium. On the other hand, several strategies to synthesise supramolecular polymers have been developed and explored, such as: (i) step-growth, (ii) chain-growth, (iii) ring-opening, (iv) enzymatic polymerization, (v) biocatalytic polymerization.[56] One of the main issues with synthetic polymer growth, both ordinary (covalent) and supramolecular, is the control over the outcome polymer structure. Pollutants or impurities present in the synthesis environment may drastically affect the polymerisation conditions, altering the final results. In nature these strategies have evolved throughout the years in order to maximize the yield of specific structures, often involving catalytic reactions and irreversible consumption of “fuel” molecules. A typical example of these types of polymerization strategies is reported in Figure1.7. Initially, the equilibrium state of the building block makes it inert towards polymerization. In order to initiate the process, the monomer needs to be activated, usually *via* conformational changes or small chemical reactions. Once initiated, the polymerization process runs until the activation conditions hold, hence providing an extra layer of control to the whole polymerization mechanism. The artificial counterpart of this phenomenon is called *living* supramolecular polymerization (LSP), a wide terminology that spans many concepts referring to the pathway complexity of this activated phenomena.[57, 58] Takeuchi and Sugiyasu proposed one of the first examples of LSP[59], in which a porphyrin derivate aggregated into metastable nanoparticles that could be converted to thermodynamically more stable nanofibers by the addition of reactant seeds and the subsequent kinetically controlled chain-growth. A lot of applications arose right away, exploiting a wide variety of functional building blocks,[60–62] emphasizing the flexibility and the versatility of this synthetic strategy toward diverse implementation, such as, in solar cell manufacturing[63], nanophotonics[64] and biomedicine[65].

1.3.2 2-D and 3-D supramolecular assemblies

The category of the 2-D and 3-D soft aggregates includes all self-assembled materials that form respectively two or three dimensional architectures. The former includes principally all the micelles and membranes based structures (Fig. 1.8a,b), where the 2-D surface can be wrapped around different kinds of shapes, sphere, cylinders, and parallel planes. The building blocks are characterised by an amphiphilic nature, a polarised symmetry of interaction with the solvent and with other sub-units, forcing an intrinsic order of aggregation where all the solvophilic ends point towards the solvent and the solvophobic ones away from it, so that, the outer part of the aggregate is populated only by solvophilic moieties. The resulting structural complexity of the 2-D soft aggregate is highly dependent on the monomer *packing parameter*, which is widely invoked in the literature to explain, rationalize and even predict self-assembly in surfactant solutions. This parameter is defined as the ratio $V_{tail}/\ell_0 A_{mol}$, where V_{tail} is the tail volume of the monomer, ℓ_0 is the tail length, and A_{mol} is the equilibrium area occupied by each monomers in the aggregate surface.[66] A determined value of the packing parameter can often be directly translated into a specific shape and

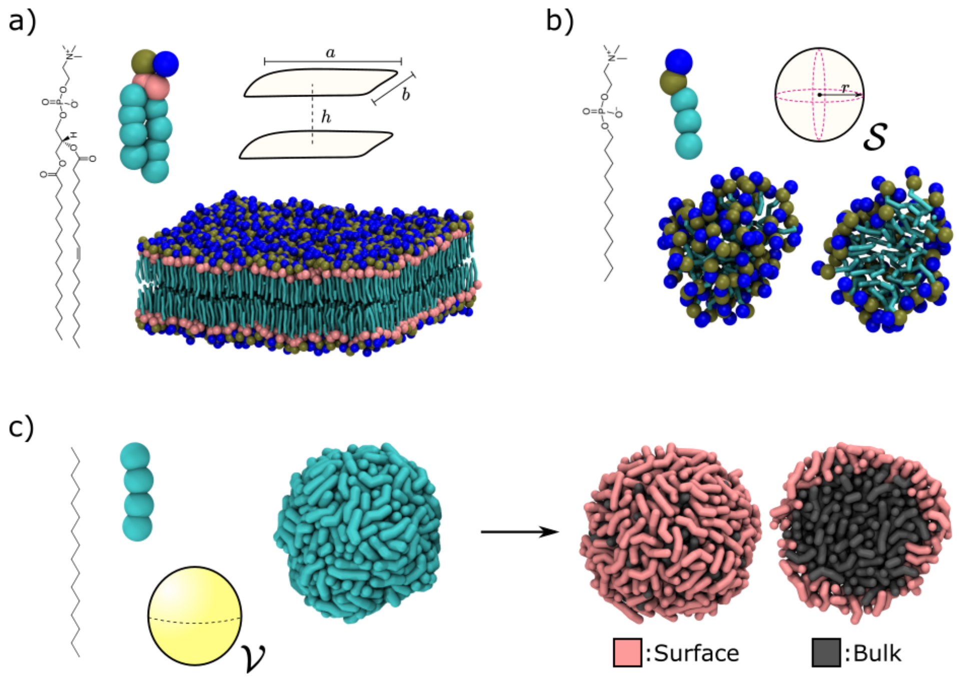


Fig. 1.8 Examples of 2-D and 3-D soft-assembled architectures. a) Dipalmitoylphosphatidylcholine (DPPC) lipid molecule: chemical formula and Coarse-Grained (CG, see Sec. 2.2.1 for more details) representation of the single molecule and bilayer sheet section, with geometrical sketch of its topology, commonly found in lipid bilayers. b) Dodecylphosphocholine (DPC) surfactant molecule: chemical formula and CG representation of the single molecule and micelle structure, with geometrical sketch of its topology, the surface of a spheare. The monomer is very similar to the DPPC one, but having just one apolar tail makes bilayer structure unstable, favoring a micelle arrangement. c) Hexadecane (Hex) alkane hydrocarbon molecule: chemical formula and CG representation of the single molecule and aggregate, with with geometrical sketch of its topology. The complete apolar molecule in presence of a polar solvent (*e.g.*, water) is insoluble, and it creates globular aggregates with monomers occupying the entire surface and volume (dark shade of yellow).

size of the equilibrium aggregate. Figure 1.8a,b reports an example of the effect of two different packing parameters: the two molecules are very similar but one (the DPPC molecule, panel a) crucially differ from the other (the DPC molecule, panel b) having two apolar tails, thus significantly increasing the tail volume V_{tail} (as well as the total area A_{mol}), causing the two equilibrium structures to be very different.

In contrast, 3-D soft aggregates are typically characterised by building blocks that interact isotropically with themselves and the solvent molecules. The easiest example, of such category, would be a long apolar organic molecule (*i.e.*, an oil, like in Fig. 1.8c) forming an emulsion in a polar solvent (for example, water): the oil molecules form spherical droplets, as this shape minimizes the molecular contact with the solvent.

As discussed in the sections above, the past few decades saw remarkable advancements in the field of soft-assembled materials, with an interest in synthetically reproducing biologically relevant architectures, in particular for those involved in the function and structure of cells and organelles.[67, 68] A biological cell is a container, which is kept separated from the outside environment by amphiphilic membrane walls, which act as both shield for toxic and preferred canals for nutrients compounds. Moreover, a cell organism needs to respond to stimuli, to produce and broadcast signals that are essential to maintain life's physiological activities, features that are controlled by spatial complexity (*e.g.*, presence of multiple complex peptides or proteins inside the lipid layers) and temporal regulation (*e.g.*, thickness and fluidity). Artificial membranes could be exploited as building blocks for potential medicine applications, especially structures like liposomes, DQAsomes (liposome-like vesicles containing dequalinium chloride), and polymeric membranes, which can be thought as functional "nanocarries" for drug delivery.[67]

In order to design synthetic functional materials and/or gain deeper insight into the intimate mechanisms of the natural counter parts, we first require reliable and accurate strategies for the analysis of the aggregate structures and dynamics. In the next section we will address this issue, which represents a hot topic of research, in both computational and experimental fields.

1.4 Emergence of structural defects and their relevance in dynamics

In the previous sections we briefly discussed how some topological traits, of the self-assembling monomers, have a strong impact on the supramolecular hierarchy yielded by a the aggregation process. Depending on the chemical or physical features, and also on the surrounding environment, a sample of monomers can spontaneously assemble into a linear fibre (1-D aggregate), a membrane/micelle (2-D aggregate), a droplet (3-D aggregate), or even a combination of them. However, the "influence" carried by the building blocks is not limited to the dimensionality of the overall supramolecular architecture, but, most importantly, to the presence and concen-

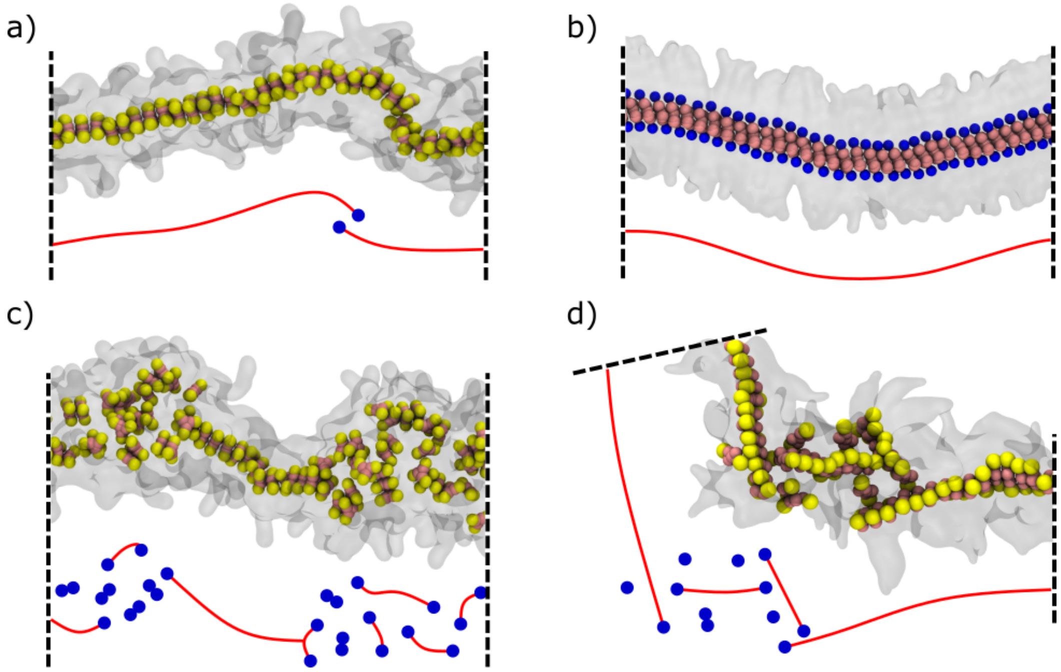


Fig. 1.9 Examples of structural motifs in supramolecular fibres; a,c) BTT based fibres and b,d) NDI based fibres. The ordering of the panels highlights how it is possible to achieve a drastic difference in structural motifs while still remaining in the same fibre family.

tration of characteristic structural motifs. In section1.3, we reported two example of differently functionalised BTA monomers that lead to divergent supramolecular architectures, one closely resembling a perfect 1-D stack of monomers (Fig. 1.6a) and the other having numerous branches growing from the internal fibre backbone (Fig. 1.6b). Figure 1.9 reports four interesting additional examples of structures based on two different monomer types: the benzotrithiophene (BTT)[69] (Figure 1.9a,c) and the core-substituted naphthalenediimide (NDI)[70] (Figure 1.9b,d) monomers. The monomers in both sets, while retaining their general chemical identity from the top panels to the bottom one in Fig. 1.9, they exhibit slightly different chemical character. Notably, these deviations are enough to imprint peculiar structural traits into the equilibrium structures of the respective fibres (more detailed information about these two families of fibres and their intrinsic structural motifs will be discussed in Chapter 4). The deviations from the ideal 1-D arrangement of the sub-units of a given supramolecular polymer define the so-called structural motifs or structural defects (Fig.1.9).

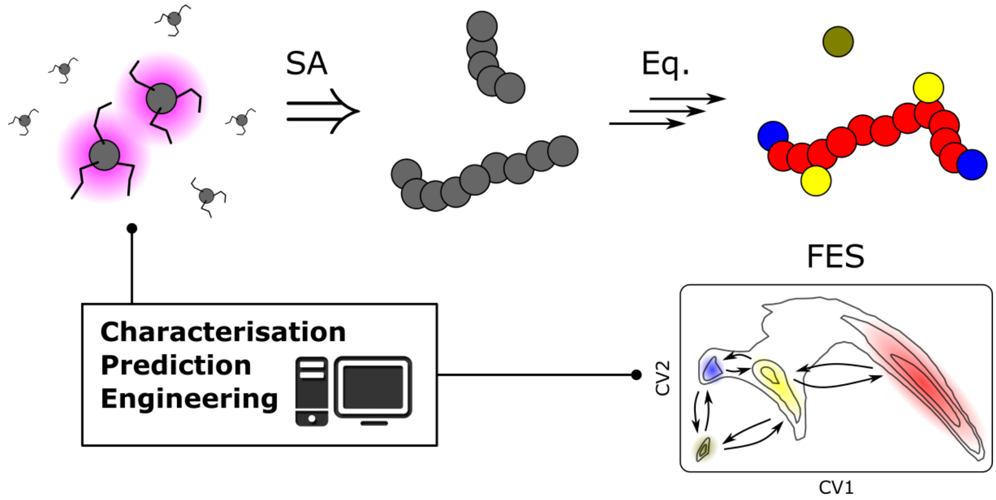


Fig. 1.10 The role of monomers identity during the SA process: the single-monomer will directly influence the features of the equilibrium aggregate. Learning how to link the two will greatly enhance the engineering possibilities of functional materials.

The types and concentration of structural motifs (or defect) represent key equilibrium properties of a given supramolecular aggregate as they are originated by complex energy balances between the monomers with themselves and the external environment (Fig. 1.10). Given the non-covalent nature of these materials, defects are not to be regarded as some static or fixed properties, but they are, in fact, in constant dynamic equilibrium with the rest of the structure.[26, 50] Recent studies[26, 29, 31, 71] reported that the existence of defects is actually invaluable and desirable, since they are the key precursors needed to open pathways of exchange between monomers, which in turn define specific macroscopic function of a soft material (*e.g.*, stimuli responsiveness, self-healing, elasticity, etc.). However, to this day, there is no universal analysis tool available to researchers to thoroughly identify, characterise, and compare structural motifs in soft materials and, most importantly, to link the presence of a particular structural motifs to concrete physical/chemical features of the building blocks that constitute the aggregate under investigation. In our opinion, such analysis tools, would be of paramount importance, as it would pave the way toward a more targeted engineering of functional materials. Figure 1.10 nicely sketches the main motives behind this research line: a starting sample of monomers will spontaneously interact (self-assembly) following their unique interaction field, eventually reaching an equilibrium aggregate structure. The emerging supramolecular structure will also show distinctive equilibrium properties, which are directly influenced by the resulting interaction field of the assembled monomers and the types of structural motifs that populate the structure. Ultimately, this suggests that by directly acting on the monomer identity it is possible to induce or influence the equilibrium properties of the final aggregate. In order to achieve this level of control we require two main tools: (i) a first tool capable of detect and classify, with high accuracy, the elusive structural motifs that populate the structure of self-assembled materials, and (ii) a second tool capable of encoding and trace back these motifs to some physical variables of the monomer structure and/or interaction field.

In this context, the contents of this thesis have been primarily motivated by the challenges posed by the difficulty of finding a general and agnostic way to represent, classify and compare the elusive molecular motifs that emerge in the complex structural behavior of supramolecular materials in equilibrium conditions. We will rely on state-of-the-art computational techniques to build, simulate, and validate many molecular models for different supramolecular structures. This will allow us to study soft-assembled material with a very high resolution, impossible to reach with other experimental analysis, however not without downsides, which will be briefly discussed in the following section.

1.5 Experimental limits and the role of computer simulations

Studying soft-assembled materials can be challenging due to their complex and often dynamic nature, although it is precisely because of such particular nature that these kind of materials gained so much attention recently. Most experimental limitations and issues in characterizing soft materials arise from a few main concepts. (i) Complex behavior: soft matter often exhibit a behavior that is difficult to predict or control, due to its distinct thermodynamic regime. For example, supramolecular polymers can exhibit non-trivial dynamic behaviors such as viscoelasticity, shear thinning, and non-Newtonian flow. (ii) Dynamic behavior: directly connected to the previous point, many materials exhibit a behavior that is difficult to follow experimentally, due to the rapid time-scales involved. For example, in membranes, the constant motion and mutation of the lipid layer allows for complex diffusion of proteins and other chemical compounds which is very challenging to follow. (iii) Sample size: keys structural properties of soft materials often originate from aggregates in the microscopic or nanoscopic scale, bringing technical challenges to the experimental sample collection and imaging/visualisation. A notable example is the ionic channels and other delivery mechanisms, usually residing within the thickness of a lipid bilayer, which is only $\sim 7-8$ nanometers in average.[68]

Until the advent of computer era, scientists could only rely on experimental techniques to probe the properties of materials. The most common techniques can be summarised by the following macro-areas: thermal analysis, mechanical testing, rheology, microscopy, and spectroscopy.[72] The first three are used to study macroscopic properties, thus dealing with macroscopic samples, not at the molecular scale. Thermal analysis techniques, such as differential scanning calorimetry and thermogravimetric analysis, allow the study of thermal properties, such as melting point and boiling point determination and phase transitions. Mechanical testing techniques, tensile testing and compression testing, allow to study mechanical properties, such as strength and stiffness, and deformation and failure under load. Finally, rheology is the study of the flow and deformation of materials, which allows the study of viscoelastic properties.

Microscopy and spectroscopy are higher resolution methods, employed to visualize the micro-structure of materials down to molecular resolution, and even to study their vibrational and electronic properties. However, the set up for these high-resolution experiments is often critical: low temperatures and pressure, and limited sizes are required.[72] These techniques are divided into single molecule techniques: circular dichroism, nuclear magnetic resonance, infrared spectroscopy, and fluorescence, and techniques to infer the morphology of material samples, like, atomic force microscopy, electronic microscopy (scanning and transmission) and X-ray diffraction.

Computer simulations, such as Molecular Dynamics (MD), brought an immense contribution on the theoretical study of molecular systems, as they grant complete control over the parameters and conditions over which the simulation runs, *e.g.*, temperature, pressure, intra- and inter-molecule interactions. Specifically, MD techniques allow us to follow the temporal evolution of every single atom of the physical system under investigation.[73] MD involves solving the classical equations of motion (a more theoretical discussion on the foundations of the method will be given in the next chapter, Chapter 2) for a system of interacting particles, and the motion of each particle is determined by the forces acting on it, which can be

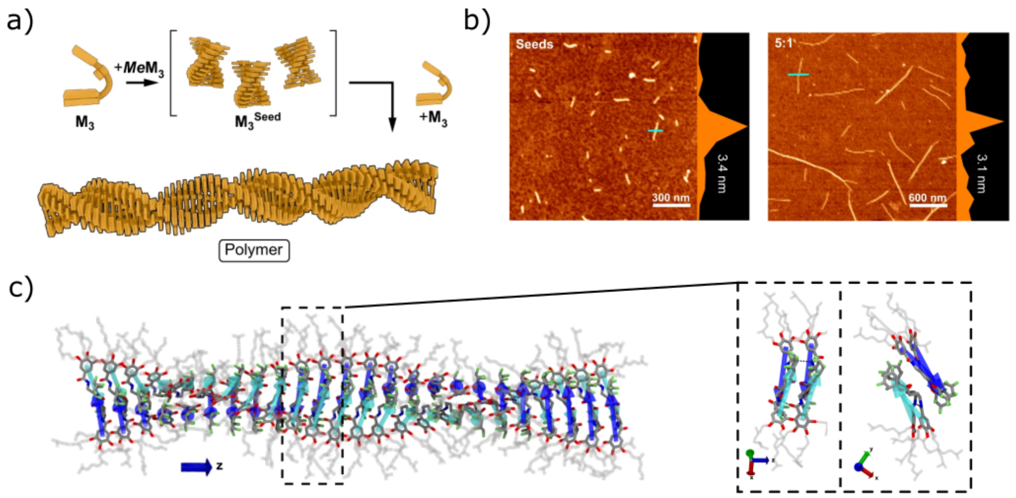


Fig. 1.11 Comparison of experimental and computational inference techniques for the living supramolecular polymerization presented and adapted from Ref. [60]. a) Illustration of the theoretical predicted mechanism for a seeded living supramolecular polymerization. M_3 represents the free monomers and M_3^{seed} represents the seeds which drive the pathways for the polymerization (Fig. 1.7). b) AFM height images spin-coated on a silicon wafer of two separate moments of the polymerization: left side, representing the M_3^{seed} seeds; right side, the supramolecular polymers obtained with an M_3/M_3^{seed} ratio of $5:1$. In both cases the cross-section was measured along the blue line. c) Atomistic resolution snapshot of the monomers arranged into a stable helical structure, obtained *via* MD simulation. The two arrows, blue and cyan, connect the benzene (tail) to the pentafluorocyclohexyl groups (head) of each monomers, underlying the fine structure of interconnections of the building blocks. The color code of the atoms is green Fluorine, red Oxygen, blue Nitrogen, and gray Carbon.

computed from the potential energy function of the system. Obviously, this does not come without downsides: a properly executed MD simulation requires overcoming two main issues, namely having an accurate physical model that is able to describe and reproduce all the interactions between atoms (*i.e.*, the force field), and having a sufficient sampling of the system dynamics to gather statistically relevant results. Over the years, there have been several advancements in MD simulations that have greatly enhanced its capabilities and efficiency.[73] For example, improvements in computer hardware and software with enhanced algorithms for solving the equations of motion and improved sampling techniques, made it possible to simulate larger and more complex systems over longer timescales. Furthermore, the development of more accurate and efficient atomic force fields, techniques for modeling non-equilibrium processes and time-dependent phenomena, enabled the study of atoms and molecule in a wide range of physical conditions. In Figure 1.11 is reported a nice example of interplay between experimental and computational techniques, adapted from Ref. [60]. Finally, over the last few years, Machine Learning (ML) techniques have snuck in every scientific field. In the context of material science ML brought useful tools to either speed up or augment analysis workflows:[74] recognition and translation of recurring pattern or properties of molecular systems,[75, 76] and training and developing of high accuracy Force-Fields[77, 78]

1.6 Aim and overview

The aim of this study is to propose a general analysis for the static and dynamic characterization of soft self-assembled architectures and, most importantly, for their comparison and classification across different material families. Data on numerous soft aggregates has been gathered through the application of state-of-the-art computer simulations: using all-atom (AA) and coarse-grained (CG) models, together with standard MD or advanced sampling techniques. A substantial effort has been put into the development and the validation of the molecular Force Fields (FF) for the chemical species constituting the studied materials, in order to accurately reproduce their behavior in our simulation environments. Furthermore, ML-based methods have been optimised and used to encode, classify and compare amongst different materials, as well as across different structural families and dimensionalities. In the following we will show and discuss how these approaches allow to draw interesting conclusions on the presence, nature and role of structural defects in self-assembled materials and how they participate in defining the material identity.

The thesis content is organised as follow. In Chapter 2 we will introduce all the theoretical tools required to understand and follow the analysis steps. In the first section (Sec. 2.1), we will cover topics of computer simulations, introducing the fundamental aspects of unbiased equilibrium MD, equilibrium CG-MD, and biased enhanced sampling MD. Subsequently (Sec. 2.2), we will establish the mathematical background and the concepts behind the so called “atomic descriptors”, which are the fundamental ingredients of our analysis workflow. Finally, in the last section (Sec. 2.3), we will present the general *in silico* workflow, introducing the algorithms that will transform the information obtained from the atomic descriptor into human-readable data (*i.e.*, dimensionality reduction and clustering). In the next three chapters we will go over all the main results obtained. In Chapter 3 we will show how combining atomic descriptors and ML-based approaches we can acquire critical information on complex soft materials, allowing us to reconstruct their structure and the dynamic interconnections by which the monomer can interact with themselves and the environment. In Chapter 4 we will introduce the main results of this thesis work, showing how we can exploit our analysis workflow to compare and classifying different soft-assembled aggregates based only on their structural variability. In Chapter 5 we will discuss a different approach toward the classification of defects, showing how we can include kinetic-based models (*e.g.*, Markov State Model) into our analysis, enriching the characterisation of these complex systems, which would be difficult to obtain otherwise. Finally, in the last chapter, Chapter 6, we will present an entirely different system of aggregating charged NPs to showcase how the same paradigms of out ML analysis can be applied to virtually any molecular systems.

Chapter 2

Theoretical background

2.1 Computer simulations

Simulation is a word that implies the concept of “imitation of a process or a situation”, and it is used in disparate contexts of our lives. In modern days, however, it acquired an additional terminology: “a computer calculation that allows the study and observation of phenomena by means of numerically reproducing the behavior of a physical model”. Computer aided simulations rapidly became essential to the scientific community, and nowadays cover a sizeable portion of its research efforts.

2.1.1 Molecular Dynamics

In literature, *classical* computer simulation of molecular systems can be divided into two main approaches: Molecular Dynamics (MD) simulations and Monte Carlo (MC) simulations.[73] While both of these techniques aim at generating an ensemble of configurations that reproduces the physical behavior of the system under study, they fundamentally differ in the way such configuration are generated. In MD, the equations of motion of the system particles (*i.e.*, Newton’s equation of motion for classical MD) are numerically solved, based on a defined a *priori* set of interaction rules, which determine the total potential energy surface of the system. In contrast, in MC simulations, the configurations are properly generated to follow a Boltzmann statistics, based on the model potential energy surface. Thus, in a MC approach the information regarding the particle dynamics is lost during the generation of the configurations ensemble, whereas MD simulates the actual dynamics of the system. Then, the appeal of MD simulations comes from their ability to investigate a wide plethora of phenomena: static, structural, and thermodynamics properties, but also dynamical properties like transport coefficients, response functions and even non-equilibrium phenomena. Furthermore, the MD framework boasts a greater variety and efficiency of algorithms.

Let us consider a generic system made of N particles (*e.g.*, atoms of a molecular system), their instantaneous space configuration at time t is described by the set,

$$A_t = \{\mathbf{r}_i(t), \alpha_i\}_{i\in N}, \tag{2.1}$$

where $(r_{i,x}, r_{i,y}, r_{i,z}) = \mathbf{r}_i \in \mathbb{R}^3$ are the instantaneous Cartesian coordinates of the i-th particle, and α_i is the identifier of said particle. For example, in the atomic case, we can define $\alpha_i \equiv Z_i$, the atomic number of the i-th atom, but in general, α_i can carry all the physical parameters associated to the particle, such as, species, mass, charge, etc. In the context of MD simulation, A_t, is usually referred as (simulation) *snapshot*, as it represents effectively an instantaneous picture of the system.

The classical time evolution dynamics of a generic set of coordinates A can be described by Newton's equations of motion:

$$m_i \frac{\partial^2 \mathbf{r}_i}{\partial t^2} = \mathbf{F}_i = -\frac{\partial \mathscr{U}_{pot}}{\partial \mathbf{r}_i}, \tag{2.2}$$

where m_i, is the mass of the i-th particle, and $\mathbf{F}_i$ the total force acting on it. Forces are derived from a potential energy function $\mathscr{U}_{pot}(A) = \mathscr{U}_{pot}(\mathbf{r}_1, \dots, \mathbf{r}_N)$, which is usually expressed as the sum of two main contributions:

$$\mathscr{U}_{pot} = U_{inter} + U_{intra}. \tag{2.3}$$

The U_{inter}, intended as the potential energy of interactions that exist between particles that do not belong to the same molecule or, for bigger molecules, for atoms that are separated by more than three connections (*i.e.*, bonds, although this restriction can change depending on the specific system parametrisation). This potential energy contribution is commonly and conveniently expressed into separate body-ordered contributions, 1-body, 2-body, 3-body, and so on,

$$U_{inter} = \sum_{i=1}^{N} u_i + \sum_{i=1}^{N} \sum_{j>i} u_{ij} + \cdots . \tag{2.4}$$

An example of 2-body term entering Equation 2.4 is the Lennard-Jones pair potential[79, 80], where the simplest form can be expressed as

$$U_{LJ}(r_{ij}) = 4\varepsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right], \quad with \quad r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|. \tag{2.5}$$

The second contribution, U_{intra}, of Equation 2.3, accounts for the interactions between atoms separated by three or less connections (bonds), and can be expressed as a combination of bonds (r_{ij}), angles (θ_{ijk}), and dihedral (ϕ_{ijkl}) terms:

$$\begin{aligned} U_{intra} = & \frac{1}{2} \sum_{bonds} k_{ij} \left(r_{ij} - r_{ij}^{eq} \right)^2 + \\ & + \frac{1}{2} \sum_{angles} k_{ijk} \left(\theta_{ijk} - \theta_{ijk}^{eq} \right)^2 + \\ & + \frac{1}{2} \sum_{dihedrals} \sum_{m} k_{ijkl}^{m} \left(1 - \cos \left(m\phi_{ijkl}^{m} - \omega_{ijkl}^{m} \right) \right) . \end{aligned} \tag{2.6}$$

A bond is defined as a single connection between two atoms, often visualised as a segment (r_{23} in Fig. 2.1), an angle is formed by two consecutive bonds and a total of three adjacent atoms (θ_{234} in Fig. 2.1), and a dihedral is formed by three consecutive bonds, *i.e.*, four adjacent atoms (ϕ_{1234} in Fig. 2.1). The first two contributions of Equation 2.6 account for the energetic cost of stretching/compressing the bonds, and bending the angles, with respect to their equilibrium values. The last contribution appears as a sum of torsional potentials, which express the energetic cost of rotating the dihedral angle, *e.g.* keeping two bonds fixed and rotating the third one maintaining its angle value untouched. An dihedral defined this way is often defined as “proper” dihedral angle. In contrast to this, one can also define a dihedral angle by listing three atoms which are all attached to one "central" atom. A group of four atoms described in this way is referred to as an “improper” dihedral, and the angles formed by this group are called improper dihedral angles. In all the terms, k represents the force constant of the mode. These bonding interactions are usually kept constant

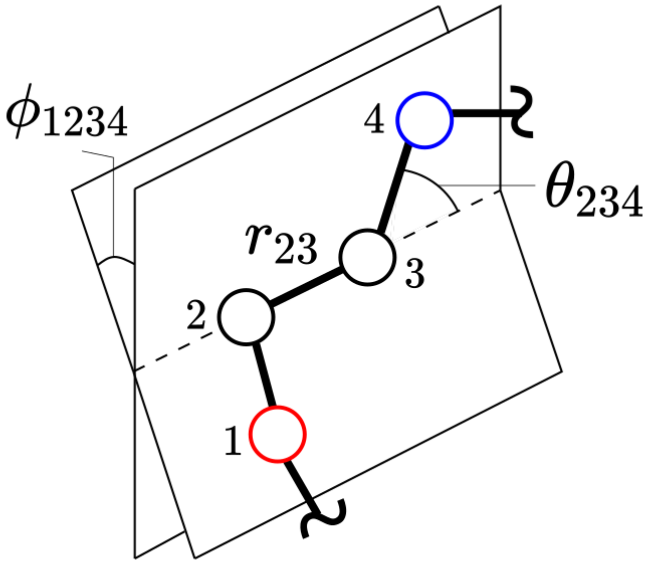


Fig. 2.1 Visualisation of the contributions of Equation 2.6. The red and blue circles highlight the two particles that lie on the two intersecting planes, defining the dihedral angle ϕ_{1234}.

during a MD simulation and are needed to simulate the covalent interactions that exist between the various atoms forming a molecule.

Equation 2.2 is discretised and solved numerically for each time interval $t_n = t_0 + n\Delta t$, where Δt is defined as the simulation *timestep* (*i.e.*, the simulation time unit). This is carried by well-tailored algorithms, which are defined to be time-reversible, symplectic, and to conserve the overall energy of the system.[73] One of the most used algorithm is the so called "Velocity Verlet"[81], known to be easily implemented with the upside of a particularly small long-term energy drift. The Velocity Verlet scheme can be summarised with the following system of equations,

$$\mathbf{v}_i^{n+\frac{1}{2}} = \mathbf{v}_i + \frac{\Delta t}{2m_i}\mathbf{F}_i^n, \tag{2.7}$$

$$\mathbf{r}_i^{n+1} = \mathbf{r}_i^n + \Delta t \mathbf{v}_i^{n+\frac{1}{2}}, \tag{2.8}$$

$$\mathbf{v}_i^{n+1} = \mathbf{v}_i^{n+\frac{1}{2}} + \frac{\Delta t}{2m_i}\mathbf{F}_i^{n+1}, \tag{2.9}$$

where, $\mathbf{r}_i^n$ and $\mathbf{v}_i^n$ represent the position and velocity at time $t^n = n\Delta t$. The numerical solution for all the particles through time outputs the so called *trajectory*, a collection of instantaneous configurations (or snapshots, Eq. 2.1), which describes the time evolution of all the simulated components. "Real-life" systems are often made up by 10^{23} atoms (*e.g.*, a mole of water is around 18 grams of water), or similar

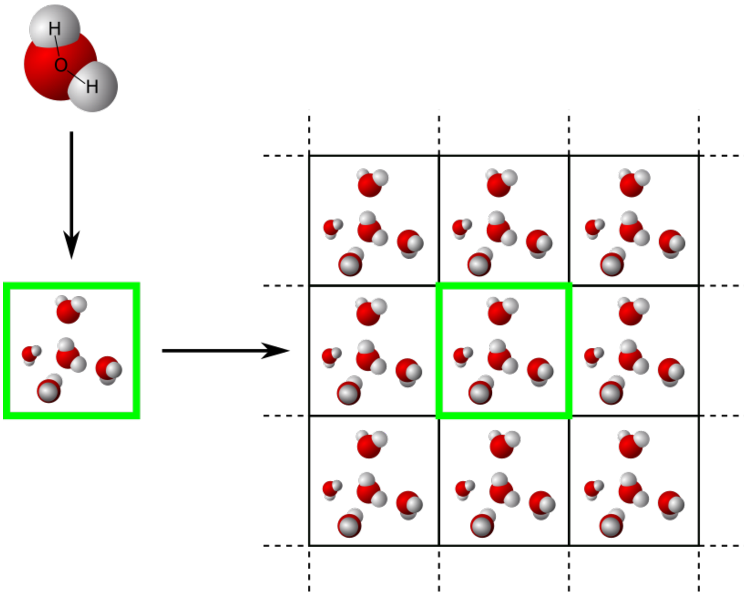


Fig. 2.2 Visual representation of the periodic boundary conditions. An idealised simulation box (green highlighted rectangle) is filled with a few water molecules. Turning on PBC conditions it is equivalent replicate of the original box in all spatial directions.

orders of magnitude. Due to limits induced by their technological implementation, simulations can rarely reproduce such amounts of atoms, but instead, are carried out for a finite and representative sample of a real system. This represents another fundamental approximation, and it typically introduces relevant finite size effects. In order to mitigate them, it is customary, to simulate the system using the so-called *periodic boundary conditions* (PBC). Under PBC, the boundary surfaces (edges of the simulation volume) of the simulation box are connected to the opposite ones in all three directions (Fig. 2.2), so that the system appears as infinitely replicated. Particles can freely cross the box boundaries, and interact with the whole system.

Usually, MD is employed to study systems close to their thermodynamic equilibrium. therefore, we usually simulate and gather relevant information on configurations, A_t, that span a bounded region of the total energy landscape for a given collection of atoms (and interactions between them),

$$E_t \sim E_{t'} \quad \forall t,t' \in T, \tag{2.10}$$

where $E_t \equiv E\left(A_t\right) \equiv E\left(\{\mathbf{r}_i(t)\}_{i \in N}\right)$ is the (potential) energy associated to a given atomic configuration A, at a given moment t of a simulation. A possible simulation condition for this exploration would be a fixed number of particles N, fixed volume

V (of the simulation box), and fixed Energy E; these conditions define the microcanonical ensemble NVE. The NVE ensemble assumes the system to be isolated, *i.e.*, it can't exchange energy (or particles) with its environment, thus its total energy is forced to remain constant with time. This specific set-up would be unrealistic for many real-life purposes.[73] Instead, if we consider a system with a fixed number of particles N, simulations are customarily run by either fixing the Temperature and the Volume (NVT canonical ensemble) or the Temperature and the Pressure (NPT isothermal-isobaric ensemble), achieved by adding a thermostat and/or a barostat during the simulation. Thermostat and barostat are numerical algorithms that modify the Newtonian MD simulation scheme introducing stochastic corrections (*e.g.*, re-scaling) to the computed atomic positions and velocities, generating energy fluctuations (thermostat) and volume fluctuations (barostat). This ensure the sampling of the correct canonical distribution at constant T and/or P.[73] Once the conditions are chosen, and the equation of motion are set, the system evolves along the potential energy landscape defined by Eq.s 2.3-2.6.

The energy landscape of a collection of atoms, usually takes also the name of Potential Energy Surface (PES).[82] A PES describes all the configurations a specific collection of atoms is expected to visit, as a function of its positional variables, $\mathbf{r}_i$, and according to its energy contributions, $\mathscr{U}_{pot}$ (Eq. 2.3). Commonly, the dynamics of a physical system can be described through a more generalised set of coordinates, typically named reaction coordinates, collective variables, or order parameters; defined to isolate the relevant degrees of freedom associated to the process of interest from all the other coordinate of the system.[82] This is achieved by averaging all the contributions from the non-relevant variables, thus obtaining a simpler and effective version of the PES. This scheme is often exploited for the development of coarse-grained resolution molecular models (Sec. 2.1.2), and in Enhanced sampling methods (Sec. 2.1.3) in order to facilitate the exploration of the energy landscape, by considering only some relevant variables of the system. Assuming that the probability distribution function of our variables of choice, $P(\mathbf{x})$, is the canonical distribution associated with the potential energy function $\mathscr{U}(\mathbf{x})$,

$$P(\mathbf{x}) \propto \exp\left(-\beta \mathscr{U}(\mathbf{x})\right), \tag{2.11}$$

where $\beta = 1/k_B T$ is the inverse of the thermal energy, with k_B the Boltzmann constant and T the temperature, we can define the Free-Energy function of the system as

$$F\left(\mathbf{x}\right) = -\frac{1}{\beta}\ln P\left(\mathbf{x}\right). \tag{2.12}$$

Following Eq.s 2.11-2.12 we can define the free-energy surface (FES) as the Boltzmann weighted PES. The terms free-energy (or potential energy) and free-energy surface (or potential energy surface) refers to similar concepts. The former is formally used when referring to a function $F\left(\mathbf{x}\right)$ where the variable $\mathbf{x}$ includes all the configurational variables, thus retrieving the true thermodynamic potential. The latter, refers to a $F\left(\mathbf{x}\right)$ where $\mathbf{x}$ represents only a representative set of collective variables.[82]

The FES (or PES) is an essential ingredient if we want to extract relevant information from the dynamics of a physical system. Effectively, the surface, indicates how the system statistically populate all the accessible regions of the phase space. Any physical observable descends from this statistical information, as its value is obtained *via* the so-called ensemble average, an average over all the phase space weighted by the Boltzmann factor. Therefore, a given physical observable can be extracted only if the FES is known. However, unless when dealing with very simple, ideal systems, the FES is a very complicated function of the system variables, and it is not known a *priori*.

The MD framework grants to thoroughly explore the free-energy landscape, allowing to estimate these ensemble averages. Nonetheless, such exploration can be itself hindered whenever the energy landscape is very rough, hence having multiple local minima corresponding to multiple (meta)stable states. If the energy barriers that separate these local minima are high enough, probably standard MD wont be able to visit all the relevant states, in a reasonable computational time, *e.g.* remaining trapped in a free-energy minimum nearby the starting configuration of the MD simulation. Ways to overcome these limitations exists, and they go under the name of enhanced sampling techniques, where the exploration of the energy landscape is made more efficient, *e.g.* by introducing an energy bias acting on the collective variables of choice (these methods will be briefly discussed in Subsec. 2.1.3).

2.1.2 Coarse Grained MD

The computational cost of MD simulations scales with the number of atoms in the simulated system, as well as with the complexity of the interactions considered (*e.g.*, long range electrostatic interaction, three-body potentials).[73] Furthermore, many interesting phenomena could be rare and/or slow, hence, they require longer simulation times to be properly sampled, increasing the computational cost even more. These computational limitations are typical problems that weigh on the *in silico* studies of complex molecular systems, such as supramolecular polymers, lipid membranes, proteins or other mesoscaled sized systems.[83, 84] A possible solution, to mitigate this problem, is employing lower resolution molecular models, *e.g.*, Coarse-Grained (CG) models. The idea behind the CG approach is to approximate entire molecules, or fragments of them, with the so-called "pseudo" atoms, or "CG-beads", achieving a considerable reduction in the number of simulated bodies. This reduction is done in such way to substantially reproduce the behavior of the molecular structure (*i.e.* certain key physical properties) at the original resolution, while critically reducing the computational cost of simulating the entire system. Amongst the available CG frameworks, the MARTINI Force Field[85], and the MARTINI "philosophy"[1], is one of the most used approaches to build CG models. Originally, it was developed to accurately reproduce the properties of lipid-based systems, but it rapidly evolved to a general-purpose CG force field, focusing on many different (bio)molecules.[86] Ref. [86] presents a nice road-map on the evolution of the MARTINI CG-FF, and on the numerous tools that have been developed to support the usage of this CG scheme.

In a more rigorous way, building a CG model involves a coordinate transformation, from our reference structural frame $A = \{\mathbf{r}\} \in \mathbb{R}^{3N}$ (Eq. 2.1), to a reduced dimensional one:

$$\Lambda = \xi(A) \in \mathbb{R}^{3n} \quad \text{with} \quad n \ll N. \tag{2.13}$$

The function $\xi(\cdot)$ can be of any kind, but is usually represented in matrix form $\Xi \in \mathbb{R}^{3n \times 3N}$, where it can linearly project our coordinates into CG ones,

$$\Lambda = \Xi A. \tag{2.14}$$

The missing ingredient, for simulating the CG system, is the formulation of a CG (potential) energy function $\mathscr{U}_{pot}^{CG}(\Lambda;\chi)$, which follows the usual definition as presented in the general case in Eq. 2.3. The variable χ contains the all the parameters for the definition of said energy function: non-bonded parameters, force constants, and partial charges. Notably, the set of unknown a *priori* parameters χ can be also

[1] Philosophy, here, means the concepts that underline the MARTINI FF architecture.

optimized by means of ML approaches (*e.g.*, *via* Neural Networks), often resulting in an improved accuracy of the CG-FF.[87]

A key requirement for a well-built CG-FF is the so called "Thermodynamic consistency"[88]: the equilibrium distribution of the CG model has to closely resemble the equilibrium distribution of the AA model (or the finer-grained model of reference), once the latter is mapped to CG coordinates. To satisfy this confition, let the CG potential energy fucntion be defined as:

$$\mathscr{U}_{pot}^{CG}(\Lambda;\chi) \equiv -\frac{1}{\beta}\ln p_{CG}(\Lambda) + \text{const.}, \tag{2.15}$$

where $\beta = 1/k_BT$ is the inverse of the thermal energy, with k_B the Boltzmann constant and T the Temperature. The term, $p_{CG}(\Lambda)$, is the probability distribution of the atomistic model, mapped onto the CG one,

$$p_{CG}(\Lambda) = \frac{\int \mu(A)\,\delta(\Lambda - \xi(A))\,dA}{\int \mu(A)\,dA}, \tag{2.16}$$

where $\mu(A) = \exp(-\beta\mathscr{U}_{pot}(A))$ is the Boltzmann weight associated to the atomistic (potential) energy model, assuming an MD simulation done in the canonical ensemble. This is a fundamental step in developing an accurate CG framework. Coarse-graining a molecular system, through the application of Eq. 2.14, leads automatically to a loss in information. Hence, while applying this process, we should preserve as much as possible the important degrees of freedom required to describe the dynamics of the system. Usually, the requirement of thermodynamic consistency is a stringent one, as enforcing the equivalence at the level of the distribution function (or phase space) is more challenging that, for example, equivalence of some collective observables. However, when dealing with complex molecular systems, there is no a *priori* knowledge of which set of degrees of freedom has an higher impact in describing a specific behavior. In this sense, generative machine learning brought immense contributions, allowing for recipes to find low dimensional representations.[89–91]

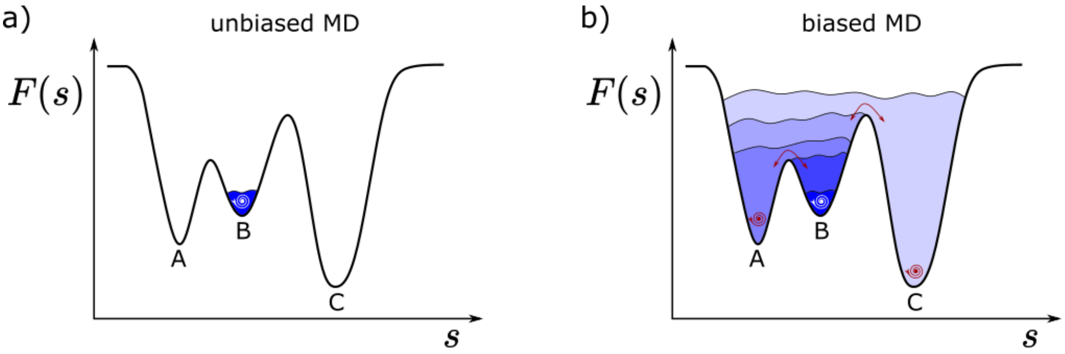


Fig. 2.3 Illustration of the metaD process, where the bias applied to the set of chosen CVs drives the exploration of energy landscape defined in said variables; here represented as a 1-D function for simplicity. a) Unbiased MD simulation: the system can explore configurations that are reachable through thermal fluctuations, hence it remain stuck in the local minimum B. b) Biased exploration of the phase space: following the addition of energy bias, the system can visit a bigger portion of the reduced phase space, eventually covering it all.

2.1.3 Metadynamics

Metadynamics (metaD)[92, 93] is an enhanced sampling technique that allows for a more efficient exploration of the phase space, resulting in a more thorough sampling of rare fluctuations, hence in a more accurate estimation of the FES associated to complex molecular systems. The most general implementation of MetaDynamics works by iteratively “adding” to the potential energy of the system an energy bias, $B(\mathbf{s},t)$, defined as a function of time and of a chosen set of Collective Variables (CVs). The added bias forces the system to explore the phase space along the directions defined by the chosen set of CVs, helping the system migrating from one (local) minimum to another.[92] The CVs are functions of the system coordinates, $\Pi(\mathbf{x})$,

$$\Pi : \mathbf{x} \to \mathbf{s} \in \mathbb{R}^m \quad \text{with} \quad m \ll 3N, \tag{2.17}$$

which select only some “relevant” degrees of freedom of the system, describing the motions across all the relevant metastable states, and possibly the relevant transition pathways.

Given a set of CVs, the probability distribution function takes the form,

$$P(\mathbf{s}) = \int d\mathbf{x} P(\mathbf{x})\, \delta(\mathbf{s} - \Pi(\mathbf{x}))\,, \tag{2.18}$$

and consequently the free energy as a function of **s** is defined as,

$$F(\mathbf{s}) = -\frac{1}{\beta} \ln P(\mathbf{s}) \,. \tag{2.19}$$

A metaD simulation is then run with a modified potential energy function,

$$\tilde{\mathscr{U}}_{pot}(\mathbf{x},t) = \mathscr{U}_{pot}(\mathbf{x}) + B(\Pi(\mathbf{x}),t) \,, \tag{2.20}$$

granting, in principle and at convergence conditions, free diffusion between the (local) metastable states.[93] The bias is “deposited” in specific time intervals, hence the time dependence of Eq. 2.20, and in such a way to not drive the system to far from its equilibrium conditions. To better understand the effect of $\tilde{\mathscr{U}}_{pot}(\mathbf{x},t)$ we can consider a simple 1-D case as reported in Figure 2.3, where three local minima are present. The system is prepared in (local) minimum B and the energy barriers which separate B from A and C are much higher than the thermal energy. Thus, by means of standard (unbiased) MD simulation, at fixed temperature, the system is bound to explore only configuration energetically close to the B minimum (Fig. 2.3a blue colored filling and white swirl symbol). During a metaD run, as time goes, bias is deposited and the system explore further the phase space. From B the natural and more convenient escape route is to pass the lowest energetic barrier and fall into the A basin (Fig. 2.3b red arrow and swirl symbol). In such state, the system will be able to cross the barrier between state B and A, visiting that portion of the phase space, until the bias is enough to overcome the next energetic barrier (and access local minimum C). As a result of this enhanced and bias-driven sampling of the phase space, we can finally reconstruct the original and *unbiased* probability distribution function of the system, granting access to energetic estimates of all the metastable states visited. Different flavors of metaD have been developed, differing by the procedure of bias $B(\mathbf{s},t)$ addition, and by the way the original (unbiased) FES profile is retrieved.[93–95]

A point that all types of metaD have in common, and need to face, is that their capability and accuracy to reproduce a specific energy landscape, crucially depends on the CVs employed. As already stated, CVs are functions of the atomic coordinates, usually arbitrarily chosen following some physical and/or chemical prior experience, and they provide a low dimensional projection of the crucial degrees of freedom in the original conformational space. In this sense, the definition of such variables

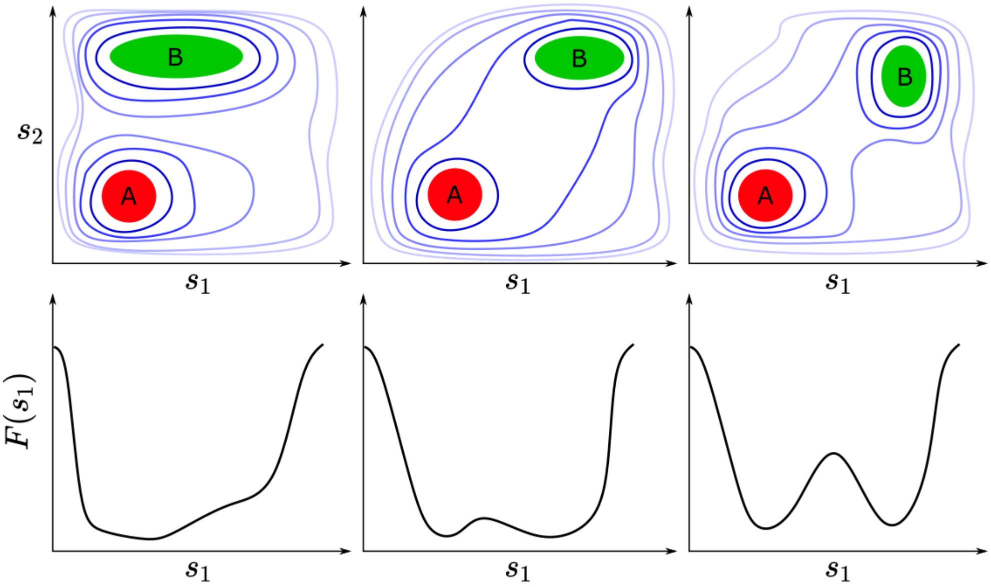


Fig. 2.4 Schematic representation of different CVs selection cases and their impact on the metaD free-energy estimation. In all three cases the original phase space is composed of two variables, s_1 and s_2 (top three panels), and the FES is projected along s_1 alone (bottom three panels). (Left) The chosen CV (s_1) cannot distinguish the two local minima (A,B). (Center) The two (local) minima are distinguished, but the transition between them is poorly sampled. (Right) The two (local) minima and the transition path are well sampled and reproduced by the free-energy profile estimated on s_1 alone.

should follow at least two requisites,[93] also visually summarised in Figure 2.4. First, different metastable states should be clearly separated in the projected low dimensional phase space (Fig. 2.4 right panel). If this condition is not satisfied, the added bias will be inefficient in driving the system across states that the CVs do not differentiate (Fig. 2.4 left panel). Second, the CVs should be able to efficiently sample the transition states, thus taking different values all along the “reaction” observed. If this is not the case, the barrier will be in general underestimated, leading to a false free energy estimation (Fig. 2.4 central panel). However, in general an accurate estimate of the barrier height is a challenging task, as the sampling of configuration along the so-called reaction path is less effective and complex, than the sampling of the local minima. Finally, it should be noted that even the absolute number of CVs has an impact on the FES landscape computation, since the biasing procedure in a multidimensional space becomes more and more expensive, and the actual exploration harder, as the dimension grows. Being one of the main problems of the method, the searching for optimal CVs is an active branch of research, in which machine learning is starting to be applied as well.[96–98]

2.2 Descriptors of atomic environments

Physical processes like self-assembly, response to external stimuli, phases transformation, annealing, as well as the evolution of dislocation networks or micro-structural domains during mechanical deformations are all phenomena that can affect a generic material and critically determine its purposes and applications. These phenomena posses a inherently multiscale nature, as the intimate nature of the local building blocks, that constitute the material, has tangible repercussions on its global properties. Atomistic simulations using empirical model FFs play an important role in bridging the gap in length scales accessible in single molecules and/or atoms simulations (atomic length scales) to continuum scale methods (macroscopic length scales). Using MD, and all its different flavors and variations, we have potential access to a huge amount of data regarding the behavior of complex self-assembled materials. Moreover, through the analysis and interpretation of such data, we would like to accurately correlate a certain atomic/molecular pattern to a specific property, and eventually even the opposite, *i.e.*, predicting atomic arrangements *given* particular target properties. These types of studies have the potential to have a great impact on society, accelerating the discovery of new functional material all over the scientific fields.[99]

In the past two decades, with the advent of machine learning, many efforts have been put towards the development of modern ML approaches that can learn, detect, and eventually predict molecular patterns. In order to achieve this goal, aside from the ML algorithm itself, all is required is a large enough data set, containing statistically relevant sampling of the phenomenon we want to investigate and an efficient and concise way of representing atomic and molecular environments. As discussed in Section 2.1, the instantaneous raw output of a MD simulation is represented by a set of tuples,

$$A = \{\mathbf{r}_i, \alpha_i\}_{i \in N}, \tag{2.21}$$

each of which contains the Cartesian Coordinates $\mathbf{r}_i \in \mathbb{R}^3$ and the identity α_i of the i-th particle (we dropped the time dependence of the positions as it is not relevant for the current discussion). The Cartesian configurations, although in principle they contain all the structural information needed, they are not enough to be used as input for any given ML approach. These spatial coordinates are not unique representations of a given atomic configuration, as they are arbitrarily ordered and often it is possible to obtain different configurations by application of a simple symmetry operation (*e.g.*, permutation, rotation, translation), which would lead to a biased or fundamentally erroneous prediction.

A "descriptor" of atomic environment is defined as a transformation of the Cartesian coordinates reference space to a "descriptor space", where all the atomic neighborhoods are mapped to unique fingerprints, carrying their structural information. Hence, the instantaneous atomic configuration can be re-written as

$$\Gamma = \{\boldsymbol{\gamma}_i, \alpha_i\}_{i\in N}, \tag{2.22}$$

with $\boldsymbol{\gamma}_i$ being the feature vector corresponding to the reference Cartesian component $\mathbf{r}_i$. This is reminiscent of the definition of some collective variables for a specific molecular environment, where the CVs reduction is needed to better isolate the relevant degree of freedom under investigation.

The concept of structural representations is actually not a novelty in the scientific community. Often defined under the name of *order parameters*, since the late 20^{th} century, structural descriptors have been used to describe and identify different phases in hard condensed matter[100, 101] and in liquid crystals systems[102]. More recently, a number of new descriptors have been proposed,[75] mainly aimed at accurately representing chemical environments of single molecules and materials, as a first step towards the definition of ML-based potential energy functions.[78, 103, 104]

2.2.1 Smooth overlap of atomic positions

The *Smooth Overlap of Atomic Positions* (SOAP)[105] is a particular descriptor, that associate to a set of selected points in a molecular system (centers), namely atoms, molecules, or generic groups of particles, a function that describes their atomic environmet. This is obtained by computing the rotational and translational invariant many-body density correlation features of each center with respect to its neighbours, inside a set cutoff sphere. Originally introduced in Ref. [103], it was developed to accurately represent chemical environments with the purpose of approximating potential energy surfaces and interatomic potentials of molecules and materials. The vectorised SOAP representation for the *i*-th local environment is built considering the local density from the neighbors of center *i*, expressed as a sum of gaussian contributions:

$$\rho_i^{(\alpha)}(\mathbf{r}) = \sum_j \exp\left[\frac{-|\mathbf{r}-\mathbf{r}_{ij}|^2}{2\sigma^2}\right] f_{\mathtt{rcut}}(|\mathbf{r}-\mathbf{r}_{ij}|). \tag{2.23}$$

The function $f_{\mathtt{rcut}}$ ensures that the summation takes into account only particles inside a cut-off radius, and all the other contributions decay smoothly to zero. The density function in Equation 2.23 is then rewritten as an expansion on orthonormal radial basis functions (*e.g.* spherical Gaussian type radial functions), and spherical harmonics, such that

$$\rho_i^{(\alpha)}(\mathbf{r}) = \sum_n \sum_l \sum_{m=-l}^{l} c_{nlm}^{\alpha} g_n(r) Y_{lm}(\theta,\phi), \tag{2.24}$$

where c_{nlm}^{α} are the expansion coefficients. Finally, the SOAP vectorised representation is obtained by combination of the elements of the expansion in Eq. 2.24. The simplest way is to form the partial power spectrum of the expansion coefficients,

$$p_{nn'l}^{\alpha\alpha',i} = \pi\sqrt{\frac{8}{2l+1}} \sum_{m=-l}^{l} \left(c_{nlm}^{\alpha,i}\right)^{*} c_{n'lm}^{\alpha',i}, \tag{2.25}$$

$$\mathbf{p}_i = \{p_{nn'l}^{\alpha\alpha',i}\}. \tag{2.26}$$

The coupling critically ensures rotational invariance, granting SOAP all the desired attributes of a "good" descriptor (*i.e.*, permutation, translation and rotation invariant). All the definitions and details reported here and further mathematical proofs can be found in the original work of Bartok, *et al.*, in Ref. [105].

In the SOAP framework, a generic MD snapshot would be represented as the collection of all the $\mathbf{p}_i$ vectors for all the components, which recalling Eq. 2.22, would read

$$\Gamma = \{\mathbf{p}_i, \alpha_i\}_{i\in N}, \tag{2.27}$$

where we replaced the $\boldsymbol{\gamma}$ symbol to stress that we are dealing with a specific representation. The number of components, *i.e.*, the dimensionality, of the vectors in Equation 2.27 depends on the extent of the radial and angular basis expansions, and on the number of different atomic species taken into account,

$$\mathcal{N}_{features} = N_\alpha \cdot n_{\max} \cdot \frac{(N_\alpha \cdot n_{\max}+1)}{2} \cdot (l_{\max}+1), \tag{2.28}$$

all parameters that can be tuned during the analysis. The total number of features can easily reach very high values, due to the quadratic dependence in both the number of elements and radial basis numbers. Therefore, the SOAP vectors associated to an MD trajectory constitute a substantially high-dimensional data-set, that can be interpreted only after further processing, as discussed in the following.

The features vectors of Eq. 2.27 can be defined as "local" descriptor, as they describe the local environment around each selected center (atom) of our system. Following the same idea, we can define a "global" descriptor, which aims at representing the atomic/molecular environment that characterize, on average, a collection of local ones. The easiest way to build such global descriptor is to directly average different $\mathbf{p}_i$ vector contributions into a single, averaged power spectrum $\bar{\mathbf{p}} = \{\bar{p}_{nn'l}\}$. The different components of the average feature vector are defined as follow,

$$\bar{p}_{nn'l} \sim \sum_{m=-l}^{+l} \left(\frac{1}{M} \sum_{i}^{M} c^{i}_{nlm} \right)^{*} \left(\frac{1}{M} \sum_{i}^{M} c^{i}_{n'lm} \right), \tag{2.29}$$

where the index i runs over the M atomic environments considered in the average. In all the examples treated in this thesis, we will apply Eq. 2.29 to compute the average environment associated to the entire structure of the simulated system, thus $M \equiv N$. Furthermore, we can compute an average fingerprint for each of the available T snapshots, which results in a set of global fingerprints, $\{\mathbf{p}_t\}_{t \in T}$, describing the general behavior of a given system. Assuming that our MD simulation spans the equilibrium conditions of the simulated structure, we can further average $\bar{\mathbf{p}}_t$ over the whole time interval,

$$\langle \bar{\mathbf{p}} \rangle = \frac{1}{T} \sum_{t}^{T} \bar{\mathbf{p}}_t, \tag{2.30}$$

obtaining a single averaged SOAP fingerprint representative of the system at the equilibrium. This unique feature vector represents the main ingredient to asses global similarities between different soft-aggregates.

2.2.2 Similarity between molecular environments

As stated in the introduction of this section, descriptors are necessary ingredients if we intend to unambiguously compare complex molecular structures. To this purpose we require tools to measure the similarity between different fingerprints, which can then be used to easily classify different materials. Most ML approaches, that compare or classify data, make use of similarity measurement, which therefore have to be solid and easily computed, even when dealing with high dimensional, highly non-linear data.[106] A similarity measure is the quantification of how close/similar are two mathematical objects. In our case we will focus on SOAP feature vectors, but the mathematical foundations are general and applicable to wide variety of situations.

Our goal is to measure how similar are two D-dimensional SOAP vectors, $\mathbf{p}_i \in \mathbb{R}^D$, that come from the output of a MD simulation run. We can define a *similarity function*, $\kappa(\cdot,\cdot)$, which takes as input two vectors and gives a bonded measure of similarity/closeness between them. The function $\kappa(\cdot,\cdot)$, normally referred to as *kernel function*, can take the simple form of the dot product between the two components that we want to measure:

$$\kappa(a,b) = a \cdot b, \quad \forall a,b \in \mathbb{R}^D, \tag{2.31}$$

but other forms are possibles (*e.g.*, Gaussian kernel, n-polynomial kernel). All the kernel function types needs to obey determined mathematical rules, which grant them a set of interesting properties.[106–108] One of them being that the definition of a kernel (similarity) distance follows naturally from Eq. 2.31,

$$D^2_\kappa(a,b) = \kappa(a,a) + \kappa(b,b) - 2\kappa(a,b) = 2\left(1 - \kappa(a,b)\right). \tag{2.32}$$

In plain words, the kernel distance between two points in the feature space, can be seen as the sum of the self-similarities of the two points, from which we subtract the cross-similarity. We will employ these concepts to quantify similarities amongst different local/global descriptor vectors throughout the results presented in this thesis.

2.3 Structural features characterisation

The previous two sections introduced the main methods by which data regarding a generic soft-material structure are generated, *via* computer simulations, and translated in high-dimensional feature vectors. This represents the starting point, to build our characterisation workflow.

In the upcoming sections we will introduce a series of methods used to process, compare, classify, and rationalize the amount of data at hand. Lastly, we will summarise all of the discussed techniques and methodologies into a general, systematic computational pipeline for the analysis of simulations of soft-assembled structures.

2.3.1 Dimensionality reduction

Dimensionality reduction is often a mandatory step in every ML-based workflow, due to the fact that, usually, the generated feature vectors are high dimensional. Reducing the data dimensionality consists in applying a transformation to such data, from its original space, into a lower-dimensional one, while retaining most of the meaningful information and descriptive properties of the high-dimensional dataset. Existing algorithms are numerous and diverse, depending on the field of application. Generally speaking they can be divided into two main classes: linear and non-linear methods, which can further divided into feature selection or projection approaches, according to how the dimension of the feature space is reduced.[76, 109] In the remaining of this section, we will quickly overview a famous example of feature projection algorithm, the Principal Components Analysis (PCA), and an extension of it, the Time-structure Independent Components Analysis (tICA).

The core idea of feature projection algorithms is to find the most efficient transformation operation to embed the original, high-dimensional data, into a lower dimensional space. This operation is formally carried out by an operator, and in matrix form can be expressed as

$$\mathbf{W} = \mathbf{X}\mathscr{P}_{D \to M}. \tag{2.33}$$

The matrix $\mathbf{X}$ is the data matrix [2] , or *feature matrix*, that contains all our data. It has dimension of $N \times D$, where N is the number of elements ($\{\mathbf{x}_i\}_{i \in N}$), and D is the original feature space dimension ($\mathbf{x}_i \in \mathbb{R}^D$). The matrix $\mathbf{W}$ is the equivalent feature matrix but in the *latent* (*i.e.*, lower dimensional) space and it has dimension of $N \times M$, with M being the dimension of the latent space with $M \ll D$ (*i.e.*, $\mathbf{w}_i \in \mathbb{R}^M$ for all i in the set). Finally, $\mathscr{P}_{D \to M}$ is the matrix that carries the operation of (forward) projection: from the space with dimension D to a space of dimension M. Notably,

[2] We use a generic X symbol here to stress that this is a general approach toward the dimensionality reduction of any kind of data.

for each of these types of matrices there exists the opposite one, which carries the backwards operation, $\mathscr{P}_{M\to D}$.

The goal for a feature projection algorithm is to find a good projection matrix that minimizes the reconstruction error, that is

$$\ell = \|\mathbf{X} - \tilde{\mathbf{X}}\|^2, \tag{2.34}$$

where $\tilde{\mathbf{X}}$ represents the reconstructed data matrix, from the latent space one, by means of the backwards operation,

$$\tilde{\mathbf{X}} = \mathbf{W}\mathscr{P}_{M\to D}. \tag{2.35}$$

If the value of Eq. 2.34 is zero it means that the projection operation is flawless, performing a perfect low-dimensional embedding of the original space, hence with zero information loss.

The *Principal Component Analysis* (PCA)[110, 111] is amongst the most used dimensionality reduction algorithms. PCA aims to find the latent space such that the projected vectors are uncorrelated and unitary, and have maximal variance.[110] This problem, written in terms of Equation 2.33, reduces to solving the eigenvalue problem[112], that reads,

$$\mathbf{\Sigma\Phi} = \mathbf{\Phi\Lambda}, \tag{2.36}$$

where $\mathbf{\Phi}$ and $\mathbf{\Lambda}$ are both $D \times D$ matrices. The column elements of the former are the eigenvectors ϕ_i, and the diagonal elements of the latter are the eigenvalues λ_i. $\mathbf{\Sigma}$ is the data covariance matrix, defined as $\mathbf{\Sigma} = \mathbf{X}^T\mathbf{X}$. From algebraic considerations, the eigenvalues (and corresponding eigenvectors) will be ordered from the largest one in a descending order. Each entry can be interpret as a measure for the amount of "data variance" retained by that component: the sum of all the eigenvalues is one, by definition, and the first have higher absolute values that the last. As a result, it can be shown, that the projection reduces to taking the truncated eigenvectors matrix, $\mathscr{P}_{D\to M} := \tilde{\mathbf{\Phi}}$, where $\tilde{\mathbf{\Phi}}$ is a matrix of dimension $D \times M$, that contains only the first M eigenvectors. In the past twenty years PCA gained a lot of popularity in the context of MD simulations, due to its highly accessible implementation, and its flexibility, since it can be used to solve both linear and non-linear problems.[30, 76, 113, 114] One of the main disadvantages of the PCA, in the context of MD simulation, is that for complex molecular systems, the low-dimensional subspace identified by PCA might not coincide with the optimal subspace of *slow* degrees of freedom of the system. Meaning that the reactions coordinates that dominate the conformational dynamics of the system are not well represented in the low-dimensional latent space.

The *Time-Structure Based Independent Component Analysis*(tICA)[115, 116], was designed to overcome the aforementioned limitation, resulting in a method that could distinguish between slow and fast equilibrating degrees of freedom. The tICA algorithm aims at finding the latent space such that the projected vectors are uncorrelated, unitary, and maximize their time autocorrelation function, expressed by the *time-lagged* autocorrelation matrix,

$$\mathbf{C}(\tau) = \mathbf{X}(t)^T \mathbf{X}(t+\tau). \tag{2.37}$$

In contrast to the covariance matrix, $\mathbf{C}(\tau)$ requires a time-ordered knowledge of the underlying data (*e.g.*, a time-series) and the τ parameter is the time "lag" between two adjacent points. This problem can also be cast in a way similar to Eq. 2.36, and it reduces to finding the solutions to the generalised eigenvalue problem,[112] which in matrix form reads

$$\mathbf{C}(\tau)\mathbf{\Phi} = \mathbf{\Sigma}\mathbf{\Phi}\mathbf{\Lambda}. \tag{2.38}$$

Once the solution is available, we can use Equation 2.33 to project our high-dimensional data on the first M slowest components. tICA was introduced in the context of MD simulations as a good method for building Markov state models, as it was demonstrated that this type of projection is an optimal approximation to the Markov operator's eigenvalues and eigenfunctions.[117, 118]

Figure 2.5 reports a simple and effective example that summarises the difference of the PCA and tICA approaches. We use as a test dataset the trajectory of a particle that moves in a two-dimensional double-well potential, transitioning between two low-energy states (the wells) separated by an energy barrier (Fig. 2.5a). In this simplified case, we do not require any descriptors as the coordinates alone are sufficient for characterizing the particle evolution in time. We collected a total amount of 1000 evolution steps, thus, the data matrix has dimension of $\mathbf{X} = 1000$, with elements $\mathbf{x}_i = (x_i, y_i)$. Since we are dealing with only a 2-dimensional dataset, we can only reduce it to a 1-dimensional one. Panels (b) and (c) of Fig. 2.5 respectively show the direction of the first and only relevant eigenvector of the projection matrix and the outcome of said projection. The tICA projection is able to retrieve with high accuracy the transit of the particle between the two low-energy states, which in this

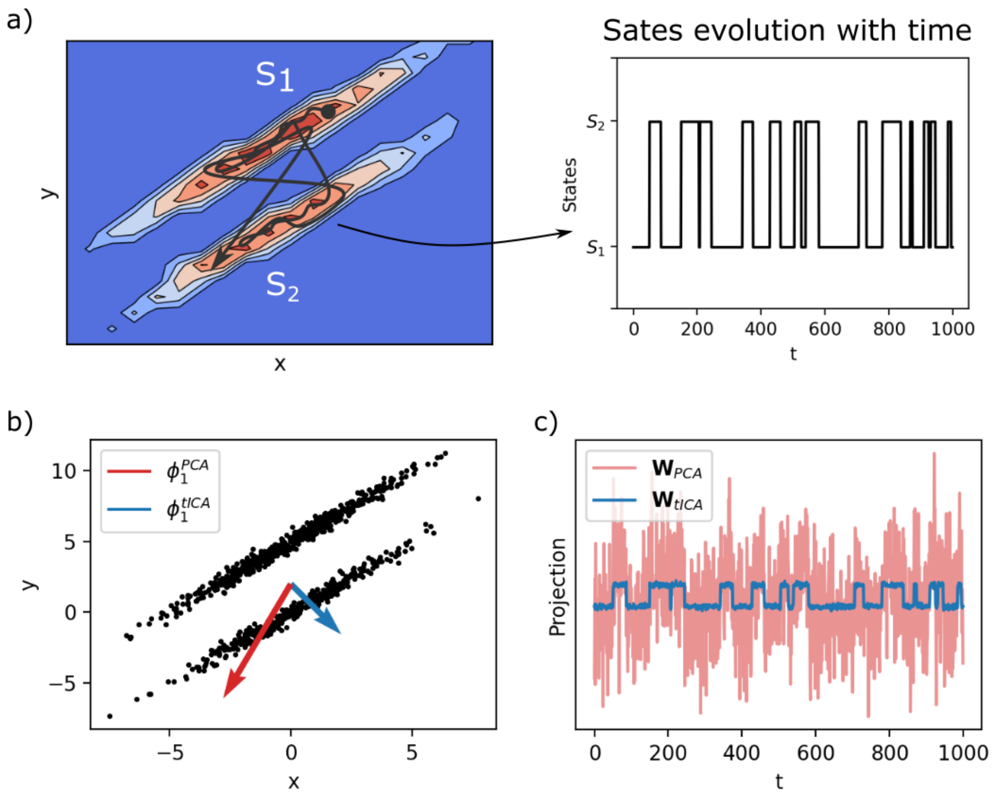


Fig. 2.5 Differences between the PCA and the tICA features projection methods on a test dataset. a) Representation of the double elliptic wells potential: warm colours represent low energy states and cold colours increasingly higher energy states. The grey line represent a sample of the particle trajectory, starting from the solid circle and evolving toward the arrowhead. On the side is plotted the states evolution with time. b) 2-D data points (black scatter) representing the instantaneous (x,y) coordinates of the particles. On top of that is plotted the direction of the principal (and only) eigenvector for the PC (red arrow) and tIC (blue arrow) analyses. c) 1-D data points obtained after applying the PCA (red) and tICA (blue) dimensionality reductions.

case coincides with the “slowest” degree of freedom of the system. On the contrary PCA fails entirely in reproducing the particle two-states dynamics.

2.3.2 Clustering algorithms

Clustering algorithms interest another big chunk of the ML methods pie as they are essential to uncover patterns and pair similar elements of a dataset.[119] In the context of molecular simulation analyses, these algorithms are often used as one of the final step in the analysis, to simplify and summarize large and complex MD data. These data, obtained through the application of CVs or atomic descriptors are representative of some important molecular behavior, hence the application of a clustering algorithm is fundamental to find recurring patterns. These patterns inevitably carry a physical meaning as they identify common structural features and identify important conformational changes observed in the molecular dynamic of the system under study.[30, 71, 120] However, it is worth to stress that a good classification result heavily depends on the underlying data and on the specific algorithm features of the clustering method, thus all the steps in the classification workflow have to work in synergy to grant the desired outcome.

Generally speaking, a clustering analysis is the task of grouping elements from a generic dataset, $\mathbf{X} = \{\mathbf{x}_i\}_{i \in N}$, into subsets $\{S_j\}_{j \in J}$, called *clusters*. These clusters are selected such that

$$\bigcup_{j}^{J} S_j = \{S_1 \cup S_2 \cdots \cup S_J\} = \mathbf{X}, \tag{2.39}$$

and each elements, $\mathbf{x}_i \in S_j$, are closer to each other than to the ones in a different $S_{j'}$, according to some predefined rules. The specific definition of said rules differentiates the clustering method and how it operates. However, we can identify a set of general approaches that account for most of the cases: connectivity-, centroid-, distribution-, and density-based clustering algorithms. Moreover, the clustering analysis can be *unsupervised* or *supervised* depending on the (hyper)parameters provided when initializing the algorithm. As well as dimensionality reduction, clustering algorithms are ubiquitous in every aspect of ML, scientific and non. In this work we will make use of a few different clustering approaches, and in the reminder of this section we will give some context on how these algorithms operate. Anyway, for a complete and in-depth derivation and discussion of these methods we refer to the appropriate literature.

K-means[121, 122] is a well-established clustering algorithm, where the S_j clusters are built in a way that minimizes the cost function (or inertia)

$$\mathscr{L}(S) = \sum_{j=1}^{J} \sum_{\mathbf{x}_i \in S_j} ||\mathbf{x}_i - \mu_j||^2, \tag{2.40}$$

where $\mathbf{x}_i$ and μ_j are respectively the points and the center of mass of the S_j cluster. The K-means algorithm, and some of its variations are easily accessible from Python libraries like Scikit-learn[123].

Spectral clustering[124, 125], represents an extension of the K-means methods. Given a set of data points, a dimensionality reduction is first performed and then the clustering algorithm is applied (*e.g K-means algorithm*) to points in the new latent space. When combined to K-means, this approach improves its clustering accuracy[125], at the cost of a worse computational scaling. Notably, if the dimensionality reduction takes into account of the data time-series, the spectral clustering takes the name of Perron-Cluster Cluster Analysis (PCCA)[126]. PCCA, and its augmented versions PCCA+/PCCA++[127], are largely employed in the Markov state models theory, to define the maximally metastable (long-lived) sets of states for a given Markov state model.

Probabilistic analysis of molecular motifs (PAMM) is a recent density-based clustering algorithm, introduced in Ref. [120]. It was developed with the purpose of analyzing (long) MD simulation data, and identifying the so called "molecular motifs"; namely, recurrent patterns in the local atomic/molecular environments of a selected atom/molecule in a complex system. PAMM algorithm combines various aspects, which make it efficient even when processing huge amounts of data. The first step is the estimation of the Probability Distribution Function (PDF) of an input dataset, $\{\mathbf{x}_i\}_{i \in N}$, obtained *via* Kernel-Density Estimation (KDE),

$$P(\mathbf{y}_k) = \frac{1}{\sum_i^N \omega_i} \sum_i^N \omega_i K_H(\mathbf{x}_i - \mathbf{y}_k), \tag{2.41}$$

$K_H(\mathbf{x}_i - \mathbf{y}_k)$ is the kernel function used in the estimation to relate the data points to the subset $\{\mathbf{y}_k\}_{k \in K}$, namely points of a grid, defined in such a way that $K \ll N$ (*i.e.*, $\mathbf{Y} \subseteq \mathbf{X}$). In this way the total computational cost of evaluating Eq. 2.41 is affected only by the extent of the grid. Once the density at the grid points, $P(\mathbf{y}_k)$, is computed, we can proceed to the identification of several distinct clusters (motifs), that effectively represent maxima in the probability distribution (corresponding *e.g.*, to recurring molecular patterns when PAMM is applied to an MD simulation of a supramolecular material). The number and position of modes in the PDF is determined by a non-parametric quick-shift algorithm[120] that assigns each grid points to the basin of attraction of the nearest maximum. Finally, the total probability distribution is fitted to a sum of Z multivariate Gaussian distributions,

$$\hat{P}(\mathbf{x}) = \sum_{z=1}^{Z} p_z G(\mathbf{x}|\mu_z, \Sigma_z), \tag{2.42}$$

where $G(\mathbf{x}|\mu, \Sigma)$ is the Gaussian distribution and p_z its associated weight. The partitioning of the PDF in Gaussians can be further improved by inferring clusters “stability” and by an additional "meta-clustering", which defines a dendrogram containing the clusters hierarchy. The stability of a cluster is inferred though a bootstrapping approach: the initialisation of the grid and the following KDE are repeated $N_{bootstrap}$ times (a parameter of the algorithm) and deviations of clusters identity are computed with a custom metric.[120] Based on those parameters it is possible to merge together the obtained clusters into larger ones that share some similarity (“macro”-clusters), obtaining a more general picture, and usually a simpler interpretation (see the results in Chap. 3 for specific examples on this procedure).

2.4 General analysis workflow

In the present thesis we extensively applied the methodologies introduced in Sec. 2.1-2.3 to thoroughly analyze the structure and dynamics of soft, supramolecular material, with a particular focus on the crucial role that structural defects play in determining their dynamics. In order to gather this information we have often combined the presented approaches. The majority of the analysis and results that will be presented in this thesis work share a common series of steps. The general workflow is depicted in Fig. 2.6, and listed as follow:

I. *Full MD trajectory*: an MD trajectory of the complete system (typically at CG resolution) is generated at the desired conditions of temperature, simulation box volume and pressure, and number of particles. Usually, we are interested in the near equilibrium behavior of the system, for this reason preliminary simulations steps might be required to equilibrate a system.

II. *“Reduced” trajectory*: the full trajectory is reduced to that of a few key (pseudo) atoms, which are fundamental to represent the intimate hierarchy of the monomers the form the self-assembled material, a sort of “skeleton” of the whole structure. This reduction is not essential, but might significantly simplify the procedure. However, it is not straightforward to generalize such selection for every system and it will be discussed in the results sections when needed.

III. *SOAP computation*: the set of Cartesian coordinates of the reduced trajectory is transformed into SOAP feature vector, with the appropriate parameters of $n_{\max}$, $l_{\max}$, and r_{cut} in order to capture the atom/molecular environments information.

IV. *Data processing*: the data from the SOAP feature vectors can be analyzed in different ways, depending on the ultimate scope of the analysis.

   IV_a. *Similarity measures*: using a global descriptor and Eqs. 2.31, 2.32.

   IV_b. *Dimensionality reduction*: using an algorithm to lower the dimension of the SOAP feature vectors (*e.g.*, PCA or others) to later clusterise the local features.

   IV_c. *Kinetic models*: using the SOAP vectors as input to build a Markov State Model (MSM).

V. *Structural motifs*: identification *via* clustering, and subsequent classification, of the main structural motifs and their dynamic relations (*i.e.*, transitions among states) at equilibrium conditions.

The list above summarises the general series steps of a soft-matter aggregate structural analysis but, of course, each of those steps could be further developed to include mathematical notation and specific details on the data handling. Moreover, almost every time aspects from each of the data processing steps (IV) will be used simultaneously. The main difference arises when dealing with multiple different systems and one would like to compare and relate the different structural motifs found in each of these systems. A first option involves training a dimensionality reduction algorithm using samples from the features vectors of all the considered systems. This way, the low-dimensional feature data of each individual system will be embedded in the same low-dimensional manifold, thus allowing well-defined similarity measurements and comparisons between motifs belonging to different systems. A second option would be applying both dimensionality reduction and clustering algorithm to each individual system separately. This produces low-dimensional, structural representations that are not directly comparable to themselves, as each dataset “lives”

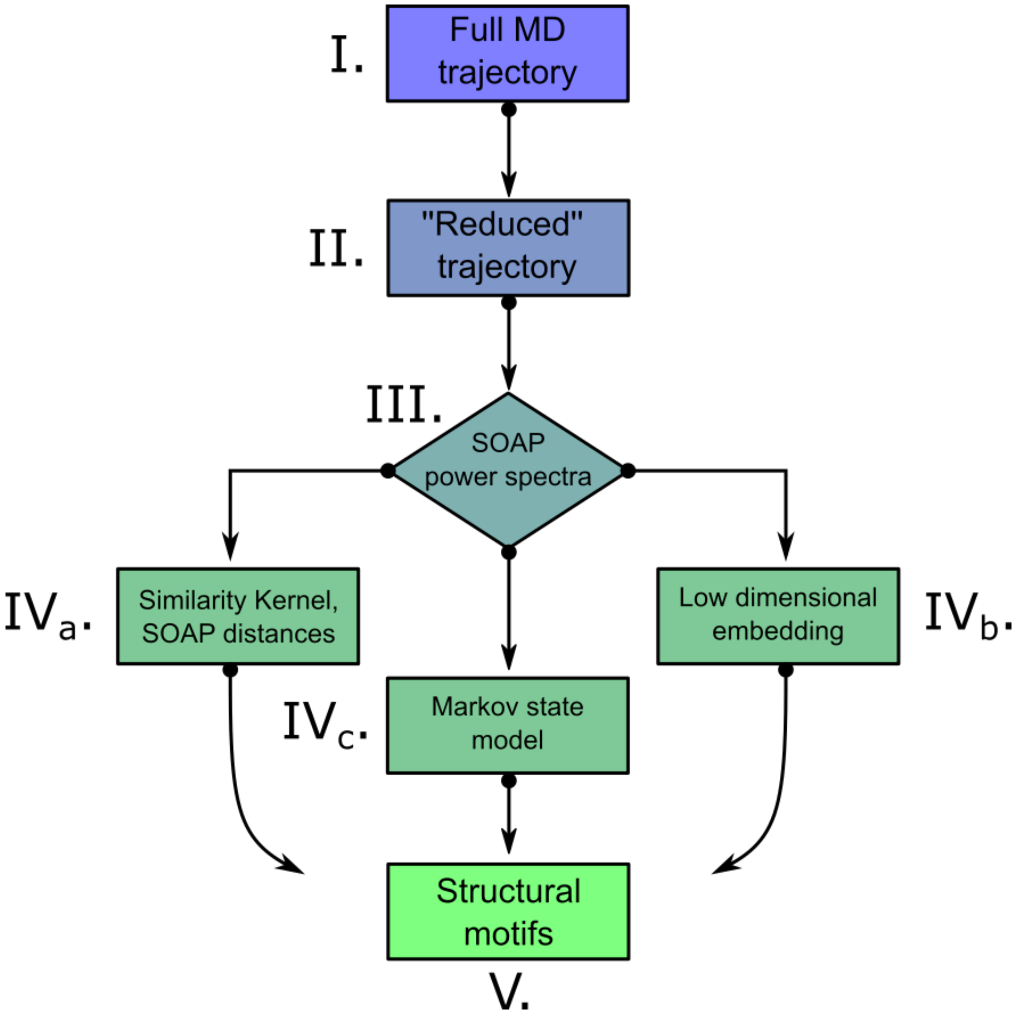


Fig. 2.6 Flowchart visualisation of the main steps of a general structure analysis of a soft supramolecular aggregate.

in a unique low-dimensional manifold. In this last case, the structural motifs can be extracted from the separate low-dimensional representations, but a comparison can be performed only after remapping the different structural states to the original, high-dimensional feature space, which is common to all the different systems.

Chapter 3

Characterization of supramolecular structures based on their molecular motifs

In the first chapter we introduced one of the central paradigms of self-assembly: the creation of ordered structures starting from a disordered collection of building blocks that spontaneously interact with each other through non-covalent interactions. A self-assembled structure is often thought of as defect-free, an ideally perfect mono- or multi-dimensional stacking of monomers, overlooking on the effect that the specific local disordered arrangement of building blocks might have on the overall dynamic character of the structure. Therefore, these structures are typically studied at a global level of statistical ensemble averages. However, MD simulations have recently shown the crucial role of local structural defects in controlling and determining complex dynamic behavior, that are absent otherwise.[29, 50] The study of these defects does not come without challenges, as the lack of a crystalline lattice in self-assembled systems makes even a simple structural analysis problematic. Recently, in Ref. [30] an unsupervised data approach for the identification of structural domains, and characterisation of their dynamics, in supramolecular polymers was obtained by means of machine-learning inspired methods. This methodology allowed to identify the various structural motifs populating the equilibrium state of different supramolecular fibres, also unravelling their persistence and dynamic exchanges. This ML approach, based on the SOAP description of atomic environments and on

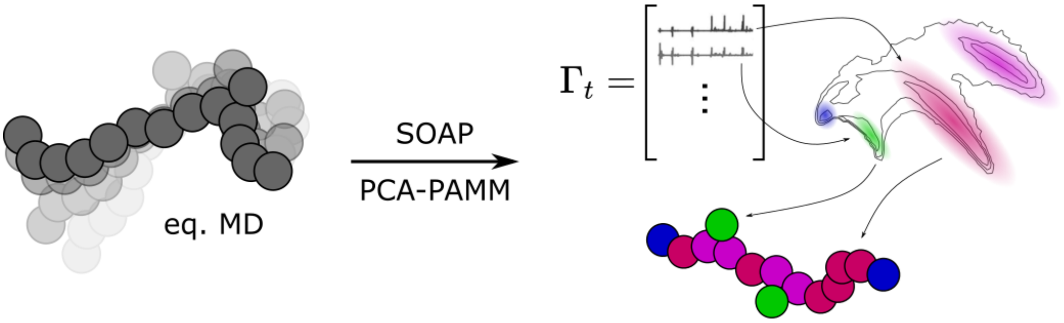


Fig. 3.1 SOAP-PCA-PAMM approach toward the structural characterisation of a self-assembled system. The set of structural configurations extracted from the equilibrium MD simulation of a target supramolecular system is represented by means of the SOAP descriptor. Each SOAP fingerprint is then projected onto a low-dimensional PC space and the data motifs are classified by means of the PAMM clustering algorithm, yielding detailed structural information of the system under investigation.

the PAMM unsupervised clustering, is further extended and generalised in this thesis work, by applying it to a wide variety of different supramolecular systems.

In this chapter we will report the results obtained through the use of the proposed ML-based approach to different model systems and at different models resolutions, following the general steps previously reported in Sec. 2.4 and visually schematised in Fig. 3.1. In most cases this ML pipeline of steps was used to support and validate other computational analysis and experimental evidences.

In the same spirit of the work reported in Ref. [30], I started by applying the ML analysis to a set of “super”-CG-FF supramolecular fibres (*i.e.*, lower resolution model compared to standard fine-CG models), with the purpose of finding the molecular determinants for the exchange pathways that exist in supramolecular polymers and explore possible ways to exploit them to achieve a deeper control over the fibres structural dynamics. In a second work, I tested the ML strategy on a supramolecular polymer obtained through living supramolecular polymerisation (LSP), where we could support the previously obtained computational and experimental evidences, providing a deeper understanding on the structural hierarchy of the monomer and on the equilibrium dynamics of the fibres under investigations. Overall, this ML-based approach proved to be efficient in fully characterizing the different structural features and in following their behavior throughout MD simulations. All the works presented in this chapter contributed to a specific publication, which will be appointed in the text.

3.1 Finding exchange pathways in supramoleculer polymers by characterizing their defects

The work presented in this section resulted in a publication in *ACS Nano*, Ref. [31], with the title "Controlling Exchange Pathways in Dynamic Supramolecular Polymers by Controlling Defects". Supramolecular polymers, fibre-like structures that self-assemble thanks to non-covalent directional interactions, are starting to gather great interest in chemistry, biology and other related scientific fields. In these dynamical structures, the individual building blocks continuously exchange in-and-out the assembly according to complex energy balances of the interactions in play. A full characterisation of the exchange pathways and their molecular determinants is often not an easy task. In Ref. [31], we tackle these problems by combining CG molecular modeling, enhanced sampling, and machine learning approaches to investigate the key factors controlling the pathways of monomer exchange. We demonstrate that most of the dynamic behavior can be associated to a competition of directional and non-directional interactions between the monomers, and that, such competition critically controls the creation/annihilation of defects in the supramolecular structure. Finally, thanks to our general models we start to investigate the single monomer feature that might contribute to control the exchange pathways in these dynamic assemblies. In the framework of my thesis work, I contributed to this article by applying the ML-based analysis to a set of supramolecular fibres in order to discover their structural motifs and extract their equilibrium exchange pathways, as it will detailed in the following.

3.1.1 Scientific context

Supramolecular polymers, fibres composed of monomers that self-assemble unidirectionally (1-D) *via* non-covalent interactions, are very popular in nature, and often play a fundamental role. Notable examples are the microtubules (MTs), fibre-like assemblies built by protein units, usually tubulin, whose dynamic polymerization and depolymerization are key factors that regulate different fundamental functions of the cell. All these systems share the unique physical trait of having a complex network of dynamic pathways of monomer exchange in-and-out the assembly, which is responsible in regulating the overall properties of the material.

Recently, the synthesis and control of supramolecular polymers attracted great interest in the perspective of designing artificial materials possessing similar complex dynamic behaviors to biological materials, as this would enable the rational design of material tailored for specific tasks. However, despite the massive advancements in technology, it is still very challenging to experimentally monitor such complex dynamic network of exchanges in supramolecular structures, at the necessary spatial and temporal resolution. In particular, unambiguously linking certain global behaviors of the assembly to monomers structural features often remain a daunting task.

In general, the dynamic exchange of monomers between two model fibres can be summarized into three fundamental steps: (i) monomers jumping out from a fibre, (ii) monomer freely diffusing in the solvent, and (iii) monomers adsorption onto another fibre. Step (ii) is mainly controlled by the laws of molecular diffusion, and it is not influenced by tiny changes in the chemical structure of the monomers. Steps (i) and (iii) are the most important, which directly define the complexity of a fibre. Both phenomena highly depend on local interactions with other specific parts of the fibre, which can either enhance or quench the kinetic of the processes.

In this work, computational investigations on 1,3,5-benzenetricarboxamide (BTA) based supramolecular polymers, across different model resolutions, revealed that the monomer exchange, out from the fibre body, can only originates from specific "hot-spots" in the supramolecular structure, *i.e.*, terminal monomers (tips) or structural bulk defects.[26] These are less ordered, weaker, and energetically higher spots in the assembled structure from which monomers exchange is most likely to proceed. The physical origins and kinetic properties of such "hot-spots" was thoroughly investigated in this work by means of supervised analysis and biased MD simulation (infrequent metaD) using some tailored CV to excite the motion of the interesting monomers. The key feature regulating the type and concentration of structural motifs was found to be the competition between the directional and non-directional forces between monomers inside the assembly. The first is determined by the cohesive, non-covalent, interaction strength that keeps the monomer together (and drive the SA process), while the second is regulated by the interaction of the monomer with the surrounding environment (*i.e.*, the solvent). Most importantly, it was shown that by directly altering these two monomer properties it is possible to control the presence and absence of specific structural motifs on the fibre backbone. Finally, despite the limitations imposed by the use of a supervised analysis it was possible to rationalise

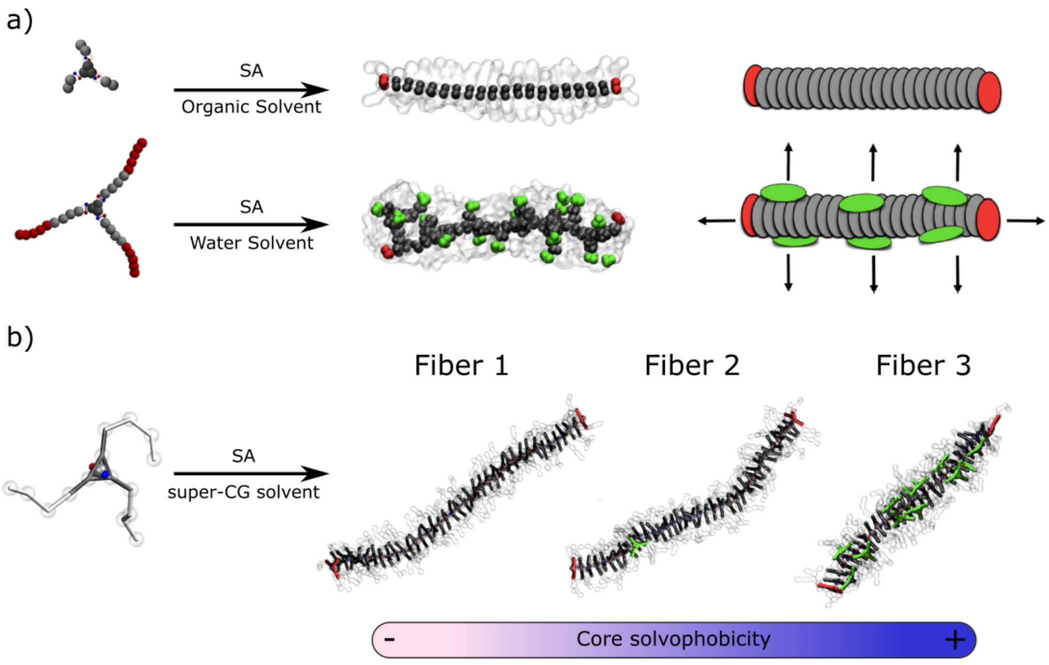


Fig. 3.2 Representation of different BTA monomer CG-FF models and their supramolecular aggregates studied in this work. a) Two fine MARTINI CG models for an organic solvent soluble BTA-C6 molecule (top) and a polar solvent (*e.g.*, water) soluble BTA-C12-PEG molecule (bottom) b) "Super"-CG BTA model. The model solvophobicity can be tune by changing the specific core CG-beads and it results in a different structural behavior of the aggregate: Fibres 1 to 3 have monomers with increasingly higher core solvophobicity.

important trends on the monomer exchange kinetics, proving how the presence of structural defects highly enriches the dynamic pathways and kinetic exchanges of the monomers.

In the framework of the present thesis, I have contributed to this project employing ML-based tools to detect in a unsupvervised way the defects that emerge in supramolecular polymers. The idea was to rationalise the dynamic pathways, that defines those assemblies, proving the tight relationship that exists between global defects and dynamics of a supramolecular polymer and the local features of its building blocks. The study is focused on MD simulations of the fibres based on the "super"-CG BTA molecular scaffold,[31] a highly editable physical model, well suited to our purposes. We will represent each fibre as a set of high dimensional feature vectors, by means of the SOAP descriptor of molecular environment. Dimensionality reduction and clustering algorithm will be then used to gather all data required to characterise the internal dynamics of such aggregates.

3.1.2 Methods

Extensive details on the preparation of the physical models, unbiased and biased simulations and supervised analysis can be find in Ref. [31] and related supporting information. Here we report the details regarding the application of the SOAP-PCA-PAMM unsupervised analysis towards the detectionn and dynamic characterisation of structural motifs in supramolecular polymers.

Unsupervised identification of defects and of defect dynamics

A complete characterization of the fibres structural motifs populations, their nature, and their equilibrium dynamics can be retrieved by application of atomic descriptors and unsupervised clustering. We focus on a set of three fibre models, referred to as fibre **1** to **3** for which a total of 30 μs equilibrium MD simulation were prepared for each system. The snapshots collected from the MD production runs were analyzed, focusing on the last 10 μs with a temporal stride of 10 ns. For all the collected snapshots, the SOAP atomic descriptor[105] was computed for the centre of each monomer, a position coinciding with the CG-bead carrying the dipole (Fig. 3.2b, central bead) This allowed us to have a computationally efficient way of capturing the structural variability of a target fibre, in terms of the structural/dynamic reorganization of its monomers. The SOAP descriptor was computed using the python library DScribe[128], setting *rcut*= 8 Å, *nmax*= 8, *lmax*= 8 as parameters, and leaving the rest as default. This kind of setup was already proven to be efficient in detecting fibres structural features.[30]. Given the high dimension and rich information of the resulting SOAP feature vectors we employed PCA to reduce the dimensionality and ease the computational cost of treating the complete dataset. The data was reduced to the first three principal components, retaining more than 90% of the total variance. Linear PCA dimensionality reduction was carried out using the python packaged Scikit-Learn.[123] The first two PC components where used as input to build low dimensional free-energy profiles (FES), while all the retained components were used as input for the PAMM[120] unsupervised clustering algorithm (Section 2.3.2). The clustering step allowed us to classify all structural motifs assigning a different label to each one for each of the three different "super"-CG fibres (fibre **1**-**3**). The parameters used for the PAMM training were kept identical, fspread = 0.30, quick-shift = 1, bootstrap-runs = 73, merger-threshold = 0.005, apart for the grid-size sample used for the density estimation which was 1000 points for fibre **1**, and 2000 points for fibres **2** and **3** (further details on the parameters of the PAMM clustering algorithm can be found in Ref. [120]). Lastly, to complete the analysis, we built a probability transition matrix for each fibre variant, by counting the labels transitions during the equilibrium MD simulation, and we used them to estimate the frequency of exchange between the structural motifs of the fibres.

3.1.3 Results

Defects and monomer exchange in BTA-based supramolecular polymers

Figure 3.3a reports preliminary results obtained *via* supervised clustering projected on two tailored CVs: the coordination number of each monomer core (a value of 2 indicates the coordination of perfectly stacked monomers) and the minimum distance from the other monomer cores (c being the stacking distance between two perfectly parallel neighboring cores). By setting to 3 the number of total clusters it is possible to recognise different physical states for each fibre (Fig. 3.3a).

In fibre **1**, we identify in black those monomers belonging to the bulk of the fibre, having coordination number 2 and unitary minimum distance c. In red those monomers populating the tips of the fibre, which have coordination 1 and minimum distance c. Finally, in blue those monomers that are not part the fibre body during the CG-MD run, having coordination zero and minimum distance $> 2c$. Interestingly, thanks to the sampling granted by the lower resolution CG model, it was possible to observe monomer exchange events even during an unbiased CG-MD simulation. The trajectory shows that for fibre **1**, the dominant exchange pathway originates mainly from the pre-existing fibre tips, or once in a while, from backbone cleavage that might occur during the dynamics. The formation of any bulk structural defect or disordered domain along the fibre backbone is not observed during the whole MD trajectory.

In fibre **2** and fibre **3** a common set of molecular states was observed: bulk ordered/stacked monomers (in black), monomers in the fibre tips (red), and bulk defects along the fibre (in green). The latter are present in both cases, but with a critical difference in number density, being way higher in the fibre **3** case. The average number of green monomers present in fibre **2** during the CG-MD simulation

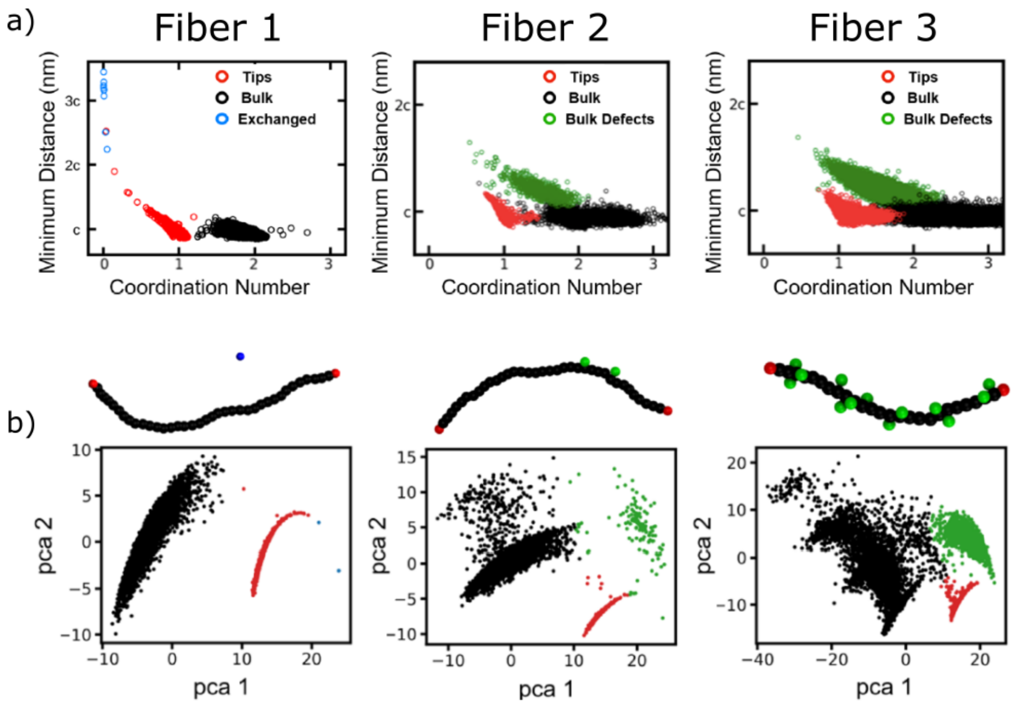


Fig. 3.3 Clustering results for the three "super"-CG fibre models, Fibre 1, 2, 3 (from left to right). a) Supervised spectral clustering, projected onto 2 main CVs, identifying different structural motifs: bulk monomers in black, fibre tips in red, bulk defects in green, and exchanged monomers (free in solution) in blue. b) Unsupervised PAMM clustering results, projected onto the first 2 principal components, obtained from the PCA of the SOAP feature vectors. Scatter plots are colored according to the main macro-clusters obtained from the PAMM analysis: bulk monomers in black, fibre tips in red, bulk defects in green, and exchanged monomers (free in solution) in blue.

is slightly less than 1 (~ 0.8), while ~ 12 for fibre **3** (this number was obtained by averaging the total number of clusters for each frame of the simulation). From a visual inspection of the trajectory it was observed that in fibre **2**, the green motifs, does not have a persistent nature and is not always populating the backbone of the fibre, but it rather dynamically emerge and disappear during the whole CG-MD. In the same way, for fibre **3** it was observed a much higher and persistent number of structural defects all along the fibre body, during the CG-MD. Notably, in these two low resolution models a cluster corresponding to the blue motif in fibre **1** is absent, as the increased solvophobicity of the monomers lowers the chances of spontaneous monomer exchange out from the fibre during the time accessible from the CG-MD simulations. These results fit well with the overall qualitative findings of previous enhanced sampling simulations,[31] showing that the event of a

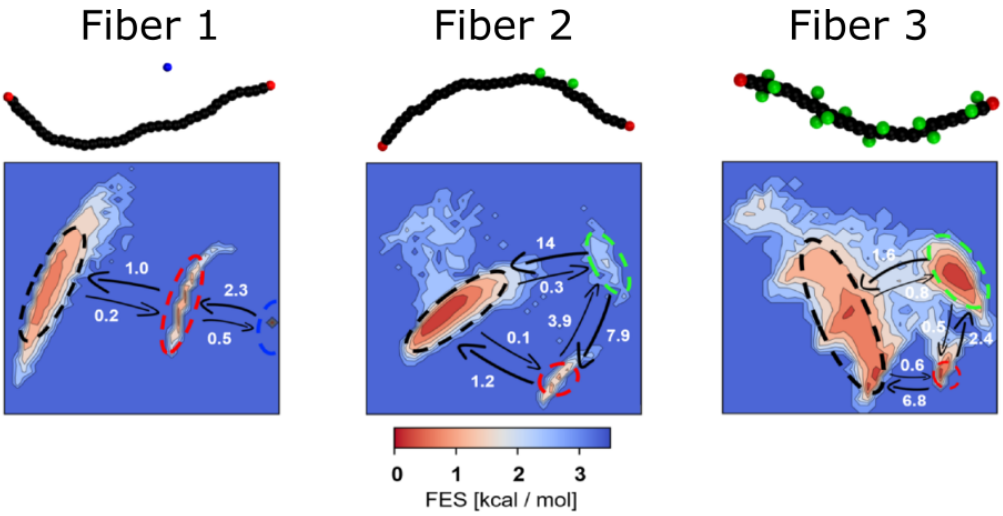


Fig. 3.4 Low dimensional FES built on the first two PC of the SOAP dataset for **fibre 1-3**, from left to right respectively. The area occupied by the main molecular motifs is represented as dashed circles, coloured as the respective motif. Black arrows represent the transition rates between said motifs, expressed in μs^{-1}. Above each plot is shown an MD snapshot for the fibre reduced representative structure coloured following the PAMM molecular motifs.

monomer exchanging out from the fibre becomes much slower as the non-directional interactions between the monomers are increased. Such results were obtained with an approach that assumes the possibility of defining an ensamble of CVs able to describe the defected states within the structural complexity of such assemblies. However, (i) this is not always the case and (ii) "human"-selected CVs are typically unique to a given system (*i.e.*, they can't be re-used to study different system having different topology). For these reasons and as a way to generalize our findings, here I have employed a computational recipe of descriptors and ML-based tools to classify the same fibre structures using the unsupervised SOAP-PCA-PAMM apporach. We analysed the same equilibrium CG-MD trajectories of the three fibres models following the general pipeline of steps described in Section 2.4. We computed the SOAP vectors associated to each monomer along the trajectories, employing a one-center reduced representation of the monomer (considering only the dipole carrying CG-bead of the monomers) as SOAP centers. Considering just this central single point was already proven to be enough to fully characterise the fibre backbone.[30] Lastly, the data were processed through a PCA dimensionality reduction and then clusterised using the PAMM unsupervised algorithm.

The molecular motifs (clusters) obtained *via* this unsupervised classification approach identify structural features and differences amongst fibres that are consistent with previous supervised findings (comparison between the panels a and b in Fig. 3.3), but crucially overcome the need to define specific CVs for their identification. From all the collected data, we computed a low dimensional free energy landscape for each of the fibres (*i.e.*, expressed as a function of the first two principal components), which is reminiscent of their equilibrium distribution of states, as captured by the descriptor of atomic environments (Fig. 3.4). Moreover, by following the monomer transitions between different clusters during the MD trajectories, we can estimate the transition rates between the identified molecular motifs (reported as black arrows on top of each FES in Fig. 3.4). This allows us to gain interesting information on how fast a structural domain changes during an equilibrium molecular dynamics. Although the computed rates have to be interpreted as only qualitative trends, since they result from low resolution CG models, they can be useful to compare different transitions between the states identified in a single fibre model, or to compare transitions among states across different models. Noteworthy, the green minimum of fibre 3 (right panel in Figure 3.4) clearly appears deeper, comparable with the depth of the black cluster (bulk monomers), indicating the higher statistical presence of this structural motif. Instead, in fibre **2** (central panel in Figure 3.4), the corresponding green local minima is very weak, ~ 2 kcal/mol higher in free energy compared to the global minimum, implying that bulk defects are a less energetically stable state, compared to the "perfect" fibre body. In particular, in fibre **2**, the ratio of annihilation/creation rates of bulk defects is ~ 40, whereas for fibre **3**, the two rates are very much comparable, ~ 2, as in this case they represent a more stable part of the structure identity. Finally, as for the supervised clustering results, the green domain is absent in fibre **1** and the blue domain is absent in the two others fibres. These results, once again, support the hypothesis that defects, either bulk defects or tips, act as fundamental hot spots for monomers exchange in/out, from a fibre, and that the presence of these defects can be controlled by a fine tuning of the competing non-directional and directional interactions between the building blocks that constitute the polymer (in this case by controlling the core solvophobicity).

In summary, in this work, we successfully collaborated to develop a "super"-CG molecular model for a generic supramolecular polymer, that allows for an easy accessible screening of single monomer parameters which affect the overall behavior of the polymer. We combined multiscale modeling, classical and enhanced sampling simulations, and unsupervised ML approaches to obtain a thorough characterization of the internal structure and dynamics of different types of supramolecular fibres. Furthermore, the analysis presented herein displays the agnostic and general applicability of the SOAP-PCA-PAMM computational analysis across different model resolutions. These findings demonstrate the intimate connection between structural defects and dynamic pathways, and that by controlling defects, it is possible to control the exchange of monomers in the fibres, in terms of both exchange kinetics/frequency and pathways.

3.2 Structural characterisation of a living supramolecular polymer

The work presented in this section was done in collaboration with the group of Prof. Max Von Delius (Institute of Organic Chemistry, University of Ulm, Ulm, Germany), which took care of all the experimental work. The joint work resulted in a publication in *Nature communications*, Ref. [60], with the title "Living supramolecular polymerization of fluorinated cyclohexanes".

The importance and complexity in the development of new pathways for living covalent polymerization has been discussed previously (Chap. 1). In Ref. [60] we collaborated to report a novel type of minimal molecular platform for living supramolecular polymerization based on a unique functionalizable scaffold. The considered building block posses a very large dipole moment, 6.2 Debye, which can be exploited to generate kinetically inert trapped monomeric states. Upon addition of specific seed units, the dormant monomers can engage in a kinetically controlled supramolecular polymerization, obtaining nanofibres with an unusual double helical structure and with a length that can be fine-tuned by controlling the ratio between the added seeds and the monomers concentration. All these behaviors were characterised by application of both experimental techniques and state-of-the-art computational approaches, including the discussed ML approaches. Relevant to the topics of the present thesis, I have contributed to Ref. [60] by applying the ML-based analysis to the LSP fibres simulated within this work, as detailed in the following.

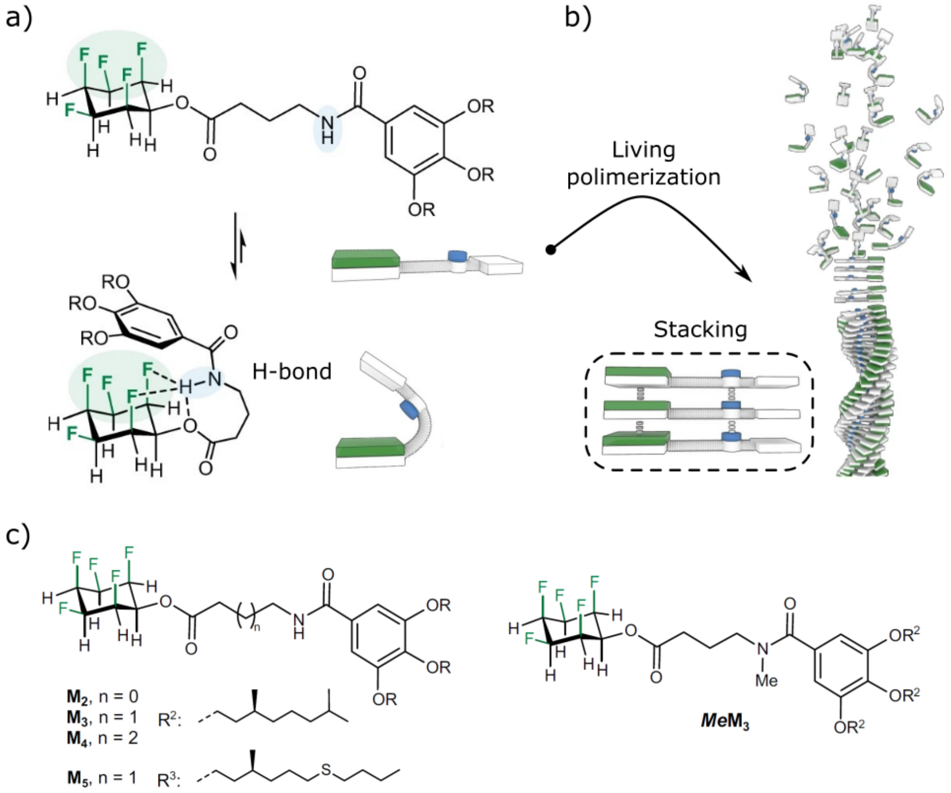


Fig. 3.5 Raffiguration of the living supramolecular polymerization for the 1,2,3,4,5,6-hexafluorocyclohexane scaffold. a) Chemical structure and schematic cartoon representation of the key monomer in a representative folded (bottom) and the unfolded (top) state. b) Illustration for the spontaneous assembly process following the addition of a seed. c) Chemical structure variants for the monomers used in our joint collaboration work.

3.2.1 Scientific context

As introduced in Chapter 1, controlled polymerization, has revolutionised polymer chemistry, endowing synthesised macromolecules with a structural complexity that can be surpassed only by natural occurring biopolymers. However, to get the most out of these artificial synthesis strategies, and making them a really versatile method, we are in need of a simple molecular base, namely: a minimalistic scaffold that can be easily functionalised, hence allowing for a fine tuning of the resulting self-assembled structure. For these purposes, simple derivatives of the all-cis hexafluorocyclohexane moiety (Fig. 3.5) have been chosen. The general monomer scaffold carries interesting properties toward the rational design of a synthetic LSP: an enormous dipole moment, 6.2 Debye,[129] a straightforward synthesis preparation,[130] and emerging potential as a supramolecular host[131]. Exploiting all these promising chemical features and by using a variety of different functional R groups, it is possible to alter the monomer ability to fold and its ability to interact with other type of substrates. This gives rise to a LSP that can successfully target both homopolymers and copolymers, with a considerable degree of control over the resulting fibre length and shape.

Initial experimental evidences on the supramolecular polymerization for different monomer variations of the all-*cis* 2,3,4,5,6-pentafluorocyclohexan-1-ol (Fig. 3.6a) suggested that one of the key factors in driving and activating the aggregation process is the ability to form intramolecular hydrogen bonds between units, which leads to the formation of a peculiar helical structure (Fig. 3.5b). The aggregation process was followed by our collaborator with time-dependent and temperature scanning circular dichroism (CD) and the final fibre aggregates visualised with atomic force microscopy (AFM), revealing that the all-*cis* $C_6H_6F_5$ motif and the amide bond are crucial for supramolecular polymerization. Moreover, the strength of such interaction was found to be heavily dependent on the length of the spacing, *i.e.*, number of carbon atoms, between the cyclohexane and the amide groups inside the monomer, as well as the availability of the amide hydrogen. Monomers having a shorter spacing showed a weaker CD signal for the aggregation, whereas the *MeM3* monomer (Fig. 3.5c, right side structure) showed no evidence of supramolecular aggregation, due to the absence of the amide hydrogen. Additional nuclear magnetic resonance (NMR) and infrared (IR) spectroscopy measurements revealed how the formation of $\mathrm{CH}\cdots\pi$ interactions between the all-*cis* fluorinated cyclohexane and the aromatic ring participate alongside the intramolecular H-bond to the folding mechanisms. Extensive details on the experimental setups and analyses can be found in Ref. [60] and supporting information.

In the article we employed computational methods to gather further evidences on the monomer folding mechanics: AA WT-MetaD simulations were used, for the **M2-4** and **MeM3** monomers, in explicit 84:16 v/v cyclohexane and chloroform solvent, to match the experimental conditions. These simulations allowed to bias the sampling for the folding/unfolding conformational change and to estimate the corresponding free energy landscape for each monomer (Fig. 3.6). The energy landscape, computed as a function of the 2 CVs, shows that the more stable configuration for **M3** is the folded state (Fig. 3.6, top panel, dark blue region characterized by low CV2 values, IV conformer). In this state, the hydrogen atoms of the pentafluorocyclohexyl moiety collapse onto the benzene ring, stabilised by the formation of h-bonds. Other local minima in the landscape are found to be favored either by the interaction of the

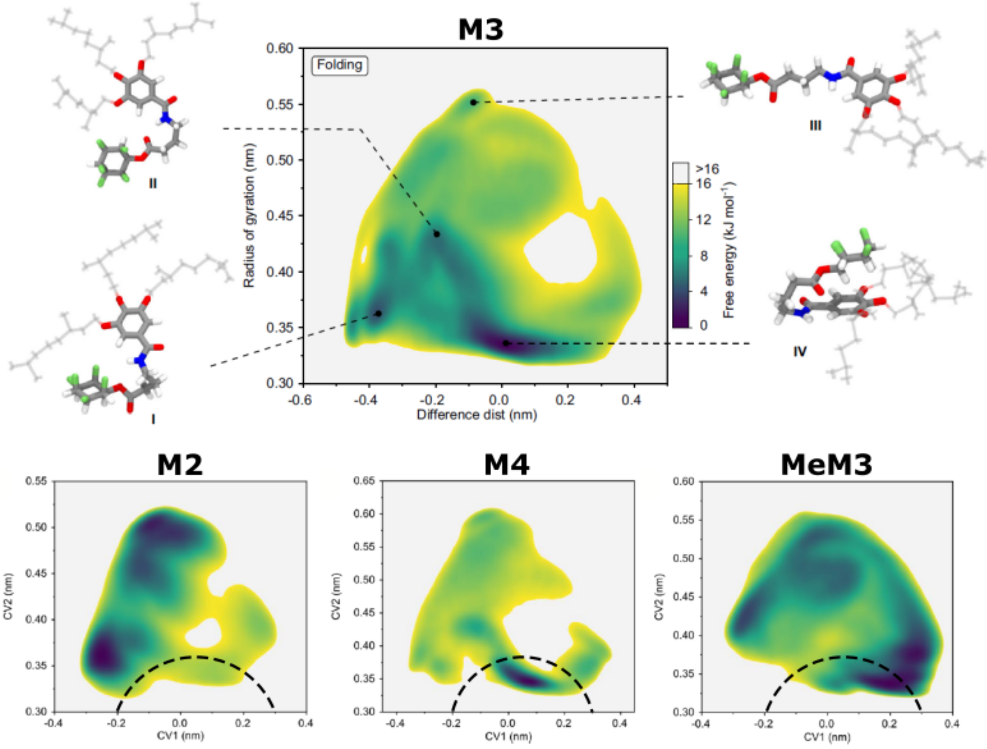


Fig. 3.6 Free energy surfaces (FES) estimated from AA-WT-MetaD simulation of some representative monomers variants studied in the work (for detail on the definition of the CVs and the simulation setups see the methods section of Ref. [60]). Blue colored areas represent the low free energy domains while the gray colored background is high in free energy or not accessible. The dotted lines highlights the portion of FES where the monomer is expected to be in a folded state.

fluorine atoms with the amide-H (I conformer) or by the formation of a hydrogen bond with the carbonyl group (II conformer). The same analysis was performed on other relevant monomers (Fig. 3.6, bottom panels), obtaining results in line with our expectation: the **M2** and **MeM3** monomers unfolded state is much more energetically favourable, due to the differences in length of the spacing and the occlusion of the amide hydrogen, both preventing the formation of stable h-bonds.

Once our knowledge on the folding dynamics of the monomer scaffold was consolidated, it was decided to test whether it could be possible to exploit those molecular properties to drive the polymerization process. A specific synthesis protocol was developed by our collaborators, that makes use of a well-defined molecular seed initiator to trigger the LSP (details of which are reported in Ref. [60]). By carefully mixing solutions of monomers and seed initiators, they were able to synthesise many different fibres, including homopolymers and block-copolymers variants, all with a surprisingly high efficiency in controlling the size distribution of the obtained fibres. The fibres structures, arrangement and size distributions were investigated by our collaborators by means of atomic force microscopy (AFM), CD spectroscopy measurements and transmission electron microscopy (TEM). Results suggests that the molecular stacking appears to be driven by a series of dipole-dipole and H-bond interactions, crucially arranging pairs of neighbouring stacks into an antiparallel fashion, which presumably is enough to cancel out the large macro-dipoles carried by the single monomers.

We then turned to MD simulations to gather more information on the supramolecular monomer stacking and to study the single fibre monomer arrangements. First, the two possible monomer configuration were tested: where the **M3** monomers are arranged into a parallel or antiparallel fashion, comparing the energetic stability of these two structures (detail on the MD simulation are available in Ref. [60] and supporting information). Result showed that a parallel axial stacking would be stabilized by the H-bond and the dipole-dipole interactions between the pentafluorocyclohexane dipoles. However, the assembly shows an implicit amphiphilic character, with all solvophilic tails pointing toward one side and all the solvophobic heads on the other side, causing an unbalanced situation. On the other hand, an antiparallel axial stacking would still allow for favorable H-bonding between the monomers, while guaranteeing a more uniform solvophilicity of the fibre. In both cases, the single filament was found to be unstable during a MD run, suggesting that a hierarchical self-assembly of multiple filaments is a likely event, consistent with the obtained experimental evidence. Secondly, the equilibrium stability of different types of fibres were tested. In particular, the aim was to investigate which configuration of monomers gave the higher stability in both homopolymers and copolymer fibre arrangements. Two homopolymer fibre of **M3** and **M5** monomer and three different copolymer models where compared *via* equilibrium MD simulations. The homopolymers, showed high stability during MD simulations and the spontaneous formation of an helical pitch, resembling the experimental evidences. The three copolymer configurations showed interesting results in comparison to the homogeneous counterparts: the blocks configuration showed an energy difference very close to the ones of the two system separated, while higher grade mixing (columns or random fashion) appears to be more energetically unfavorable. A complete segregation in separated **M3** and **M5** homopolymers, as well as 1:1 homogeneous intermixing appear to be

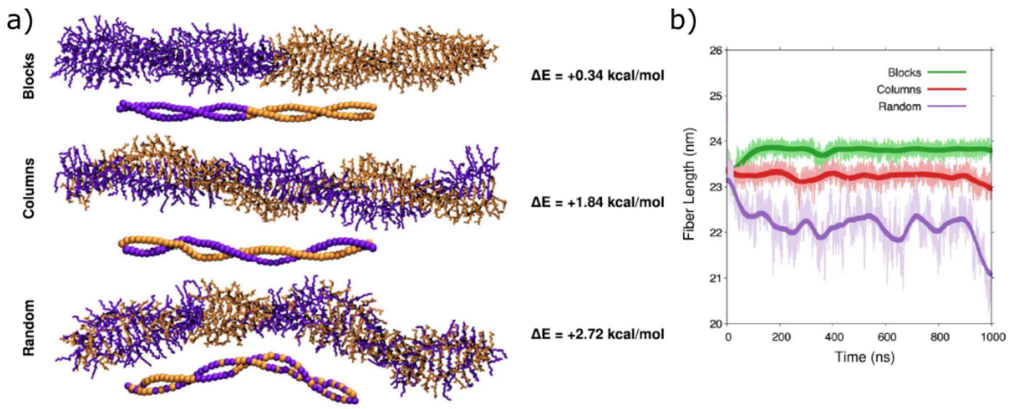


Fig. 3.7 MD snapshots for the equilibrated structures of **M3**-**M5** copolymers models, with the **M3** (in orange) and **M5** monomers (violet) arranged in blocks (top), intertwining columns (middle), and randomly mixed (bottom). Below each fibres are shown how the fluorinated cyclohexane centers are arranged in each copolymer. The energies of the different **M3**-**M5** mixing schemes relative to completely segregated **M3** and **M5** homopolymers (ΔE) are reported beside each equilibrated copolymer model. b) Length of the Blocks (in green), Columns (red) and Random (violet) copolymer models as a function of MD simulation time.

very unlikely from an entropic point of view. Finally, we compared the stability of these fibres during unbiased MD runs and we observed that the blocks model is the more persistent, more than the columns or random ones, as it preserves the fibre length (Fig. 3.7b green curve) for the whole duration of our runs, while the other two arrangements evolve towards a more distorted/bent conformation (Fig. 3.7b red and violet curves).

In the context of the present thesis work, I have directly contributed to the characterisation of these supramolecular fibre structures by employing our unsupervised SOAP-PCA-PAMM ML approach, which proved to be succesful in the characterisation of supramolecular polymers.[30, 31]

3.2.2 Methods

In this section we will focus only on the details of the application of the unsupervised SOAP-PCA-PAMM ML approach to the five total supramolecular fibres simulated in Ref. [131], *i.e.*, two **M3** and **M5** homopolymers and three copolymers configurations of the same (as decribed before and depicted in Fig. 3.7).

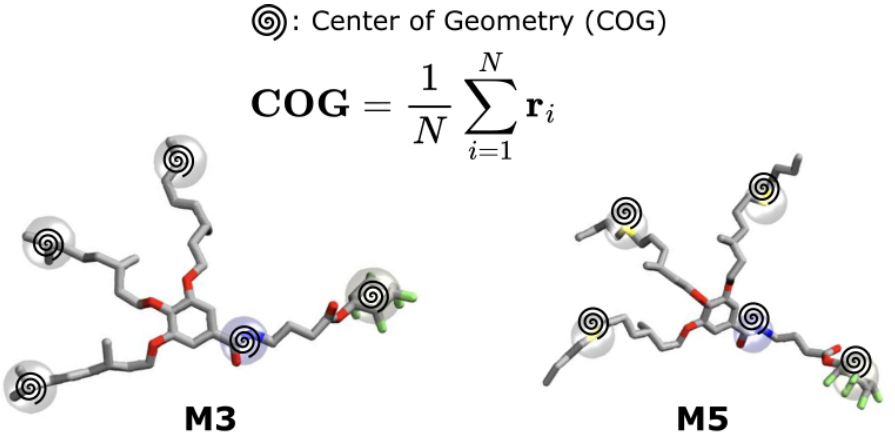

: Center of Geometry (COG)
$\mathbf{COG} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{r}_i$
M3
M5

Fig. 3.8 Definition of the descriptor centers for the SOAP analysis on the **M3** monomer (left side) and **M5** monomer (right side). The spiral represent taking the COG of the enclosed atoms.

Structural motifs dynamics of the fibres

We employed our unsupervised SOAP-PCA-PAMM approach to investigate the structure and dynamics of the **M3** and **M5** monomers inside the corresponding supramolecular polymers. The configuration and molecular environments of each monomer, along a MD trajectory, were encoded into a feature vector by means of the SOAP[105] descriptor. The SOAP vectors of each molecule were computed considering a total of five centers per monomer, namely: the center of the cyclohexane group, the center of the amide group, the alkyl centers in the three tails in **M3**, and the sulfurs atoms on **M5**, which occupy an equivalent position of the alkyl moieties in **M3** (Fig. 3.8). The SOAP analyses were carried out with the Python package DScribe,[128] setting the input parameters as *rcut*= 60 Å, *nmax*= 5, *lmax*= 5, and leaving the other parameters as default. The dimensionality of the SOAP dataset was reduced by means of PCA, keeping only the first three principal components, which account for up to 86% of the total dataset variance. Linear PCA dimensionality reduction was performed using the Python3 package Scikit-Learn.[123] The PAMM[120] unsupervised clustering was used to classify the low dimensional dataset, extracting the main molecular motifs of each monomer during the MD simulations. Finally, the analysis was completed by estimating the frequencies and rates of exchange between the clusters, by counting the number of transitions of each monomer along each snapshots of the trajectory and dividing them by the total number of transitions (for the frequencies) or the time window between each snapshots (for the rates, in our case $\Delta t = 0.1$ ns).

3.2.3 Results

In depth structural analysis of the fibres

We now review the results of applying the unsupervised SOAP-PCA-PAMM approach to this supramolecular polymer system. As introduced in Sec. 2.2, the SOAP[105] analysis allows us to classify the local environments that surround each monomer site (defined as the centres depicted in Fig. 3.8) based on their spatial configurations. We then used PCA to reduce the dimensionality of the SOAP feature vectors and we classified the low-dimensional data according to their main "molecular motifs" using the unsupervised clustering algorithm PAMM[120]. The results of the clustering are reported in Figure 3.9, were the cluster labels are projected onto the first two PCs of the SOAP data each monomer. By transferring the obtained cluster memberships onto the original atomic structure (Fig. 3.9a), we observe that the red and blue molecular motifs correspond to the cyclohexane ring and to the amide respectively, and the other three clusters (gray, cyan and green) are linked to the monomer tails, and to the surface of the fibres. Finally, monitoring the label transitions along the trajectory we could follow the dynamic evolution of the atomic environments in form of transition probabilities, which highlight the most probable motifs exchange events (Fig. 3.9d).

This analysis demonstrates how in the time spanned by the MD simulations the core section of the aggregate appears substantially static compared to the surface parts, *i.e.*, represented by the absence of transitions between the red and the blue macro-clusters. On the other hand, the dynamic interconnections between the gray, cyan and green clusters suggest the presence of exchange pathways happening on the surface of these fibres. Moreover, the interconnection between those clusters increases going from left-to-right (Fig. 3.9d) indicating that the surface of these fibres become more dynamic and disordered increasing the degree of mixing of the two building blocks.

In conclusion, in Ref. [60] we demonstrated the potential toward supramolecular polymerisation of a new interesting monomer scaffold. We were able to demonstrate

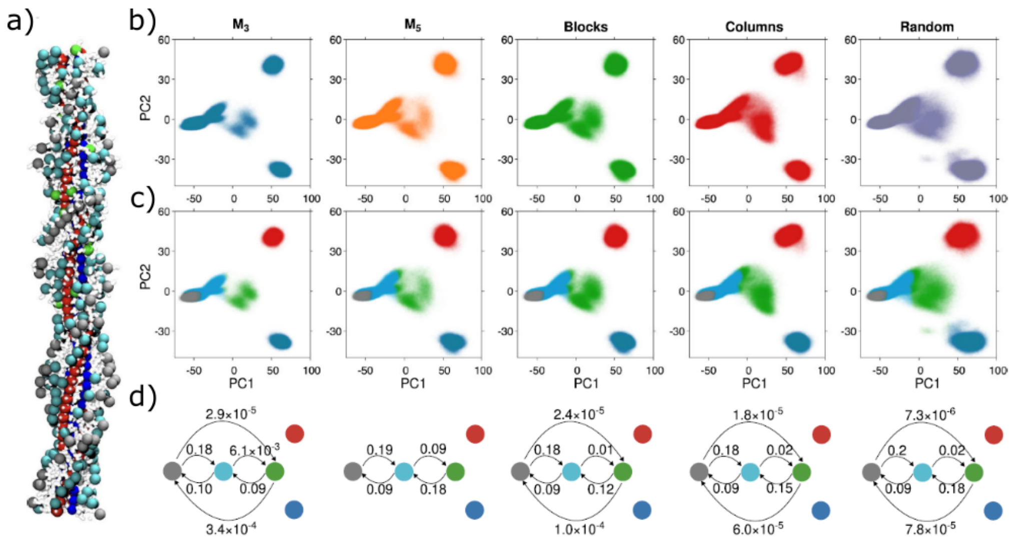


Fig. 3.9 SOAP-PAMM structural analysis of supramolecular block copolymers. a) Structure of the blocks fibre colored based on the molecular motifs identified by the SOAP-PAMM analysis. The fluorinated cyclohexane and amide groups in blue and red, the alkyl side chains of the monomers in gray, green and cyan. b) PCA data, obtained from the SOAP vectors, for **M3** and **M5** homopolymers, and for Blocks, Columns and Random copolymers (left-to-right). c) Unsupervised PAMM clustering of the PCA data. d) Dynamic interconversions for the different structural motifs identified by the PAMM unsupervised clustering. Arrows shows the transition pathways with the relative event probability reported on top.

how the monomer configurational changes (*i.e.*, folded and unfolded states) enable the polymerisation process, and how structural chemical changes enable higher degrees of control over the outcome fibre structure. Our computational approaches proved to be valuable in describing both the single monomer and the final aggregate behaviors, results which were often backed up by experimental evidences. Through the use of the SOAP-PCA-PAMM approach I contributed providing a highly detailed structural characterisation of the fibres, without having the need to defined special CVs or compute many quantities, as all the required information is capture by the SOAP descriptor alone.

Chapter 4

Classifying self-assembled materials *via* machine learning of defects

In Chapter 3 we have presented and discussed the potential of a novel computational approach, based on unsupervised ML, to detect and characterize structural motifs in soft-assembled materials. That approach was used to showcase the role of defects and the importance of having a way to detect them in supramolecular aggregates. However, as it is, it does not allow for a direct comparison or quantification of the (dis)similarity between different materials or families of materials. In the current chapter, we introduce a step forward in our computational analysis that allows us to classify and compare supramolecular self-assembled structures in an objective way, based on the structural motifs identified *via* unsupervised ML (Fig. 4.1).

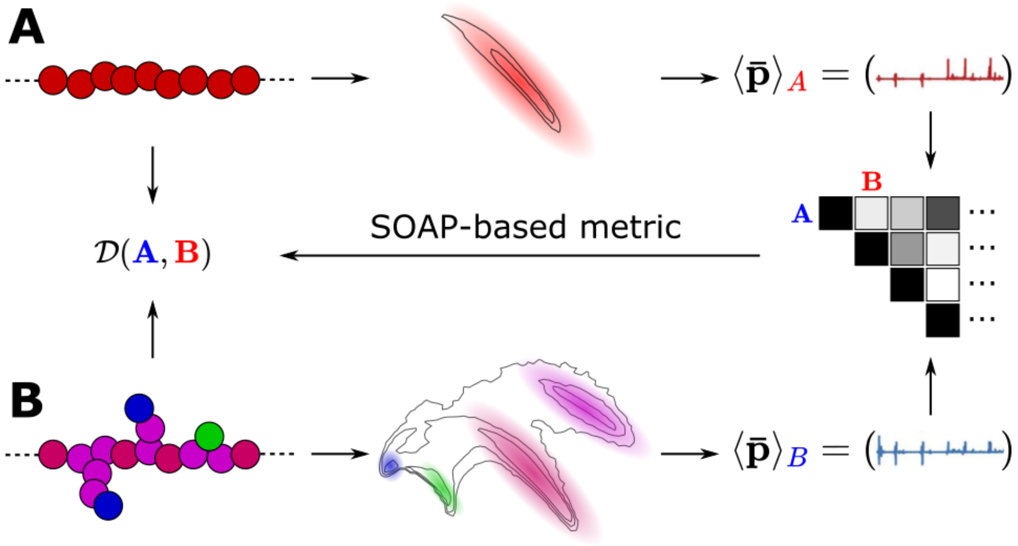

A
$\langle\bar{\mathbf{p}}\rangle_A = (\ )$
B
A
SOAP-based metric
$\mathcal{D}(\mathbf{A},\mathbf{B})$
B
$\langle\bar{\mathbf{p}}\rangle_B = (\ )$

Fig. 4.1 Distance-based classification of two generic supramolecular assemblies. The two structures **A** and **B** are first completely characterized by mean of the SOAP-PCA-PAMM unsupervised approach, which yields the structural motifs as cluster memberships (sketched isolines in the low dimensional PC space with coloured membership). Through the application of a SOAP metric we can effectively measure the distance between the two supramolecular assemblies, based on their average population of structural motifs.

Using the local SOAP spectra gathered from the SOAP-PCA-PAMM unsupervised analysis, we can compute a global descriptor that is able to capture the essential features of the structural environments that populate a target structure. Then, different global fingerprints will be used as inputs for a SOAP-based metric, which can efficiently compare between various structures regardless of their specific physical or chemical origin. This approach represents a natural improvement from our previous work and opens up new possibilities in the classification of soft-assembled materials.

In the next section we will introduce one of the main results of this thesis work, presenting the so-called data-driven "defectometer", a structure-based ranking system that allows the comparison and characterisation of (dis)similarities that might exist amongst generic materials, across various families and dimensionalities. Finally, in the last section, we will present a different data-driven approach, which always makes use of said ranking approach, but is adapted to compare and classify the accuracy of many known lipid FFs, highlighting some key intrinsic limitations in implicit versus explicit solvent models.

4.1 Development of a general purpose data-driven “defectometer”

The work presented in this section resulted in a publication in *Nature Communication Chemistry*, Ref. [71], with the title "Classifying soft self-assembled materials *via* unsupervised machine learning of defects". Soft self-assembled materials, like fibers, micelles, vesicles, and other, all exhibit both ordered and high variable structural domains, with defects that form and repair continuously, conferring to them unique adaptive properties. However, comparing and classifying such materials based on their complex internal dynamics proved to be challenging. This study introduces a data-driven analysis that makes use of high-dimensional SOAP descriptor fingerprints, collected from equilibrium MD simulations, to compare various families of soft supramolecular assemblies. This approach yield a so-called “defectometer”, a computational analysis tool, that can classify different types of supramolecular materials based on the statistical emergence of ordered and disordered molecular environments.

4.1.1 Scientific context

As we saw in details in the previous chapters, supramolecular structures, molecular aggregates composed of building blocks that self-assemble *via* non-covalent interactions, represent the key substrate for biological systems (membranes, micelles, protein fibers, etc.), thanks to their distinctive physical and chemical traits. A major goal in the study of self-assembled materials is rationalising how specific changes in the structure of the self-assembling building blocks (input) can affect the global properties of their supramolecular architecture (output). Gaining such knowledge would pave the way toward the creation of new functional materials, possessing tailored properties for specific purposes. In the last two decades, many efforts have been made in this direction, exploiting the enormous technological progress of both experimental and computational techniques.[5, 132–134] However, given the complexity and the dynamical nature of such systems, a direct link between the global properties of a material and the local features of the constituent monomers remains often daunting to attain.[135–138]

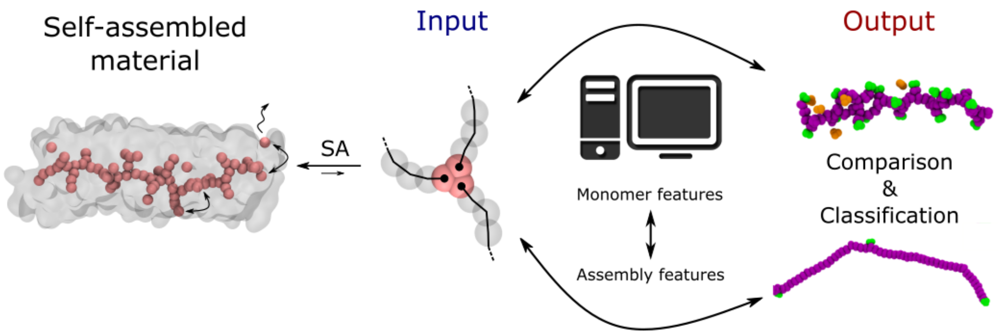


Fig. 4.2 Visual representation of a ML-based approach toward the characterisation of supramolecular soft-assembled materials. A given soft material is spontaneously created by the SA of monomers, which possesses a set of local chemical/physical properties that will imprint a set of global properties on the self-assembled material. By means of ML techniques it is possible to gather information on the effect of such local properties, comparing and classifying different structures, and eventually learn how to engineer a specific monomer to a specific material property.

"Human-based" analyses often rely on simple observables or low-dimensional descriptors, usually defined based on the experience and on those features directly readable by visual inspection. This might lead to fundamentally biased predictions, where the chosen low-dimensional variables overlook important degrees of freedom of the system. To overcome these limitations, data-driven ML approaches, such as, unsupervised clustering, dimensionality reduction, etc., proved to be particularly useful, allowing for an almost completely unbiased characterization of both structural and dynamic properties of any given molecular system. In Chapter 3 we saw how the SOAP power-spectra proved to be a very efficient structural descriptor, as it encodes the information of an atomic environment into a rich, high-dimensional and agnostic feature vector, ready to be used as input for other analysis algorithms.[30, 139, 140]

In this study, we push forward this data-driven characterisation to allow comparing a wide range of various soft-matter systems. Using the SOAP descriptor and unsupervised clustering, as well as a SOAP-based metric, we build a data-driven analysis workflow, that effectively acts as a "defectometer", namely: an analysis tool by which is possible to measure and compare various structures based on their molecular structural environments, allowing us to compare between fundametally different assemblies and obtain an unbiased classification of supramolecular soft-assembled materials taking into account of their structural and dynamic features. Notably, this approach differs significantly from our previous SOAP-PCA-PAMM analysis: by focusing directly on a SOAP-based comparison we avoid the possible distortion

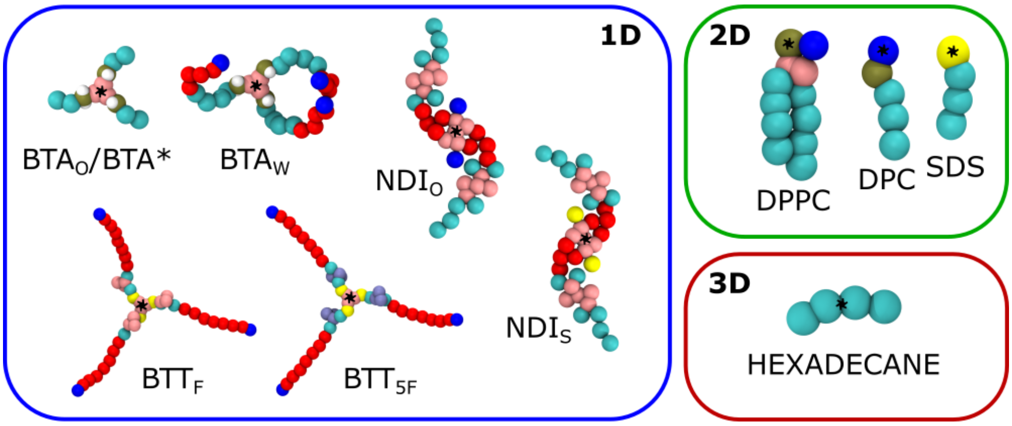


Fig. 4.3 Visualisation of the CG-FF monomers models of the studied self-assembled materials, grouped according to the effective dimensionality of the corresponding soft aggregate they produce. The black asterisk, in each molecule, represents the single center chosen to represent the structure for the SOAP calculation.

brought on the data by the low-dimensional projection, granting a more unbiased comparison between different dataset. More in details, throughout the analysis, we have simulated, characterised and compared supramolecular polymers such as fibres of benzenetricarboxamides (BTA), benzotrithiophene (BTT), or naphthalenediimide (NDI), but also micelles and bilayers made of lipids and surfactants, as well as spherical assemblies of hexadecane (see Fig. 4.3).

4.1.2 Methods

Supramolecular material families

All the CG-FF molecular models (Fig. 4.3) were parametrised using literature available models or following the Martini[85] CG-FF standard parameters, if not stated otherwise.

- *Supramolecular polymers or fibres (1-D assemblies)*. The supramolecular polymers studied belong to three main families, distinguished by a specific chemical structure of the monomer functional cores, for which we additionally considered few variants each.

  (i) 1,3,5-Benzenetricarboxamides (BTA) based monomers. We considered three variants: the water soluble BTA$_W$, the organic solvent (*e.g.*, octane) soluble BTA_{C8}, and an intermediate case, BTA^* derived from BTA_{C8} by altering the directional interaction of the monomers. The Martini-FF-based parametrisation of the three variants is the same as the one reported in Ref. [30]. The two explicit CG-FF solvents used in the simulations were water and octane, parametrised according to the Martini-FF standards.[141] For the calculation of the SOAP descriptor we reduced each of the three monomer structures to a single center, located in its COG (Fig. 4.3).

(ii) Core-substituted naphthalenediimide (NDI) based monomers. We studied two different variants of NDIs, which differ in the identity of the atoms on the side of the core structure: NDI_O (oxygen) and NDI_S (sulfur). The parametrisation of the two monomers was taken from the one presented in Ref. [70], based on the Martini-FF parameters. In both cases the explicit solvent used was cyclohexane, at the CG-level of description. For the calculation of the SOAP vectors we reduced each of the two monomer structures to a single center, located in its COG (Fig. 4.3).

(iii) Benzotrithiophene (BTT) based monomers. We studied two variants of these monomers, differing for the nature of the three amino-acids attached to the aromatic core of the molecule: BTT_F (L-phenylalanine) and BTT_{5F} (pentafluoro-L-phenylalanine). The parametrisation of the two monomer models was taken from the one available in Ref. [69]. In both cases the solvent used was explicit CG-FF water, parametrised following the Martini-FF standards. For the calculation of the SOAP vectors we reduced each of the two monomer structures to a single center, located in its COG.

- *Micelles and membranes (2D assemblies).* The lipid molecule selected as reference case for the two-dimensional aggregate is the dipalmitoylphosphatidylcholine phospholipid (DPPC), for which we prepared homogeneous membranes at three different temperatures, across its phase transition, *i.e.*, $273, 293, 323$ K. Moreover, we selected as reference case for micellar aggregates, we selected two surfactant molecules: the Dodecylphosphocholine (DPC) and the Sodiumdodecylsulfate (SDS). The explicit CG solvent used for these molecules is water. For all the monomers we employed a single-center approach to compute the SOAP vectors, choosing a CG-bead located in the head portion of each amphiphilic molecules: the “PO4” (DPPC), “PO4” (DPC) and “SO3” (SDS) beads (Fig. 4.3).

- *Spherical nanoparticles (3D assemblies).* The molecule chosen as representative case for the spherical 3-D aggregate is the hydrocarbon Hexadecane (HEXA). The explicit solvent used for the simulations was CG water. As center for the SOAP representation we chose the COG of the two central beads of its CG (Fig. 4.3).

MD simulations

All the simulations, for all the systems studied, were performed using the MD software package GROMACS[142], version *2018.6*, from the setup of the simulation boxes, to the equilibration and production runs. PBC conditions were enforced during all MD runs to both limit the finite-size effects, and also to simulate seemingly infinite aggregate when required (*e.g.*, supramolecular polymers and lipid membranes). In fact, the supramolecular polymers were prepared from perfectly ordered fibres (following the structures provided by the appropriate literature), crossing the periodic boundaries og the simulation box so that the two extremities were effectively bound together. After this initial configuration was prepared, a standard minimisation and equilibration MD protocols were applied. The minimisation run was performed with a steepest descent algorithm until energy and force convergence to machine accuracy. The equilibration was performed in NPT condition for a few of μs. The lipid membranes, instead, were prepared by pre-arranging perfect bilayers of lipid (again, through the periodic boundaries of the simulation box) following the official equilibration protocol given by the CHARMM-GUI[143, 144] software interface and also described in Ref. [114]. Lastly, the micellar and nanoparticles structures were prepared *via* spontaneous self-assembly from the respective dispersed monomers. Once we obtained the equilibrium structure for each system, we proceeded with the production CG-MD runs, simulating every system for an additional 2 μs of CG-time. For all the CG-MD runs (equilibration and production) we used a 20 fs timestep and a sampling window of 1 ns. All the system were simulated at 300 K, except for the lipid membrane cases, where we simulated the system at different temperatures, across the gel-to-liquid transition. The temperature was maintained constant through the V-rescale thermostat[145] with a coupling parameter of 1.0 ps. The pressure was maintained constant at 1 atm by application of the Berendsen barostat[146], with a coupling parameter of 2.0 ps. When dealing with finite-sized aggregates, like the micelles and the nanoparticles, the pressure scaling was isotropic in all directions of

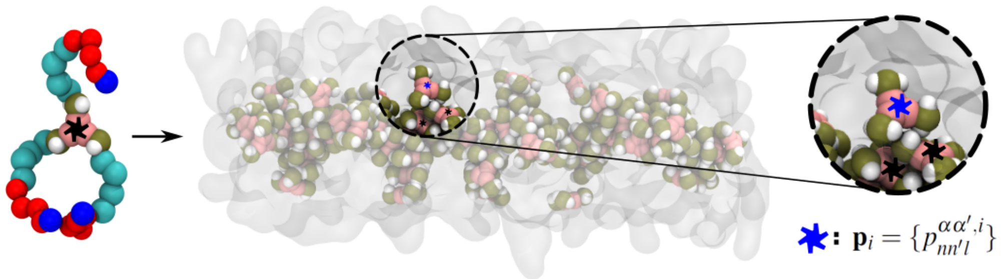


Fig. 4.4 Visual representation for the SOAP calculation on the structural reduced assembly. From left to right, first, the position of the centre is define on the chose monomer: the COG of the core part of the BTA$_W$ monomer (black asterisk, see Fig. 4.4 for a summary of all the defined SOAP centres). Thus, each monomer, in the assembly, will be represented by the position of the chosen single point, instead of its total amount of CG-beads. The SOAP power spectra of the *i*-th monomer (blue asterisk) will only take into account of the other surrounding monomers inside the radial cutoff (other black asterisks).

the simulation box. For the "infinite" aggregates, supramolecular polymers and lipid membranes, we adopted a semiisotropic scaling of the barostat, decoupling the x/y dimensions from the z one.

Description of molecular environments

We employed the SOAP descriptor of atomic environments to gather structural information of the studied structures (see Chap. 2 for a theoretical introduction on descriptors). The descriptor was computed for each of our systems in a "reduced" configuration, *i.e.*, only a representative centre for each monomer was accounted in the SOAP computation (Fig. 4.4). The SOAP power spectra were computed using the python package DScribe[128]. Identical parameters were used for all the calculations: `nmax,lmax`$= 8$, three different cutoff values `rcut`$= 0.8, 1.6, 3.0$ nm, and all the other parameters left as default options. The three different cutoff values are necessary as we are dealing with various hierarchies of self-assembled monomers. Figure 4.5 reports the relationship between the cutoff radii and the radial distribution functions (RDF) of all the systems considered. All the RDFs are computed from the same reference centers used in the SOAP calculation. From Ref. [30] we know that a value of `rcut`$= 0.8$ (consistent with the first neighbours peak, Fig. 4.5) is an optimal radius for the characterization of supramolecualr polymers at CG resolutions, and an increase of said distance does not have a sizeable impact on the output of the

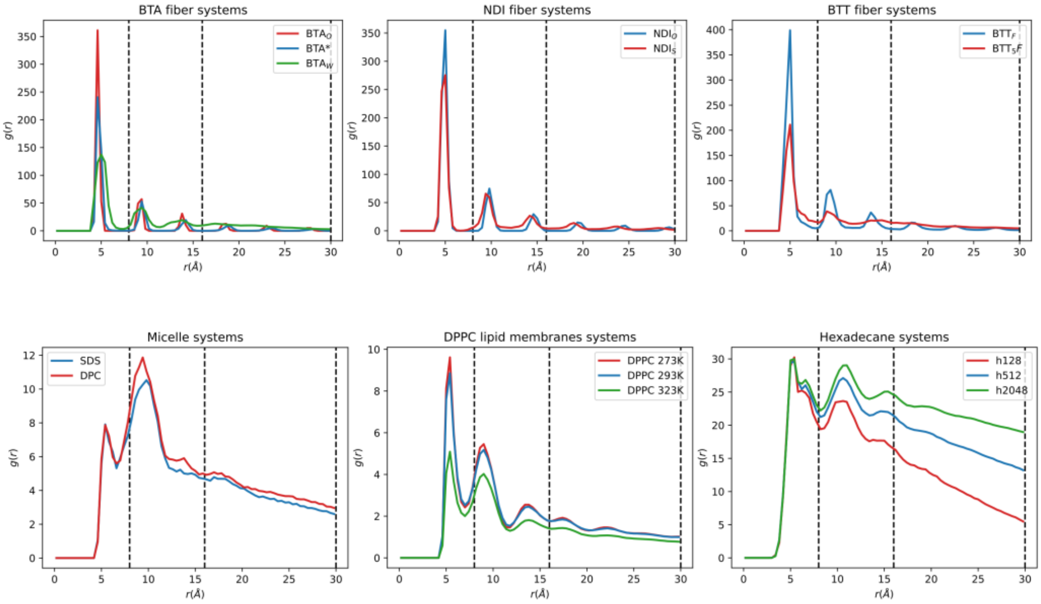
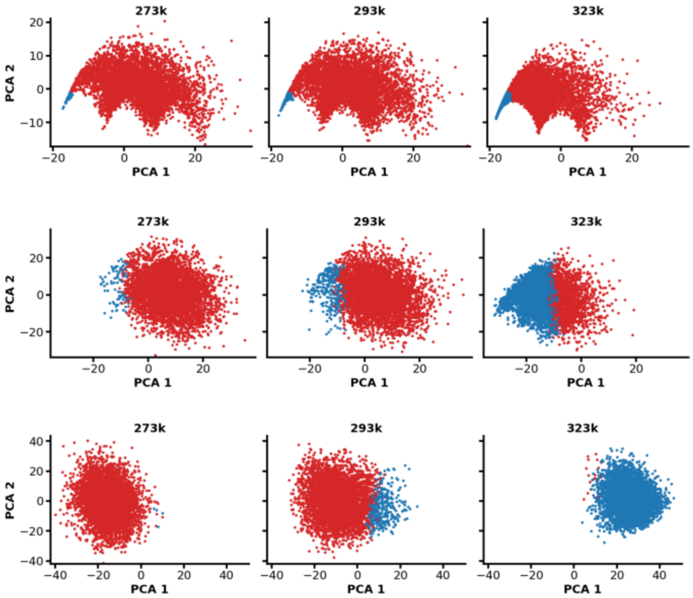


Fig. 4.5 Relationship between the radial distribution function and the SOAP cutoffs used in our analysis. The RDFs of all the different systems simulated are reported in the different panels.

SOAP analysis. We can expect this trend to hold also for the other supramolecular polymers, as their FF resolution is the same (*i.e.*, Martini CG-FF) and their structure is reduced in the same way; their RDFs are very similar to each others (Fig. 4.5 top raw). However, when considering higher order assemblies this does not necessarily hold. As a key example, Figure 4.6 reports the effect of changing the cutoff radius on a DPPC lipid bilayer, and the consequent ability to capture the phase transition phenomenon of the SOAP descriptor. In this case, a `rcut`= 3.0 nm appears to be necessary to obtain a faithful representation of the two phases, *i.e.*, the SOAP-PCA-PAMM analysis can distinguish the gel lipid organisation from the liquid one. Overall the choice of the cutoff radius is an important step in the preparation of the analysis, as different values impact both the accuracy and the computational scaling of the workflow.

Building of a training set

Crucial to our comparative analysis is the construction of a “shared” dataset, that comprises SOAP feature vectors from all the individual systems datasets we want to include in the comparison. This shared dataset will be then used as input set for the



$(r_{cut} = 8\AA)$	DPPC_{273K}	DPPC_{293K}	DPPC_{323K}
Red	0.993	0.983	0.831
Blue	0.007	0.016	0.169

$(r_{cut} = 16\AA)$	DPPC_{273K}	DPPC_{293K}	DPPC_{323K}
Red	0.985	0.943	0.665
Blue	0.015	0.056	0.335

$(r_{cut} = 30\AA)$	DPPC_{273K}	DPPC_{293K}	DPPC_{323K}
Red	< 0.999	0.946	0.004
Blue	> 0.001	0.054	0.996

Fig. 4.6 Effect of the cutoff on the 2-D hierarchical assembly studied in our comparison.

dimensionality reduction algorithm, that will project all the data points onto the same low-dimensional space. The usage of comprehensive data coming from multiple systems allows the data-driven method to take into account the overall data diversity and compare the molecular environments identified across the different systems (as they are projected in the same low dimensional space). Given the high amount of data we collected from our MD simulations, the dataset is built by taking a standard and representative sample from each individual system.

Dimensionality reduction

To project the high-dimensional SOAP vectors onto a low-dimensional representative space we chose the PCA[110, 111] approach (already introduced in Sec. 2.3 and discussed throughout the previous results). We here performed the dimensionality reduction *via* the python class `sklearn.decomposition.PCA()` from the python library Scikit-learn[123]. In all the occurrences, we kept up to the first three PCs, which in all cases account for more than 90% of the total data variance.

SOAP-based metric

The SOAP-based metric (introduced in Sec. 2.2.2) is used to directly measure the (dis)similarity between two given SOAP feature vectors. To this end, we compared the different systems by measuring the SOAP distance of each pair of SOAP simulation-averages, $\langle \bar{\mathbf{p}} \rangle_i$ (Sec. 2.2.1), for each of the considered systems. The final results are then arranged in matrix form to better appreciate the relationships of the compared structures, as simplified in figure. 4.1 at the start of the chapter. In our cases the distance value takes the form

$$d^2_{i,j} = 2(1 - \langle \bar{\mathbf{p}} \rangle_i \cdot \langle \bar{\mathbf{p}} \rangle_j), \tag{4.1}$$

where the “SOAP” subscript was omitted for clarity. This distance represents a measure of the overlap of the SOAP simulation-average power spectra, which is unitary if the two spectra are perfectly identical (*e.g.*, $d^2_{i,i} = 1$) and goes to zero if the two spectra are completely different. Effectively, given that we are going to use data taken from equilibrium simulations, the distance outputs a measure of how similar are, in average, two populations of structural motifs.

Unsupervised clustering

Once reduced to a low-dimensional space, to classify the underlying data patters we use the PAMM[120] unsupervised clustering algorithm (see Sec. 2.3 for an introduction on the method). All the clustering analyses were conducted using a python custom wrapper of the original PAMM presented in Ref. [120]; the code is available online at https://github.com/GMPavanLab/gmplabtools.

Moreover, we also used a basic hierarchical clustering algorithm to further classify the data obtained from the SOAP-based matrix approach. This step allowed us to group systems based on their measure of (dis)similarity, resulting in a tree-like plot showing the hierarchy of the various connections. The hierarchical clustering step was performed using the open source Python library Scikit-learn[123], specifically the class `sklearn.cluster.AgglomerativeClustering()` with the `single` linkage method.

4.1.3 Results

In the following we will present the results gathered by employing the so-called "defectometer" comparison tool. First, we will build a sizable dataset made up by 14 different molecular structures in total, collecting different families and dimensionalities. For each of the different aggregate family we will characterise the equilibrium behavior of its components using the well-tested SOAP-PCA-PAMM approach. In parallel, each individual system SOAP dataset will be exploited for comparison purposes. We will compute a single global SOAP feature vector (the simulation-average) for all the structure under study and we will compare these new general fingerprints by means of a SOAP-base metric approach. Thus, measuring their pair distances and uncovering their (dis)similarity relationships. Notably, this approach differ quite significantly from the SOAP-PCA-PAMM one as it focuses directly on the global (average) descriptor fingerprints, avoiding possible distortions that the PCA projection might causes, and it is used primarily as a comparison tool and not as a pure characterisation tool. Nonetheless, we believe that these two approaches, when used together, define a complete analysis tool providing accurate information on the relationships that exists between a single or many different molecular systems.

Comparing variants of supramolecular polymers

We start by reproducing a recent result on the equilibrium behavior of BTA based supramolecular polymers reported in Ref. [30]. To this end, we simulated, at equilibrium conditions, three fibres variant made by a polar solvent soluble BTA_W monomer, an organic solvent soluble BTA_{C8} monomer, and a slightly mutated version of the latter monomer with weakened directional interactions, BTA^*. We then computed the local SOAP feature vectors for all the monomers along the equilibrium MD trajectories of each fibres and we projected each system onto the same, shared, low-dimensional space *via* PCA. Finally, we classified the dominant molecular motifs by means of the PAMM unsupervised clustering algorithm. The results obtained are shown in Fig. 4.7.

As expected, the BTA_W fibre possesses a very rich and diverse internal dynamics (*i.e.*, the graph of Fig. 4.7 has more connections), with numerous structural motifs that coexist on the fibre body. Conversely, the BTA_{C8} variant, produces substantially defect-free fibres with negligible internal dynamics. The BTA^* monomer variant,

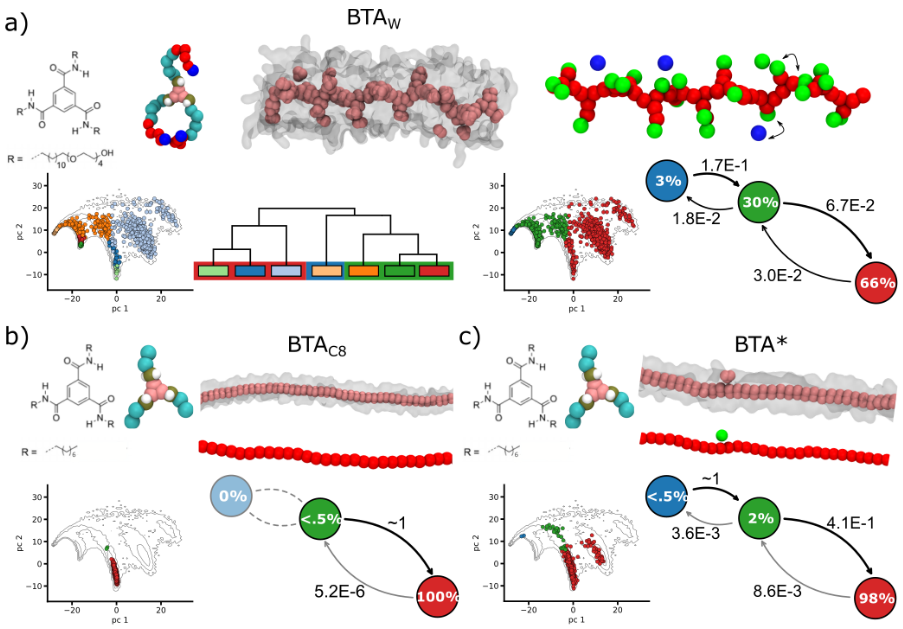


Fig. 4.7 SOAP-PCA-PAMM analysis on a family of BTA monomer variants: (a) water soluble BTA$_W$, (b) apolar organic solvent BTA$_{C8}$ and (c) its artificially mutated version with weaker directional inter-monomer interactions BTA*. Showing, monomer chemical formula and CG model, the fibre equilibrium structure (showing as solid beads only the center and all the rest as transparent surface for clarity), and the results of the PAMM clustering analysis organised as follow: equilibrium reduced structure coloured and PC scatter-plots, both coloured according to the individual PAMM motifs. For the BTA$_W$ case (panel a) we report two scatter-plots, depicting the PAMM clustering workflow (micro- and macro-cluster identification, see Sec. 2.3 or Ref. [120] for further details). On the bottom right, the interconversion graphs computed from counting the transitions of the individual clustering labels for each BTA variants.

with artificially lowered directional interactions compared to BTA$_{C8}$, shows emerging structural dynamics, with defects that are continuously created-and-repaired during the MD simulation (an effect already observed in the super-CG BTA fibre models, Chap. 3). It is worth noting, that such a direct comparison between fibre variants is now possible since all the SOAP feature vectors are projected in the same low-dimensional space and the resulting clustering is done on a "shared" dataset of the three structures.

As it was already proven in the previous chapter, this type of analysis is applicable to virtually any molecular structure, as long as we can provide a sufficient equilibrium MD sampling to characterise its structural behavior. We therefore extended our available dataset on supramolecular polymers including two additional types, made

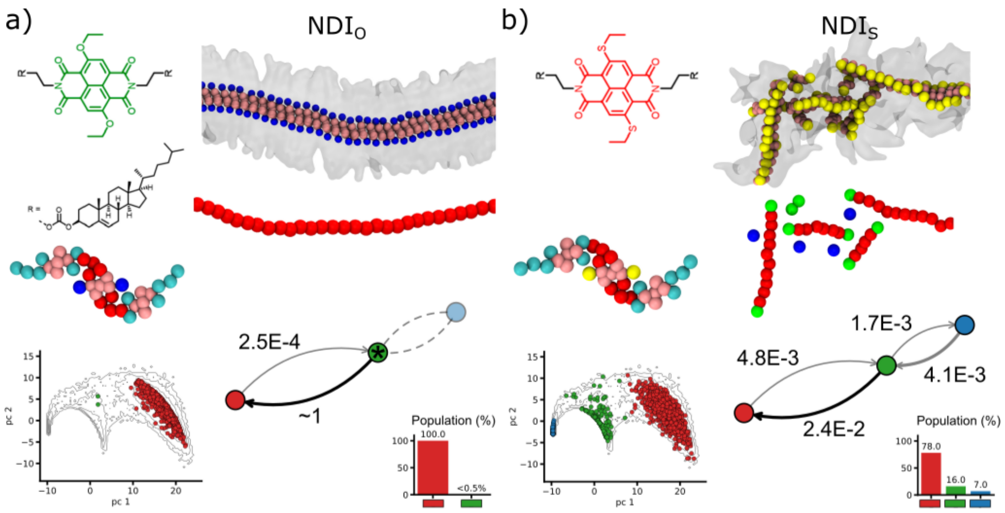


Fig. 4.8 SOAP-PCA-PAMM analysis on the NDI polymer variants: (a) NDI$_O$ (b) NDI$_S$. Each panel summarizes the results of the analysis for a single system variant, showing: chemical structure and CG model of the monomer; equilibrium CG-MD snapshot of the assembly, and view of the same with monomer centres colored according to the identified molecular motifs; scatter-plot showing the first two PCs of SOAP descriptors, again colored according to the motifs; interconversion diagram and population histogram.

by monomers based on the naphthalene diimide (NDI)[70] and the benzotrithiophen (BTT)[69] chemical moieties (Fig. 4.3). We compared two variants per-family, namely, the NDI$_O$ and the NDI$_S$, and the BTT$_F$ and the BTT$_{5F}$, for which we employed reliable CG models having an analogous resolution of the BTA models. We repeated the analysis just performed, following the same steps and building a unique SOAP dataset for each family, which contains all the descriptors from all the members of said families. The results are reported in Figure 4.8 (NDI variants) and in Figure 4.9 (BTT variants). Overall, they appear visually similar to the ones obtained for the BTA cases, although they cannot be directly compared to each other. NDI$_O$, visually resembles BTA$_{C8}$, as it is populated by a single and persistent cluster, it appears defects-free and its dynamic character is negligible (the transition graph does not show significant transition probabilities). The other three models, instead, are all populated by three main clusters in different populations and they all shows a rich internal dynamics character.

The proposed analysis allowed us to qualitatively detect the statistical formation of relevant molecular motifs, such as structural defects, in different models of supramolecular polymers. However, at this stage, the approach does not allow us to unambiguously compare between the different motifs populating different families.

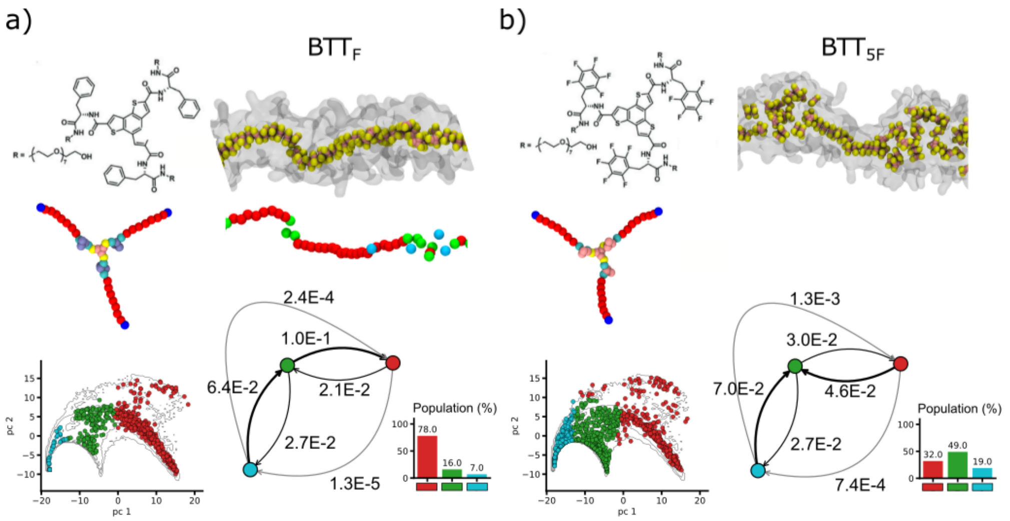


Fig. 4.9 SOAP-PCA-PAMM analysis on the BTT supramolecular polymer variants: (a) BTT$_F$ and (b) BTT$_{5F}$. Each panel summarizes the results of the analysis for a single system variant, showing: chemical structure and CG model of the monomer; equilibrium CG-MD snapshot of the assembly, and view of the same with monomer centres colored according to the identified molecular motifs; scatter-plot showing the first two PCs of SOAP descriptors, again colored according to the motifs; interconversion diagram and population histogram.

Especially, since the definition of defects emerges from the data contained in separate SOAP datasets, it is not clear to what extent defects in the BTA fibres are comparable to the BTT or NDI ones. The SOAP dataset used to compare the BTA variants, in fact, does not contain information on the monomer states in the NDI and BTT fibres. To compare these supramolecular fibres in a more objective and quantitative way, a further step is required.

Comparing different types of 1-D supramolecular polymers

As shown above, by building a comprehensive dataset containing SOAP feature vectors sampled from multiple systems, it is possible to unambiguously compare different supramolecular systems. The core idea is to retain in the SOAP analysis only those relevant features that are common to all the systems considered, *e.g.* a “common molecular denominator”. Throughout our analyses, we have always reduced each monomers structure to a single center before computing the SOAP descriptors. This comes with two main advantages: (i) the internal monomer arrangement of a supramolecular fibre can be efficiently represented by just a single point per monomer. The point is usually located at the centre of the monomer, effectively acting as the

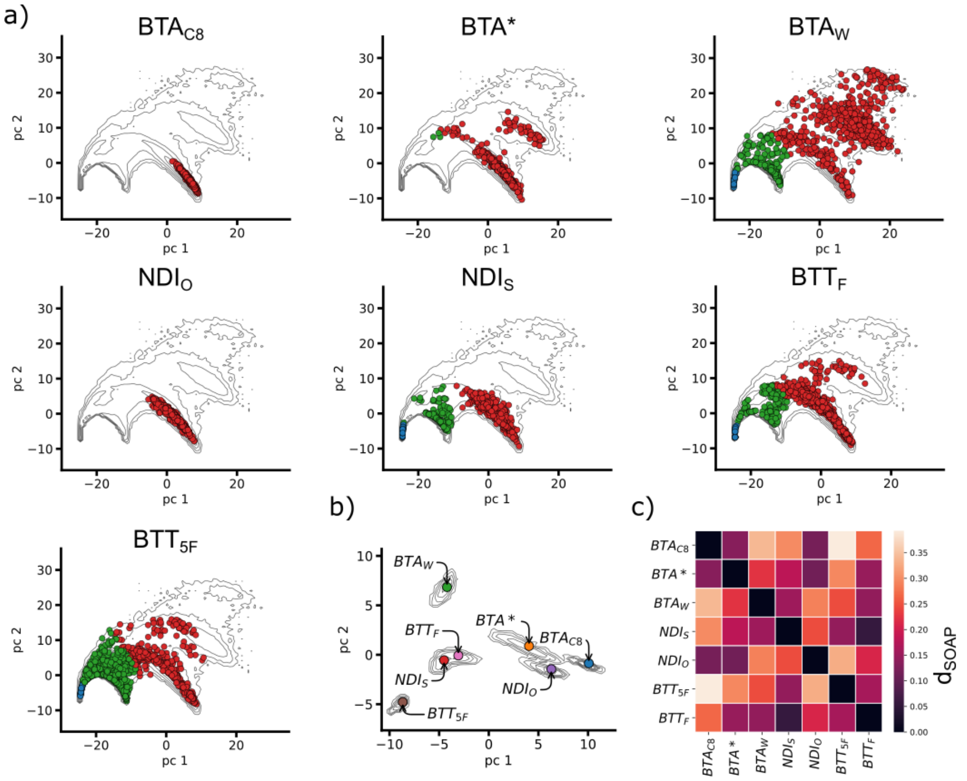


Fig. 4.10 Comparison of supramolecular polymers variants. a) Scatter plots of the PCA projections for each fibre SOAP dataset, coloured by the molecular motifs identified by the PAMM algorithm. The scatter plots are embedded in a global contour plot of the complete dataset (SOAP vectors of all the seven 1D supramolecular fibres data). b) Contour plot of the *frame-average* distributions computed from the PCs of the SOAP vectors. The colored dots represent the PC projection of simulation-averages for each 1-D fibre, manifesting the qualitative adjacency of each system in the PC space. c) SOAP distance matrix showing the quantitative mutual distance of all the systems.

fibre backbone. (ii) It makes the analysis very general and applicable to any kind of monomers, regardless of their structural complexity, thus opening the possibility to compare, not only variants of a supramolecular fibre, but also widely different assemblies.

To this end, we built a "shared" dataset containing all the SOAP vectors sampled from the equilibrium MD simulations of all the considered supramolecular polymers: 3 BTA, 2 NDI and 2 BTT monomer variants. We projected all the data onto the same low-dimensional space *via* PCA and we identified the main (shared) molecular motifs using the PAMM unsupervised clustering algorithm. The results are reported in Figure 4.10a. As for the single cases, we can observe three distinct general clusters:

fibre backbone (in red), structural defects (in green) and adsorbed/surface-diffusing monomers (in blue). These scatter-plots represent each monomer instantaneous local environment, hence they represent a sort of characteristic fingerprint of the supramolecular structures, showing which monomer states are mostly populated in each system. Qualitatively, such fingerprints provide an information of similarity between systems, in terms of structural arrangement and dynamicity of the monomers. For example, it is easy to see how the BTA_{C8} and NDI_O variants appear very similar, as they occupy almost overlapped areas of the PC space in which they are projected.

To reach a more quantitative insight, and to avoid possible distortions given by the low-dimensional projection, we turn to a SOAP-based metric, that allows the direct comparison in the original SOAP feature space. For each of the seven fibre systems we compute the *frame*-average SOAP global descriptor $\bar{\mathbf{p}}_t$ (see Sec. 2.2.2), and further average $\bar{\mathbf{p}}_t$ across all the collected frames, obtaining the *simulation*-average SOAP global descriptor, $\langle\bar{\mathbf{p}}_t\rangle$ for each of the systems. Thus, $\langle\bar{\mathbf{p}}_t\rangle$ represents a SOAP feature vector containing the average molecular environments populating the equilibrium structure of a given fibre. Figure 4.10b reports the results for such calculation of global descriptors projected onto a new PC-space, where, once again, it is possible to qualitative appreciate how some systems occupy adjacent areas of the PC-space. We can now assess the similarity amongst the different 1-D assemblies in a more rigorous way, by employing a SOAP-induced metric[139], *i.e.*, a high dimensional metric defined directly in the SOAP space, to compute the distance (d_{SOAP}) between the simulation average SOAP descriptors of each system. The result of such comparison is expressed as a distance matrix in Figure 4.10c. The off-diagonal values in the d_{SOAP} matrix grant a classification of the assemblies under investigation in the SOAP descriptor space. The darker the colour of the entry, the lower is the d_{SOAP} value between the respective assemblies, measuring their similarity (in terms of average molecular environments).

The d_{SOAP} matrix confirms the qualitative picture provided by the scatter-plots of Figure 4.10a,b. The most ordered supramolecular polymers, namely, BTA_{C8}, BTA^*, and NDI_O, are very close (*i.e.*, similar) and mainly populated by ordered structural domains (red cluster). The remaining 1-D systems have higher concentration of defected structural domains (blue and green clusters), and as such they are more or less distant from the previous three fibres. NDI_S and BTT_F SOAP spectra are nearly superimposed, *i.e.* $d_{\mathrm{SOAP}} \sim 0$, suggesting that the structural environments that emerge during the equilibrium MD are on average very similar. Conversely,

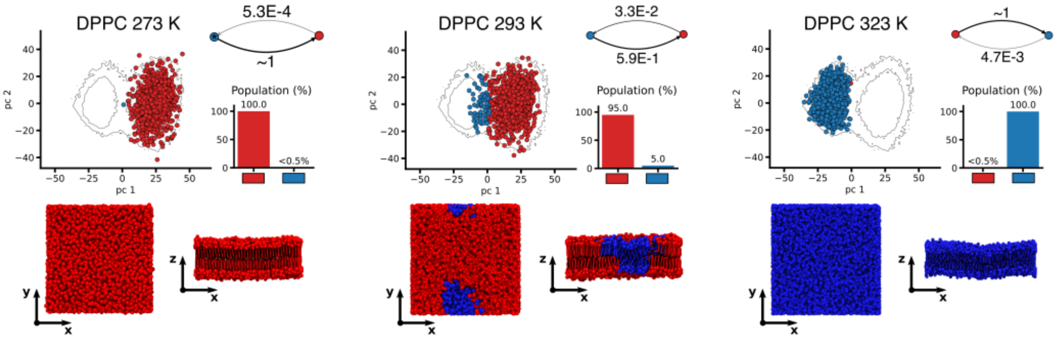


Fig. 4.11 Comparison of lipid bilayers across the gel-to-liquid phase transition. For each membrane we report a scatter-plot of the SOAP descriptors in the first two PC components, the interconversion graph and the population histogram, and lastly an equilibrium CG-MD snapshot of the bilayer at a given temperature, 273K to 323K from left to right. All the colours refer to the structural motifs identified by PAMM.

BTT$_{5F}$ and BTA$_W$ show unique features, distancing themselves from the other fibres. Overall these results demonstrate how the proposed approach is effective in comparing and classifying different types of supramolecular polymers. This is made possible by the application of an unbiased, high-dimensional, SOAP-based metric, that quantitatively compares the average molecular environments emerging in the various assemblies.

Comparing 2-D dynamic assemblies

Our goal is to build a database of various structures and characterise the possible relationships that may arise when comparing their structural environments. As a natural next step, we extend extend further our approach by testing it on the comparison of some 2-D supramolecular system.

We chose as principal case study the DPPC lipid molecule, which self-assembles into 2-D lipid bilayers, and additionally the DPC and SDS surfactant molecules, that form spherical micellar aggregates. For all these systems, we employed validated Martini-based CG-FF models (same resolution of the previously studied cases) to simulate in equilibrium conditions sample supramolecular structures, specifically: a DPPC lipid bilayer at three different temperatures $T = 273, 293, 323$ K and a SDS and DPC micelle at $T = 300$ K. DPPC bilayers are known to undergo a gel-to-liquid transition around $\sim 300 - 320$ K of temperature, which is well-captured by DPPC Martini models (expected at $\simeq 295$ K).[147] We thus repeated the same

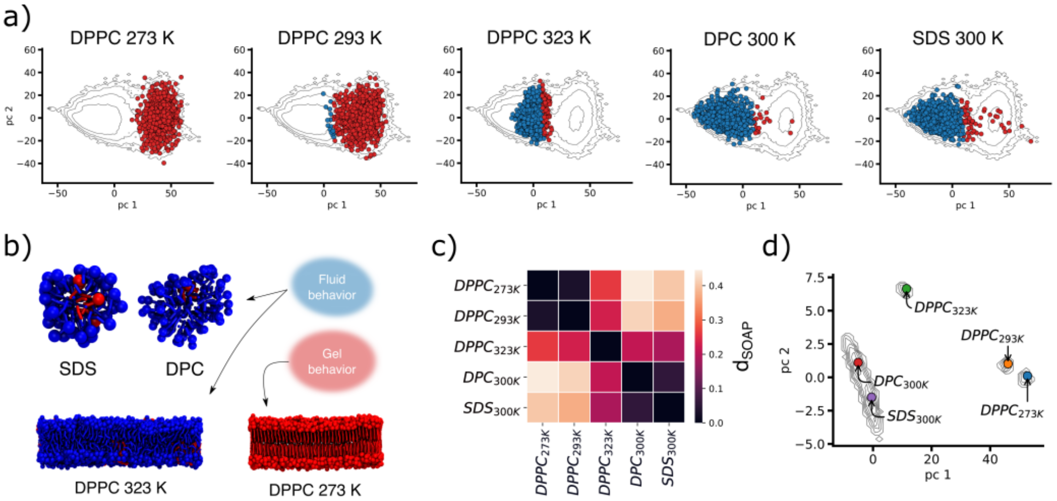


Fig. 4.12 Comparison of 2-D assembly variants. a) Scatter plots of the PCA projections for each 2-D system SOAP datasets, coloured by the molecular motifs identified by the PAMM algorithm. The scatter plots are embedded in a global contour plot of the complete dataset. b) Correlation between gel-fluid phase and the structural environments of single monomers in the different systems. c) SOAP distance matrix showing the quantitative mutual distance between all the systems. d) Contour plot of the *frame-average* distributions computed from the PCs of the SOAP vectors. The colored dots represent the PC projection of simulation-averages for each 2-D system.

steps of our previous analysis on this new set of structures. We computed the local SOAP descriptors for each monomer of each structure along the respective MD simulation, always referring to a reduced structure of one center per monomer (located in the polar head group, Fig. 4.3), and we extracted both the frame- and simulation-averages. Figure 4.11 summarises the relevant results obtained for the comparison of lipid bilayers alone, while Figure 4.12 reports the collective results for all the 2-D supramolecular assemblies we considered.

Immediately we can appreciate how the local SOAP-PCA-PAMM analysis is able to capture correctly the bilayer gel-to-liquid transition, solely based on how the environment surrounding each molecule changes with the temperature (Fig. 4.11). At low temperature (~ 273 K, Fig. 4.11 left) the bilayer is entirely in the gel phase (red cluster). As the transition temperature approaches, around $\sim 300-320$ K, nucleation of the liquid domains starts and the structure begins to melt (emerging blue cluster in dynamic equilibrium with the red cluster, $\sim 5\%$ at ~ 293 K, Fig. 4.11a left center). Finally, as soon as the temperature raises to ~ 323 K, the bilayer appears entirely liquid (blue cluster), with residual ($< 5\,\%$) gel-like environments (Fig. 4.11). In summary, our analysis built on the SOAP data, correctly distinguishes gel and liquid domains in planar lipid bilayers, without prior knowledge on the lipid arrangement in each phase. Nonetheless, interesting questions may arise on how similar the identified monomeric environments are to, *e.g.*, those present on the curved surface of a micelle.

We thus, extended our analysis including two additional sets of SOAP data coming from the MD simulations of two micellar aggregates, made from DPC and SDS surfactants (Fig. 4.12a). The scatter-plots in Figure 4.12a shows, when projected on the same PC-space, a significant overlap of both micelles local fingerprints with that of the lipid membrane in the liquid phase ($T = 323$ K). This suggests that the structural features of the environments characterizing these micelles are closer to those of a liquid bilayer, rather than to those of a gel-like one (Figure 4.12b). Some differences arise due to the intrinsic globular and more disordered arrangement of the surfactants inside the micellar aggregate (visible in the left portion of the scatter-plots in Fig. 4.12a). To gain more quantitative insights we computed the SOAP *frame*- and *simulation*-averages of these systems, as well as the mutual d_{SOAP} distances (Fig. 4.12c,d). The DPPC$_{273K}$ and DPPC$_{293K}$ bilayers appear relatively close to each other, and separated from the other three systems. Similarly, DPC$_{300K}$ and SDS$_{300K}$ micelles have $d_{\text{SOAP}} \sim 0$, and as such, close in the scatter-plot (Fig. 4.12c,d). Interestingly, the liquid DPPC$_{323K}$ shows “intermediate” features, being closer to fluid micelles (SDS/DPC), than to the DPPC bilayer at lower temperature.

Comparing 3-D dynamic assemblies

Lastly, we include in our analysis examples of a 3-D dynamic soft-assembly. In particular, three examples of homogeneous spherical aggregates made from a simple linear saturated alkane composed of 16 Carbon atoms, hexadecane (HEXA). In water, HEXA molecules, given their apolar character, undergo aggregation forming spherical assemblies, droplets or nanoparticles. We tested our method in comparing assemblies of variable sizes, respectively composed of 128, 512 and 2048 HEXA monomers. For each variation, we collected the equilibrium MD trajectory in explicit water and analysed it with the SOAP-PCA-PAMM workflow. The results for the comparison of HEXA droplets are reported in Figure 4.13. The analysis identifies two main structural motifs, essentially corresponding to the monomers that are on the surface of the droplet and in contact with the solvent (pink cluster) or the bulk

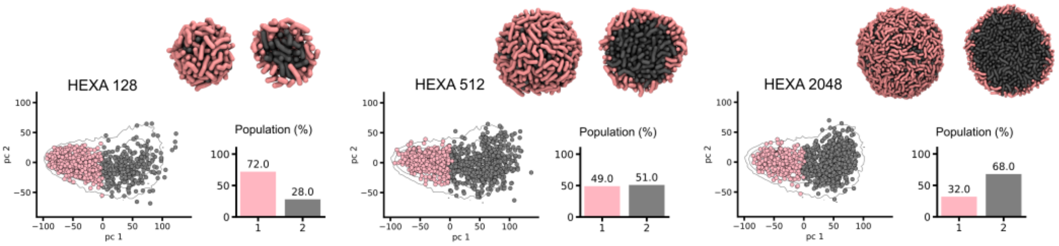


Fig. 4.13 Comparison of 3-D supramolecular assemblies of different sizes. For each droplet we report a scatter-plot of the SOAP descriptors in the first two PC components, the population histogram, and lastly an equilibrium CG-MD snapshot of spherical structure highlighting the PAMM structural motifs.

monomers (gray cluster) of the HEXA assemblies. Not surprisingly, the molecular environments do not change radically across different system sizes, where the only change is the relative population of them, as the bulk core gets bigger. The analysis also highlights a persistent dynamic trait between these two molecular motifs in all three cases, proving that the bulk monomers are always freely diffusing in the entirety of the aggregate volume, without distinction between bulk and surface dynamics. While these cases are rather simple, they show that our analysis provides robust and reasonable results also in the case of 3-D assemblies of variable size, enriching our structures database with additional data.

A data-drive "defectometer" to compare and classify different types of soft dynamic assemblies

As a final step, we processed the simulation data of all the studied systems, assessing to what extent completely different assemblies can be compared to each other. As we saw for each individual family of supramolecular aggregates this information is readily obtainable from the SOAP distance, which gives a bonded measure of the average structural diversity of each system, without the need to process the descriptor data with PCA, PAMM and/or other post processing approaches. It is worth stressing out that all systems share a common feature: the SOAP data refers always to a reduced structure that only accounts for a single centre of interaction. Hence, from a descriptor point of view, these systems differ only for the fingerprint that they carry, regardless of their actual chemical nature.

We thus gathered in a single dataset all the SOAP vectors obtained in the previous analyses and, using the high-dimensional SOAP metric, we computed the d_{SOAP}

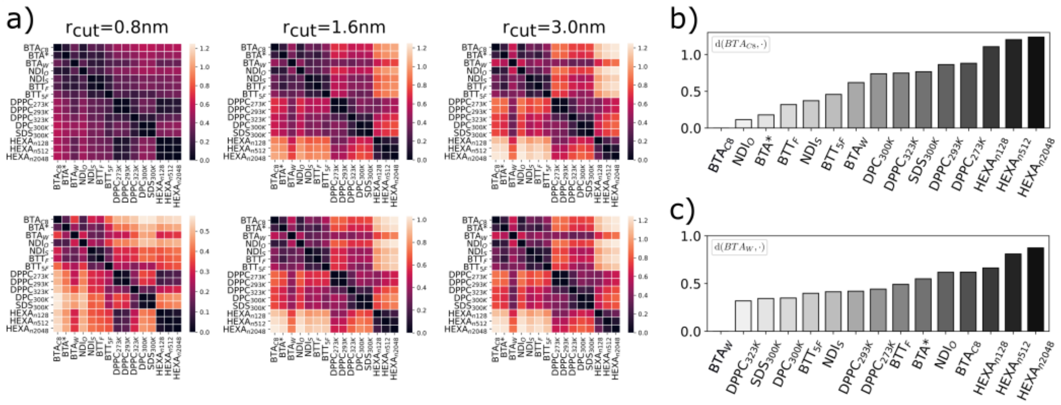


Fig. 4.14 Results for the comparison of different types of soft dynamic supramolecular materials. a) SOAP distance matrices of the simulation-averages for each system, computed at different SOAP cutoff values. SOAP distances of the systems taken setting BTA$_{C8}$ (b) and BTA$_W$ (c) as references, computed at cutoff= 3.0 nm.

matrix to compare the *simulation*-averages associated to each assembly, highlighting the similarities and differences between them.

An important parameter for the SOAP analysis is the cutoff radius (`rcut`), which determines the size of the neighborhood considered in characterising the molecular environment of each SOAP centre. Throughout our investigation, we have used three different optimal `rcut` values: 0.8 nm for the 1-D systems, 1.6 nm for the 3-D systems, and 3.0 nm for the 2-D systems. Different values become necessary when dealing with different hierarchies of assembly, as a small cutoff becomes less efficient in capturing relevant information from more complex systems. However, it is known that increasing an already optimal cutoff radius usually has little effect on its descriptive power, aside for a increased computational cost.[30, 114] Because of these reasons, and to test the stability of our “defectometer” tool, we computed the global descriptors for all the supramolecular families at the three levels of cutoff radii and compared the results. Figure 4.14a shows the results for the three different distance matrices, where we normalised the distances on the highest one (top raw) and on the highest of each individual system (bottom row). These two plot variants emphasise the effect of using a different cutoff radius when comparing different aggregates. A bigger cutoff is followed by a higher absolute value for the pair distances (increase in color brightness going from left to right, top raw of Fig. 4.14a). Nevertheless, the ratio between them seems to be overall constant with the different cutoffs (same general colours pattern, bottom raw of Fig. 4.14a). Meaning that,

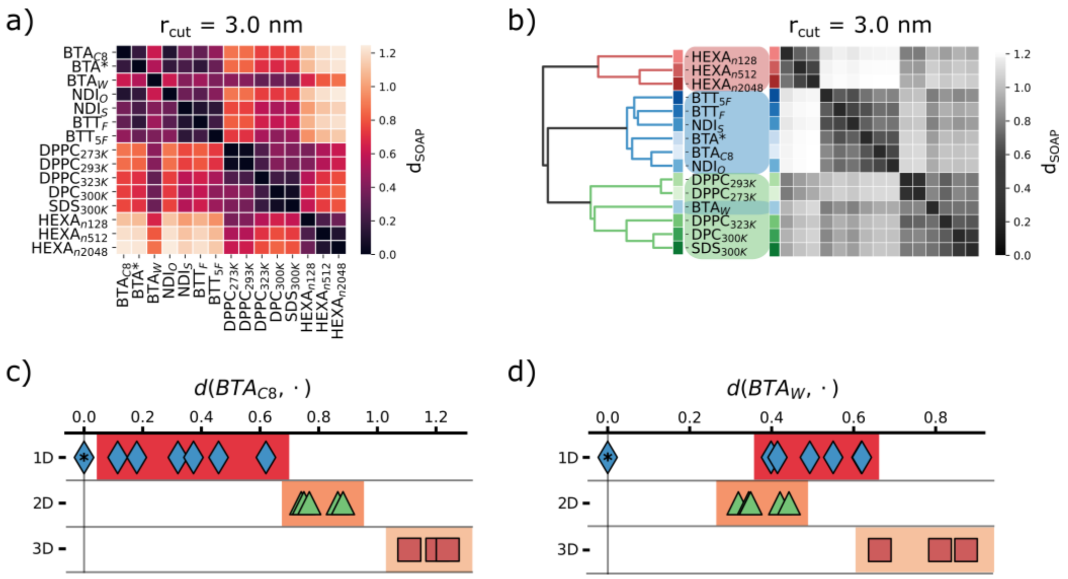


Fig. 4.15 Visual representation of the “defectometer” tool to compare and rank different soft supramolecular assemblies. a) SOAP distance matrix computed at cutoff radius 3.0 nm. b) Same matrix as (a) reshuffled after the application of a hierarchical clustering algorithm to pair similar systems together. Visualisation of the defectometer rating for the BTA$_{C8}$ (c) and BTA$_{W}$ (d) systems.

aside from subtle differences in the lower cutoff range, the global picture in terms of similarity between the compared assemblies remains consistent in all cases.

In general, in the d_{SOAP} matrices we observe four main dark areas (*i.e.*, a dark colour means low mutual distance): that of fibres (1-D assemblies), two neighboring/entangled ones for the flat and spherical 2-D assemblies, and a third one for the 3-D aggregates separated from all the rest (bottom-right in the matrix). An interesting exception is the BTA$_{W}$ fibre variant, which is found more similar to 2-D assemblies (DPPC and surfactant micelles) rather than to the other 1-D assemblies of its kind. Another interesting system is the DPPC bilayer at 323 K, which is found closer to highly dynamic SDS and DPC micelles rather than to DPPC at lower temperature (293 or 273 K).

In the d_{SOAP} matrix, one can select one assembly and rank all the others with respect to it, *i.e.* equivalent of simply selecting the corresponding matrix raw (or column) of a specific system. For example, selecting the ordered BTA$_{C8}$ fibre, the plot of Fig. 4.14b shows a high similarity (small d_{SOAP} values) with the other 1-D ordered assemblies, lower similarity with disordered 1-D fibres (*e.g.*, $d_{\mathrm{SOAP}} \simeq 0.6-0.7$ vs. BTA$_{W}$), while d_{SOAP} increases further considering 2-D and 3-D assemblies. As anticipated above, an interesting result is obtained when focusing on the BTA$_W$ fibre (Fig. 4.14c). In terms of d_{SOAP} ranking, the closer assemblies to this water-soluble, disordered fibre, are indeed highly dynamic, planar or spherical 2D assemblies. Surprisingly, all the other 1-D fibres are less similar to BTA$_W$ than all the 2-D studied systems. This suggests that the solvophobic component of the BTA$_W$-BTA$_W$ intermonomer interaction in water, the key factor in controlling defects formation in such 1-D assemblies,[30, 31] can imprint a molecular environment in the surrounding of the monomers that appears closer to the environment of 2-D micelles or liquid-like lipid bilayers, than to those of the environments of other ordered BTA or fibre variants (*e.g.*, BTA$_{C8}$). Furthermore, this is known to produce a dynamic surface adaptability in this specific fibre that might resemble the surface fluidity seen, *e.g.*, in lipid bilayers.[26, 28]

The relative d_{SOAP} distances computed at `rcut`= 3.0 nm (Fig. 4.15a) were then processed *via* agglomerative clustering, in order to assess the hierarchical tree among all considered systems. This led to the dendrogram tree shown in Figure 4.15b, which clearly underlines how the disordered BTA$_W$ fibre is closer to the ensemble of 2-D structures (highlighted in green) rather than to those of ordered 1-D fibres (highlighted in blue). All the others "leaves" in the dendrogram highlight the remaining relations between the structures, completing the global comparison picture.

In summary, our approach demonstrates how a purely data-drive descriptor analysis is capable of handling the classification of complex molecular systems, without the need to devise *ad-hoc* collective variables and to know their chemical/physical nature. This crucial difference between "standard" human-based classifications and the data-driven method proposed herein is further underlined in Figure 4.15c and d. These two schematic plots report the dimensionality of each system as a function of its distance d_{SOAP} from the BTA$_{C8}$ (c) and BTA$_W$ (d) systems, showing how a correlation between microscopic similarity and *a priori* theoretical dimensionality is not obvious when dealing with soft dynamic assemblies, as those studied herein.

It is worth noting that such similarity measurements rely solely on data extracted from equilibrium MD simulations, with no major assumption on the structure/features of the studied materials. While this approach is flexible, and can be used to compare in an unbiased way different types of materials, it also creates the important opportunity to classify assemblies based on the molecular environments that populate them, which we believe is a crucial step towards the rational design of supramolecular materials with programmable dynamic properties.

4.2 A data-driven approach to compare and classify lipid force fields

The work presented in this section contributed to a publication in *The Journal of Physical Chemistry B*, Ref. [114], with the title "A Data-Driven Dimensionality Reduction Approach to Compare and Classify Lipid Force Fields". A crucial point in the MD simulations of lipid systems concerns the performance and accuracy of the classical FFs, either AA or CG, used to model the single lipid molecules. To date, assessments on the reliability of such FFs are mostly based on the comparison with experimental observables, which typically are average properties. Thus, they tend to overlook the impact of local (molecule-wise) differences of the structure and dynamics, which can have a big impact, but are notoriously elusive to characterize. In Ref. [114], we address this problem by proposing an agnostic way to compare different FFs at different resolutions (atomistic, united-atom, and coarse-grained), by means of a high-dimensional similarity metrics built on the framework of the SOAP descriptor. Through this SOAP-based metric we compare and rank a total of 13 FFs, modeling 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) bilayers, exposing differences between the FFs in reproducing non-average events (*e.g.*, originating from local molecular environments). Lastly, as a proof-of-concept for the reliability of the metric, we report the fluid-to-gel phase transition in dipalmitoylphosphatidylcholine (DPPC) bilayers, for which we clearly identifiy nucleation centers, highlighting some intrinsic resolution limitations in implicit versus explicit solvent FFs. In the framework of the present thesis work I have contributed with the development of the SOAP metric analysis and the analysis of the bilayer fluid-to-gel phase transition.

4.2.1 Scientific context

Lipid bilayers belong to the family of 2-D soft self-assembled materials. They are ubiquitous in nature and play a very important role, directly impacting the regulation

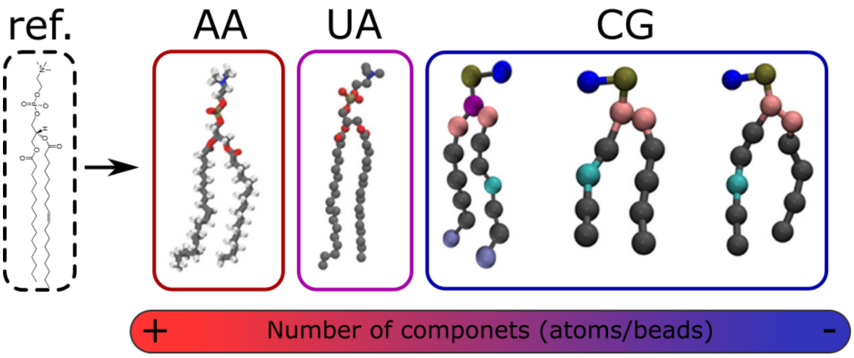


Fig. 4.16 Representation of a POPC molecule across different molecular model resolutions. The chemical structure is used as base for the All-Atom (AA) model, which has the higher details. Coarser versions of the AA model, like United-Atom (UA) and Coarse-Grained (CG), can be defined by joining together cluster of atoms into the so-called pseudo atoms.

of human-body cells through their specific chemical and mechanical characteristics. For example, membranes constitute a barrier between the cell and its external environment, and are involved in transport processes[148], signaling[149], and regulators in protein interactions[150]. To study these phenomena a plethora of experimental techniques, such as nuclear magnetic resonance (NMR), calorimetry, small angle X-ray scattering (SAXS) and others, have been applied to lipid systems to obtain relevant structural and dynamics information. The large amount of experimental data gathered paved the way for the creation and validation of reliable computational models, which in turn can be exploited in computational-based approaches, like MD simulations, for an in depth characterisation of such materials. However, despite the advances in computational hardware and software, classical unbiased atomistic MD simulations are still unable to cover all the time scales of interest in biological systems.[151] To solve this problem various models with a reduced number of degrees of freedom were developed, from united atom (UA) representations, to coarse-grained (CG) and super-CG models, where atoms are joined together and represented by pseudo-atoms or “beads”. As introduced in Section 2.1, the reduction of the number of degrees of freedom accelerates the dynamics of the simulated systems at the cost of some accuracy loss, which is not straightforward to asses. The precision and performance evaluation of a force field (FF) is usually obtained by comparison of average equilibrium observables computed from simulation data with the experimental available counterparts. However, the general assessment of the performance of a FF remains a challenging task, requiring the consideration of multiple parameters at the same time. Moreover, the comparison becomes particularly

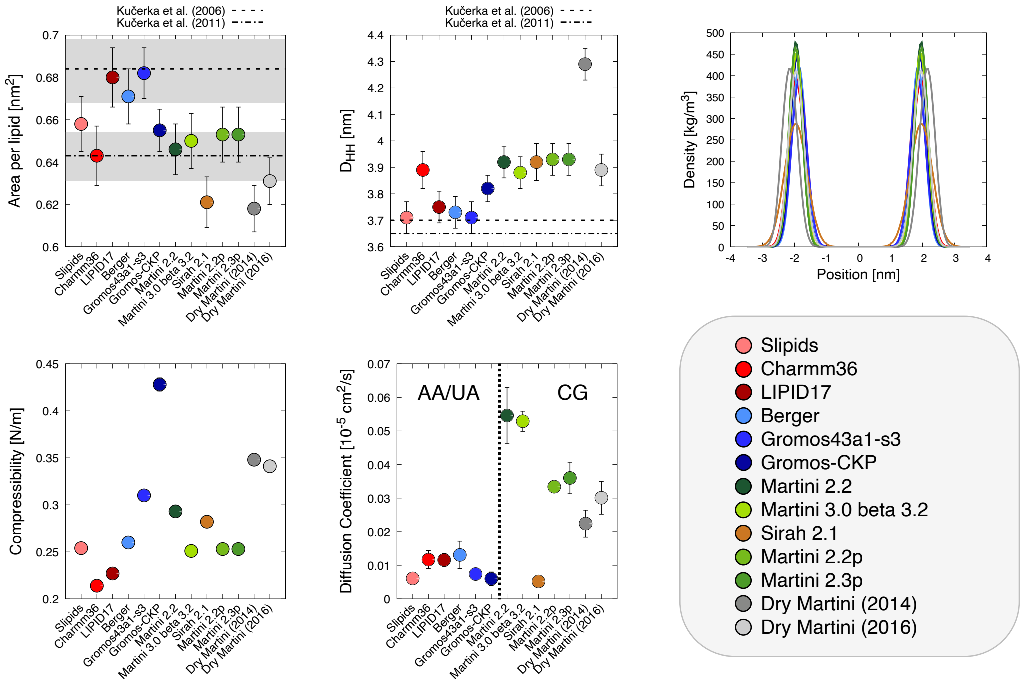


Fig. 4.17 Comparison of the average observables extracted from the production run of the POPC lipid membranes for all the studied FFs.

awkward if different levels of FF resolution are considered in the same analysis (*e.g.*, comparing between AA and CG descriptions), as local properties in the organisation of the lipids inside the membrane are reproduced differently. In Ref. [114] we tackle the FFs comparison problem by exploiting the previously introduced data-driven analysis of molecular environments (for more details see the theory in Chapter 2 and the results throughout Chapters 3,4). The 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphocholine (POPC) lipid molecule (Fig. 4.16) was taken as reference and it was modeled in a total of 13 different FFs across various resolutions (*e.g.*, AA, UA, CG). For each FF a representative POPC membrane was built and simulated, with the purpose of extracting equilibrium and dynamic average physical observables from each of the systems (a complete description of the methodologies and tools used to build, simulate such lipid bilayers and compute the physical obsevables is available in Ref. [114] and supporting information). Figure 4.17 summarises the results obtained for the POPC at different resolutions. Overall the results shows that the lower resolution models are affected by high data variability, often overestimating the experimental available values. This is probably imputable to differences in the definition of the various CG-FF parameters, *e.g.*, slightly different approximations in the enthalpy/entropy balance in the models can make thermal vibrations larger/smaller, broadening/narrowing the density distributions.

In the following we will show the main contribution that the methodologies presented in this thesis brought to the study of Ref. [114], namely how SOAP power spectra and the SOAP metrics can be used to unveil intimate relationships between different resolution FFs, without the need of prior knowledge of physical observables.

4.2.2 Methods

SOAP descriptor analysis

Local data on the lipid environments is extracted by means of the SOAP[105] descriptor. For all systems, the SOAP power spectra were calculated in periodic boundary condition along the xy dimensions, *i.e.*, the periodic dimensions of the lipid bilayer. We first reduced the lipid molecule structures down to 4 SOAP-centres each: two for the two main head functional groups and one for each tail (Fig. 4.18 step i) ensuring an uniform treatment of all our molecule at different levels of resolution. Depending on the FF, the different centres used to represent the lipid molecules are placed according to the coordinates of the relevant atoms (in AA models) or pseudo-atoms (in CG models). Then, SOAP power spectra were computed by taking as output only the fingerprints related to the phosphate head centres (Fig. 4.18 steps ii-iii). The resulting dataset will therefore contain a number of SOAP feature vectors equal to the number of lipid molecules, but still retaining residual collective information coming from all the four centres that are found in the neighborhood of the phosphate heads (inside the cutoff range). The SOAP calculation was carried out setting `nmax,lmax`$=8$ and a radial cutoff of 3 nm, and leaving other parameters as default. The cutoff value was selected to be large enough to include all particles intended to be characterized by the descriptor, *i.e.*, the 4 centres defined, and to be smaller than the total width of the membrane to exclude the opposite surface beads. Hence, smaller than 3.7 nm, but large enough to include the two tails centers, located in the mid-part of the membrane, thus larger than about 2 nm. All the SOAP calculations were done using the DScribe[128] python package.

The global SOAP descriptor, used in the metric evaluation, was obtained, as already introduced in previous chapters, by averaging the local SOAP contributions

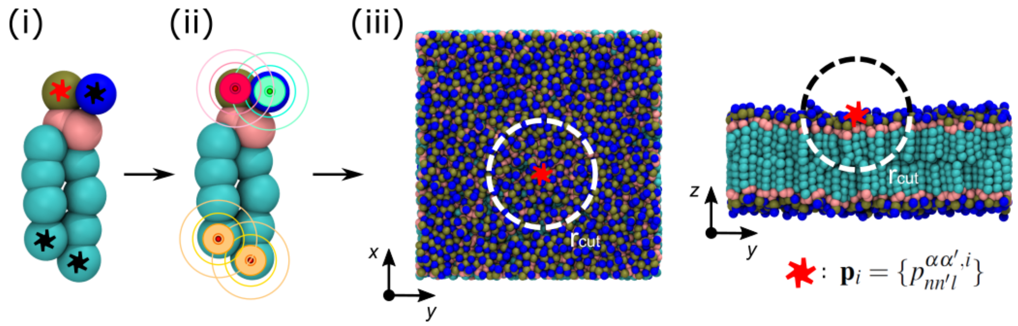


Fig. 4.18 The “4-centers” SOAP calculation on lipid molecules inside a bilayer: (i) for each lipid we select 4 points to represent the overall lipid structure inside the membrane. (ii) Following the SOAP mathematical framework (Chap. 2 and Ref. [105]) the punctual structure is smeared with a gaussian base. (iii) The SOAP power spectra is computed with respect to the phosphate head group (red asterisk).

across all components of a frame (*i.e.*, the number of lipid molecules populating a bilayer) and then across all the snapshots of the MD trajectory. This allows us to obtain a single global feature vector for each of the FFs under investigation. The global features will be then used as input for the comparison by means of the SOAP metric.

DPPC analysis across the phase transition

Regarding the comparison of the explicit- and implicit-solvent FFs in modeling the DPPC fluid-gel phase transition, all the DPPC bilayers were further simulated for a total of 1 μs of simulation time at the selected temperatures for both the Marini 2.2 and Dry Martini FFs. SOAP local descriptors were calculated from the production MD run trajectories considering one snapshot every 10 *ns* (for a total number of 101 frames). The training dataset was composed by merging all the data coming from the same FF simulations and then projected onto a 5-dimensional space by mean of PCA. In this way, the identified molecular motifs in each simulation (at different T) are embedded in the same parameter space and can be directly compared. The PAMM[120] unsupervised clustering was used to get insight on the phase transition of the DPPC lipid bilayer. We thus applied the PAMM clustering on this low-dimensional space and we obtained the distinction of different molecular motifs.

4.2.3 Results

The results obtained by applying the presented SOAP-based analysis to the study of different lipid FFs are presented here and are divided in two main sections. First, we compare between different lipid FFs using both average observables and SOAP-based analysis, to highlight strengths and weaknesses of the molecular models across various resolutions. All the data handled in this section is extracted from a representative lipid bilayer containing 128 POPC lipid molecules, simulated for 1 μs of MD at 303 K for each of the FF under investigation (more details on the setups and simulation protocols can be found in Ref. [114] and related supporting information). Secondly, we attempt to rationalise the information obtained from the comparison to more human-comprehensible data. In this regard, we report proof of the intrinsic difference that exist between explicit and implicit CG-FFs, and the reasons why they show very different levels of accuracy in modeling a lipid bilayer. We will here study the phase transition behavior of a representative DPPC bilayer simulated using the Martini 2.2 (explicit solvent) and the Dry Martini (implicit solvent) CG-FFs. Data on the phase transition is recovered by means of the already tested unsupervised SOAP-PCA-PAMM approach.

SOAP metrics to compare lipid FFs

The initial comparison based on average global observables reveals how the performances of different FFs are dramatically dependent on the specific quantity of interest, which in turn is highly affected by the specific resolution of the used model. While such parameters may be useful for a first and qualitative comparison, it is not easy to obtain a complete picture of the FFs relationships, which would require a more general and high-dimensional approach. To attain a general and unbiased criterion to compare lipid FF models having different resolutions (AA, UA and CG), we turn to the SOAP[105] descriptor and the SOAP-based metric. We start by considering a unified SOAP representation of our lipid molecules well-suited for such comparison. For this reason, we chose four relevant beads in the lipids structure that are in common between all considered lipid models, and that we use as centres for our SOAP analysis. One centre for the lipid charged head group, one for the phosphate group, and one for each of the two alkylic lipid tails, for a total of four centres, that will define the "reduced" structure representation of our lipids molecules. A SOAP

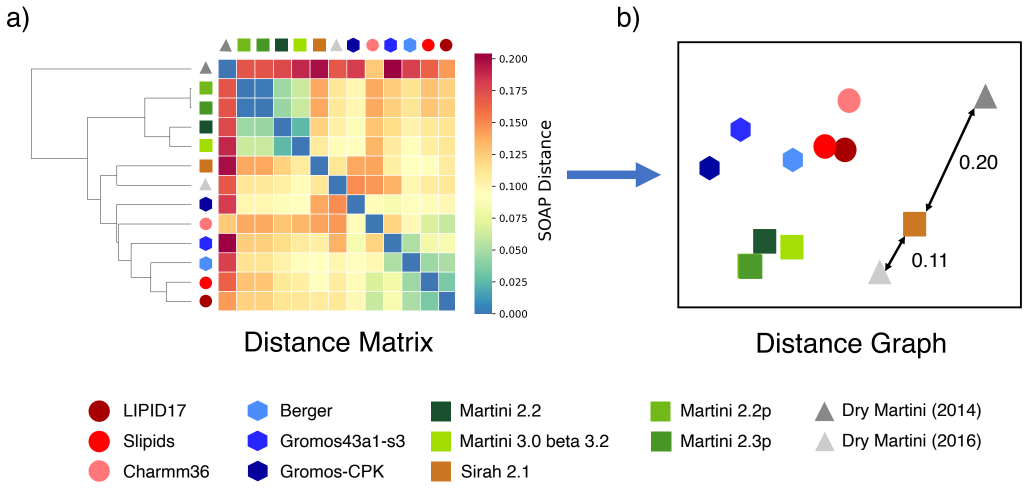


Fig. 4.19 Representation of the FFs relationships *via* SOAP metrics. a) SOAP distance matrix built from the mutual pair distances of every FF global simulation-average descriptor. b) Projection of the matrix in panel (a) on a representative 2-D plane.

power spectrum is computed for each phosphate centre, considering the contribution from all the four centres that are found in their environment (for further details see Fig. 4.18 and related discussion). Hence, we obtain a SOAP power spectrum that is indicative of the local environment that surrounds each phosphate group in the bilayer, which accounts of *e.g.* levels of order/disorder, spatial displacements and packing of the lipids heads, phosphate groups and the lipid tails. The choice of the phosphate groups, as the main SOAP centres, is due to their central position, both in geometric terms in the lipids structure, and in chemical terms, as the phosphates are at the interface between the hydrophobic and hydrophilic parts of the lipid molecules. From the resulting dataset we computed for each FF the simulation-average SOAP feature vectors $\langle \bar{\mathbf{p}} \rangle$ (see Sec. 2.2.2), a global descriptor, that in this case, contains information on the average complexity of a lipid bilayer, at a given level of FF accuracy considered. Finally, to quantify the (dis)similarities in the SOAP space of each of the studied FFs, we computed their pair distances and represented them in matrix form. Figure 4.19a shows the obtained distance matrix and the hierarchical clustering of the corresponding pair distances of all the FFs we considered in the work. In particular, cold colours identify short distances between the FFs, indicating that these FFs behave similarly, *e.g.* the diagonal terms are all zero. Warm colours identify larger SOAP distances, and increased discrepancy between the FFs behavior. From Figure 4.19a, it appears evident that all wet and polarizable CG Martini FFs

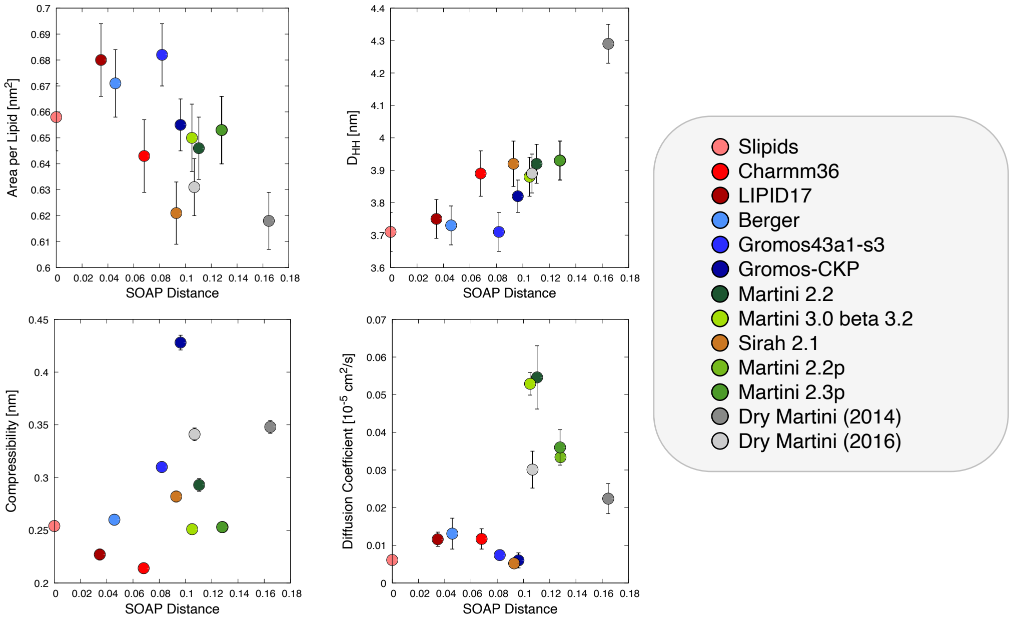


Fig. 4.20 Relationships between the SOAP distance and various physical observables extracted from the MD production runs.

account for the lipid behavior in a similar way. It is worth noting that Martini 2.2p and Martini 2.3p CG-FFs have zero mutual distance (*i.e.*, they are superimposed), due to the representatio of the POPC molecule being identical. Another interesting point is the improvement of the Dry Martini performances between the initial 2014 version (light grey triangle) and the 2016 one (dark grey triangle), as the latter show overall a better agreement with all the other FFs considered, and most importantly, with the the wet CG Martini counterparts. AA and UA FFs show a substantial proximity between all of them in the SOAP metrics space.

As a final proof-of-concept, we selected the Slipids lipid FF and compared its SOAP distances from the other FFs with all the scalar average observables previously reported in Figure 4.17. The results are reported in Figure 4.20. We can observe that the physical observables which are most correlated with the selected SOAP distance are the area per lipid (APL) and thickness (D_{HH}) of the lipid bilayers. This is reasonable, and could be expected to some extent, as SOAP is in fact a high-dimensional way to represent the spatial displacement of atoms/beads along the MD trajectories (an information that is strongly connected with, and to a considerable extent captured by, the APL and D_{HH} parameters).

Capturing the gel-to-liquid phase transition on a local Level with the SOAP descriptor

The approach discussed above allowed us to compare between different FFs in a rather comprehensive way. However, it is not straightforward to link such information extracted from high-dimensional analyses to human-comprehensible data. At the same time, the comparison with only a few experimental data available (Figure 4.17a) may not be enough to obtain a clear and exhaustive picture. As a test case we focus on two widely used CG-FFs: the Martini 2.2 and the Dry Martini 2016. Comparing their data using the obtained results, the two models appear rather similar to each other, *e.g.* the values of the physical observables and the mutual distance from AA-FFs are close. However, in Figure 4.19 they are separated from each other, suggesting the existence of an intrinsic difference between the two. Recently, it was shown that, on average, Martini 2.2 and Dry Martini CG FFs tend to respectively over-structure and under-structure the bilayers compared to AA FFs.[152] All these analyses and comparisons provide evidences that are limited to the average characterization of the bilayers, while, on the other hand, it has been shown that the behavior of complex supramolecular assemblies (*e.g.*, lipid bilayers) may be strongly controlled by local events, or fluctuations, that cannot be captured with average evaluations.[30, 31, 50]

To move our investigation to a deeper level, we decided to investigate the phase transition, from gel to liquid, of a lipid bilayer using a high-dimensional SOAP-based approach (*i.e.*, the SOAP-PCA-PAMM analysis) to detect and characterize local nucleation processes underpinning the formation a new rearrangement of the bilayer. Furthermore, we believe that such analysis can be useful in identifying limitations in the FF representation of the lipid assembly. The reported experimental melting temperature T_m for POPC is approximately 273 K[153], which is not ideal when operating with classical MD simulations. Thus, we decided to turn to another well-studied lipid molecule, the dipalmitoylphosphatidylcholine (DPPC), whose experimental transition from gel to liquid phase occurs at a temperature of ~ 315 K[154]. For the wet Martini FF, T_{melting} has been reported in semi-quantitative agreement at $\sim 295 \pm 5$ K[147], while for Dry Martini T_{melting} has been estimated at ~ 333 K[155]. We repeated our local SOAP calculation (using a lipid reduced structure of 4 beads and centering a SOAP power spectra on the phosphate head) on a reference DPPC membrane, containing 1152 lipid molecules, and simulated for 1 μs at five different temperature for both CG-FFs. We then projected our dataset

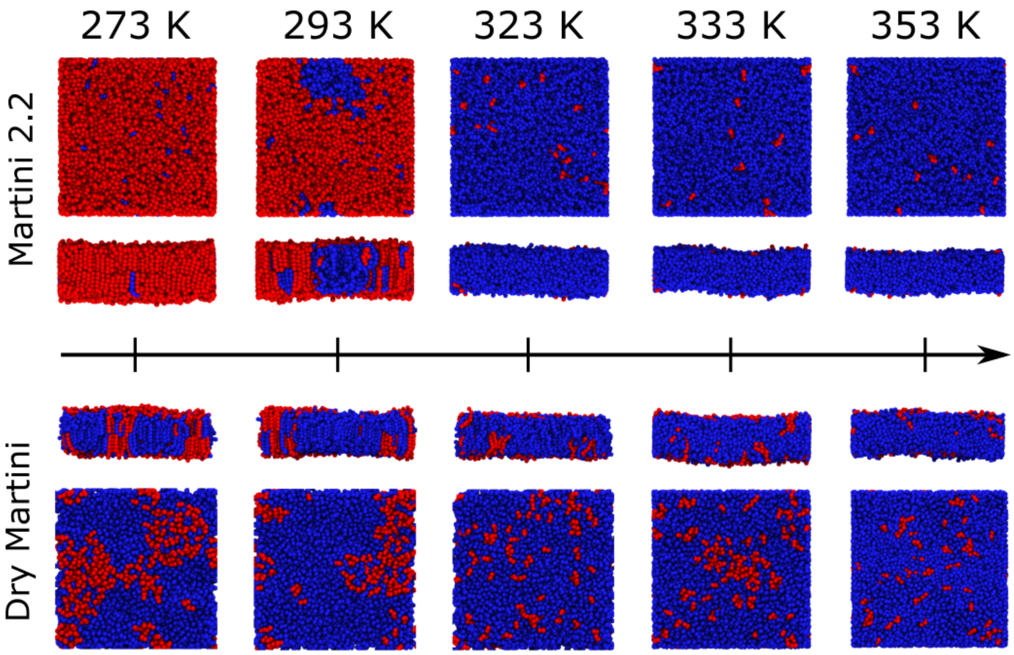


Fig. 4.21 Local lipid environments across the gel-to-liquid phase transition detected by the SOAP-PCA-PAMM approach. The two colours represent the PAMM clusters membership: red for the gel phase and blue for the liquid phase.

on a low-dimensional space by means of PCA and clusterised it with the PAMM unsupervised algorithm, extracting the molecular motifs. The results of the PAMM analysis at different temperatures for the Martini 2.2 and Dry Martini DPPC bilayers are reported in Figure 4.21.

In the case of Martini 2.2 (top half of Fig. 4.21), the PAMM analysis is able to discriminate between lipids belonging to an ordered phase (gel, red colour) and a disordered phase (liquid, blue colour). As expected, the relative populations within the two molecular motifs show an inversion from $T < T_{\text{melting}}$ to $T > T_{\text{melting}}$ validating the interpretation given to the two identified clusters, as the two distinct phases. Below 293 K the system is dominated by red lipids, while above 323 K by blue ones. We can clearly observe nucleation events along the MD trajectory at $T \simeq T_{\text{melting}}$ (see snapshot at 293 K in Figure 4.21), thus the phase transition, and its local origin, appear to be well-reproduced.

The implicit-solvent Dry Martini FF provides a different picture (bottom half of Fig. 4.21). The PAMM analysis again identifies the same two clusters: red lipids in gel phase, and blue lipids in liquid phase. However, although the gel population (red cluster) remains somehow inversely proportional to T, we could not observe a sharp phase transition in the simulations performed across the temperatures. The bilayer remains always in a liquid-like phase, although becoming on average more static by lowering the temperature. This is probably due to the unavoidable approximations resulting from encoding both solute-solute and solute-solvent interactions (in the explicit-solvent model) into solute-solute equivalent interactions in the implicit-solvent model.

These results highlight the role of having explicit solvent molecules in the system in reproducing locally-triggered events that are poorly replicated when the effect of the solvent is averaged in the system. Moreover, our results indicate that the main difference between Martini 2.2 and Dry Martini (indicated by the SOAP pair distance of Figure 4.17b) has a local nature, namely how local environments in the lipids are modeled in the two FFs. Finally, they further display how comparison approaches, whose rely only on average data/obervables, may be insufficient, and demonstrate the potential of high-dimensional data-driven analyses in providing detailed information in this sense.

Chapter 5

Markov state model analysis of supramolecular polymers

In the previous chapters, we presented and extensively discussed the role of unsupervised ML-based approaches to process MD simulation data of complex and dynamic molecular systems. A detailed description of a system during any given time evolution mechanics, like in the MD case, usually requires the most complete knowledge of the conformational changes that the system undergoes when transitioning between the relevant stable (and metastable) states populating the reference phase space of the system under investigation. Our results showed how combining descriptors of atomic environments and dimensionality reduction algorithms, we can first identify and then classify such conformations; usually through the help of a clustering algorithm. To this end, the SOAP[105] descriptor proved to yield a very suitable representation of the structural features of virtually any kind of supramolecular aggregate, offering a nice compromise between descriptive power and computational cost.

The appeoaches discussed in the previous chapters are based on the concept of projecting high-dimensional SOAP data on "static" low-dimensional spaces (*e.g.*, by using PCA), where the keyword *static* means that no time-dependent knowledge is required when projecting the data. For example, the PCA algorithm is trained without any information on the evolution history of the original high-dimensional data points, even though they come from time-ordered MD snapshots, so that it can to reproduce as much as possible the data spatial variance (for more information on PCA and dimensionality reduction in general see Sec. 2.3). In general spatial and time data

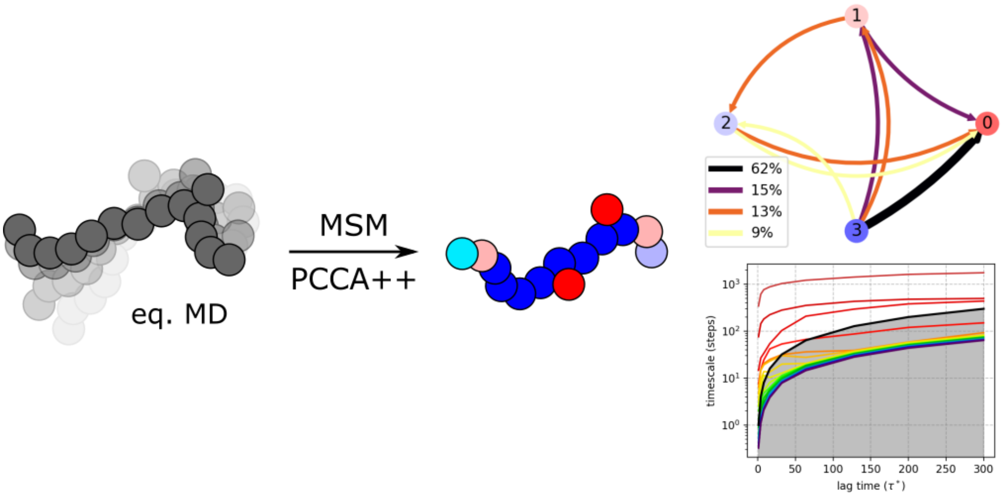


Fig. 5.1 Schematisation of the application of the MSM theory to MD simulation data of a supramolecular polymer.

variances do not correlate with each other, hence, by enforcing this “static” approach, we could overlook and/or badly represent some important time-dependent features of the system, as shown in the example of Figure 2.5 in Chapter 2. Despite this downside, our SOAP-PCA approach proved to be quite effective in capturing all the necessary structural information, yielding a detailed static and kinetic picture, once it is paired with a clustering algorithm and the notion of time-history is reintroduced (*e.g.*, by counting the transition between different clusters).

In the following we introduce a different approach towards the study of the kinetic of conformational changes inside a supramolecular aggregate, namely the application of the Markov State Model (MSM) theory on data obtained *via* MD simulations. Starting from the SOAP description of a given system, we will reduce its dimensionality using the tICA[115, 116] algorithm, achieving a projection of our data on a low-dimensional space that critically takes into account the time evolution of each monomer (*e.g.*, the time-series) inside a given structure. Then, through the application of the MSM theory we will extract the number of relevant kinetic states that participate and define the equilibrium dynamics of a given supramolecular system, the transition probability matrix that controls their evolution with time, and other related important dynamic quantities.

In the next section we will briefly introduce the fundamental theoretical framework of MSM and present the general workflow of steps needed to use those concepts to study MD simulation data. Finally, we compare our tested PCA-PAMM approach with the new tICA-MSM one, taking as test case the minimalistic CG-FF BTA fibres previously studied (Chapter 3.1) highlighting the main differences, pros and cons of the two approaches.

5.1 Markov state models

Markov state models (MSMs)[117, 126, 156] represent a class of stochastic physical models used for characterising the long-timescale dynamics of molecular systems. The dynamics of a given system is represented by a series of "memoryless", probabilistic jumps between a set of states. Thus, the physical ingredients of a general MSM are just a set of relevant conformational states and a set of rules on how these states evolve in time. The theoretical framework and methodological progresses on MSMs are based on the mathematical theory of conformation dynamics introduced by Schütte *et al.*[126] and further developments by many contributors, amongst them, by Noé *et al.*[117, 118]. In this scheme the time evolution of a molecular system is modeled by the so called Markov propagator $\mathscr{T}(\tau)$, a mathematical operator, that effectively set the rules on how the distribution of states of a given molecular system, $\rho_0(x)$, evolves with time:

$$\rho_\tau(x) = \mathscr{T}(\tau)\rho_0(x), \tag{5.1}$$

where τ represents the unitary step of the time evolution and $\rho_\tau(x)$ is the "updated" distribution of states. For example, if we consider a discretised ideal dynamical system, composed of N particles that can be found in only three unique and separated states, $\mathbf{q} = \{A,B,C\}$, then $\rho_0(\mathbf{q})$ would represent the initial, $t = 0$, distribution of the N particles in said states. At time $t' = t + \tau$ instead we would have a different distribution, $\rho_\tau(q)$, which is obtained through the application of the transition rules set by the Markov operator. Interestingly, the same holds when considering a more complex molecular system undergoing MD evolution (Fig. 5.2 left side). However, in this case the description of the possible available states, their evolution and their relative distribution at a given time is usually more complex. A possible solution, to simplify the problem, is to follow the dynamics of the system as a function of some CV, effectively transforming the original phase space into a feature space Ω, that is (possibly) easier to represent. For example, the dialanine molecule (L-Alanyl-L-alanine, $C_6H_{12}N_2O_3$) is fully represented in AA-MD by a configuration

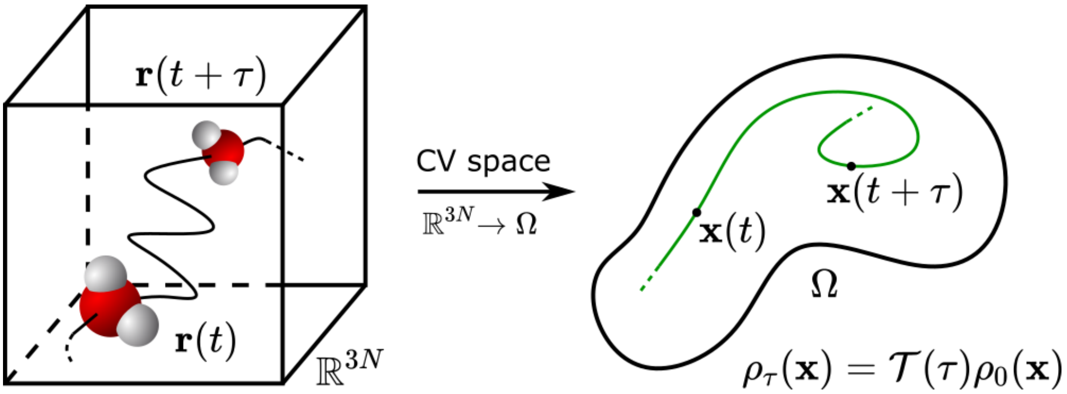


Fig. 5.2 Representation of the time evolution, *i.e.*, the trajectory, of a molecular system in its original Cartesian space and in a given CV representation.

vector $\mathbf{r} \in \mathbb{R}^{69}$ (*i.e.* three Cartesian components for each of the 23 atoms), but we can successfully follow its configurational changes by focusing on a much simpler representation made by just two angular variables, $\mathbf{x} = \{\phi, \psi\}$, *i.e.*, the two backbone dihedral angles.

The key goal, when formalising a MSM, is to find an approximation to the exact Markov operator $\mathscr{T}(\tau)$, which, as said, sets the rules for the time evolution of the system under investigation. Once a form of $\mathscr{T}(\tau)$ is known, it is possible to formally extract all the information about the system dynamics from the spectral decomposition of said operator,

$$\mathscr{T}(\tau) \circ \phi_i(\mathbf{x}) = \lambda_i \phi_i(\mathbf{x}), \tag{5.2}$$

where ϕ_i and λ_i are the eigenvectors and eigenvalues respectively. From theoretical considerations, the first eigenvalue is guaranteed to be unitary, $\lambda_i = 1$, and the related first eigenvector coincides to the equilibrium distribution of states, $\phi_1 \equiv \mu$. The others eigenvalues have values strictly less than 1, and their associated eigenvectors represent the dynamical modes that describe the transient evolution of the system toward its equilibrium state. Moreover, and most importantly, the individual eigenvalues, $\lambda_{i>1}$, are directly connected to the so-called *implied timescales*, through the expression,

$$t_i = -\frac{\tau}{\ln \lambda_i} \quad \text{with} \quad i > 1. \tag{5.3}$$

Equation 5.3 relates a dynamical mode to its relaxation timescale, a quantity that can correlate with experimentally measurable correlation functions (*e.g.*, fluorescence

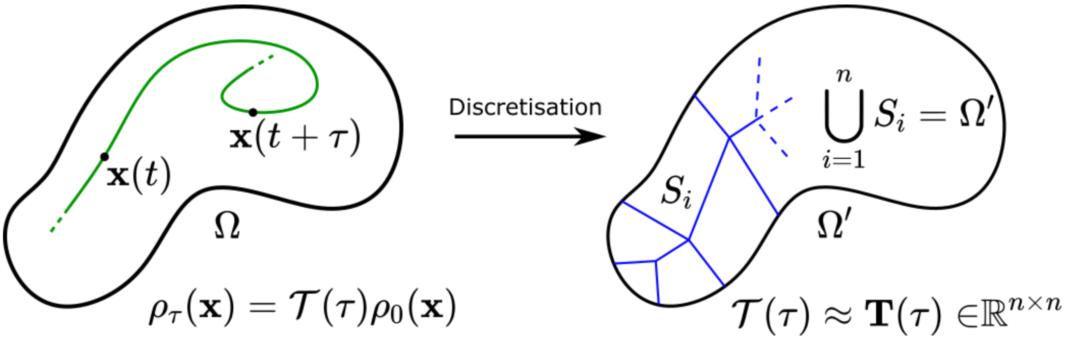

$\mathbf{x}(t+\tau)$
$\mathbf{x}(t)$
Ω
$\rho_\tau(\mathbf{x}) = \mathcal{T}(\tau)\rho_0(\mathbf{x})$
Discretisation
$\bigcup_{i=1}^{n} S_i = \Omega'$
S_i
Ω'
$\mathcal{T}(\tau) \approx \mathbf{T}(\tau) \in \mathbb{R}^{n\times n}$

Fig. 5.3 Approximation of the Markov operator through discretisation of the phase space.

correlation, scattering functions in neutron or X-ray scattering).[157, 158] Finally, the magnitude of the eigenvalues allow us to partition the total dynamics of the system into an set of *slow* (dynamical) modes, $\{\lambda_i\}_{\lambda_i \geq \lambda_m}$, and a set of *fast* (dynamical) modes, $\{\lambda_{i'}\}_{\lambda_{i'} < \lambda_m}$, where λ_m is a chosen relaxation timescale threshold relevant for the physics of the system.

At this point, we require to define an explicit and usable expression for the Markov operator $\mathscr{T}(\tau)$. For most systems, this is often impossible, as each point in the continuous phase space of interest should represent a different metastable state, adding up to an infinite amount of possible configurations, hence translating to an indefinite form for $\mathscr{T}(\tau)$. One straightforward solution is to force a discretisation of the phase space, effectively creating a finite amount of *n macro*-states, by pairing together similar states (Fig. 5.3). Then, the time-evolution operator can be approximated by computing the corresponding probability transition matrix that describes the time-evolution of the particles in the *n macro*-states, $\mathscr{T}(\tau) \approx \mathbf{T}(\tau) \in \mathbb{R}^{n\times n}$. The discretisation operation is often accomplished by simple clustering algorithms, like K-means. Although there is no set rule for choosing the number of clusters, the tessellation must be dense enough to optimally approximate the Markov operator (*i.e.*, the expected values of the corresponding relaxation timescales approache the real values) but not too high to compromise the meaning of the transition probability matrix (*i.e.*, if the transition between the defined clusters are too rare, the computed transition probabilities will be numerically instable and will not have statistical significance).[157, 159] Anyway, the total number of discrete states is usually on the order of hundreds or at maximum few thousands.[118]

Once again, from the probability transition matrix (*i.e.*, the approximation of the Markov operator) we can compute its spectral decomposition and all the relevant related quantities, which are the core of the Markov State Model. Finally, it is possible to obtain an even coarser representation of the dynamic states, easing the interpretation of the kinetic model. This is usually done by means of an additional clustering step, which directly acts on the matrix $\mathbf{T}(\tau)$, creating an "effective" transition matrix $\tilde{\mathbf{T}}(\tau) \in \mathbb{R}^{m\times m}$, with $m \ll n$. One of the most used algorithms is know as PCCA+[127] and its enhanced version PCCA++[127], which acts on the space of the eigenvectors of the matrix $\mathbf{T}(\tau)$ and creates linear combinations of them, pairing those that are similar and reducing their number to a total of m. Notably, since the number of eigen-pairs is equivalent to the number of entries in the matrix (*i.e.*, the number of discrete states of Ω), the same rules used in creating combination of the eigenvectors will be used to obtain a coarser representation of the discrete states of the Ω space.

In the following paragraph we will discuss a simple application of the MSM concepts on an idealised system. We will go through all the practical steps, which we will later use to characterise the minimalisti CG-FF BTA fibres.

5.1.1 MSM of a particle in an asymmetric quadruple well 2-D potential

Most numerical implementations for MSM estimation, follow a defined amount of steps,[118] which take in input the *time-series*, *e.g.*, the trajectory of the particles in the system under investigation, define an approximation for the Markov operator $\mathscr{T}(\tau)$, and find the optimal dynamic states which characterise the system. If applied to these output of a MD simulation the steps can be summarised as follow:

- *Featurisation*: the particles coordinates, if necessary, are represented as function of a few CVs, which effectively describe the system kinetic evolution during the MD, and form the input time-series for the MSM (*e.g.*, using the two dihedral angles of the dialanine peptide instead of the atoms positions).

- *Dimension reduction*: if the *featurisation* results in an excessive number of dimensions (CVs) we need to reduce the input to just a few slow CVs (typically 2-100).

- *Discretisation*: the space is discretised by clustering the projected data, typically resulting in 100-1000 discrete microstates.

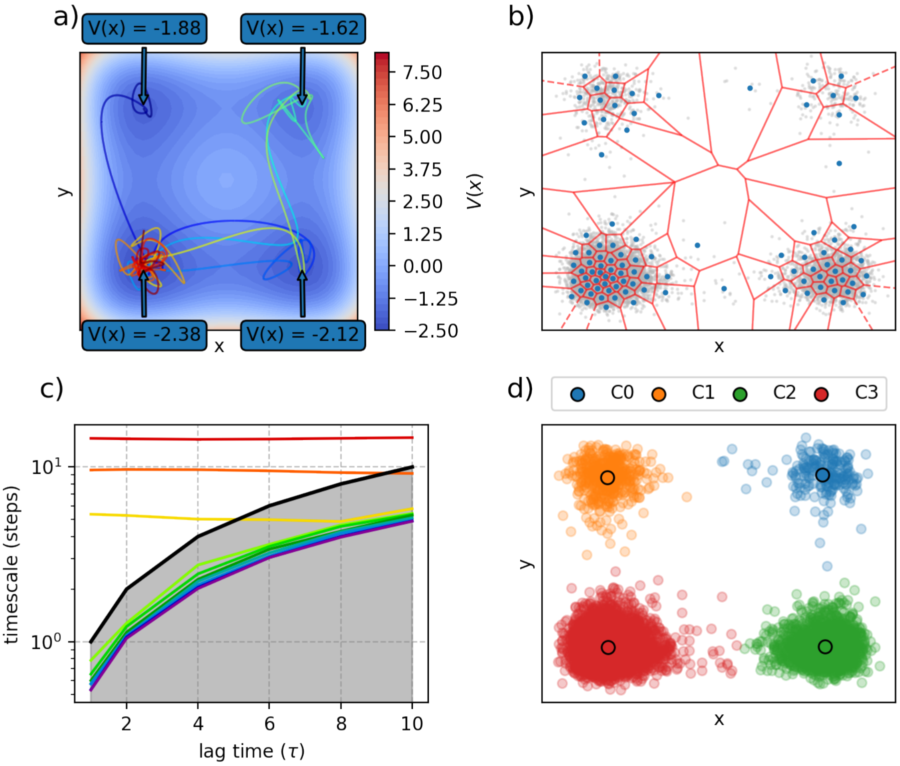


Fig. 5.4 Complete MSM analysis of a particle in a asymmetric quadruple well 2D potential. a) Example of trajectory (first 1000 steps) depicted by a solid coloured line on top of the PES. The values of the well depths are in arbitrary energetic units. b) Voronoi tessellation of the total space visited by the particle in a total of 100 clusters. c) Implied timescales plot (Eq. 5.3) computed from the transition probability matrix obtained from the labels of panel (b). d) PCCA++ assignment of relevant dynamic states for a given lag time ($\tau = 2$).

- *MSM estimation*: the transition probability matrix is estimated by counting the transitions between the defined micro-states (discretisation), given a specified time window τ, also called lag-time. The matrix effectively represents the Markov operator and it can be decomposed to compute the state-dependent implied timescales (Eq. 5.3).

- *Coarse-graining*: the estimated MSM is often coarse-grained to a few number of states, granting an easier interpretable kinetic model.

In this example, and in all the further applications, the MSM and related analysis will be computed using general python scripts and the python library *Deeptime*[160].

The chosen test case is a point particle moving in a 2-D potential energy surface with four minima, for which we collect a total of 10000 steps of its MD trajectory (Fig. 5.4a). In this simple test case it is not necessary to apply a featurisation or a dimensionality reduction as the tuples of particle positions $\{x(t), y(t)\}$ are already descriptive enough. Following the analysis workflow, the space spanned by our particle in those 10000 steps is divided into 100 discrete clusters by K-means clustering (Fig. 5.4b, red lines, equivalent of a Voronoi tessellation of space). At this point, we have to compute the probability transition matrix $\mathbf{T}(\tau)$, which for any given lag time τ will be of size 100×100, and the related implied timescales (Eq. 5.3). Notably, those relaxation timescales are physical quantities, which define the physics of the system under investigation, and as such they should not depend on the specific lag time chosen, since the latter is just a parameter that we can freely set.[118] Moreover, it should be easy to understand that not all of those 100 states are relevant for a kinetic description of the system, or rather only four should be relevant, given how the system is built (*i.e.*, the particle bounces among the four potential wells). The relevant number of states can be retrieved from computing the relaxation times for increasing lag times and plotting them against the lag time variation (Fig. 5.4c, implied timescales plot). This particular plot is usually represented in a log-log scale of the lag time τ *versus* the implied timescale values and it is divided in two main regions delimited by a solid black line, that represents the theoretical timescales with the increase in lag time (computed using Eq. 5.3). This serves as a limit boundary, as all the data that lies below the black line (in the grey area of Fig. 5.4c) will have an implied timescale that is faster than the lag time used to sample the trajectory, making its value not statistically relevant. Thus, those lines that are flat (*i.e.*, independent of the choice of τ) and do not intercept the black solid line, will be the m relevant slow (dynamical) modes of the system, $\{\lambda_i\}$ with $2 \leq i \leq m$. In our example, from lag time $\tau = 1$ to $\tau = 4$ the number of relevant states are always four. Thus, we can use this number as input for the PCCA++ supervised clustering, which groups the 100 original clusters into the four relevant ones. Finally, we have access to $\tilde{\mathbf{T}}(\tau = 2) \in \mathbb{R}^{4\times 4}$ from which we can compute all sort of related dynamic quantities to give a complete description of the system dynamics between the four potential wells (*e.g.*, equilibrium distribution, dynamical modes, physical timescales, flux pathways).

5.2 MSM of representative supramolecular polymers

In this last section we will go over the application of MSM theory toward the structural and kinetic characterisation of the crucial defect dynamics in a representative family supramolecular polymer. We choose the "super"-CG BTA[31] model to be our test case, as it is a relatively simple, yet functional, model making the results more straightforward to interpret (*e.g.*, we can easily obtain a wide range of defects related behavior).

5.2.1 Methods

MD parameters and SOAP-PCA-PAMM analysis

All MD simulations were carried out with the GROMACS 5.1.2 software[161] in NPT conditions. We used the same monomers models as presented in Chapter 3.1, with the same MD simulations setup and parameters. We simulated the three fibre structures, here referred as $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\omega}$ (respectively **Fibre 1,2,3** in Ref. [31]/Sec. 3.1) for an additional 10 μs of unbiased MD, sampling the trajectories every 1 ns.

We repeated the structural characterisation through the application of our SOAP-PCA-PAMM analysis, again using the previously introduced setups (see Ch. 2 for the theory behind it and Ch. 3-4 for relevant results). The SOAP descriptor was applied using a custom python script using the DScribe[128] package, setting the parameters *rcut*= 8 Å, *nmax*= 8, *lmax*= 8, and leaving the rest as default. The total dimension of the SOAP feature vectors was reduced by mean of PCA, keeping the first 3 principal components, which account for more than 90% of the total data variance. The PAMM clustering algorithm was applied to the three minimalistic fibres, using once again the same parameters as in the previous case (Ref. [31]/Sec. 3.1): fspread = 0.30, quick-shift = 1, bootstrap-runs = 73, merger-threshold = 0.005, and grid-size samples 1000, 2000, 2000 for $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\omega}$ respectively. It is worth pointing out, that in this case we kept the PC spaces separated, thus obtaining an individual description of each fibre structural motifs, not directly comparable between them (see Ch. 4 for discussions on comparability and classification).

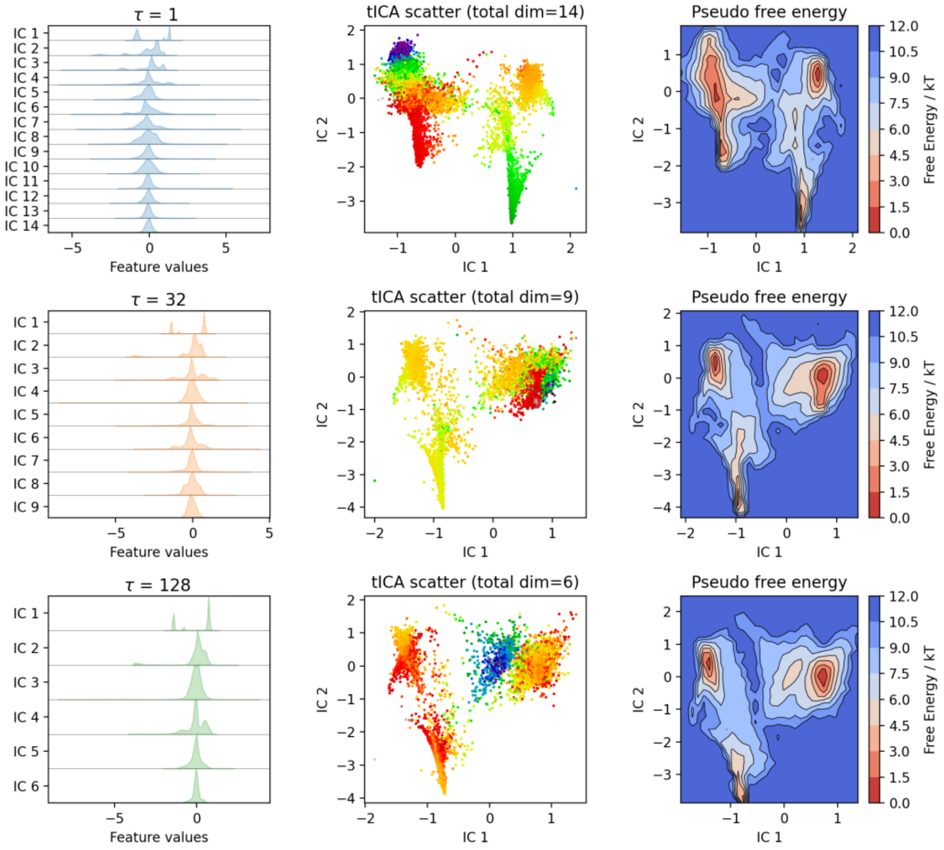


Fig. 5.5 tICA dimensionality reduction on the supramolecular polymer $\boldsymbol{\omega}$ across three different lag times, $\tau = 1, 32, 128$ (top to bottom). In detail on the left, the distributions of the single ICs values; the total number of ICs reported is the number needed to retain 0.95 of kinetic variance for that system at that specific lag time. In the centre, scatter plot of the first two ICs, coloured according to the third IC (hot colour means positive values and cold colours means negative ones). On the right, pseudo FES computed from the first two ICs of the reduced dataset.

MSM analysis

The MSM analysis was carried out by mean of python scripts using the library Deeptime[160]. The tICA dimensionality reduction algorithm was applied to the SOAP data of each of the three fibres, setting a kinetic variance threshold of 0.95; the sum of the eigenvalues of the projection matrix, in increasing order, amounts to 0.95. Thus, this means that for all the applications of tICA, we retained as many independent components as required to reach the set kinetic variance value. Additionally, we repeated the tICA reduction for a set of different lag times, $\tau^* = \{1, 4, 8, 16, 32, 64, 128, 200, 300\}$ for all fibres $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\omega}$. The adimensional lag time

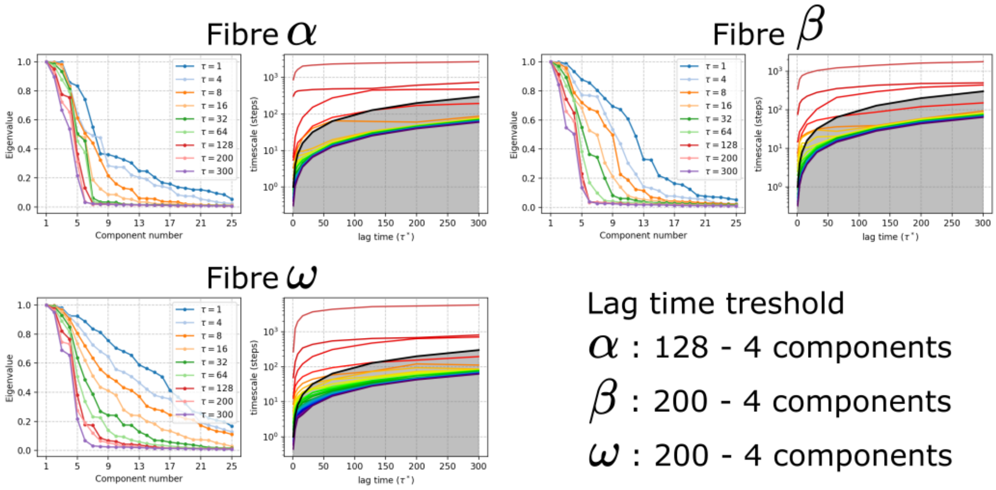


Fig. 5.6 MSM spectral decomposition and implied timescale plots for Fibres $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\omega}$. Each system shows a Markovian regime with 4 total relevant kinetic states.

parameter is defined as $\tau^* = \tau/\Delta t$, where Δt is the native simulation time window, which for our case is 1 ns. In the following, to simplify the notation, we will just refer to τ, without specifying its units, when referring to the lag-time parameter. Figure 5.5 reports three samples tICA dimensional reduction done on the Fibre $\boldsymbol{\omega}$ supramolecular polymer for three different increasing lag times. For each of the tICA reduced datasets, we computed 50 discrete clusters using the KMeans clustering algorithm, and used those to compute a probability transition matrix representative of each supramolecular polymer system. Finally, we built the MSM for each fibre using the Deeptime python library, *via* the `MaximumLikelihoodMSM()` estimator class and fitting directly from the discrete timeseries obtained through the discretisation step. We computed the implied timescales plots, selected the number of relevant kinetic states and coarse-grained the model with PCCA++; in all cases the number of relevant states was found to be four (Fig. 5.6).

SOAP metric comparison

Given that we treated our system as separated, *i.e.*, each has its own low-dimensional projection (PCA or tICA) and its own clustering, we need a way to compare between their individual clustering motifs. We approached the problem in a different way: once the clusters are defined we computed for all of them a global simulation-average fingerprint, by averaging each frame the local fingerprint of the monomers that shares

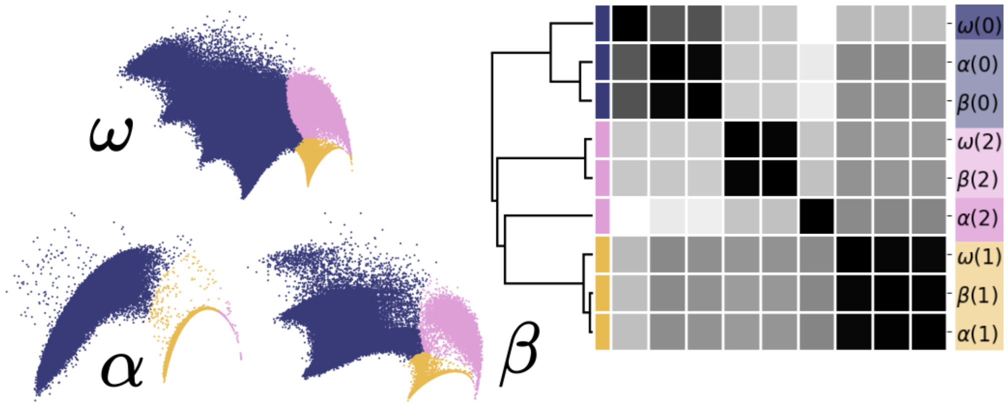


Fig. 5.7 SOAP-PCA-PAMM individual analysis for Fibres $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\omega}$. Left side, the main molecular motifs for each fibre are projected as colours on the scatter plot data. The three plots show the distribution of points of the SOAP data in the first two principal components. Right side, SOAP distance matrix between the motif-wise simulation-averages global SOAP descriptors; dark colours mean low distance values (black is zero) and lighter colours mean higher distance values.

a same cluster index, and then averaging again those values to obtain a single global for each of the cluster in each of the fibre.

5.2.2 Results

Molecular motifs-centered metric comparison

As said, we started by applying our SOAP-PCA-PAMM analysis to the trajectories of the three supramolecular fibre "super"-CG models, extended by 10 μs of simulation time. We here employ a slightly different analysis scheme, that make use of the definition of motifs-centered averages to compare and classify the molecular motifs emerging in different systems. As expected, we obtained equivalent results as those presented in Section 3.1. Each fibre is dominated by three main molecular motifs, that represent the backbone (or fibre body) monomers, labeled as cluster zero, the terminal monomers or tips, labeled as cluster one, and either the freely diffusing monomers (in Fibre $\boldsymbol{\alpha}$) and the defect monomers (in Fibres $\boldsymbol{\beta}, \boldsymbol{\omega}$), all labeled as cluster two (Fig. 5.6 left side). However, with the current available data, we cannot quantitatively compare the clusters of different fibres to each other, aside from a qualitative visual inspection of resemblance (as also explained in previous chapters).

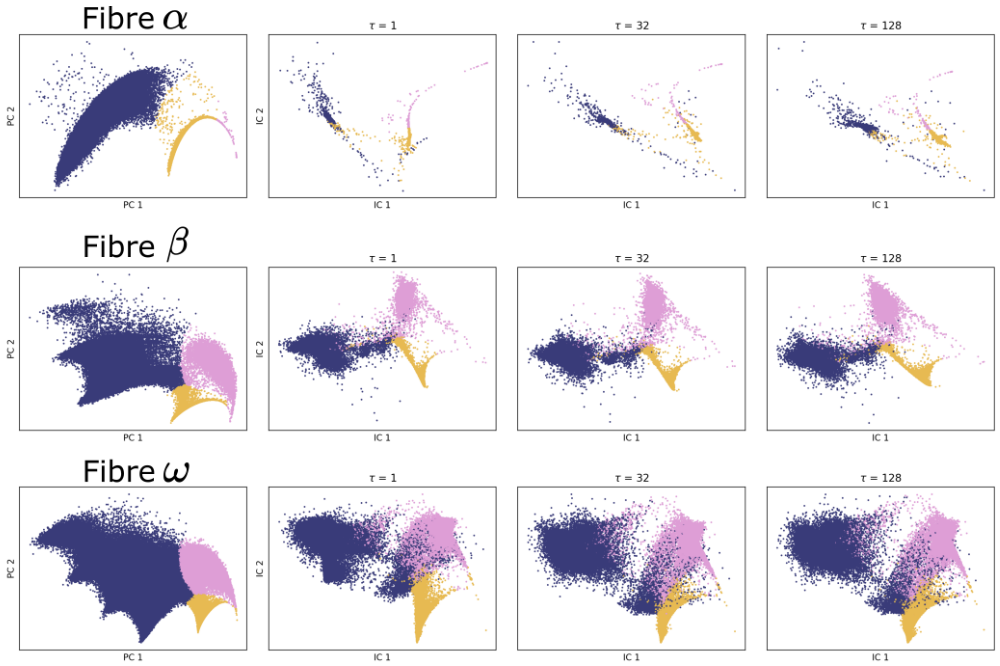


Fig. 5.8 Comparison of the PCA and tICA embedding. From left to right it is reported the PCA data, and the tICA data obtained with lag time $\tau = 1, 32, 128$. The colours reflect the PAMM clustering labels and is kept the same for all the plots shown.

In order to assert the relationships between the clusters, we extended the SOAP metric approach presented in Chapter 4. For each of the different fibre molecular motifs we computed the SOAP global fingerprints representative of that motifs, obtaining a single feature vector for each motif. Then, we built a SOAP distance matrix, representative of the motifs mutual pair distances, and we rearranged it using a hierarchical clustering, which highlights the similarities, rearranging the matrix (Fig. 5.6 right side). The matrix underlines some interesting points: the fibre tips (cluster one) are all identical and equally separated by all the other motifs (except from the $\omega(0)$ cluster which is slightly more separated). The defect motifs, cluster two, are identical for $\beta(2), \omega(2)$ and very different from $\alpha(2)$. This is expected, as they represent very distinct molecular environments, and they should be effectively colored with two different colours (as done in the previous analysis of Sec. 3.1) Lastly, the backbone monomers motifs, or body motifs, show how the presence of an higher concentration of structural defects (fibre $\boldsymbol{\omega}$ case) causes the fingerprint $\omega(0)$ to fall apart from the other two, $\alpha(0), \beta(0)$, which have none or little persistent structural defects. Overall these results follow closely those we would have obtained by considering a shared dataset containing all three fibres examples, as in that case clusters are directly comparable from the start. This approach can be used as a nice alternative to that explained in Chapter 4 for the quantitative comparison of a given set of unknown supramolecular structures, where the single motif fingerprints could be even stored and used for future comparisons, without the need to re-train the clustering algorithm on an enhanced dataset. Furthermore, this small variation on the SOAP metric analysis allows us to directly compare different motifs fingerprints obtained through different approaches, by simply updating the distance matrix with the new set of descriptor fingerprints. Our next goal, is to assert the differences that might arise in the overall structural patterns by application of a MSM to the same systems.

MSM of supramolecular polymers

We start the MSM approach by computing the tICA dimensionality reduction of the fibres SOAP dataset at different and increasing lag times. Figure 5.9 shows a sample of the tICA results, with the scatter plots coloured according to the PAMM clustering labels. Starting from the first value of lag times, $\tau = 1$, we observe a completely different spatial distribution of points. This is expected, as the two dimensionality reduction algorithms differ heavily on the way they embed the data on the low dimensional space (see Chap. 2). The colour distribution of the structural motifs appears distorted, indicating that in all three cases, but mostly for Fibre $\boldsymbol{\omega}$, there could be structural details that are not consistent between PCA and tICA predictions. From the tICA data we compute the initial transition probability matrix, by dividing each space in 50 discrete states and counting the transitions between them. To complete the MSM estimation, we computed the spectral decomposition of each transition matrix and plotted the correlated implied timescale trends with respect to the lag-time (Fig. 5.6). The three fibres shows a similar behavior, with a total of 3 timescales that reach plateau inside the lag-times range considered, Fibre $\boldsymbol{\alpha}$ from $\tau = 128$ and $\boldsymbol{\beta}, \boldsymbol{\omega}$ from $\tau = 200$. Furthermore, in all cases, two timescales have smaller and similar values compare to the other one, which is one order of magnitude bigger. Finally, to gain a better insight on the intrinsic kinetic of the systems we used the supervised clustering PCCA++ to reduce the number of states from the initial 50 to the 4 relevant from a kinetic point of view.

Figure 5.9 summarises the main results obtained through the application of the PCCA++ clustering algorithm. All of the structural motifs found are coloured from

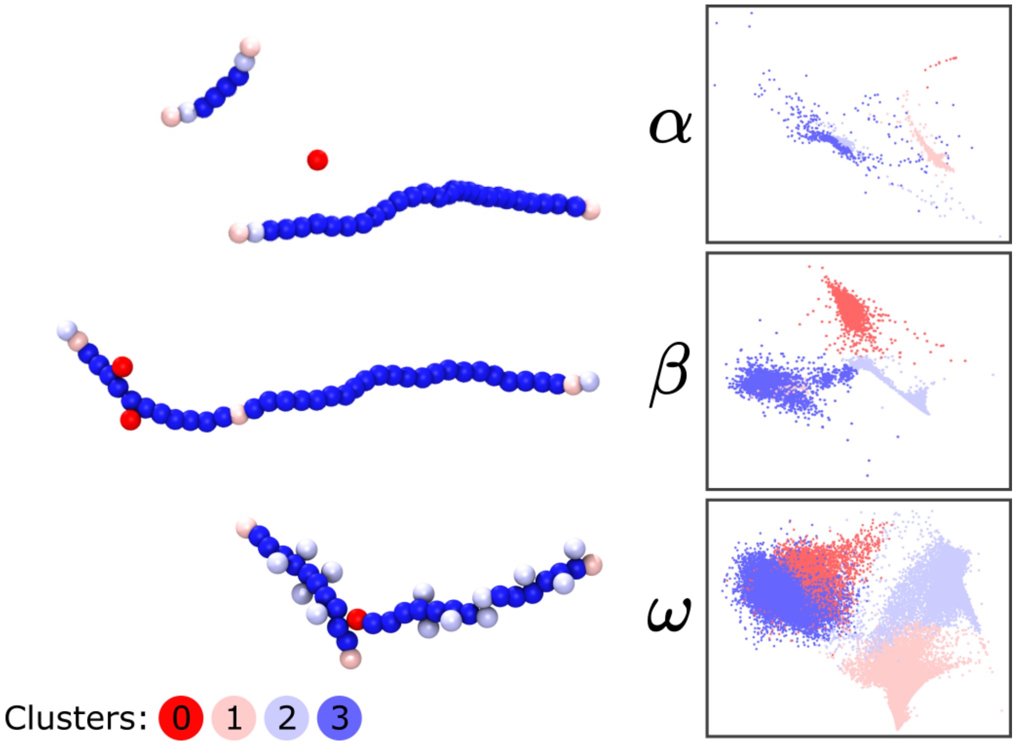


Fig. 5.9 MSM results on the three sample fibres. Markov state based clustering. The scatter plots are coloured according to the cluster membership obtained through application of the PCCA++ algorithm.

red to blue, *via* automatic colour assignment by the algorithm. Once again, it is worth stressing out, that each clusters set is computed individually, hence they cannot be directly compared.

Fibre $\boldsymbol{\alpha}$ is divided in four relevant structural motifs. Visual inspection confirms that these motifs resemble closely the one found by the SOAP-PCA-PAMM analysis: fibre body (blue cluster), fibre tips (pink cluster) and freely diffusing monomers (red cluster). Additionally, the MSM identifies a fourth motif (light blue cluster), absent in the previous analysis. This new structural environment consistently appears as the monomer between the tips and the body of the fibre (Fig. 5.9 top). Although being almost identical to a "body" monomer, from pure structural considerations, the separation suggests that this new motifs might have some kinetic relevance. Furthermore, the cluster labelling (from red to blue) matches with a qualitative observation of the monomer mobility: the red cluster being the most mobile (freely diffusing monomers) and the body monomers being the most constrained ones.

Fibre $\boldsymbol{\beta}$ is divided in four relevant structural motifs, and again such division follow closely the previously obtained structural motifs: fibre body (blue cluster), fibre tips (light blue cluster) and structural defects (red cluster). This time the additional structural motif is represented by the pink cluster, which appears sporadically along the fibre body. Once, again the qualitative connection between cluster labels and monomer mobility appears to be preserved. In the absence of freely diffusing monomers, the most mobile are the structural defects, which are known to be able to undergo a "surfing"-type motion, moving along the fibre body.[31, 50] The structural identity of the pink motif can be explained considering the physical features of the monomers that constitute fibre $\boldsymbol{\beta}$ compared to fibre $\boldsymbol{\alpha}$. In the former fibre, the monomers have an increased core solvophobicity, making the perfect stacking configuration of monomers slightly unfavorable from an energetic point of view, which in turn, allows the fibre body to spontaneously twist and bend till a structural defect is formed.[31] Because of this reason the pink monomer can be considered more mobile compared to the other body-type motifs.

Lastly, fibre $\boldsymbol{\omega}$ is also divided in four relevant structural motifs, which are in line with the previous obtained ones: fibre body (blue cluster), fibre tips (pink cluster) and structural defects (light blue cluster). Also here, an extra structural monomer (red cluster) is identified, representing a monomer which acts as a tip for a whole fragment of the main fibre body and, at the same time, as a structural defect for the remaining part of the fibre (Fig. 5.9 bottom). This is a completely new and undefined motif in the PCA-PAMM approach, where it would be paired either with the fibre body or with the tips motifs (Fig. 5.8). Surprisingly, a visual inspection of the MD trajectory suggest also an enhanced mobility of the red motif compared to the other states populating the fibre, as it allows for a slight surfing of the ordered fibre fragments. Overall the MSM approach is able to detect all the structural motifs previously obtained, providing in some case some further structural detail.

To better quantify the individual and visual information obtained from the MSM-based clustering, we compare the different motifs using the motif-centered SOAP metric approach just introduced. We thus computed the simulation-average SOAP global descriptors of each MSM identified motif, we built the respective distance matrix using the SOAP metric and computed their overall hierarchy with a hierarchical clustering algorithm (Fig. 5.10a). Overall, the results follow our previous visual inspection adding more details on the specific relationships that exist between the clusters. We can identified four main dendrogram leaves: one containing all

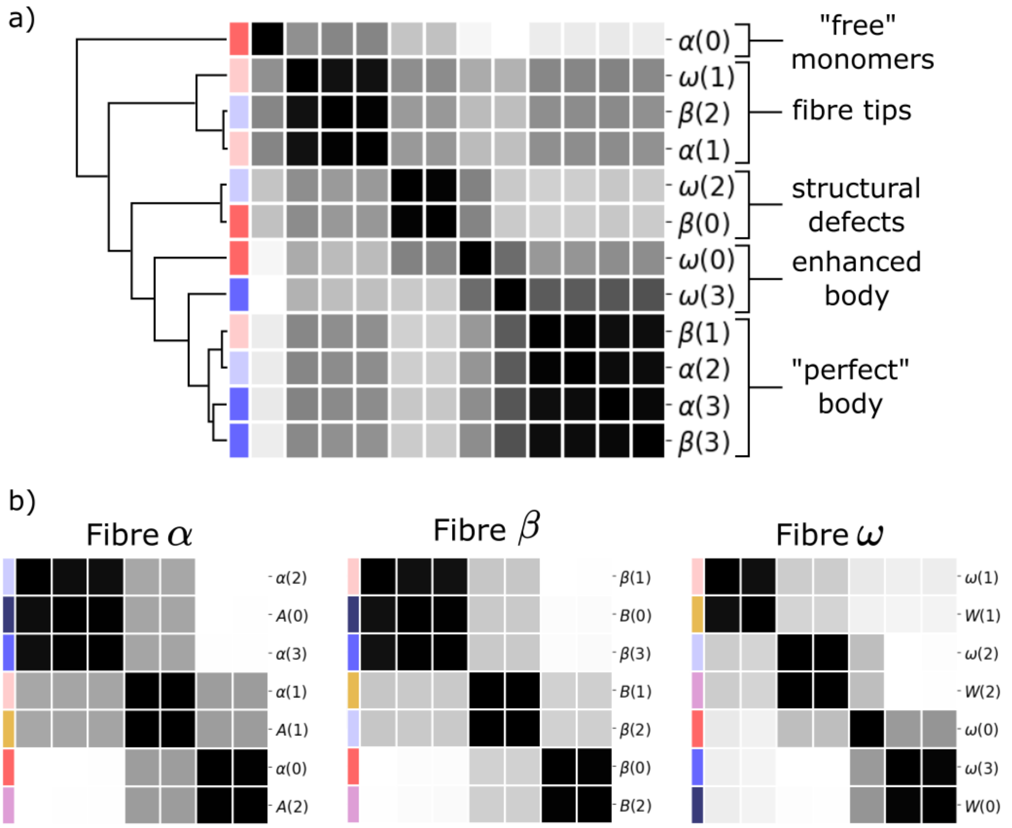


Fig. 5.10 MSM motif-centered metric comparison. a) Distance matrix computed from the simulation-averages of each of the structural motifs computed through the MSM. b) Comparison between the individual structural motifs of each fibre obtained with both PCA-PAMM and tICA-MSM approaches. To avoid confusion, the three fibres, when referring to the PCA-PAMM classification method, are indexed as $\boldsymbol{\alpha} \rightarrow \mathbf{A}$, $\boldsymbol{\beta} \rightarrow \mathbf{B}$, and $\boldsymbol{\omega} \rightarrow \mathbf{W}$.

the structural motifs that describe the polymer backbone, one containing all the structural defects that can grow on it, one containing the terminal monomers or tips, and one for the freely diffusing monomers. The body dendrogram leaf is in turn divided into two additional separates ones containing the $\omega(3)$ and $\omega(0)$ motifs. These appear close to each other, but exhibit critically different relationships towards the other motifs. The $\omega(3)$ motif shares more similarities with the other backbone motifs distancing itself equally from the remaining structural ones. Instead, $\omega(0)$ is found to be closer to the structural defects than to the other body motifs. Finally, we compared all the individual fibre motifs obtained *via* PCA-PAMM and tICA-MSM by computing fibre-specific distance matrices (Fig. 5.10b); for clarity we changed the labels indexing the PCA-PAMM identified motifs to $\boldsymbol{\alpha} \rightarrow \mathbf{A}$, $\boldsymbol{\beta} \rightarrow \mathbf{B}$, and $\boldsymbol{\omega} \rightarrow \mathbf{W}$. Interestingly, almost all the tICA-MSM identified motifs share perfect similarity with one of the PCA-PAMM motifs, with the exception of $\omega(0)$, that positions itself in the middle between a body and a defect environment. This last evidence suggests that the tICA-MSM approach might in fact capture motifs that are relevant from a purely dynamic point of view, which appear to be overlooked otherwise.

Kinetic characterisation based on the MSM

In order to assert whether the additional structural motifs might be relevant from a kinetic point of view we characterised some of the key aspects of the fibres dynamics. Starting from the newly defined coarser transition probability matrix, $\tilde{\mathbf{T}}$, we computed its spectral decomposition and observed the behavior of the different eigenvectors. The first eigenvector, $\phi_1 \equiv \mu$, is a static property, and it represents the equilibrium distribution of states for a given system. From the other three eigenvectors we focus on the second, ϕ_2, which represent the slowest dynamic mode that contributes to relaxation toward the equilibrium distribution states of a given fibre. Figure 5.11a reports a visualisation of the ϕ_2 modes plotted on the tICA embedding of the three fibres and coloured by the value of the specific eigenvector components. In plain words, ϕ_2 is a vector containing a total of four contributions (*i.e.*, equals to the number of states of the matrix it has been computed from), which effectively represents the dynamic relationship that exist between those states and that is characterised by the relaxation times t_2 (Eq. 5.3). The explicit numbers composing the vector (ϕ_2) act as a weight factor highlighting which are the most relevant states that contributes to the selected dynamical mode, which have a value of either $+1$ or -1, or zero if they do not participate. Other intermediate values accounts for secondary contributions.

In Fibre $\boldsymbol{\alpha}$ the slowest mode (ϕ_2) is represented by the reaction of a monomer in the body of the fibre (blue cluster) that becomes a tip (pink cluster). This transition is reminiscent of a cleavage of the fibre body into two smaller fragments. Conversely, the detachment of a single monomer from a fibre tip, ϕ_3, is a faster process, having an implied timescale one order of magnitude smaller. The last mode, ϕ_4, mainly represents fluctuations on the fibre body that creates the light blue domain. Fibre $\boldsymbol{\beta}$ slowest mode (ϕ_2) is also represented by the structural transition that brings a monomer in the body of the fibre (blue cluster) to become a tip of the fibre (light blue cluster). The other two represent respectively the fibre body/tip turning into a structural defect (ϕ_3) and the fluctuation of the body monomers into pink structural domains (ϕ_4). Finally, for fibre $\boldsymbol{\omega}$ the slowest mode (ϕ_2) is represented by the

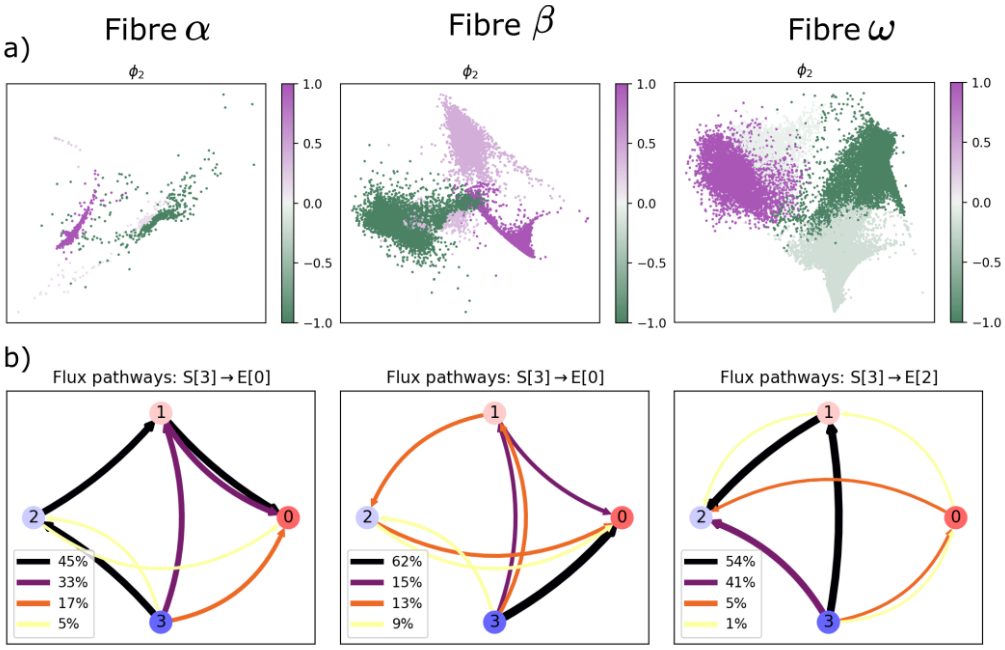


Fig. 5.11 Detailed information of the three fibre kinetic behavior within the MSM. a) Slowest mode (second eigenvector ϕ_2) visualised on the tICA scatter plot of each fibre. The mode describes the mutual transition of the states having values close to ±1. b) Dominant structural transformation pathways computed from the MSM for the creation of a freely diffusing monomer, in fibre $\boldsymbol{\alpha}$, and structural defects, in fibres $\boldsymbol{\beta}, \boldsymbol{\omega}$ starting the respective generic body structural motifs of each fibre (blue cluster).

reaction of a monomer in the body of the fibre (blue cluster) that becomes a structural defect (light blue cluster). The two remaining represent respectively the mutual transitions between tips and structural defects (ϕ_3) and between body monomers and the red structural environment (ϕ_4). The qualitative analysis of the eigenvectors trend allows us to retrieve some additional information on the key dynamical modes that eventually drive the system under study toward the equilibrium state, but they do not give any indication of the actual reaction paths that occur between the defined structural domains.

As a final proof of concept we computed the qualitative reaction path flux for the creation of a defected state starting from the fibre backbone, in all three fibres. The so-called “reactive flux”[162] can be computed directly from the discrete transition probability matrix exploiting the formalism introduced by Transition Path Theory[163] and implemented in the python library *Deeptime*. Figure 5.11b reports the reactive paths of a monomer starting (S) from state i and ending (E) in state j, *i.e.* the reaction $S[i] \rightarrow E[j]$. We investigated the pathways of a monomer starting from the fibre backbone structural environment ($S[3]$) to respectively a freely diffusing monomer in fibre $\boldsymbol{\alpha}$ ($S[3] \rightarrow E[0]$), a body structural defect in fibre $\boldsymbol{\beta}$ ($S[3] \rightarrow E[0]$) and fibre $\boldsymbol{\omega}$ ($S[3] \rightarrow E[2]$). The first graph (fibre $\boldsymbol{\alpha}$, paths $S[3] \rightarrow E[0]$) provides a confirmation of the fact that the monomer escape originates from the fibre tips. In fact, the two main reaction pathways computed, accounting for roughly 80% of all the total pathways, have a step involving the tip structural environment. The remaining 20% might be due to abrupt fibre cleavage and/or numerical imperfections of the clustering algorithm. The creation of structural defects in fibre $\boldsymbol{\beta}$ (paths $S[3] \rightarrow E[0]$) appears to be dominated by a direct "reaction" from the fibre body to the defected state, 62% of the total pathways. This suggests that the physical origin for a defect might be a sudden, sharp bending of the fibre, creating an elbow-like structure from which the structural defects can separate itself. The other relevant paths contain all different structural environments, reminiscent of step-wise processes. The picture of fibre $\boldsymbol{\omega}$ is it slightly different (paths $S[3] \rightarrow E[2]$), with the creation of a structural defect being dominated (54%) by a step-wise process that pass first by the creation of a fibre tip (pink structural environment). However, the second most probable pathway it is still represented by a direct conversion of the body monomer to defect. The shift in behavior between fibre $\boldsymbol{\beta}$ and $\boldsymbol{\omega}$ could be appointed to the differences in reaction speed, critically determined by the different physical nature of the two monomers (*e.g.*, different solvophobicity).

In summary, in this chapter we have presented a nice alternative on the comparison of different supramolecular aggregates based on their intrinsic structural motifs. In this case we centred our focus on the molecular motifs identified by the clustering analysis of our choice (PCA-PAMM or tICA-MSM) applied to different system separately, and we asserted their (dis)similarity by application of the SOAP metric, that is used to compare the resulting motifs-wise averages in the SOAP space. We believe that this approach, although it yields very similar results to the previous approach (*i.e.*, the simulation-average centred one, presented in Chap. 4), it brings notable improvements to the comparison analysis. It gives direct access to an in-depth comparison of each of the structural motifs and it opens the possibility of comparing between different clustering algorithms. Moreover, we extended our computational investigation of soft-assembled material introducing the MSM approach, usefull to detect and characterise important aspects of the aggregate dynamics. This approach demonstrated that the "static" PCA-PAMM analysis might indeed overlook some details about the structural dynamics due to the PCA inefficiency in representing dynamical features of molecular environments (*e.g.*, red motif, $\omega(0)$, in fibre $\boldsymbol{\omega}$). However, for simple systems, as either fibre $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$, the overall picture does not change significantly. Finally, we believe that the MSM theory naturally lends itself to a proper in-depth investigation of the kinetic pathways that might lead to the formation of a specific defected state or even the study of the kinetic pathways that occur during the SA process, which are all very important aspects worth exploring in the future.

Chapter 6

Other related modeling works on complex supramolecular structures

6.1 Supramolecular assembly of charged nanoparticles into static and dynamic superstructures

The work presented in this section was performed in collaboration with the group of Professor Rafal Klajn (Department of Organic Chemistry, Weizmann Institute of Science, Rehovot, Israel), which took care of all the experimental synthesis and techniques to investigate the supramolecular co-assembly of charged gold nanoparticles. Our group took care of the computational studies. The joint work resulted in a publication in *Nature chemistry*, Ref. [8], with the title “Electrostatic co-assembly of nanoparticles with oppositely charged small molecules into static and dynamic superstructures”. Coulombic forces between charged nanoparticles can be exploited to drive their self-assembly into higher-order lattice structures. Such process usually requires oppositely charged partners that are similarly sized (*i.e.*, nanoparticles), with the function of mediating the self-assembly between the nanoparticles. The possibility to use structurally simple small molecules as mediators, would greatly facilitate the fabrication and control of these nano-sized structured materials, widening their possible applications in catalysis, sensing and photonics. In Ref. [8] we report how small charged molecules, with as few as three electric charges, can effectively induce attractive interactions between oppositely charged nanoparticles in water, obtaining high quality colloidal crystals. The findings of this study demonstrate an approach for the facile fabrication, manipulation and further investigation of static and dynamic nanostructured materials in aqueous environments. In the following we will quickly go over the main experimental findings and then cover the *in silico* results, with particular focus on the analyses that were performed using the methodologies presented in this thesis work.

6.1.1 Scientific context

Size plays an important role in the chemical and biological functions of a given system, spanning from single atoms and molecules, to nanoscopic and mesoscopic aggregates, to entire organisms. Nanoparticles (NPs) systems fall into the category of nanochemistry[164], a sub-discipline of chemical and material sciences, that deals with the development and study of nanoscale materials. A popular topic in nanochemistry is the examination of the intriguing and diverse analogies that exist between molecular-scale and nano-sized species. Evidences of super-paramagnetic and super-ferromagnetic NPs systems, behaving in many ways like atoms of paramagnetic and ferromagnetic metals, were reported in the literature.[165] Single NPs can also be assembled into one-dimensional (1D) structures,[166] in a process similar to the polymerization of small molecules, as well as more complex 3-D lattice structures.[167] Moreover, spherical NPs functionalised with charged chemical groups, typically COO^- and NMe_3^+, show a behavior very similar to simple (atomic) ions analogues, hence, being able to attract each other and precipitate from solutions as crystalline solids. However, these nanosized species fundamentally differ from simple atomic ions: the aggregation of NPs continues until the electroneutrality is reached, but further addition of the NP "superions" causes a dissolution of the precipitate and the reduction in size of the NP aggregates.

The ability of small molecular multi-charged ions in driving the aggregation of oppositely charged NPs was extensively experimentally investigated by our collaborators, the group of prof. R. Klajn. They synthesized different sizes of spherical gold NPs, functionalised with the TMA ligand (between 2.8 and 13.1 nm of core diameter). Irrespective of the core size, the NPs exhibited an excellent solubility in water in presence of their original monovalent counterions. The colloidal stability was tested by the addition through titration of various aqueous solution of doubly and triply charges ions in different concentrations. Only in the latter case the system was reported to aggregate, confirmed by Surface Plasmon Resonance (SPR) emission and

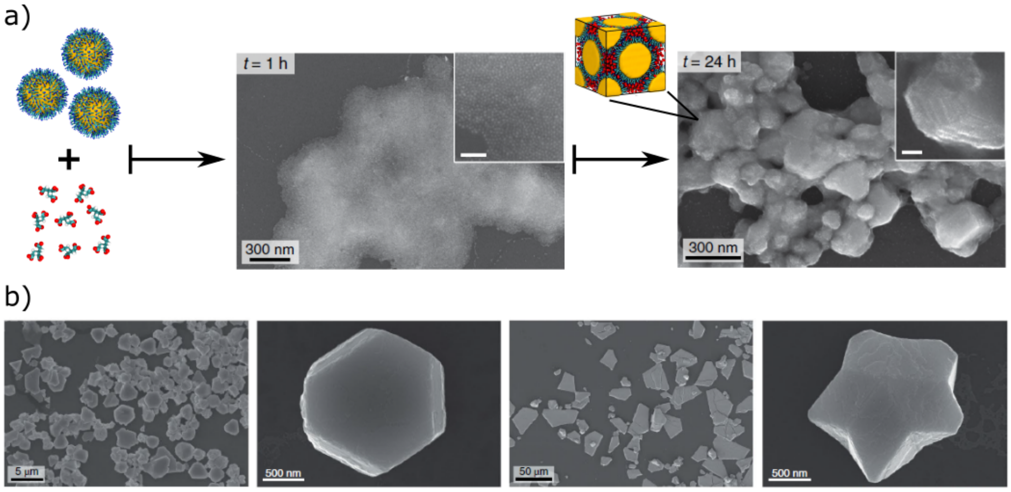


Fig. 6.1 Au·NP-TMA aggregates evolution towards superlattice structures. a) Schematic example and representative SEM images for the evolution of the amorphous NPs aggregate, driven by the citrates ions, to a colloidal crystal, at constant fixed temperature. b) Representative SEM images for colloidal crystal obtained using different multivalent anionic species (*e.g.*, ATP^{4-} and $P_3O_{10}^{5-}$) in different experimental conditions.

Transmission Electron Microscopy (TEM) imaging, which showed the formation of mostly amorphous NP aggregates (Fig. 6.1a). Moreover, it was hypothesised that the attractive electrostatic interactions, between positively charged NPs and oligoanions could be exploited to construct large colloidal crystals, and ultimately engineer nano-structured materials. To test them, our experimental collaborators did extensive trials employing a well defined synthesis strategy[168] to generate ordered colloidal assemblies. The obtained colloidal crystals exhibited a typical morphology of atomic crystals with a face-centred cubic (fcc) lattice, including hexagonal plates and truncated decahedra, supporting the initial suppositions (Fig. 6.1b).

Numerous other experiments to extensively probe the properties and the capabilities of this Au·NP-TMA system were conducted by our experimental collaborators: probing the interactions between negatively charged NPs and small cations, and out-of-equilibrium crystallisation behavior of such NP-TMA charges systems (*e.g.*, by titration of a charged initiator species). Although these represent interesting and valuable results we will not discuss them in details here, as they deviate from the main topic of this thesis work; a complete description is available in Ref. [8]. In the following, we will focus on the complete computational characterisation of the charge NPs, showing how I developed the molecular modes and the various analysis, which follows the guidelines presented in the present thesis work.

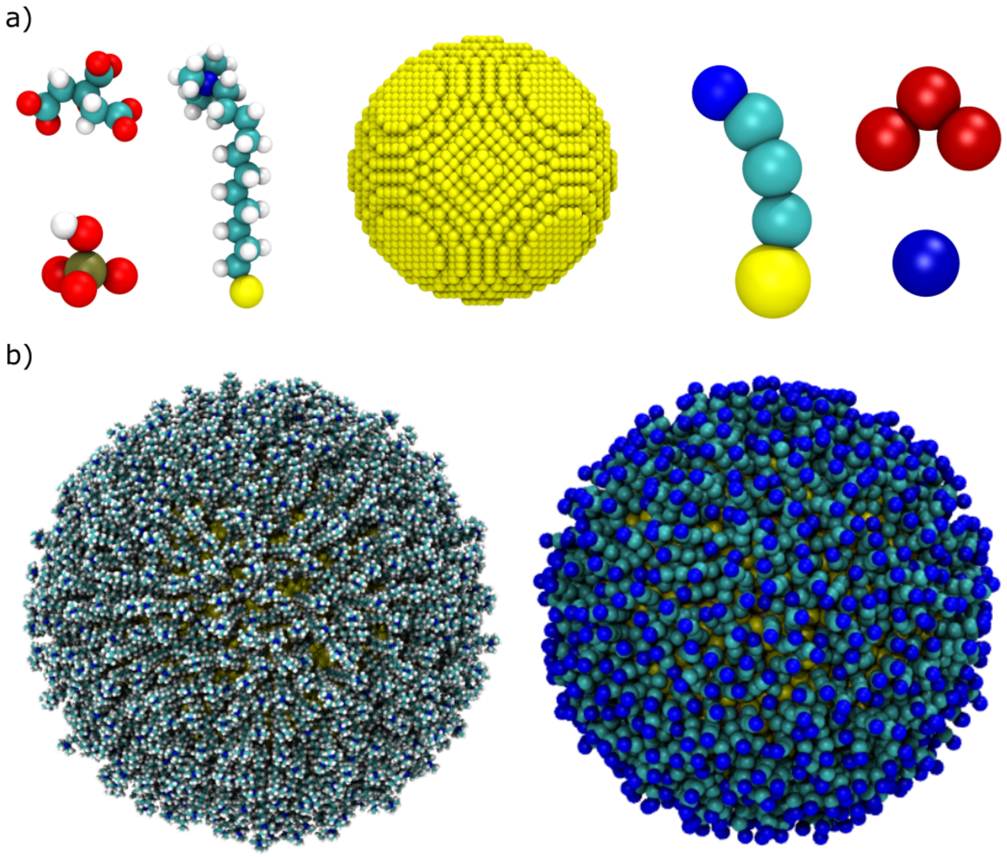


Fig. 6.2 AA and CG molecular models for the Au·NP systems. a) Left side, AA models for the citrate^{3-}, HPO_4^{2-} and TMA^{1+} ions. Right side, respective CG models for the citrate^{3-}, HPO_4^{2-} and TMA^{1+} ions. In the center is visualised the gold NP core, which remains the same at both level of accuracy. b) Representation of the fully coated (804 TMA ligands) AA NP (left) and CG NP (right).

6.1.2 Methods

We developed models and performed MD simulations at two different levels of molecular resolution: All-Atoms (AA) and explicit/implicit solvent Coarse-Grained (CG) (Fig. 6.2).

All-Atom (AA) simulations

The gold NP was parametrised following the recipe described in Ref. [169]. The diameter of the its metallic core in our simulations is ~ 7.4 nm, in agreement with a typical experimental size. The ligand ((11-mercaptoundecyl)-N,N,N-trimethylammonium, in short TMA) was parametrised using the general AMBER Force-Field (GAFF)[170]. Water molecules were treated explicitly in the AA models, using an explicit TIP3P water model.[171] All the other small ions, Cl^-, HPO_4^{2-}, citrate^{3-}, were also treated explicitly in the AA models and parametrised consistently.

First, a AA model of one NP hemisphere was built and coated it with 347 TMA ligands (Fig. 6.3) in order to cover that NP portion with the appropriate ligands density, in agreement with experimental findings. Each of the TMA moieties carries a 1+ electrostatic charge, hence the same number of mono-valent counterions (Cl^-) was added to the simulation box in order to guarantee the system charge neutrality. To study the effect of ions binding we employed WT-MetaD simulations adding an extra single Cl^-, HPO_4^{2-} or citrate^{3-} anions to the system, together with the corresponding number of Na^+ counterions to ensure charge neutrality. All the AA WT-MetaD simulations were conducted in NPT ensemble, under periodic boundary conditions, and with a timestep of 2 fs. The temperature was maintained at 300 K using the V-rescale thermostat[172] and the pressure was kept at $P = 1$ atm using the Parrinello-Rahman barostat[173] with anisotropic pressure scaling (x and y decorrelated from z, where z is the direction of ion binding/unbinding biased during the AA WT-MetaD simulations). Long-range electrostatics were treated by means of Particle Mesh Ewald.[174] We emplyed two different CVs to bias the system dynamics: the distance of the centre of mass (COM) of the Cl^-, HPO_4^{2-} or citrate^{3-} ions from (i) the charged heads of the TMA ligands (CV1), and (ii) the centre of the NP (CV2). To run the WT-MetaD simulation we set a gaussian height of 1.0 $kJmol^{-1}$, σ values of 0.025 nm and 0.1 nm (for CV1 and CV2, respectively), and a bias factor of 8. Finally, we reconstructed the 1-D FES for the three different ionic species, as a function of CV1 alone.

To study, at AA resolution, the efficiency of either HPO_4^{2-} or citrate^{3-} ions in driving the NPs aggregation in a competitive ionic environment (*i.e.*, competition against the Cl^-) we built a system composed of two NP hemisphere replicas facing each along the Z axis of the simulation box (Fig. 6.4a). Between the two half NPs we inserted 10 citrate^{3-} in one case and 15 HPO_4^{2-} in another, always neutralized by the equivalent positive counterions. We performed AA-MetaD simulations biasing the binding/un-binding using the distance between the half-NP centres as the CV (Fig. 6.4b) To run these MetaD simulations we employed a gaussian height of 0.25 $kJmol^{-1}$, σ values of 0.1 and a bias factor of 15. During the runs, we imposed a

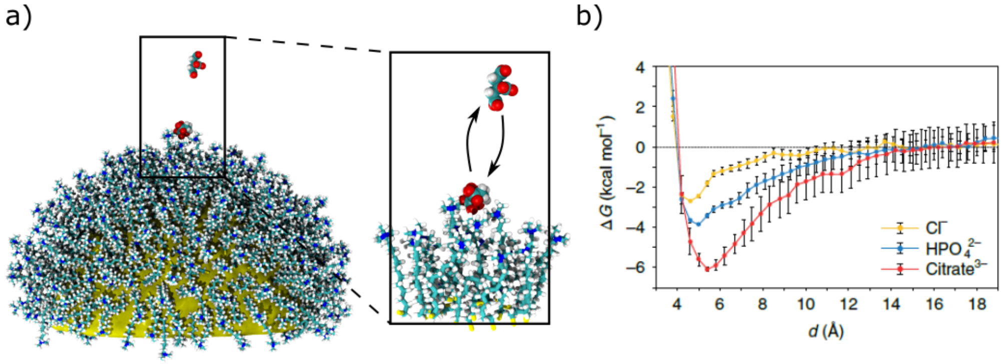


Fig. 6.3 Single ion free-energy landscapes of interaction with the NP surface. a) MetaD simulations setup, as described in the methods section. The ion is biased to bind and unbind to the charged surface of the NP. b) FES for the three anionic species, Cl^-, HPO_4^{2-} and citrate^{3-}.

cylindrical restraint on the HPO_4^{2-} / citrate^{3-} ions in such a way that they would remain localised between the two NP interface.

Coarse-grained (CG) simulations

To simulate multiple NPs systems we developed CG molecular models for all the components. We tuned two CG model variants: one based on the standard MARTINI FF[85] description, with explicit solvent, and the other with an implicit solvent description, based on the "dry"-MARTINI FF[175]. After verifying that the two models behave qualitatively the same, we focused on the latter one, to ensure more computational efficiency due to the absence of the solvent. To simplify the simulation even further, we did not include any monovalent ionic species and run simulation only with enough HPO_4^{2-} and citrate^{3-} to neutralise the charged NPs. The CG model of the NP was decorated with a total of 804 positively charged ligands, as it was in the AA case. The HPO_4^{2-} and citrate^{3-} ions were modeled as a single 2- pseudo-atom and a triangular shaped collection of 3 pseudo-atoms respectively, each carrying 1- charge (Fig. 6.3 right side). We then built two simulation boxes, each containing two NPs and the corresponding HPO_4^{2-} / citrate^{3-} to ensure the overall charge neutrality. Both systems were then simulated by means of classical CG molecular dynamics (CG-MD) simulations. We used the leap-frog stochastic dynamics integrator with a timestep of 20 fs, and simulated the binary aggregation process of the two systems.

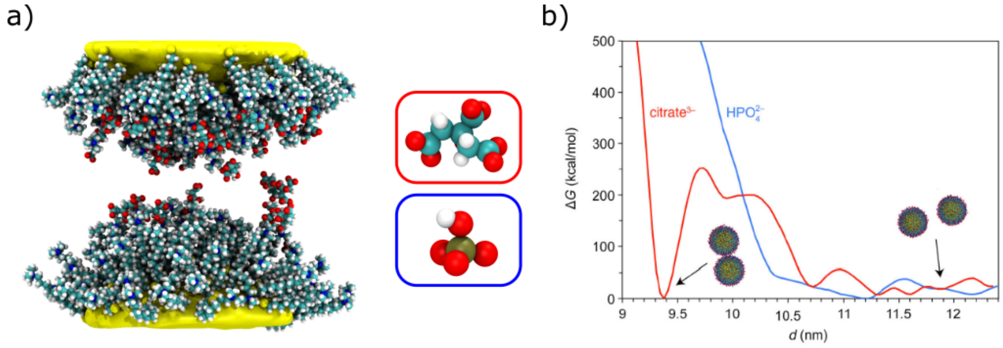


Fig. 6.4 AA metaD experiments for the binding of two Au·NP-TMA in presence of HPO_4^{2-} or citrate^{3-}. a) Definition of the system starting position: two hemispheres of the Au·NP-TMA. b) Qualitative estimation of the energy landscape for the binding/unbinding biased dynamics.

Finally, we extended our simulations up to 4 NPs and run CG-MD simulations to study the aggregation process. However, these simulations appeared to be very instable, and only the citrate^{3-} ions case was run for few hundreds nano-seconds of implicit solvent CG simulation time.

Supervised and unsupervised analysis of ion environments

In order to study the dynamics of citrate^{3-} ions on the surface and interface of a binary NP system we analysed the trajectory of a CG-MD simulation of two already aggregated NPs. Starting from $t_0 = 900$ns (*i.e.*, shortly after the two NPs were permanently attached), up to $t = 8.4\mu$s, every 25 ns, we analysed the configuration of the ions by means of the SOAP[105] descriptor and PAMM[120] clustering algorithm. Using the python library *DScribe*[128] we computed a SOAP feature vector for all the 536 CG citrate ions in the system, setting the SOAP input parameters: *rcut*= 35 Å, *nmax*= 8, *lmax*= 8, and leaving the other as default. The data dimensionality was then reduced by means of PCA, keeping only the first 5 principal components, accounting for more then 95% of the total data variance. The PC analysis was carried out using the python library *Scikit-Learn*[123]. Finally, the molecular motifs were extracted from the low-dimensional dataset by means of the PAMM clustering algorithm, using as parameters: fspread = 0.15, quick-shift = 1, bootstrap-runs = 73, merger-threshold = 0.005, and grid-size = 1000 points.

To support the results obtained through the unsupervised clustering, we carried out an additional analysis by specifying two properly defined CVs, in order to

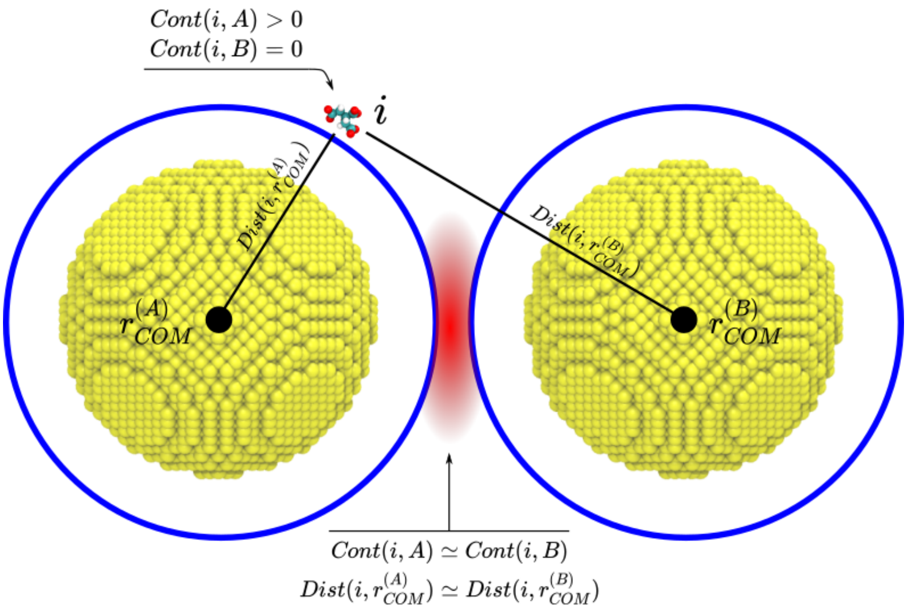


Fig. 6.5 Visual representation of the two CVs used for the supervised citrate analysis. The blue circumference represents the extent of the TMA coating around the NP, and the red shaded area is representative of the interface area, *i.e.* where a citrate would be bound to both NPs.

capture the effective dynamics of the citrate^{3-} ions around the two NPs (Fig. 6.5). Labeling the two NPs with A and B the first CV is defined as the absolute value of the difference between the number of contacts of a citrate ion with respect to the charged ligands heads of the two NPs inside a defined cutoff. For the i-th citrate ion the expression reads:

$$CONT(i) = |Cont(i,A) - Cont(i,B)|, \tag{6.1}$$

where in this case A and B represent the set of all charged head positions for the two NPs. For each evaluation of the function $Cont(\cdot,\cdot)$ we chose a cutoff radius of $R_0 = 10$Å. The value of said function goes continuously from 1 to 0 for each counted lingad heads, following a smooth rational switching function, that ensure to take into account of the fluctuations of the citrate ion on the charged NP surface. If the citrate is localised only on one of the NPs, one of the two $Cont(i,\cdot)$ contribution will be zero and the other will be a positive value, thus giving a positive $CONT$ value. Whereas, if the citrate is located at the interface (red zone of Fig. 6.5), the two $Cont(i,\cdot)$ contributions will be comparable, thus $CONT$ will be close to zero.

The second CV is defined as the absolute value of the distance difference between a citrate centre of mass and the centres of the two NPs gold cores. For the *i*-th citrate ion, the expression reads:

$$DIST(i) = |Dist(i, r_{COM}^{(A)}) - Dist(i, r_{COM}^{(B)})|, \tag{6.2}$$

where $r_{COM}^{(\cdot)}$ represent the position of the center of mass of the gold core of one NP. These values follow the same trend as for CV1, with the difference that when the citrate is relegated on one NP surface one $Dist(\cdot,\cdot)$ value corresponds to the contact distance (~ 5Å) and the other is larger.

Finally, to further characterise the dynamics of citrate ions in the two different spatial environments (*i.e.*, NP surface and interface), we computed the diffusivity of the anions. To this end, we defined two ensembles, based on the SOAP-PAMM clustering: one containing citrate ions that only diffuse around a NP surface (avoiding the interface region), and the other containing ions that only remain inside the interface region for the time spanned by our simulation. The diffusivities within the two citrate ensembles were computed by means of the mean square displacement (MSD), and then the diffusion constant was estimated by the angular coefficient of the linear fit of the MSD profiles.

6.1.3 Results

Interactions between positively charged NPs and small anions

To better understand the experimental findings provided by our experimental collaborators, we performed molecular simulations of TMA-coated NPs, interacting with HPO_4^{2-} and citrate^{3-} as representative di- and tri-valent anions. As a first step, we probed the interaction of the positively charged NP with each individual anions, namely Cl^- (as the "native" mono-valent ion), HPO_4^{2-} and citrate^{3-} (Fig. 6.3a). We performed multiple WT-metaD simulations biasing the bound and un-bound states of the different ions, allowing us to compute the free-energy associated to the binding/un-binding. All three free-energy profiles show a minimum at around 0.5nm from the positively charged heads, indicative of the attractive electrostatic interaction, and as expected, the depth of each minimum is found to be proportional to the ionic multi-valence. Within the accuracy of the AA models and MetaD simulations the estimated values are: ~ 2.7 kcal mol^{-1} for Cl^{-}, ~ 3.9 kcal mol^{-1} for HPO_4^{2-} and ~ 6.1 kcal mol^{-1} for citrate^{3-} (Fig. 6.3b). Next, we studied the ability of HPO_4^{2-} and citrate^{3-} to mediate an attractive interaction between the two TMA-coated NPs. To this end, we carried out different AA WT-MetaD simulations of two NP halves facing each other and each decorated with 347 ligands (Fig. 6.3a). We prepared two different configurations placing between the two NP halves 10 citrate^{3-} in one, and 15 HPO_4^{2-} in the other, and we estimated the free-energy profiles of the NP-NP interaction as a function of the NP-NP separation (Fig. 6.3b). Despite the complexity of the system, we could observe that with the citrate^{3-} ions the free-energy profile exhibits a sharp and deep minimum at distances consisten with those of an assembled configuration ($\sim$ 9.4nm). Whereas, in the presence of HPO_4^{2-} ions the overall energetic minimum is found at higher distances, suggesting that the system does not aggregate. However, these results have a qualitative meaning, due to the restrictions and approximations imposed on the system to be simulation viable.

To model the anion-mediated aggregation of positively charged NPs on a larger scale we switched to a CG molecular model, due to the complexity of the system at hand. Taking advantage of the benefits brought by this lower resolution model, we performed CG-MD simulations in implicit solvent on two parallel systems, each containing two NPs, each decorated with 804 TMA CG-ligands, one neutralised with HPO_4^{2-} and the other with citrate^{3-} ions (804 and 536, respectively and all at CG level of description). Whereas both anions exhibited a strong affinity to the NPs, due to fundamental electrostatic interactions, it appears that only citrate ions have the ability to bring the particles together during the CG-MD runs (Fig. 6.6). Our qualitative MD results were in line with the evidences of the titration experiments, done by our collaborators, of aqueous suspensions of TMA-coated NPs with additional trianions species, which all yielded some sort of NPs aggregate.

All these findings are consistent with the Schulze–Hardy rule[176], which states that the ability of an electrolyte to mediate and drive the coagulation of colloidal suspensions is proportional to the ionic valence z of the charged species used. Specifically, the critical coagulation concentration (CCC) scales as $\propto z^{-6}$, suggesting a sharp increase in the ion potency to trigger the coagulation just after small changes of z. In our cases, compared to citrate^{3-}, HPO_4^{2-} would decrease the CCC by $\sim$ 11 times, requiring much higher concentrations of phosphate ions, but at those concentrations the whole experiment would be strongly hampered due to the extremely high ionic strength of the resulting solution.

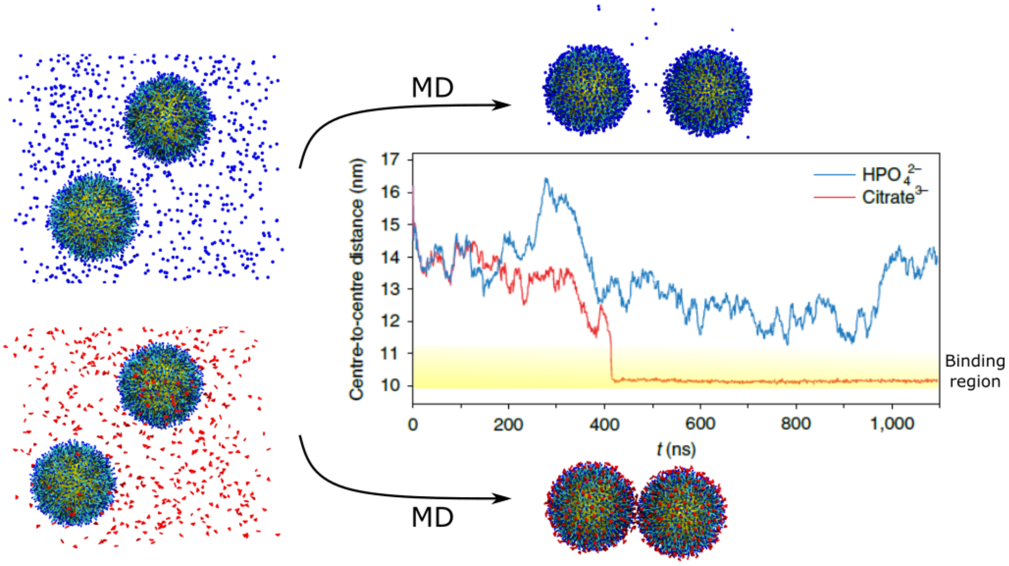


Fig. 6.6 MD behavior of two CG NP systems in the presence of HPO$_4^{2-}$ (top) and citrate^{3-} (bottom) ions. The plot reports the distance between the two gold NP cores against the simulation time.

Dynamics of small ions within NP–molecule ionic aggregates and their structural transformation into crystalline assemblies

Once it was proved that tri-valent citrate ions can readily mediate attractive interactions between two positively charged TMA-coated NPs, we investigated how these small ions behave on the surface and interface of these NPs. To this end, we extended the CG-MD simulation of the citrate NP dimer and analysed the ions local environments, through time, using the SOAP-PCA-PAMM approach. Thus, allowing us to follow local variations and fluctuations in the citrate states and also to reconstruct the dynamic mobility of the citrates in different regions of the system. The analysis classifies the citrates into three main molecular motifs (Fig. 6.7a,b): (i) citrates located at the interface between the two NPs, (ii) citrates that are at the boundaries of said interface, and (iii) citrates that interact only with a single NP. This classification pattern was also confirmed by a parallel analysis where we estimated a 2-D FES of the citrates, represented as a function of two CVs, specifically defined to follow their dynamics (as defined in the methods section of this study). The energy landscape showed 3 main local minima, which coincided qualitatively with the PAMM molecular motifs (Fig. 6.7c).

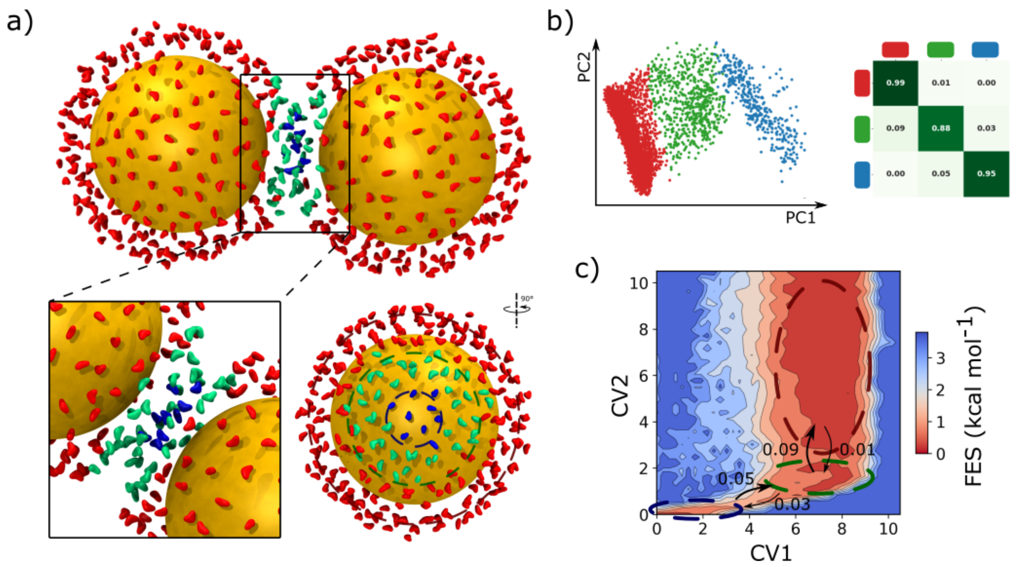


Fig. 6.7 Local analysis of the citrate^{3-} ions environments in a NP dimer. a) Representation of the molecular motifs (clusters) detected by the SOAP-PCA-PAMM analysis projected onto a sample snaptshot from the unbiased MD simulation of a NP dimer. The NP gold core is represented as a golden solid sphere and the ligands are not shown for clarity. The single citrate ions are colored according to their belonging to different PAMM clusters: NPs interface (blue), boundary of the interface (green), and interacting with a single NP (red). b) Plot of the first two principal components of the SOAP-PCA-PAMM analysis, colored following the cluster membership, and matrix of the probabilities of transitioning between the same clusters registered during the MD simulation. c) Free-energy landscape as a function of the two CVs, *CONT* (CV1) and *DIST* (CV2), defined in the text, with the corresponding probabilities of transition between the three states.

The transitions between the clusters during the CG-MD were collected and used to define a transition probability matrix, which highlights the dynamic exchanges of the citrate ions in each of the three different clusters (expressed as probabilities). It is worth pointing out that these transition probabilities are extracted from a simplified CG model, and thus have only a qualitative value. However, they can be compared to one another and are proportional to the rate constants of citrates in one state transitioning to another. In addition to that, we computed the MSD of citrates comparing their value at the interface (blue cluster) and in the “open” areas (red cluster), to get further insight into the internal dynamics of the ions. These results clearly demonstrate the mobility difference of the ions in different spatial positions: the citrates at the interface between two bound NPs are less dynamic than those freely bound to the surface of a single NP, although not completely static. This peculiar trait can be thought as a “ionic” glue that keeps the NPs together but still allowing for subtle motions and configurational rearrangements of these. Indeed, the appearance of small ordered crystalline domains was found in NP-TMA/citrate solutions over the course of 24 hours after titrations experiments of multivalent ionic species by means of TEM and SEM imaging by our collaborators (Fig. 6.1).

In summary, we investigated electrostatic interactions across different molecular resolutions, of a model system of positively charged NP and negatively charged small molecules. We found that small anions with three (or more) charges can mediate attractive interactions between oppositely charged NPs, hence driving their aggregation. MD simulations combined with the unsupervised SOAP-PCA-PAMM analysis revealed that the oligoanions that hold NPs together behave like a "dynamic ionic glue", and thus facilitate structural transformations of the NP aggregates. These results open up several avenues for further follow-up researches in the development of interesting colloidal structures that can be possibly used as catalysts or substrate for scientific and/or technological applications.

6.2 Nanoparticle superlattices with supramolecular semiconductive behavior

The work presented in this section contributed in a publication in ACS Nano, Ref. [177], with the title "Supramolecular Semiconductivity through Emerging Ionic Gates in Ion–Nanoparticle Superlattices". We here present the results of this article, by stressing the contributions obtained in the framework of the present thesis work, namely by employing the SOAP-PCA-PAMM analysis to study the structural configurations and dynamics of the system of interest.

In the previous section we learned how the self-assembly of charged NPs, driven by small oppositely charged molecules, may produce superlattice systems. However, to what extent the properties of such supramolecular crystals actually resemble those of atomic materials often remains unclear. In Ref. [177] we propose a computational study on the response induced by an external and constant electric field on a supramolecular FCC lattice constituted by Au·NP-TMA particles and mediated by citrate^{-3} ions. Using the previously validated implicit solvent CG molecular models, a simulation box containing 4 Au·NP-TMA was prepared, with the corresponding citrate^{-3} counterions, arranged in such a way as to produce, when replicated through

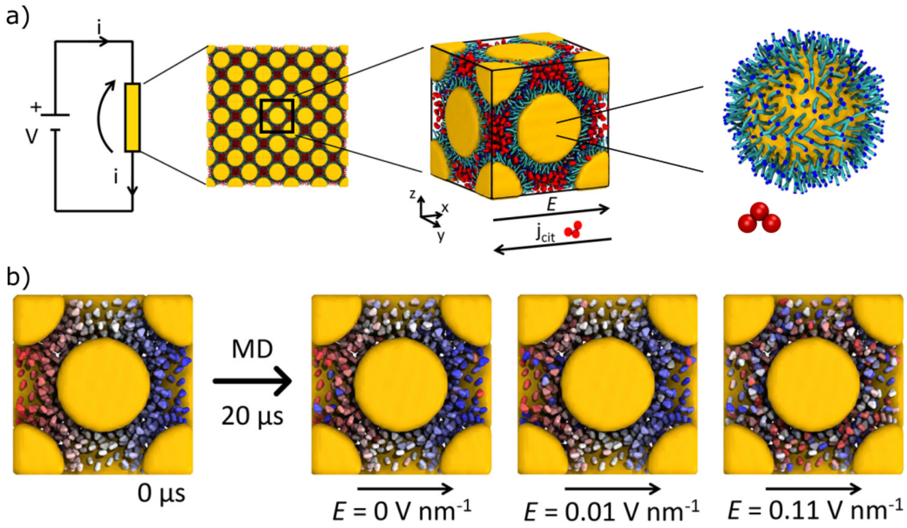
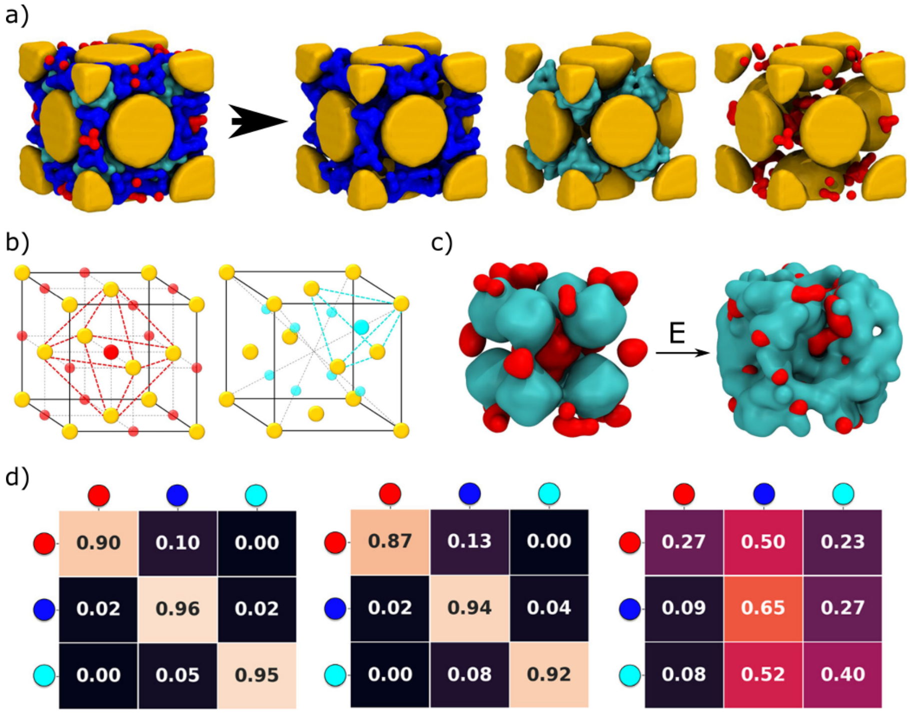


Fig. 6.8 Au·NP-TMA superlattice subjected to an external electrical field. a) Scheme and molecular details of the *in silico* resistivity experiment to study the ionic conductivity of the NP superlattice. b) Visual demonstration of the reshuffling and diffusion of the citrate ions along a MD simulation in different condition of applied electric field.

periodic boundary conditions (PBC), an infinite FCC super-lattice (Fig. 6.8a). Further details on the molecular models and on the simulation protocols are provided in the methods section of Refs. [8, 177].

As pseudo-atoms, the NPs in such supercrystals are surrounded by smaller particles, which keep them together as pseudo-electron equivalents. In response to a constant electric field, these electron equivalents might break their distribution symmetry around the atom equivalents and start moving across the superlattice, resembling the behavior of real electrons in metal lattices (Fig. 6.8b). It is possible, by monitoring the overall citrate ions dynamics at different E intensities, to compare the effect of the electrostatic stimulus on the internal dynamics of the citrate ions, referred as electron-analogues (EAs) in our experiments. Figure 6.8b reports visually these results, as the initial set of colours of the citrates appears mixed after 20 μs of MD simulation, with a intermixing intensity related to the strength of the applied electric field. While these data are extracted from an approximated CG model and as such should be considered as qualitative, they provide useful insights to rationalise the global response behavior of these superlattices.

Interesting questions arise on the microscopic dynamics of such ions, as well as on the local mechanisms controlling their conductivity. For example, if the anions are moving in a uniform way, and if not, which anions are more prone to diffuse



	🔴	🔵	🩵
🔴	0.90	0.10	0.00
🔵	0.02	0.96	0.02
🩵	0.00	0.05	0.95

	🔴	🔵	🩵
🔴	0.87	0.13	0.00
🔵	0.02	0.94	0.04
🩵	0.00	0.08	0.92

	🔴	🔵	🩵
🔴	0.27	0.50	0.23
🔵	0.09	0.65	0.27
🩵	0.08	0.52	0.40

Fig. 6.9 SOAP-PCA-PAMM analysis on the NPs superlattices. a) Detection of dynamic ionic environments on a sample MD snapshot with $E = 0$ Vnm^{-1} and $T = 300K$ The three clusters occupy the main cavities of the FCC lattice: NP direct interfaces (blue), octahedral (red), and tetrahedral (cyan) cavities. b) Reference geometric construction of the cavities in the FCC superlattice: the octahedral cavities are identified in red (red dots, cavity centers; dashed red lines, cavity sides), and the tetrahedral ones are in cyan. c) Effect of the applied electric field ($E = 0.11$ Vnm^{-1}) on the original distribution of red and cyan clusters, depicting the ions diffusion activated by the external field. d) Transition matrices indicating the probability of citrate transition between the different clusters. The matrices, from left to right, represent three different values of external electric field E (0, 0.03, 0.07 Vnm^{-1}).

and which to be static. To answer these questions we turned to our SOAP-PCA-PAMM unsupervised ML approach, applying the computational analysis to the citrate positions in the same way as it was done previously, in the case of the NPs dimer[8]. The PAMM clustering algorithm identifies three main molecular motifs of ion arrangements inside the supramolecular lattice. Figure 6.9a shows how the three detected SOAP environments correspond to well-defined positions within the FCC cell: citrate ions between direct NP–NP interfaces (blue cluster), citrate ions in octahedral cavities of the FCC lattice (red cluster), and those in tetrahedral cavities (cyan cluster). In particular, the analysis shows how, in the unperturbed superlattice ($E = 0$ Vnm^{-1}), the two ionic environments corresponding to the tetrahedral (cyan) and octahedral (red) cavities are not in direct contact with each other, being separated by the blue environment. The situation changes drastically with the application of a strong enough electric field ($E > 0.05$ Vnm^{-1}) as the two cluster tend to come together into a single, undefined, aggregate, activating the “conduction” of ions (Fig. 6.9c). This is further clarified if we consider the transitions between the different cluster labels as the external field is increased (Fig. 6.9d, from left to right $E = 0, 0.03, 0.07$ Vnm^{-1}). The direct transition exchange of CITs between tetrahedral and octahedral cavities, which is prevented in absence of an electric field, is enabled when a certain field threshold is overcome. The appearance of these transitions opens conductive-like gates, allowing the establishment of the CIT current and the conductive superlattice regime.

In summary, an *ad-hoc* computational analysis have been used to investigate the supramolecular conductive behavior of colloidal crystal lattices co-assembled from Au·NP-TMA and citrate ions. These builds on a relatively simple, and computationally efficient, CG molecular model,[8, 177] which allows for facile access to the study of the supramolecular lattice in its complexity, while retaining its essential physical features, thereby providing us with fundamental insights on the behavior of such materials. The obtained data demonstrate how these Au·NP-TMA crystals possess a supramolecular semiconductive character, where the conduction of ions happens only above a certain (electric field) intensity threshold, hence emulating the presence of a “band-gap”, which the electron-alternatives have to overcame.

In this work, I supervised the application the novel analysis methods discussed in the present thesis. The ions arrangements were completely characterised by means of our SOAP-PCA-PAMM unsupervised approach. This allowed us to unveil the microscopic origin of semi-conductive behavior represented as dynamics transitions between trapped ionic states between the tetrahedral and octahedral cavity domains of the superlattice (Fig. 6.9). Exhaustive details are available in Ref. [177] and related supporting information. Finally, this application proves once again the general applicability of our data-driven computational analysis which can provide valuable structural insight on virtually any type of molecular system.

Chapter 7

Conclusions

In this thesis, the equilibrium structural properties of a wide variety of self-assembling supramolecular materials have been investigated using computer simulations in together with state of the art machine learning approaches. In these kind of materials the lack of a persistent long-range order on one side grants unique structural properties (*e.g.*, response to external stimuli, self-healing), but on the other side, it greatly complicates the characterisation of such structures, hence hindering the ability of understanding, targeting or even predicting a specific property when engineering synthetic materials. In the context of the rational design of new functional soft materials, the contents of this thesis have been primarily motivated by the difficulty of finding a general and agnostic way to represent, classify and compare the elusive molecular motifs that emerge in the complex structural behavior of supramolecular materials in equilibrium conditions. Specifically, taking as model the data-driven analysis introduced in Ref. [30], we extended and generalised it, demonstrating its potential in processing multiple datasets coming from very different molecular systems.

First, the problem of finding a universal and transferable way to represent a given supramolecular structure was addressed. The rise and spread of ML-based approaches, across all scientific fields, was followed by the necessity of finding ways to encode atomic/molecular information into abstract high-dimensional feature vectors, which are better handled by ML-based algorithms. In the last two decades a plethora of so-called "descriptors"[78] were introduced as a way to represent atomic coordinates in a compact and agnostic way. Depending on the particular flavor of each representation, a descriptor is better suited for reproducing either structural (*e.g.*, coordination, geometry) and/or connectivity (*e.g.*, bond orders, functional groups) or other information. In this thesis, we adopted the *Smooth Overlap of Atomic Positions* (SOAP)[105], a density based descriptor, that encode the density of neighbouring particles, referred to a chosen central one, into an expansion of Spherical Harmonics and Gaussian functions; the resulting SOAP feature vectors proved to be a useful and powerful general structural representation tool.[30, 71, 114] The choice of the number and positions of the centres to be incorporated in the SOAP representation is a critical step in our analysis. In this regard, we demonstrated how a very simple reduction of the molecular structure of the system under investigation is often enough to capture the essential structural information of monomers hierarchy inside the supramolecular architecture. This was achieved by expressing each monomer, of a given material, as a single point, located in some *hot-spot* of the monomer structure, *e.g.*, the centre of geometry or the centre of a specific interaction defining the assembly identity (*e.g.*, h-bond, solvophobic or solvophilic interactions). We considered examples of supramolecular materials coming from 1-D systems (*e.g.*, polymers, fibres), 2-D systems (*e.g.*, bilayers, micelles) and 3-D system (*e.g.*, droplets, nanoparticles), all having monomers that assemble in different criteria, but through the application of the SOAP description, to their respective reduced structure, we were able to efficiently encode different materials into feature vectors in a complete comparable. Finally, since the SOAP representation can be applied to molecular data coming from MD trajectories, this allows the building of datasets containing the overall structural behavior of any given system at the simulation conditions.

The process of embedding atomic data into high-dimensional vector spaces constitutes the foundation of many ML-based approaches[75, 78, 118], often followed by either dimensionality reductions, clustering and/or neural networks processing, which altogether they define the complete workflow of a specific analysis. In this sense we developed a modular analysis consisting of a series of steps with the specific aim of extracting, classifying and comparing structural information coming from the MD of a generic molecular system. The complete workflow is summarised in Section 2.4. The first three steps deal always with obtaining the optimal representation of a molecular system, using the concepts just discussed (reduced structure and descriptor), and it yields the SOAP dataset. The data is then processed in different ways to extract and translate to human-readable variables the high-dimensional

SOAP features according to whether we want to classify the system relevant motifs or compare it to other ones. For example, a combination of dimensionality reduction and clustering algorithm can be used to collect and visualise information on the structural motifs captured by the descriptor. We made use of the general purpose PCA[110, 111] dimensionality reduction and the PAMM[120] clustering algorithm, the latter being a recently developed method to analyse long molecular trajectories data with a contained computational cost. Moreover, the identification of the cluster memberships allows for the qualitative computation of the interconversion of such labels along the trajectory, uncovering the dynamic relationships between the structural motifs of a given material.

In the second part of the thesis, the ways to compare the global structural properties of an ensemble of materials, across different families and dimensionalities, were investigated. Our studies lead to the definition of the so-called "defectometer" (Chap. 4). This particular analysis tool is based on a global definition of the SOAP descriptor, the SOAP *frame-* and *simulation*-averages (Sec. 2.2), where the local (monomer-wise) fingerprints are averaged first along the number of monomers inside an aggregate (frame-average) and then along all the MD trajectory frames (simulation-average). The double average yields a single and comprehensive fingerprint for each system, which can then be compared to each other exploiting a descriptor-based metric. The SOAP metric[178] follows directly from the mathematical definition of the descriptor and gives a measure of how close two feature vectors are in their original high-dimensional space. Through this special metric, we were able to assert how the presence of defects affect the behavior of a material, altering its nature and granting properties that can effectively make it very different from other example of the same material family. It is the case of BTA$_W$, a water soluble BTA-based supramolecular polymer (1-D), that thanks to its complex equilibrium of structural defect it is raked closer to a 2-D aggregate than other 1-D fibres. From the literature,[26, 50] it is known, that supramolecular polymer based on the BTA$_W$ architecture posses highly dynamic properties, like stimuli-responsive and self-healing, that are triggered by the peculiar dynamics of the defects present in these fibres (*i.e.*, the "surfing" motion of the least coordinated monomers along the fibre backbone). We believe that the ability to recognise and understand the effects that a certain type of structural defects induce in a material is the first step towards the development of more efficient ways to design a specific material possessing task-tailored dynamics and dynamical properties.

In summary, the ML-driven reconstruction of the internal dynamics of a supramolecular assembly (as discussed in Chap. 4) is essentially based on three phases: (i) extraction and representation of structural features (*e.g.*, SOAP), (ii) dimensionality reduction and identification of the main structural motifs, which populate the assembly, and finally, (iii) reconstruction of the dynamic interconnections between them. In Chapter 5 we tested an alternative approach based on the concept of building a Markov State Model, where (ii) and (iii) are effectively merged together, and the dominant structural motifs (of an assembly) are determined directly from the transitions within them. We showed how the PCA embedding might overlook some details in reproducing the complete dynamic behavior of the structural features under investigation, as no time-ordered information is required to build the low dimensional space. We decided to adopt a different approach explicitly including the knowledge of the time-series of structural events into the data processing of our analysis. The SOAP dataset is processed using the tICA[115, 116] dimensionality reduction and then a Markov State Model (MSM)[117, 126, 156] is estimated to fully characterise the kinetic of the system. The tICA stands for "time-structure independent components analysis", a dimesionality reduction algorithm that finds the low-dimensional space that better reproduce the time autocorrelation of the dataset to which is applied. In contrast to PCA, which maximises the "spatial" variance of a dataset, the low-dimensional space spanned by tCIA is dominated by the slowest degree of freedom that characterises the dataset under investigation, making the algorithm perfect as a baseline processing for studying reaction paths or structural changes. This last part is formalised by the construction of a Markov State Model[117, 118] (Chap. 5) from which it is possible to extract a wide variety of important dynamic and kinetic information.

In conclusion, the work presented in this thesis highlights how exploiting ML-based analysis it is possible to get structural/dynamical information out of virtually any kind of material, granted the availability of reliable molecular models and trajectories. Moreover, the analysis lends itself perfectly to comparison and ranking of supramolecular structures, based on the molecular motifs recorded. We believe that the workflow employed through the studyes is powerful for multiple reasons. It does not require any prior knowledge on the structure of the different assemblies that are compared. It is centred on the concept of "defectivity" (*i.e.*, the local levels of order/disorder), proposing the structural defects as a common ground feature to compare different materials, which is an angle that holds great potential toward the unification of supramolecular structures. Finally, such data-driven “defectometer” allows to quantitatively classify dynamic assemblies that are different from each other (e.g., fibers vs. micelles vs. layers vs. nanoparticles). This provides us with a precious tool toward the rational design of self-assembled materials with controllable dynamic properties, which is key to conceive complex systems where multiple assembled entities can effectively communicate with each other in a dynamic way.

References

(1) I. W. Hamley, *Angewandte Chemie International Edition*, 2003, **42**, 1692–1712.

(2) *Soft Matter Physics: An Introduction*, ed. M. Kleman and O. D. Lavrentovich, Springer New York, 2004.

(3) *Intermolecular and Surface Forces*, Elsevier, 2011.

(4) G. M. Whitesides and B. Grzybowski, *Science*, 2002, **295**, 2418–2421.

(5) S. C. Glotzer and M. J. Solomon, *Nature Materials*, 2007, **6**, 557–562.

(6) P. F. Damasceno, M. Engel and S. C. Glotzer, *Science*, 2012, **337**, 453–457.

(7) S. Yang, Y. Yan, J. Huang, A. V. Petukhov, L. M. J. Kroon-Batenburg, M. Drechsler, C. Zhou, M. Tu, S. Granick and L. Jiang, *Nature Communications*, 2017, **8**.

(8) T. Bian, A. Gardin, J. Gemen, L. Houben, C. Perego, B. Lee, N. Elad, Z. Chu, G. M. Pavan and R. Klajn, *Nature Chemistry*, 2021, **13**, 940–949.

(9) M. Dijkstra and E. Luijten, *Nature Materials*, 2021, **20**, 762–773.

(10) J.-F. Lutz, J.-M. Lehn, E. W. Meijer and K. Matyjaszewski, *Nature Reviews Materials*, 2016, **1**.

(11) N. C. Seeman and H. F. Sleiman, *Nature Reviews Materials*, 2017, **3**.

(12) A. Levin, T. A. Hakala, L. Schnaider, G. J. L. Bernardes, E. Gazit and T. P. J. Knowles, *Nature Reviews Chemistry*, 2020, **4**, 615–634.

(13) C. Tanford, *The Journal of Physical Chemistry*, 1974, **78**, 2469–2479.

(14) J. N. Israelachvili, D. J. Mitchell and B. W. Ninham, *Journal of the Chemical Society, Faraday Transactions 2*, 1976, **72**, 1525.

(15) T. L. Hill, *An introduction to statistical thermodynamics*, Dover Publications, Mineola, NY, 1987.

(16) D. Frenkel, *Physica A: Statistical Mechanics and its Applications*, 1999, **263**, 26–38.

(17) D. Frenkel, *Nature Materials*, 2014, **14**, 9–12.

(18) B. Rocha, S. Paul and H. Vashisth, *Entropy*, 2020, **22**, 877.

(19) L. Onsager, *Annals of the New York Academy of Sciences*, 1949, **51**, 627–659.

(20) G. J. Vroege and H. N. W. Lekkerkerker, *Reports on Progress in Physics*, 1992, **55**, 1241–1309.

(21) A. Haji-Akbari, M. Engel, A. S. Keys, X. Zheng, R. G. Petschek, P. Palffy-Muhoray and S. C. Glotzer, *Nature*, 2009, **462**, 773–777.

(22) M. A. Boles, M. Engel and D. V. Talapin, *Chemical Reviews*, 2016, **116**, 11220–11289.

(23) T. P. Silverstein, *Journal of Chemical Education*, 1998, **75**, 116.

(24) P. M. Chaikin and T. C. Lubensky, *Principles of condensed matter physics*, Cambridge University Press, Cambridge, England, 2000.

(25) E. Cubuk, S. Schoenholz, J. Rieser, B. Malone, J. Rottler, D. Durian, E. Kaxiras and A. Liu, *Physical Review Letters*, 2015, **114**.

(26) D. Bochicchio, M. Salvalaglio and G. M. Pavan, *Nature Communications*, 2017, **8**.

(27) S. H. Jung, D. Bochicchio, G. M. Pavan, M. Takeuchi and K. Sugiyasu, *Journal of the American Chemical Society*, 2018, **140**, 10570–10577.

(28) A. Torchi, D. Bochicchio and G. M. Pavan, *The Journal of Physical Chemistry B*, 2018, **122**, 4169–4178.

(29) D. Bochicchio, S. Kwangmettatam, T. Kudernac and G. M. Pavan, *ACS Nano*, 2019, **13**, 4322–4334.

(30) P. Gasparotto, D. Bochicchio, M. Ceriotti and G. M. Pavan, *The Journal of Physical Chemistry B*, 2020, **124**, 589–599.

(31) A. L. de Marco, D. Bochicchio, A. Gardin, G. Doni and G. M. Pavan, *ACS Nano*, 2021, **15**, 14229–14241.

(32) M. F. Hagan and G. M. Grason, *Reviews of Modern Physics*, 2021, **93**.

(33) C. A. Kerfeld, S. Heinhorst and G. C. Cannon, *Annual Review of Microbiology*, 2010, **64**, 391–408.

(34) M. G. Mateu, *Archives of Biochemistry and Biophysics*, 2013, **531**, 65–79.

(35) J. D. Perlmutter and M. F. Hagan, *Annual Review of Physical Chemistry*, 2015, **66**, 217–239.

(36) R. O. Prum, E. R. Dufresne, T. Quinn and K. Waters, *Journal of The Royal Society Interface*, 2009, **6**.

(37) R. C. McPhedran and A. R. Parker, *Physics Today*, 2015, **68**, 32–37.

(38) P. Fratzl, *Current Opinion in Colloid & Interface Science*, 2003, **8**, 32–39.

(39) D. Popp and R. C. Robinson, *Cytoskeleton*, 2012, **69**, 71–87.

(40) V. Foderà, A. Zaccone, M. Lattuada and A. M. Donald, *Physical Review Letters*, 2013, **111**.

(41) T. D. Nguyen, B. A. Schultz, N. A. Kotov and S. C. Glotzer, *Proceedings of the National Academy of Sciences*, 2015, **112**.

(42) M. M. van Schooneveld, V. W. A. de Villeneuve, R. P. A. Dullens, D. G. A. L. Aarts, M. E. Leunissen and W. K. Kegel, *The Journal of Physical Chemistry B*, 2009, **113**, 4560–4564.

(43) S. Cameron, L. Kreplak and A. D. Rutenberg, *Soft Matter*, 2018, **14**, 4772–4783.

(44) L. E. Hough, H. T. Jung, D. Krüerke, M. S. Heberling, M. Nakata, C. D. Jones, D. Chen, D. R. Link, J. Zasadzinski, G. Heppke, J. P. Rabe, W. Stocker, E. Körblova, D. M. Walba, M. A. Glaser and N. A. Clark, *Science*, 2009, **325**, 456–460.

(45) R. L. B. Selinger, J. V. Selinger, A. P. Malanoski and J. M. Schnur, *Physical Review Letters*, 2004, **93**.

(46) R. Ghafouri and R. Bruinsma, *Physical Review Letters*, 2005, **94**.

(47) R. Sakhardande, S. Stanojeviea, A. Baskaran, A. Baskaran, M. F. Hagan and B. Chakraborty, *Physical Review E*, 2017, **96**.

(48) L. Brunsveld, B. J. B. Folmer, E. W. Meijer and R. P. Sijbesma, *Chemical Reviews*, 2001, **101**, 4071–4098.

(49) T. F. A. de Greef and E. W. Meijer, *Nature*, 2008, **453**, 171–173.

(50) D. Bochicchio and G. M. Pavan, *ACS Nano*, 2017, **11**, 1000–1011.

(51) T. Aida, E. Meijer and S. Stupp, *Science*, 2012, **335**, 813–817.

(52) S. H. M. Söntjens, R. P. Sijbesma, M. H. P. van Genderen and E. W. Meijer, *Journal of the American Chemical Society*, 2000, **122**, 7487–7493.

(53) A. Sarkar, R. Sasmal, C. Empereur-mot, D. Bochicchio, S. V. K. Kompella, K. Sharma, S. Dhiman, B. Sundaram, S. S. Agasti, G. M. Pavan and S. J. George, *Journal of the American Chemical Society*, 2020, **142**, 7606–7617.

(54) R. Dominguez and K. C. Holmes, *Annual Review of Biophysics*, 2011, **40**, 169–186.

(55) P. Binarová and J. Tuszynski, *Cells*, 2019, **8**, 1294.

(56) B. K. Mandal, *Polymer synthesis*, Covalent Press, 2010.

(57) R. D. Mukhopadhyay and A. Ajayaghosh, *Science*, 2015, **349**, 241–242.

(58) J. Matern, Y. Dorca, L. Sánchez and G. Fernandez, *Angewandte Chemie International Edition*, 2019, **58**, 16730–16740.

(59) S. Ogi, K. Sugiyasu, S. Manna, S. Samitsu and M. Takeuchi, *Nature Chemistry*, 2014, **6**, 188–195.

(60) O. Shyshov, S. V. Haridas, L. Pesce, H. Qi, A. Gardin, D. Bochicchio, U. Kaiser, G. M. Pavan and M. von Delius, *Nature Communications*, 2021, **12**.

(61) G. Ghosh, P. Dey and S. Ghosh, *Chemical Communications*, 2020, **56**, 6757–6769.

(62) A. Aliprandi, M. Mauro and L. De Cola, *Nature Chemistry*, 2016, **8**, 10.

(63) V. Mitchell and D. Jones, *Polymer Chemistry*, 2018, **9**, 795–814.

(64) M. Stefik, S. Guldin, S. Vignolini, U. Wiesner and U. Steiner, *Chemical Society Reviews*, 2015, **44**, 5076–5091.

(65) H. Cabral, K. Miyata, K. Osada and K. Kataoka, *Chemical reviews*, 2018, **118**, 6844–6892.

(66) R. Nagarajan, *Langmuir*, 2001, **18**, 31–38.

(67) C. Lionello, A. Gardin, A. Cardellini, D. Bochicchio, M. Shivrayan, A. Fernandez, S. Thayumanavan and G. M. Pavan, *ACS Nano*, 2021, **15**, 16149–16161.

(68) S. Tang, Z. Davoudi, G. Wang, Z. Xu, T. Rehman, A. Prominski, B. Tian, K. M. Bratlie, H. Peng and Q. Wang, *Chemical Society Reviews*, 2021, **50**, 12679–12701.

(69) N. M. Casellas, S. Pujals, D. Bochicchio, G. M. Pavan, T. Torres, L. Albertazzi and M. García-Iglesias, *Chem. Commun.*, 2018, **54**, 4112–4115.

(70) A. Sarkar, R. Sasmal, C. Empereur-mot, D. Bochicchio, S. V. K. Kompella, K. Sharma, S. Dhiman, B. Sundaram, S. S. Agasti, G. M. Pavan and S. J. George, *Journal of the American Chemical Society*, 2020, **142**, 7606–7617.

(71) A. Gardin, C. Perego, G. Doni and G. M. Pavan, *Communications Chemistry*, 2022, **5**.

(72) S. R. Nagel, *Reviews of Modern Physics*, 2017, **89**.

(73) D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, Academic Press, Inc., USA, 1st, 1996.

(74) M. Ceriotti, C. Clementi and O. A. von Lilienfeld, *Chemical Reviews*, 2021, **121**, 9719–9721.

(75) F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi and M. Ceriotti, *Chemical Reviews*, 2021, **121**, 9759–9815.

(76) A. Glielmo, B. E. Husic, A. Rodriguez, C. Clementi, F. Noé and A. Laio, *Chemical Reviews*, 2021, **121**, 9722–9758.

(77) J. Behler, *Chemical Reviews*, 2021, **121**, 10037–10072.

(78) V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti and G. Csányi, *Chemical Reviews*, 2021, **121**, 10073–10141.

(79) J. E. Jones, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 1924, **106**, 441–462.

(80) J. E. Jones, *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 1924, **106**, 463–477.

(81) L. Verlet, *Physical Review*, 1967, **159**, 98–103.

(82) M. Tuckerman, *Statistical mechanics: Theory and molecular simulation*, Oxford University Press, London, England, 2010.

(83) H. I. Ingólfsson, C. A. Lopez, J. J. Uusitalo, D. H. de Jong, S. M. Gopal, X. Periole and S. J. Marrink, *Wiley Interdisciplinary Reviews: Computational Molecular Science*, 2013, **4**, 225–248.

(84) S. Kmiecik, D. Gront, M. Kolinski, L. Wieteska, A. E. Dawid and A. Kolinski, *Chemical Reviews*, 2016, **116**, 7898–7936.

(85) S. J. Marrink, H. J. Risselada, S. Yefimov, D. P. Tieleman and A. H. de Vries, *The Journal of Physical Chemistry B*, 2007, **111**, 7812–7824.

(86) S. J. Marrink, L. Monticelli, M. N. Melo, R. Alessandri, D. P. Tieleman and P. C. T. Souza, *WIREs Computational Molecular Science*, 2022.

(87) J. Wang, S. Olsson, C. Wehmeyer, A. Pérez, N. E. Charron, G. de Fabritiis, F. Noé and C. Clementi, *ACS Central Science*, 2019, **5**, 755–767.

(88) W. G. Noid, J.-W. Chu, G. S. Ayton, V. Krishna, S. Izvekov, G. A. Voth, A. Das and H. C. Andersen, *The Journal of Chemical Physics*, 2008, **128**, 244114.

(89) F. Noé, S. Olsson, J. Köhler and H. Wu, *Science*, 2019, **365**.

(90) C. Empereur-Mot, L. Pesce, G. Doni, D. Bochicchio, R. Capelli, C. Perego and G. M. Pavan, *ACS Omega*, 2020, **5**, 32823–32843.

(91) J. Köhler, Y. Chen, A. Krämer, C. Clementi and F. Noé, *Journal of Chemical Theory and Computation*, 2023.

(92) A. Laio and M. Parrinello, *Proceedings of the National Academy of Sciences*, 2002, **99**, 12562–12566.

(93) G. Bussi and A. Laio, *Nature Reviews Physics*, 2020, **2**, 200–212.

(94) A. Barducci, M. Bonomi and M. Parrinello, *WIREs Computational Molecular Science*, 2011, **1**, 826–843.

(95) M. Invernizzi and M. Parrinello, *The Journal of Physical Chemistry Letters*, 2020, **11**, 2731–2736.

(96) G. A. Tribello, M. Ceriotti and M. Parrinello, *Proc. Natl. Acad. Sci. U. S. A.*, 2012, **109**, 5196–5201.

(97) M. M. Sultan and V. S. Pande, *The Journal of Chemical Physics*, 2018, **149**, 094106.

(98) G. Piccini and M. Parrinello, *The Journal of Physical Chemistry Letters*, 2019, **10**, 3727–3731.

(99) B. Sanchez-Lengeling and A. Aspuru-Guzik, *Science*, 2018, **361**, 360–365.

(100) P. J. Steinhardt, D. R. Nelson and M. Ronchetti, *Physical Review B*, 1983, **28**, 784–805.

(101) H. Eslami, P. Sedaghat and F. Müller-Plathe, *Physical Chemistry Chemical Physics*, 2018, **20**, 27059–27068.

(102) P. G. de Gennes, *The physics of liquid crystals*, Clarendon Press, Oxford, England, 2nd edn., 1994.

(103) A. P. Bartók, M. C. Payne, R. Kondor and G. Csányi, *Physical Review Letters*, 2010, **104**.

(104) D. P. Kovács, C. van der Oord, J. Kucera, A. E. A. Allen, D. J. Cole, C. Ortner and G. Csányi, *Journal of Chemical Theory and Computation*, 2021, **17**, 7696–7711.

(105) A. P. Bartók, R. Kondor and G. Csányi, *Physical Review B*, 2013, **87**.

(106) C. M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer-Verlag, Berlin, Heidelberg, 2006.

(107) L. Zhu, M. Amsler, T. Fuhrer, B. Schaefer, S. Faraji, S. Rostami, S. A. Ghasemi, A. Sadeghi, M. Grauzinyte, C. Wolverton and S. Goedecker, *The Journal of Chemical Physics*, 2016, **144**, 034203.

(108) A. Goscinski, G. Fraux, G. Imbalzano and M. Ceriotti, *Machine Learning: Science and Technology*, 2021, **2**, 025028.

(109) L. Van Der Maaten, E. Postma, J. Van den Herik et al., *J Mach Learn Res*, 2009, **10**, 13.

(110) K. Pearson, *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 1901, **2**, 559–572.

(111) H. Hotelling, *Journal of Educational Psychology*, 1933, **24**, 417–441.

(112) B. Ghojogh, F. Karray and M. Crowley, *Eigenvalue and Generalized Eigenvalue Problems: Tutorial*, 2019.

(113) S. Doerr, I. Ariz-Extreme, M. J. Harvey and G. De Fabritiis, *Dimensionality reduction methods for molecular simulations*, 2017.

(114) R. Capelli, A. Gardin, C. Empereur-mot, G. Doni and G. M. Pavan, *The Journal of Physical Chemistry B*, 2021, **125**, 7785–7796.

(115) L. Molgedey and H. G. Schuster, *Physical Review Letters*, 1994, **72**, 3634–3637.

(116) A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis*, John Wiley & Sons, Inc., 2001.

(117) F. Noé, I. Horenko, C. Schütte and J. C. Smith, *The Journal of Chemical Physics*, 2007, **126**, 155102.

(118) F. Noé and E. Rosta, *The Journal of Chemical Physics*, 2019, **151**, 190401.

(119) J. A. Hartigan, *Clustering Algorithms*, John Wiley & Sons, Inc., USA, 99th, 1975.

(120) P. Gasparotto, R. H. Meißner and M. Ceriotti, *Journal of Chemical Theory and Computation*, 2018, **14**, 486–498.

(121) S. Lloyd, *IEEE Transactions on Information Theory*, 1982, **28**, 129–137.

(122) D. Arthur and S. Vassilvitskii, SODA ’07, 2007.

(123) F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, *Journal of Machine Learning Research*, 2011, **12**, 2825–2830.

(124) A. Ng, M. Jordan and Y. Weiss, Advances in Neural Information Processing Systems, ed. T. Dietterich, S. Becker and Z. Ghahramani, MIT Press, 2001, vol. 14.

(125) U. von Luxburg, *Statistics and Computing*, 2007, **17**, 395–416.

(126) C. Schütte, A. Fischer, W. Huisinga and P. Deuflhard, *Journal of Computational Physics*, 1999, **151**, 146–168.

(127) S. Röblitz and M. Weber, *Advances in Data Analysis and Classification*, 2013, **7**, 147–179.

(128) L. Himanen, M. O. Jäger, E. V. Morooka, F. F. Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, *Computer Physics Communications*, 2020, **247**, 106949.

(129) N. S. Keddie, A. M. Z. Slawin, T. Lebl, D. Philp and D. O'Hagan, *Nature Chemistry*, 2015, **7**, 483–488.

(130) M. P. Wiesenfeldt, Z. Nairoukh, W. Li and F. Glorius, *Science*, 2017, **357**, 908–912.

(131) O. Shyshov, K. A. Siewerth and M. von Delius, *Chemical Communications*, 2018, **54**, 4353–4355.

(132) Y. Xia, T. D. Nguyen, M. Yang, B. Lee, A. Santos, P. Podsiadlo, Z. Tang, S. C. Glotzer and N. A. Kotov, *Nature Nanotechnology*, 2011, **6**, 580–587.

(133) R. K. Cersonsky, G. van Anders, P. M. Dodd and S. C. Glotzer, *Proceedings of the National Academy of Sciences*, 2018, **115**, 1439–1444.

(134) S. Lee, E. G. Teich, M. Engel and S. C. Glotzer, *Proceedings of the National Academy of Sciences*, 2019, **116**, 14843–14851.

(135) R. M. Capito, H. S. Azevedo, Y. S. Velichko, A. Mata and S. I. Stupp, *Science*, 2008, **319**, 1812–1816.

(136) M. C. Marchetti, J. F. Joanny, S. Ramaswamy, T. B. Liverpool, J. Prost, M. Rao and R. A. Simha, *Reviews of Modern Physics*, 2013, **85**, 1143–1189.

(137) Q. Chen, S. C. Bae and S. Granick, *Nature*, 2011, **469**, 381–384.

(138) T. Sanchez, D. T. N. Chen, S. J. DeCamp, M. Heymann and Z. Dogic, *Nature*, 2012, **491**, 431–434.

(139) S. De, A. P. Bartók, G. Csányi and M. Ceriotti, *Phys. Chem. Chem. Phys.*, 2016, **18**, 13754–13769.

(140) E. A. Engel, A. Anelli, M. Ceriotti, C. J. Pickard and R. J. Needs, *Nat. Comm.*, 2018, **9**, 1–7.

(141) S. J. Marrink, A. H. De Vries and A. E. Mark, *J. Phys. Chem. B*, 2004, **108**, 750–760.

(142) M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1**, 19–25.

(143) J. Lee, X. Cheng, J. M. Swails, M. S. Yeom, P. K. Eastman, J. A. Lemkul, S. Wei, J. Buckner, J. C. Jeong, Y. Qi et al., *J. Chem. Theory Comput.*, 2016, **12**, 405–413.

(144) P.-C. Hsu, B. M. Bruininks, D. Jefferies, P. Cesar Telles de Souza, J. Lee, D. S. Patel, S. J. Marrink, Y. Qi, S. Khalid and W. Im, *J. Comput. Chem.*, 2017, **38**, 2354–2363.

(145) G. Bussi, D. Donadio and M. Parrinello, *J. Chem. Phys.*, 2007, **126**, 014101.

(146) H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, *The Journal of Chemical Physics*, 1984, **81**, 3684–3690.

(147) S. J. Marrink, J. Risselada and A. E. Mark, *Chem. Phys. Lipids*, 2005, **135**, 223–244.

(148) J. C. Mathai, S. Tristram-Nagle, J. F. Nagle and M. L. Zeidel, *J. Gen. Physiol.*, 2008, **131**, 69–76.

(149) H. Sunshine and M. L. Iruela-Arispe, *Curr. Opin. Lipidol.*, 2017, **28**, 408.

(150) V. Corradi, B. I. Sejdiu, H. Mesa-Galloso, H. Abdizadeh, S. Y. Noskov, S. J. Marrink and D. P. Tieleman, *Chem. Rev.*, 2019, **119**, 5775–5848.

(151) S. J. Marrink, V. Corradi, P. C. Souza, H. I. Ingólfsson, D. P. Tieleman and M. S. Sansom, *Chem. Rev.*, 2019, **119**, 6184–6226.

(152) Z. Jarin, J. Newhouse and G. A. Voth, *J. Chem. Theory Comput.*, 2021, **17**, 1170–1180.

(153) J. Zhao, J. Wu, H. Shao, F. Kong, N. Jain, G. Hunt and G. Feigenson, *Biochim. Biophys. Acta, Biomembr.*, 2007, **1768**, 2777–2786.

(154) R. L. Biltonen and D. Lichtenberg, *Chem. Phys. Lipids*, 1993, **64**, 129–142.

(155) D. Stelter and T. Keyes, *J. Phys. Chem. B*, 2017, **121**, 5770–5780.

(156) W. C. Swope, J. W. Pitera and F. Suits, *The Journal of Physical Chemistry B*, 2004, **108**, 6571–6581.

(157) G. Pérez-Hernández, F. Paul, T. Giorgino, G. D. Fabritiis and F. Noé, *The Journal of Chemical Physics*, 2013, **139**, 015102.

(158) B. G. Keller, J.-H. Prinz and F. Noé, *Chemical Physics*, 2012, **396**, 92–107.

(159) B. Trendelkamp-Schroer, H. Wu, F. Paul and F. Noé, *The Journal of Chemical Physics*, 2015, **143**, 174101.

(160) M. Hoffmann, M. K. Scherer, T. Hempel, A. Mardt, B. de Silva, B. E. Husic, S. Klus, H. Wu, J. N. Kutz, S. Brunton and F. Noé, *Machine Learning: Science and Technology*, 2021.

(161) M. J. Abraham, T. Murtola, R. Schulz, S. Páll, J. C. Smith, B. Hess and E. Lindahl, *SoftwareX*, 2015, **1-2**, 19–25.

(162) F. Noé, C. Schütte, E. Vanden-Eijnden, L. Reich and T. R. Weikl, *Proceedings of the National Academy of Sciences*, 2009, **106**, 19011–19016.

(163) W. E. and E. Vanden-Eijnden, *Journal of Statistical Physics*, 2006, **123**, 503–523.

(164) G. B. Sergeev, *Nanochemistry*, Elsevier, 2nd edn., 2014.

(165) S. Mørup, M. F. Hansen and C. Frandsen, *Beilstein Journal of Nanotechnology*, 2010, **1**, 182–190.

(166) K. Liu, Z. Nie, N. Zhao, W. Li, M. Rubinstein and E. Kumacheva, *Science*, 2010, **329**, 197–200.

(167) E. V. Shevchenko, D. V. Talapin, N. A. Kotov, S. O'Brien and C. B. Murray, *Nature*, 2006, **439**, 55–59.

(168) S. Julin, A. Korpi, N. Nonappa, B. Shen, V. Liljeström, O. Ikkala, A. Keller, V. Linko and M. A. Kostiainen, *Nanoscale*, 2019, **11**, 4546–4551.

(169) H. Heinz, R. A. Vaia, B. L. Farmer and R. R. Naik, *The Journal of Physical Chemistry C*, 2008, **112**, 17281–17290.

(170) J. Wang, R. M. Wolf, J. W. Caldwell, P. A. Kollman and D. A. Case, *Journal of Computational Chemistry*, 2004, **25**, 1157–1174.

(171) W. L. Jorgensen, J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, *The Journal of Chemical Physics*, 1983, **79**, 926–935.

(172) G. Bussi, D. Donadio and M. Parrinello, *The Journal of Chemical Physics*, 2007, **126**, 014101.

(173) M. Parrinello and A. Rahman, *Journal of Applied Physics*, 1981, **52**, 7182–7190.

(174) U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, *The Journal of Chemical Physics*, 1995, **103**, 8577–8593.

(175) C. Arnarez, J. J. Uusitalo, M. F. Masman, H. I. Ingólfsson, D. H. De Jong, M. N. Melo, X. Periole, A. H. De Vries and S. J. Marrink, *J. Chem. Theory Comput.*, 2015, **11**, 260–275.

(176) H. Schulze, *Journal für Praktische Chemie*, 1882, **25**, 431–452.

(177) C. Lionello, C. Perego, A. Gardin, R. Klajn and G. M. Pavan, *ACS Nano*, 2022, **17**, 275–287.

(178) D. H. de Jong, G. Singh, W. D. Bennett, C. Arnarez, T. A. Wassenaar, L. V. Schäfer, X. Periole, D. P. Tieleman and S. J. Marrink, *J. Chem. Theory Comput.*, 2013, **9**, 687–697.