Identifying and Tracking Defects in Dynamic Supramolecular Polymers

Identifying and Tracking Defects in Dynamic Supramolecular Polymers

Piero Gasparotto,[*,†,¶] Davide Bochicchio,[‡] Michele Ceriotti,[*,†] and Giovanni M. Pavan[*,‡,§]

†*Laboratory of Computational Science and Modeling, Institute des Materiaux, Ecole polytechnique fédérale de Lausanne, CH-1015 Lausanne, Switzerland*

‡*Department of Innovative Technologies, University of Applied Sciences and Arts of Southern Switzerland, Galleria 2, Via Cantonale 2c, CH-6928 Manno, Switzerland*

¶*Thomas Young Centre and Department of Physics and Astronomy, University College London, Gower Street London WC1E 6BT, United Kingdom*

§*Department of Applied Science and Technology, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy*

E-mail: p.gasparotto@ucl.ac.uk; michele.ceriotti@epfl.ch; giovanni.pavan@polito.it

Abstract

A central paradigm of self-assembly is to create ordered structures starting from molecular monomers that spontaneously recognize and interact with each other via noncovalent interactions. In the recent years, great efforts have been directed towards perfecting the design of a variety of supramolecular polymers and materials with different architectures. The resulting structures are often thought of as ideally perfect, defect-free supramolecular fibers, micelles, vesicles, etc., having an intrinsic dynamic character, which are typically studied at the level of statistical ensembles to assess their average properties. However, molecular simulations recently demonstrated that local defects that may be present or may form in these assemblies, and which are poorly captured by conventional approaches, are key to controlling their dynamic behavior and properties. The study of these defects poses considerable challenges, as the flexible/dynamic nature of these soft systems makes it difficult to identify what effectively constitutes a defect, and to characterize its stability and evolution. Here, we demonstrate the power of unsupervised machine learning techniques to systematically identify and compare defects in supramolecular polymer variants in different conditions, using as a benchmark 5Å-resolution coarse-grained molecular simulations of a family of supramolecular polymers. We show that this approach allows a complete data-driven characterization of the internal structure and dynamics of these complex assemblies and of the dynamic pathways for defects formation and resorption. This provides a useful, generally applicable approach to unambiguously identify defects in these dynamic self-assembled materials and to classify them based on their structure, stability and dynamics.

Introduction

Defects and their dynamics play a fundamental role in controlling both equilibrium and time-dependent properties of materials.[1–5] Extended crystallographic defects, such as stacking faults and dislocations, underpin the mechanical plastic behavior of metals.[6–8] Atomic impurities that enter the lattice of semiconducting materials can be used as dopants to fine-tune their electronic properties.[9,10] While the concept of defects is typically associated to crystals, or in general to highly

ordered structures, they play a key role also in softer materials,[11] although their identification in such dynamic and disordered systems can be way more elusive.[12]

Supramolecular polymers are generated by directional self-assembly of monomers that interact via non-covalent interactions such as hydrogen bonding, $\pi - \pi$ stacking, etc.[13,14] While in these systems the interactions between the monomers are generally weak compared to covalent bonds, their reversible character imparts interesting dynamic properties to these materials that are similar to those of natural compounds and impossible to achieve with covalent polymers.[15,16] In these systems, the self-assembled monomers obey a well-defined supramolecular equilibrium with the monomers present in solution. This makes these structures innately dynamic, while the constitutive monomers are continuously exchanged depending on their interactions with each other and with the external solvent.[17,18] Many approaches have been designed to characterize the structure and dynamics of supramolecular polymers.[19,20] However, the limited resolution that can be achieved in the experiments makes it a challenge to resolve unambiguously the molecular factors that control the behaviour of these complex systems.

Recently, advanced molecular simulations allowed us to study the intrinsic dynamics (monomer exchange within and in-and-out of the assembly) of various types of supramolecular polymers at a submolecular resolution ($<$5 Å).[21,22] These works demonstrated that defects, which may be intrinsically present in these assemblies, are key in controlling any dynamic transition in the supramolecular structure. In particular, such defects were found to control the supramolecular dynamics of these assemblies,[21] how quickly they adapt and change in response to specific stimuli[23] and even how these complex supramolecular systems behave and evolve when they are far from the equilibrium.[24] Therefore, despite the wide and general effort toward reaching perfection in the self-assembly of ordered architectures, we now have clear indications that the dynamic behavior and properties of some supramolecular polymers are intimately connected to defects, and may even originate from them.[21,24] However, the complex statistical nature of these systems poses the non-trivial challenge of identifying and classifying such defects in their soft and dynamic supramolecular structures in an unambiguous, general and transferable way.

Here we show how machine learning approaches can be applied to tackle this problem and achieve a complete and automatic characterization of defects in a prototypical class of supramolecular polymers, based on 1,3,5-benzenetricarboxamide (BTA - Figure 1a), without prior assumptions of what exactly a defect is in these assemblies. BTA monomers self-assemble into one-dimensional fibers due to a combination of π-π stacking of the aromatic cores, threefold hydrogen bonding between the amide groups that surround the cores and solvophobic effects. Here we use a recently-developed fine coarse-grained (CG) model that guarantees satisfactory sampling of these complex systems and good accuracy in the treatment of all the key interactions controlling the assembly.[25] While allowing to study the structure and dynamics of BTA supramolecular polymers at sub-molecular resolution (<5 Å), this CG model recently provided some compelling preliminary evidence of the presence of defects in water-soluble BTA fiber variants.[21] Nonetheless, the elusive and fleeting nature of defects in these dynamic systems demands for a more general and unambiguous procedure to identify and compare them. Our agnostic data-driven analysis relies on the Smooth Overlap of Atomic Potential (SOAP)[26–28] framework in combination with the Probabilistic Analysis of Molecular Motifs (PAMM)[29,30] to process trajectories (<5 Å resolution) obtained from unbiased molecular dynamics simulations of BTA supramolecular polymers and to describe the structural and dynamic motifs present in the supramolecular polymers in an automatic way. We obtain a complete description of how defects are continuously and spontaneously created and re-adsorbed in the structure of supramolecular fiber variants, along with qualitative mechanisms and pathways. This greatly improves our understanding of the internal structure and dynamics of supramolecular polymers, demonstrating how these assemblies cannot be simply seen as static structures or studied in terms of average properties. We provide detailed insights into the complex dynamic nature of these supramolecular fibers and demonstrate the capabilities of a general analysis tool that is also applicable to study other classes of dynamic self-assembled materials.

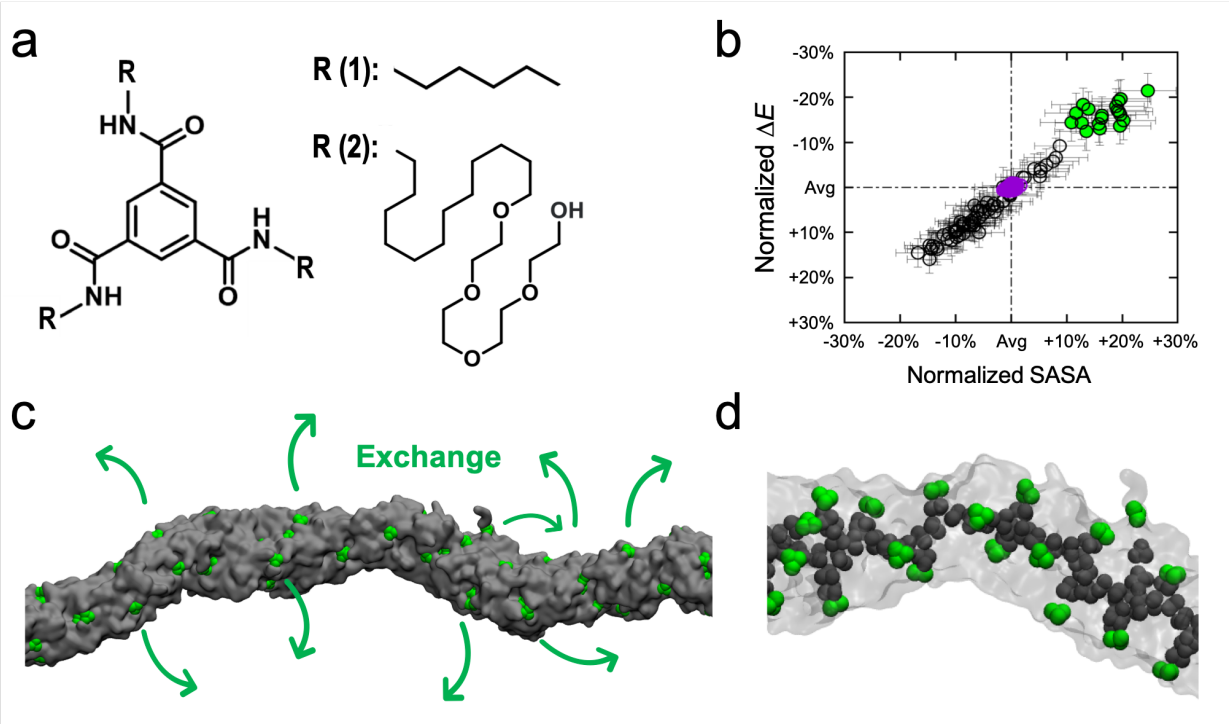


Figure 1: Defects in BTA supramolecular polymers. (a) Chemical structure of the BTA supramolecular polymers studied herein - BTA fibers soluble in an organic solvent (1) (i.e. octane) and in water (fiber (2), amphiphilic side chains R(2)). (b) Heuristic analysis of CG-MD trajectories for fiber (2): classification of monomers (black and green) based on their incorporation energy into the assembly (ΔE) and their exposure to the external solvent (SASA) normalized with respect to the average.[21] This gives a preliminary qualitative evidence of defects. Data for fiber (1) are reported in purple for comparison (no defects appearing in this fiber). (c,d) Exterior (c) and interior (d) snapshots of a section of equilibrated fiber (2) from CG-MD simulation.[21] The fiber surface is colored in grey, green spots (monomer cores) are clearly visible from the outside. These are defects in the stack (in green) that constitute "hot spots" from which monomer exchange originates in these fibers.[21]

Computational Details

Molecular models and MD simulations

The CG molecular models for the BTA fiber variants (1) and (2) simulated herein have been developed and tested by our group in recent works[21,25] Briefly, these models are based on the popular MARTINI force field,[31] while they also include an explicit treatment of hydrogen bonding between the amide CG beads in the form of rigidly rotating dipoles, which can interact with each other mimicking the directionality of hydrogen bonds. These models have been extensively

validated, and were proven reliable in capturing the effect of slight variations in the structure of the monomers or in the external conditions on the structure and dynamics of the supramolecular polymers .[21,25] Fiber (3) has been constructed taking fiber (1) and reducing the value of the charges forming the dipoles contained in the amide beads (from 0.8 e to 0.65 e).

All the CG-MD simulations have been performed with the GROMACS (2016 series) software,[32] using the *md* leap frog integrator with a time step of 20 fs. The fibers have been constructed starting from 80 initially extended (1) and (2) monomers perfectly stacked, grazing the terminal BTA cores along the main fibre axis to effectively model via periodic boundary conditions the bulk of infinite fibers. After that, the simulation boxes have been filled with standard water MARTINI beads in the case of fiber (2) or with MARTINI octane in the other cases to solvate the fibers, which have been equilibrated via classical CG-MD simulations in NPT conditions (constant N: number of particles, P: pressure and T: temperature). After equilibration, production runs have been conducted long enough to sample the internal fiber dynamics in each case (20 μs for fiber (2), 2 μs for fiber (1) and (3)). Temperature has been kept at 27 °C using the v-rescale thermostat[33] with a coupling constant of 1.0 ps. Consistent with the directional nature of these infinite CG fibers, semi-isotropic pressure scaling has been used to keep the pressure in the system at 1 atm (coupling constant of 4 ps). To be able to compare the different fibers dynamics and structural complexity, we saved the trajectories with the same stride: 1 frame every 1 ns. Such CG-MD trajectories have then been analyzed via our PAMM-SOAP approach.

Probabilistic Analysis of Molecular Motifs

PAMM is a pattern recognition framework designed to identify molecular motifs based on their meta-stability. The core idea behind the PAMM scheme consists into partitioning the probability distribution function (PDF) of structures sampled from an atomistic simulation or collected from experiments.[29] The input is simply a series of N (preferably high-dimensional) vectors representing local or global environments. This could be any function of the atomic coordinates, such as distances, angles or more sophisticated descriptors. Then one needs to perform a kernel density estimation

(KDE) on a grid extracted using farthest point sampling (FPS), which has proven to be well suited for selecting the most widely spread set of landmarks from the initial set.[34] Combining FPS with KDE allows for linear scaling with the number of samples and for the use of very large datasets, which is usually a bottleneck for most clustering algorithms. Extensive details on PAMM are available in the SI.

All the analysis were done using the PAMM implementation available at `https://github.com/cosmo-epfl/pamm`. The KDE was computed on a sparse grid counting 2000 points. The kernel bandwidth and local scale factors were determined automatically as discussed in Ref. 30. The automatically determined bandwidth was scaled by a factor of 0.7, while we used for the quick-shift distance the automatic choice. Furthermore, to estimate the overlap between clusters and determine whether or not clusters should be merged, we performed a bootstrapping analysis resampling the training data 71 times. This was used to generate a cluster adjacency matrix from which we built the tree-like plot illustrating the results of a hierarchical clustering procedure performed on the micro-clusters in Figure 4. The connected clusters are those that are likely to be merged in the bootstrap samples.

Smooth Overlap of Atomic Positions

Capturing the structural complexity with simple descriptors is not trivial and could lead to results affected by strong bias. SOAP is a general state-of-art atom-centered, density-based representation of the atomic environment which has been proven very powerful for both properties prediction[27,28,35] and structural classification.[36] A sum of Gaussians centered on each surrounding atom produces a smooth representation of the atomic density around a given atom. Extensive details on SOAP are provided in the SI.

In our analyses, we centered a SOAP in a fictitious atom at the center of the aromatic ring of each monomer core and we considered only the CG atoms representing the ring and the amide beads to compute SOAP vectors. A cutoff radius of 8Å is large enough to incorporate information on several neighboring monomers (we also tested the effect of enlarging the cutoff to 12Å or 16Å

obtaining analogous results – see SI). In the analysis we set as parameters: `n_max` = 8, `l_max` = 8, `cutoff` = 8.0 Å. The width of the Gaussian functions was set to 1 Å. All other parameters were set to their default values, as implemented in the `quippy` library (`https://libatoms.github.io/QUIP/descriptors.html`).

The analyses of Figure 2 were performed using as a dataset the trajectories of the monomer cores obtained from the CG-MD simulations of the individual fibers (analysis not directly comparable between the fibers). Then, in order to compare between the different fibers, in the analyses of Figures 3 and 4 we used the CG-MD trajectories of all monomers of all fibers as merged into a single large dataset.

Dimensionality reduction

Given the high dimensionality of the SOAP vectors we use dimensionality reduction to simplify the interpretation of the results. The use of these algorithms has become commonplace in computational science and has proven to provide a great qualitative understanding of complex structures in different type of materials. These algorithms seek to find an optimal low-dimensional embedding capable of capturing the shape of a high-dimensional manifold, which would be impossible to visualize otherwise. PCA is probably the simplest statistical algorithm among all the dimensionality reduction schemes, yet it performed very well once combined with PAMM clustering in projecting in low-dimensions all the different high-dimensional motifs. The PCA linearly transforms a set of correlated data into a new set of values of variables linearly uncorrelated, called principal components. All the two dimensional maps discussed in the text were obtained using the PCA implementation available in SciKit-Learn,[37] by projecting points from the high-dimensional SOAP feature space to two dimensions using the first two principal components. Two dimensions are clearly not enough to capture all the complexity of the initial manifold, however for our systems this is already quite informative since the first two principal components alone explain $\sim$74% of the full covariance (see SI for further details).

Results and Discussion

Traces of defects in dynamic supramolecular polymers

In order to demonstrate how an automated analysis can shed light onto the presence, formation and evolution of defects in supramolecular polymers, as a case study here we focus on systems based on the BTA architecture, for which our fine CG models recently suggested that the presence of local defects is of key importance.[21,23,25] In particular, we start by considering two variants of BTA supramolecular polymers that are soluble in organic solvent and in water (Figure 1a), and we will refer to these two systems respectively as (1) and (2). The (1) monomers differ from the (2) monomers only in the side chains, that are lipophilic in (1) and amphiphilic in (2). We consider BTA supramolecular fibers composed of 80 (1) or (2) monomers modelled using the CG models of Ref. 21 and Ref. 25, equilibrated respectively in octane and water by means of coarse-grained molecular dynamics (CG-MD) simulations. Fibers were constructed applying periodic boundary conditions to mimic the structure of much longer supramolecular fibers (see Computational Methods for further details). These CG-MD simulations make it possible to obtain high-resolution insights into the structural and dynamic features of BTA supramolecular polymers in the different environments. In particular, visual inspection of converged CG-MD simulation trajectories of BTA supramolecular polymers in an aqueous medium reveals a complicated, non-uniform structure: these fibers contain branches in their backbone (or defects in the stacking of cores) as well as free monomers that are adsorbed on the fiber surface (Figure 1).

Attempts to identify monomers involved in a defect can use simple structural and energetic parameters (or fingerprints), such as, for example, the solvent accessible surface area (SASA), the interaction energies (ΔE) of the individual monomers in the assembly or the coordination between the cores in the stacking (assessing somehow the stacking order).[21] Such analyses allow to single out monomers that are more exposed to the solvent than the average (SASA>avg, Figure 1b), and that also belong to domains of the fiber where the assembly is less stable than the average (ΔE <avg, Figure 1b). However, this approach has at least two crucial weaknesses. First, it is based

on heuristic descriptors, that rely on prior chemical intuition, making the analysis somewhat circular: it is obvious that monomers with high energy or high solvent accessible area will be associated with defective stacking, and indeed the two features are highly correlated and therefore redundant. Second, it yields an over-simplified picture that does not differentiate between the different kinds of motifs (main strand, branches, adsorbed monomers) that are apparent in the trajectory, and as such it is unlikely to be able to quantitatively characterize in complete way defect structure and dynamics. Namely, as it can be seen in Figure 1b, the continuous distribution of the black points on the ΔE-SASA plot suggests a dynamic diversity and complexity in the assembly of monomers, which makes it very difficult to identify in rigorous way what exactly is a defect and what is not in the supramolecular polymer.

Automatic detection of molecular motifs in supramolecular polymers

To overcome the aforementioned limitations, we apply a combination of machine-learning techniques (i.e., high-dimensional agnostic descriptors of the molecular/atomic environments, unsupervised density-based data clustering and dimensionality reduction - see later on), aiming to obtain a characterization of the BTA fibers based on generally-applicable structural indicators and on an identification of defect states that relies exclusively on their (meta)stability. To minimize the risk of introducing prior bias in the choice of geometric descriptors,[38] we decided to adopt the SOAP feature vectors,[26] that characterize the local surrounding of each monomer core in a general and agnostic fashion, based on a discretized representation of three-body correlation functions (see Computational Methods). Once the structural sampling of the simulated systems is converged, the next step involves analyzing CG-MD simulations and mapping all the possible metastable structural patterns explored by the systems in terms of SOAP features. This allows to track all possible states visited by the monomer cores in the model fibers during the simulations. To do so, we choose a density-based clustering scheme, the probabilistic analysis of molecular motifs (PAMM).[29] PAMM estimates the probability density with which different motifs appear, and identifies local probability maxima as significant (metastable) patterns. Thus, it allows to partition all the configurations for the

monomer cores into a small number of clusters and to identify them in different simulations. While assessing the stability of clusters with respect to noise, PAMM can also clarify the hierarchical relations between motifs.[30]

In our analysis, SOAP vectors are computed by attaching one fictitious atom to the center of each BTA monomer. As better clarified below, by explicitly considering in the analysis only the CG atoms corresponding to the aromatic rings and amide groups in the monomers we make the representation transferable across the systems. This is important to measure defects in BTA supramolecular polymer variants in a comparable way (see next section). Preliminary tests showed clearly that the core of the monomer alone is a key discriminant of the variance in terms of order/disorder within these supramolecular fibers, which is well captured considering only the stacking of the aromatic rings. Further discussions on the minimum amount of structural details necessary to rationalize order/disorder in supramolecular polymers can be found in the Supporting Information (SI).

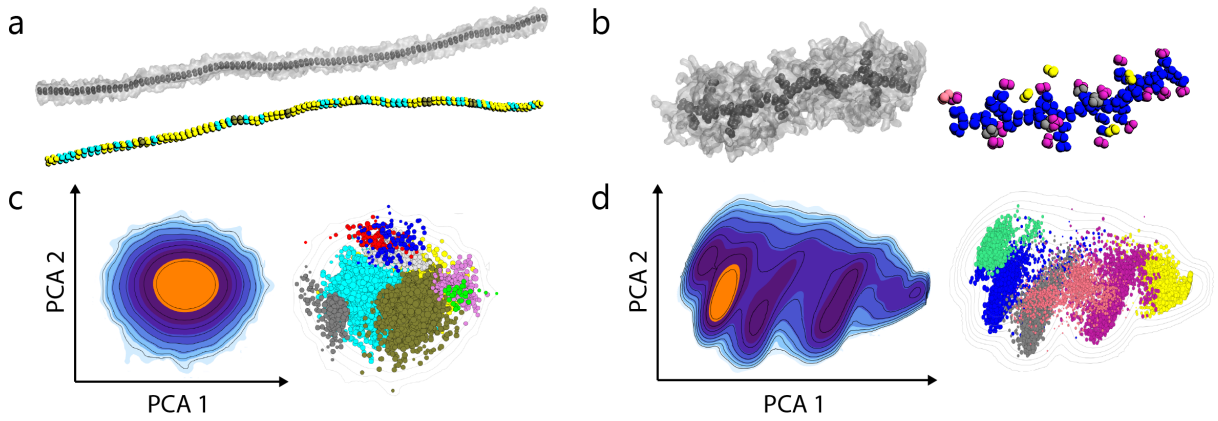


Figure 2: Automatic detection of ordered/disordered molecular motifs in BTA supramolecular polymers. (a,b) CG-MD snapshots and PAMM analysis on CG models of BTA fibers (1) (a) and (2) (b). In the CG-MD snapshots, the BTA monomer cores are shown in dark grey, while the side chains are in transparent grey (explicit solvent molecules not shown for clarity). The fibers are also shown having the monomer cores colored based on the PAMM analysis - different colors identify differences in the molecular environment surrounding the monomer cores. (c) Analysis for BTA fiber (1). Left: density distribution (with contour plots and color scale that indicates increasing density - from cyan to blue to orange) of SOAP vectors projected on the first two PCA components. Right: PCA projections of 2000 SOAP vectors randomly sampled from the full CG-MD trajectory of fiber (1): points colored according to PAMM classification of the identified miscrostates (see Computational Methods). (d) Same analysis as in (c) for fiber (2).

Figure 2 reports the results of a separate automatic PAMM-SOAP analysis for fiber (1) (Figure 2a,c) and fiber (2) (Figure 2b,d). The combination of SOAP features and PAMM clustering provides a robust and agnostic scheme to detect and classify all the different states for the individual monomers in the dynamic supramolecular structures. Given the (very) high-dimensional nature of SOAP vectors, a PAMM analysis based on the full descriptors would be computationally demanding and also too susceptible to insignificant structural features. Therefore, we decided to reduce the dimensionality of the dataset using Principal Component Analysis (PCA) to transform SOAP vectors from 10585 to 8 dimensions. As discussed in the SI, 8 components are enough to cover more than 93% of the overall variance of the original dataset. Projecting on the first two principal components, one can produce simple low-dimensional maps useful to visualize the otherwise hard to interpret SOAP HD feature space (Figure 2c,d). Each point in the map represents the configuration of a single monomer at given time, and is colored according the 8*D* PAMM clustering. The combination of clustering and low-dimensional representations is useful to rationalise the relationship between different molecular motifs.

The colors in Figure 2a-d identify multiple metastable *structural states* in both fiber (1) and (2). A cursory inspection of the maps, however, suggests that the diversity within fiber (2) is larger, and more clear-cut, than in fiber (1). In the case of fiber (2), clusters are clearly distinct, and one can easily recognize multiple modes in the distribution, separated by lower-population transition regions, even in the 2D projection. On the other hand, the clusters in fiber (1) are quite similar and, although PAMM is capable of separating them based on the 8 PCA features, there are no clear separate modes visible in the two principal components visualized in Figure 2c. One of the main objectives of this analysis is to compare structural motifs between different fibers. However, at this stage the analyses of Figure 2 are not comparable between the different fibers. Namely, we cannot determine whether different clusters in fiber (1) (Figure 2a, bottom) identify relevant differences within the assembly when compared to the differences observed in fiber (2), and whether there is a relationship between the motifs observed in the two fibers. Indeed, it could be that the structural diversity observed in (1) turns out to be simply noise, or fluctuations around the same class of

patterns, when compared to (2). Being able to perform a systematic comparison and classification of different assemblies according to a common metric would be a first important step toward setting order in such complex dynamic systems.

Automatic comparison of defects in different supramolecular polymers

Above all, to make the PAMM-SOAP analysis comparable between different fibers, one needs to choose a representation of motifs that takes into consideration the common structural features between the different systems. To this end, we focused only on the packing of BTA cores, as otherwise the analysis would detect, as the most prominent feature, the fact that different systems have different side chains or solvents. Then, one needs to choose a common set of features to perform clustering and to obtain a low-dimensional representation. An option would be to pick one of the systems as a common reference and to project all the others into this one. However, projecting the data of other systems on a reference one would automatically imply that one already knows what is the system that is richest in terms of microstates (the least rich ones should be projected on the map obtained for the most diverse, as doing the opposite would lead to loss of information). This could be useful if one had a baseline system, and wanted to investigate how small perturbations affect the motifs that are prominent in such reference (see SI for an example of such approach). Given that in our case we want to compare fibers (1) and (2) on equal footings, it is more appropriate, more general and less biased to merge all data coming from the CG-MD trajectories of the monomer cores of the fibers into a single dataset, and perform a single PCA decomposition, followed by a combined PAMM clustering.

Figure 3 shows the results of this unified PAMM-SOAP analysis, that we then use to classify structural motifs in the different fibers. The 2D map for fiber (2) provides a rich, diverse picture, where the yellow and red clusters identify respectively monomers that are connected to a single other core or not connected at all (Figure 3b,left). On the other extreme, the defect-free fiber (1) (Figure 3a,b, right) shows a single cluster corresponding to the perfect stacking of cores in the polymer. This comparison allows to evaluate the internal diversity/richness in these assemblies

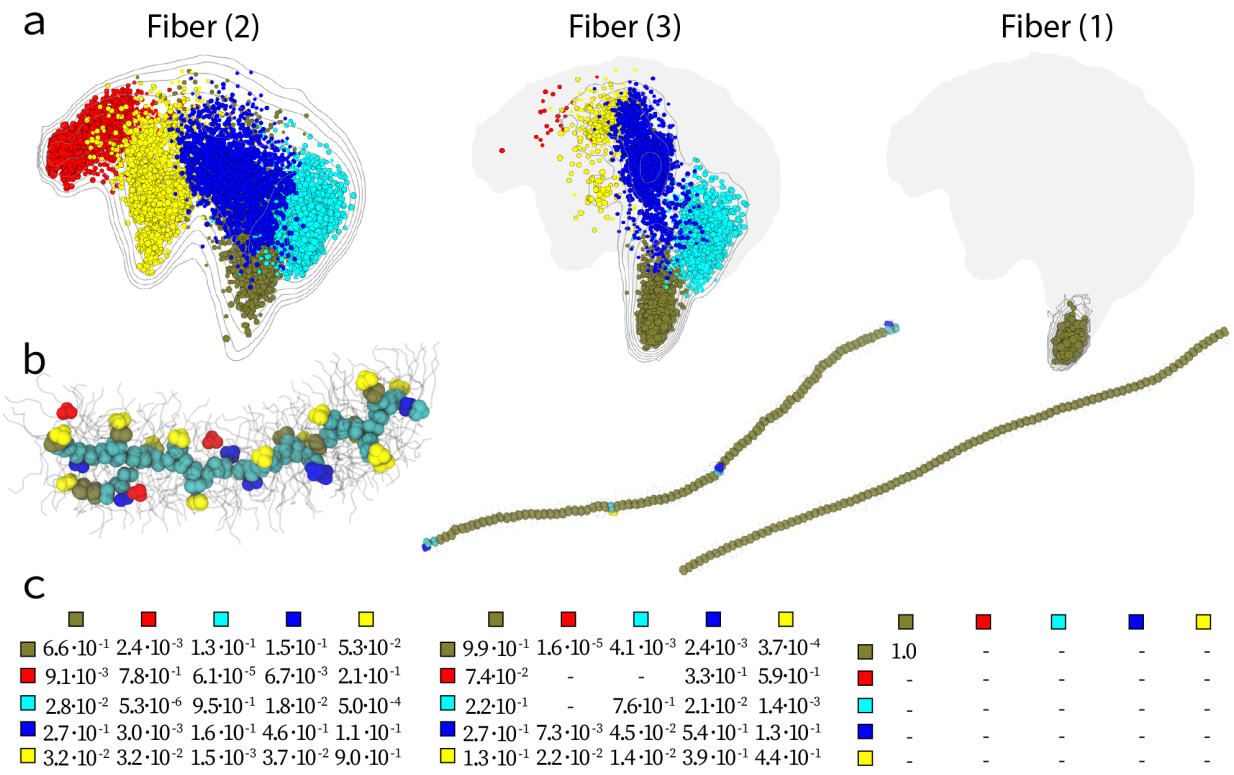


Figure 3: Comparing molecular motifs in different supramolecular polymers. Comparable results from PAMM analysis are shown for fibers (2), (3) and (1) (left to right). (a) PCA of PAMM micro-clusters (identified by different colors) of the individual fibers projected on the complete configurational space of these supramolecular polymers. Isolines correspond to the probability distribution of motifs in each fiber, while gray shadows represent the contour of the complete configurational space of the three merged systems. (b) CG-MD snapshots where the monomer cores are colored based on the PAMM analysis. (c) Interconversion probability matrix for the residence and exchange of monomers between micro-states, calculated directly from the PAMM analysis of the CG-MD trajectories, with a time interval Δt=1ns. Numbers are unitless, and represent normalized transition probabilities (see Equation 1).

on the same scale, and provides a first answer to the questions related to the results of Figure 2. Indeed, it is now much clearer that the water-soluble fiber (2) is richer in terms of monomer states and molecular motifs compared to fiber (1). Such diversity within fiber (2) is consistent with recent experimental and computational observations that revealed that these water-soluble assemblies are way more dynamic and less ordered than expected.[21,39] On the other hand, fiber (1) appears as a nearly perfect stack of monomers (similar to the cartoons of ideal fibers often used in the literature).

To enrich our analysis, we also considered one additional system in this comparison between the fibers, referred in the following as fiber (3). This is a variant of fiber (1), composed of identical

monomers, but having decreased interactions between the cores. In our CG models, this is obtained by lowering the partial charges present in the dipole mimicking the hydrogen bonding between the monomers (see Computational Methods). While this system is not meant to model a real chemical system, it is useful to investigate the role of molecular interactions in determining the presence and the stability of defects. The behaviour of fiber (3) – which possesses artificially weakened directional interactions between its monomers – is somewhat intermediate between fibers (1) and (2) (see Figure 3, middle). The concentration of defects and the diversity of motifs that are identified in the analysis is increased in fiber (3) compared to fiber (1). The majority of monomers still belong to a highly-ordered state, but defects that are structurally similar to defect states in the more dynamical fiber (2) can be formed. This result highlights the fact that thinking to supramolecular polymers only as extended and ordered fibers/stacks of monomers can be misleading, as they can be also quite far from this ideal model and very different one from the other as a consequence of tiny variations in the structure of the monomer.

Given that our data originates from a sequence of successive time frames along the CG-MD simulations – with a spacing of Δt=1ns –, we can also extend our analysis to incorporate information about the transitions between the different clusters (see also Computational Methods and the SI for details). By monitoring the state of each monomer in consecutive frames along the trajectory – that is thoroughly sampled thanks to the speed-up provided by the CG scheme – we can extract information on the relative probabilities for a monomer in a certain state (or color) to remain in that state or to transform into a different configuration (i.e., to change color) in the following frame. We can thus introduce the transition probability:

$$P(n', \Delta t | n, 0) = \frac{P(n, \Delta t \cap n', 0)}{P(n)} \tag{1}$$

that indicates the probability that a monomer that is in state n at a given time will be in state n' at time Δt.

In Figure 3c we report the interconversion probability matrices for the three fibers. The non-zero entries of these matrices provide quantitative insight into the dynamical behavior of the different

fibers. The bulk of BTA fiber (1) appears as completely static and homogeneous internally as compared to the other two fibers, being populated only by olive monomers that never change state. The situation is radically different for fiber (2), with the transition matrix indicating a high probability of interconversion between different states. For example, we observe that, between two consecutive frames, a cyan monomer will remain in the same state $\sim$95% of the times, with $\sim$2.7% probability of transforming into an olive state and $\sim$1.8% probability of undergoing a transition to blue. It is worth noting that the transition probabilities between different states depend on the rate at which we sampled the trajectory. These transition probabilities can be easily converted into transition rates between the various states by dividing (normalizing) them by the interval Δt between the frames (1 ns). However, here we are not much interested into quantitative numbers or rates (also considering that the CG dynamics is usually greatly accelerated), but rather in comparing between the pathways and the hierarchical steps underpinning the creation and resorption of defects in the various assemblies. In this sense, we find that in this specific case the data are more meaningfully discussed in terms of (sampling-dependent) normalized surviving and transition probabilities for the different states.

PAMM can also be used to extract information about the proximity/similarity between the clusters, allowing to organize them into a hierarchy of macro-clusters. As explained more in details in Ref. 30 and shown as inset in Figure 4, one can build an adjacency matrix between the various metastable states and from that define a tree representing the hierarchical interconnections. This additional analysis quantifies the proximity in SOAP space, and provides an indication on how clusters could be merged by cutting the adjacency tree at a suitable height – this builds coarse-grained regions in feature space that we will refer to as *macro-clusters*.

This procedure produces a much simpler picture of the structural complexity of these fibers. Olive and cyan micro-clusters are merged into one macro-cluster (Figure 4, in purple), which from the physical point of view identifies the ordered domains (more stable and static) within the fibers. Blue and yellow micro-clusters, merged in the green macro-cluster in Figure 4, represent in a broad sense the domain of defects in the fibers. The orange macro-cluster of Figure 4 is exactly the same

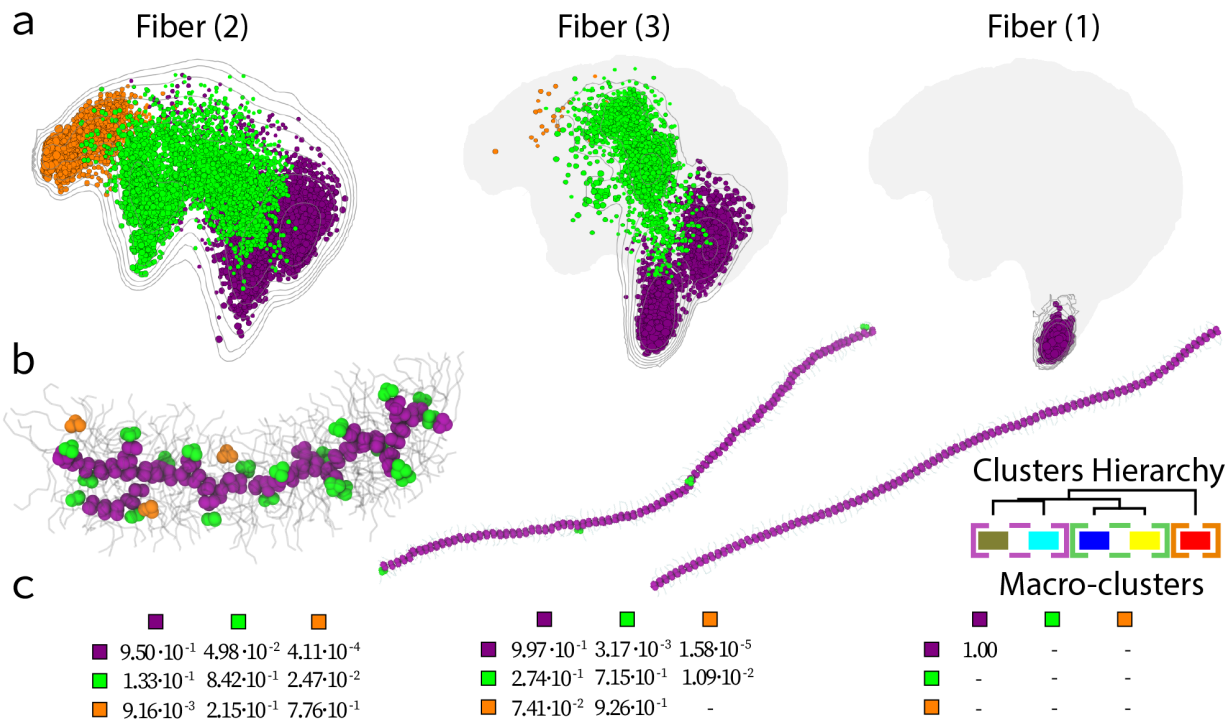


Figure 4: Construction of macro-clusters starting from the PAMM clusters depicted in Figure 3. Results are shown for fibers (2), (3) and (1) (left to right). (a) PCA of PAMM macro-clusters of the individual fibers projected on the complete configurational space of these supramolecular polymers. Isolines correspond to the probability distribution of motifs in each fiber, while gray shadows represent the contour of the complete configurational space of the three merged systems. Purple macro-clusters (stable/ordered domains in the fibers) come from merging cyan and olive micro-clusters, green macro-clusters (defects) from yellow and blue micro-clusters, while orange ones correspond to red micro-clusters (monomers diffusing between defects on the fiber surface). (b) CG-MD snapshots where the monomer cores are colored based on the PAMM macro-clusters. (c) Interconversion probability matrix for the residence and exchange of monomers between the macro-states, calculated directly from the PAMM analysis of the CG-MD trajectories, with a time interval Δt=1ns. Numbers are unitless, and represent normalized transition probabilities (see Equation 1).

of the red micro-cluster of Figure 3. This identifies those monomers whose core is not stacked to any other one in the assembly, and which are absorbed on the fiber surface spontaneously moving from one defect to another. At this higher level of hierarchy, the PAMM-SOAP analysis for fiber (2) (Figure 4b, left) looks quite similar to the heuristic analysis of Figure 1b. However, at the same time, this new one is significantly better. First, this is a completely ”bottom-up” analysis, where the distinction between the clusters are automatically identified from the data (i.e., from the CG-MD trajectories). Second, it provides a deterministic rationale to group, for example, blue

and yellow micro-clusters into a macro-cluster of ordered monomers in the fibers based on the similarity and proximity between the micro-states. Third, it is more complete, as for example we automatically capture the difference between orange and green monomers, which was neglected in Figure 1b. The latter is far from being a trivial consideration. In fact, orange and green states are drastically different, as the orange diffusing monomers are not exactly part of the core stacking in the supramolecular polymer, but they are rather monomers having high mobility, which are separated from the stack of cores and are absorbed on the fiber surface. The ability to distinguish automatically the difference between such states is crucial. In fact, for example, the spontaneous surface diffusion of such orange (or red) absorbed monomers was recently demonstrated to be key in controlling the dynamics and dynamic adaptive properties of the skin of fiber (2).[21,23]

It is worth noting that this comparison has not only a structural meaning - e.g. showing that fiber (2) has more defects (green) than fiber (3) - but also a very important dynamical meaning. For example, we can observe that, between each time frame, in fiber (2) green monomers (defects) remain green $\sim$84% of the time, with a probability of $\sim$13.5% of transforming into purple (repaired defect) and $\sim$2.5% of exchanging and diffusing along the fiber. In fiber (3), the same probabilities are $\sim$71%, $\sim$27% and $\sim$1% percent. This comparison indicates that defects (green) are repaired faster in fiber (3) compared to fiber (2), and thus that they are less persistent. When looking at the orange monomers (absorbed, out-of-stack), in fiber (2) these have a surviving probability of $\sim$78%, while with a $\sim$21% probability these are reincorporated into the stack (transformed into green: reincorporated into a defect). In fiber (3) the situation is completely different. Orange monomers appear as a purely statistical short-lived state. In the rare cases where these are generated in fiber (3), coming from a defect (green) with a probability of $\sim$1%, these do not survive as orange absorbed monomers, but they are immediately reincorporated into the fiber stack (re-transformed into green monomers) with a probability as high as $\sim$93%. Thus, in fiber (3), monomer diffusion on the surface is substantially absent. In this sense, fiber (3) is a straight fiber that, from time to time, dynamically breaks forming intermittent defects that are continuously created and repaired, while, statistically, fiber (2) possesses a number of constitutional defects that are present along its length.

These results suggest that questions based on a purely structural interpretation of the assemblies - e.g., whether these supramolecular fibers possess defects or not - have little meaning. In fact, these assemblies are characterized by an intrinsic supramolecular dynamics which makes defects more or less persistent and their creation and annihilation more or less probable in their structure.

A structural and dynamical map of supramolecular polymers

As shown above, the combination of SOAP descriptors with PAMM provides a complete characterization of the structural and dynamic diversity of these supramolecular fibers. The clusters identified by our automated analysis contain information on the density of the various monomer states within the supramolecular structure, their similarity and probability of interconversion. In an equilibrium ensemble, the probability to find a monomer in a given state is related to the free-energy of that state. In this way, provided that the CG-MD trajectory is well converged, we can directly calculate the free energy surface (FES) of monomer states within the fibers based on the probability distribution in feature space ($F = -K_B T \log(P)$), using the two largest principal components obtained from the combined set of fibers.

Figure 5a shows the FES calculated for fiber (2) (analogous ones for fibers (1) and (3) are reported in Supporting Figure S1). From the FES it is evident that all micro-states lie within 2 kcal/mol of free-energy. The rather flat free-energy landscape explains, retrospectively, why this system can be efficiently sampled via an unbiased CG-MD simulation, and why fiber (2) is so rich and heterogeneous in terms of monomer states. By connecting the micro-states of Figure 5a with arrows and adding the exchange probabilities (in white) obtained from the matrices in Figure 3c, we are able to provide a complete kinetic map for the internal structure and dynamics of fiber (2). This annotated map provides a concise description of the internal structural and dynamic diversity of the supramolecular polymer. Such a detailed characterization sheds new light on the complexity of such supramolecular polymers. This makes it easier to answer questions about the nature of defects and the structural and kinetic relations between them. By using metadynamics simulations, we recently demonstrated that fiber (2) predominantly exchanges with the external solution monomers

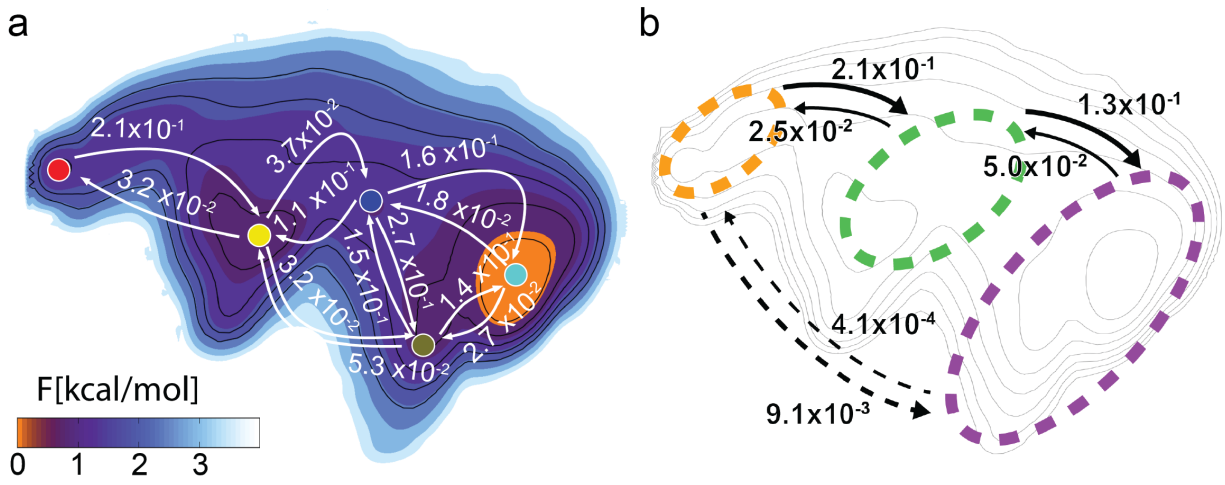


Figure 5: Thermodynamic and kinetic map of defects in supramolecular polymers. (a) FES for fiber (2), showing the differences in free energy for the various states for the monomers in the fiber. The monomer micro-states are identified by colored dots, while the probabilities for interconversion between them are represented by the white arrows. (b) A coarse-grained picture: monomer macro-states in fiber (2) (purple, green and orange areas) and relative probabilities for interconversion between them.

that are just absorbed onto the surface and not stacked/incorporated onto the fiber core.[21] These are the monomers labeled in red (or orange) in this analysis, which constitute, in fact, the monomer state having the highest free energy among all others. These absorbed monomers may be produced from the interior of fiber (2) following a variety of pathways that have a blue or yellow defective state as intermediate.

If one is not interested in the details of the pathway along different micro-clusters, it is also possible to build an analogous, yet simplified structural/dynamical map based on macro-clusters. Such map, showed in Figure 5b, provides a coarser description of the internal structure/dynamics of fiber (2), where again the transition probabilities between the higher level purple-green-orange states are obtained directly from the PAMM analysis (transition probability matrices of Figure 4c). This is a simpler scheme where monomers in the ordered interior of the fiber (purple) can dynamically become defects (green) with a probability of $\sim$5%, while such defects then have a probability of $\sim$13% to be reabsorbed/repaired (green-to-purple) and $\sim$2.5% to transform into monomers that can be exchanged and can diffuse as absorbed onto the fiber surface (green-to-orange). As a direct

consequence of detailed balance, transitions from disordered to ordered states happen with higher probabilities than transition to more disordered configurations. In more disordered and random aggregates (e.g., nanoparticles, micelles, etc.), where the cores of the monomers are not likely to be ordered/coordinated to each other, the picture would be reversed, reflecting the lower stability of ordered domains.

From a technical point of view, this analysis can also provide precious indications on how to coarse-grain further the system while preserving the correct hierarchical adjacency between the states. For example, such analysis indicates that the most rigorous way to coarse-grain the fiber model without completely losing the defects dynamics would be to do it in such a way that cyan and olive micro-clusters are merged and appear in the coarser model as a single (purple) macro-cluster. Similarly, blue-yellow motifs should merge into a single defect cluster green). The coarser model should still be able to distinguish between orange, green and purple in order to remain consistent with the general structural and dynamic features of the molecular system. This is similar, for example, to recent works by the group of Clementi where using machine learning approaches they identify dynamically coherent domains in protein models and translate them into coarse-grained models.[40] However, in our case, the use of general-purpose SOAP features as the starting point is better suited to characterize the structure and dynamics of supramolecular polymers that do not possess a predetermined backbone connectivity.

Finally, it is worth noting that the characterization obtained from such analysis could be extremely helpful to design and perform enhanced sampling simulations. Indeed, this automatic analysis can help in understanding the complex dynamic nature of these supramolecular polymers, revealing the presence of intermediate states that characterize the exchange between monomer states within the assembly. Therefore, it can provide a precious guideline to understand how to properly bias the system to sample and explore these transitions by an appropriate choice of collective variables (e.g., in metadynamics simulations), which is another direction that is currently attracting considerable interest.[41–43]

Conclusions

Defects control important properties in various types of materials, and are also key in soft self-assembled materials such as supramolecular polymers. Here we employed a data driven approach to extract directly from well-sampled coarse-grained molecular dynamics (CG-MD) simulations information on the molecular motifs and defects in variants of BTA supramolecular polymers, that we use as a demonstration and benchmark. We employed a general and high-dimensional descriptor of molecular environments (Smooth Overlap of Atomic Positions, SOAP vectors) and the Probabilistic Analysis of Molecular Motifs (PAMM) to automatically recognize and classify the prevalent structural states, developing a protocol that could be applied with minimal modifications to other classes of supramolecular materials. For example, in this case the generality of the approach allowed us to perform a transferable analysis and to compare directly between defects in different variants of these supramolecular polymers.

As an outcome, we obtained a complete picture of the rich dynamic and structural complexity/diversity of these assemblies. The results of our analysis show that supramolecular polymers of the same family (e.g., BTA) can appear both as straight nearly-perfect fibers or as more chaotic 1D filaments with a high population of defects, depending on the environment in which these are immersed. For example, the structure of the water-soluble BTA fiber studied herein is demonstrated to be a highly complex system, where the monomers in the assembly are present in multiple states that are in continuous exchange between them. Analyzing the interconversion probabilities between the different states, we could quantify the defect formation and repair pathways, and the associated probabilities (or rates), shedding new light into the dynamic nature of these complex supramolecular materials.

From a modeling point of view, the approach we present constitutes a precious tool to understand how to build rigorous coarse-grained models preserving the global structural and dynamic features of these assemblies, or how to chose critical collective variables to explore and enhance the sampling of determined transitions in the system in an accurate way during biased molecular simulations (e.g., metadynamics). From a chemistry and materials science fundamental point of view, we show

a unique data-driven approach to characterize, compare and classify self-assembled supramolecular polymers based on their structural/dynamic diversity and on their defects, setting the stage for a systematic investigation of the interplay between monomer chemistry, solvent environment, and the final properties of the supramolecular assembly.

Associated Content

Supporting Information. The Supporting Information is available free of charge at https:// . It contains additional Computational Details and five additional Figures.

Acknowledgement

GMP acknowledges the funding received by the Swiss National Science Foundation (SNSF grant number IZLIZ2_183336) and by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 818776 – DYNAPOL). MC acknowledges funding by the European Research Council under the European Union's Horizon 2020 research and innovation program (grant no. 677013-HBMAP).

References

(1) Maradudin, A. A.; Montroll, E. W.; Weiss, G. H.; Ipatova, I. P. *Theory of Lattice Dynamics in the Harmonic Approximation*, 2nd ed.; Solid State Physics; Academic Press: New York, 1971.

(2) Pertsinidis, A.; Ling, X. S. Diffusion of Point Defects in Two-Dimensional Colloidal Crystals. *Nature* **2001**, *413*, 147–150.

(3) Gao, P.; Nelson, C. T.; Jokisaari, J. R.; Baek, S.-H.; Bark, C. W.; Zhang, Y.; Wang, E.; Schlom, D. G.; Eom, C.-B.; Pan, X. Revealing the Role of Defects in Ferroelectric Switching with Atomic Resolution. *Nat. Commun.* **2011**, *2*, 591.

(4) Kim, J.; Lee, S.-H.; Lee, J. H.; Hong, K.-H. The Role of Intrinsic Defects in Methylammonium Lead Iodide Perovskite. *J. Phys. Chem. Lett.* **2014**, *5*, 1312–1317.

(5) Ding, J.; Patinet, S.; Falk, M. L.; Cheng, Y.; Ma, E. Soft Spots and their Structural Signature in a Metallic Glass. *Proc. Natl. Acad. Sci. U. S. A.* **2014**, *111*, 14052–14056.

(6) Kuhlmann-Wilsdorf, D.; Wilsdorf, H. G. F. Dislocation Movements in Metals. *Science* **1964**, *144*, 17–25.

(7) Mielke, S. L.; Troya, D.; Zhang, S.; Li, J.-L.; Xiao, S.; Car, R.; Ruoff, R. S.; Schatz, G. C.; Belytschko, T. The Role of Vacancy Defects and Holes in the Fracture of Carbon Nanotubes. *Chem. Phys. Lett.* **2004**, *390*, 413 – 420.

(8) Azizi, A.; Zou, X.; Ercius, P.; Zhang, Z.; Elías, A. L.; Perea-López, N.; Stone, G.; Terrones, M.; Yakobson, B. I.; Alem, N. Dislocation Motion and Grain Boundary Migration in Two-Dimensional Tungsten Disulphide. *Nat. Commun.* **2014**, *5*, 4867.

(9) Shklovskii, B. I.; Efros, A. L. *Electronic Properties of Doped Semiconductors*; Springer Science & Business Media, 2013.

(10) Yokoya, T.; Nakamura, T.; Matsushita, T.; Muro, T.; Takano, Y.; Nagao, M.; Takenouchi, T.; Kawarada, H.; Oguchi, T. Origin of the Metallic Properties of Heavily Boron-Doped Superconducting Diamond. *Nature* **2005**, *438*, 647.

(11) Haverkate, L. A.; Zbiri, M.; Johnson, M. R.; Deme, B.; Mulder, F. M.; Kearley, G. J. Conformation, Defects, and Dynamics of a Discotic Liquid Crystal and Their Influence on Charge Transport. *J. Phys. Chem. B* **2011**, *115*, 13809–13816.

(12) Cubuk, E. D.; Schoenholz, S. S.; Rieser, J. M.; Malone, B. D.; Rottler, J.; Durian, D. J.; Kaxiras, E.; Liu, A. J. Identifying Structural Flow Defects in Disordered Solids Using Machine-Learning Methods. *Phys. Rev. Lett.* **2015**, *114*, 108001.

(13) Brunsveld, L.; Folmer, B. J. B.; Meijer, E. W.; Sijbesma, R. P. Supramolecular Polymers. *Chem. Rev.* **2001**, *101*, 4071–4098.

(14) Yang, L.; Tan, X.; Wang, Z.; Zhang, X. Supramolecular Polymers: Historical Development, Preparation, Characterization, and Functions. *Chem. Rev.* **2015**, *115*, 7196–7239.

(15) Yan, X.; Wang, F.; Zheng, B.; Huang, F. Stimuli-Responsive Supramolecular Polymeric Materials. *Chem. Soc. Rev.* **2012**, *41*, 6042–6065.

(16) Aida, T.; Meijer, E. W.; Stupp, S. I. Functional Supramolecular Polymers. *Science* **2012**, *335*, 813–817.

(17) Davis, A. V.; Yeh, R. M.; Raymond, K. N. Supramolecular Assembly Dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 4793–4796.

(18) Lehn, J.-M. Dynamers: Dynamic Molecular and Supramolecular Polymers. *Progress in Polymer Science* **2005**, *30*, 814 – 831.

(19) Albertazzi, L.; Martinez-Veracoechea, F. J.; Leenders, C. M.; Voets, I. K.; Frenkel, D.; Meijer, E. Spatiotemporal Control and Superselectivity in Supramolecular Polymers Using Multivalency. *Proc. Natl. Acad. Sci. U. S. A.* **2013**, *110*, 12203–12208.

(20) Albertazzi, L.; van der Zwaag, D.; Leenders, C. M.; Fitzner, R.; van der Hofstad, R. W.; Meijer, E. Probing Exchange Pathways in One-Dimensional Aggregates with Super-Resolution Microscopy. *Science* **2014**, *344*, 491–495.

(21) Bochicchio, D.; Salvalaglio, M.; Pavan, G. M. Into the Dynamics of a Supramolecular Polymer at Submolecular Resolution. *Nat. Commun.* **2017**, *8*, 147.

(22) Jung, S. H.; Bochicchio, D.; Pavan, G. M.; Takeuchi, M.; Sugiyasu, K. A Block Supramolecular Polymer and Its Kinetically Enhanced Stability. *J. Am. Chem. Soc.* **2018**, *140*, 10570–10577.

(23) Torchi, A.; Bochicchio, D.; Pavan, G. M. How the Dynamics of a Supramolecular Polymer Determines Its Dynamic Adaptivity and Stimuli-Responsiveness: Structure–Dynamics–Property Relationships From Coarse-Grained Simulations. *J. Phys. Chem. B* **2018**, *122*, 4169–4178.

(24) Bochicchio, D.; Kwangmettatam, S.; Kudernac, T.; Pavan, G. M. How Defects Control the Out-of-Equilibrium Dissipative Evolution of a Supramolecular Tubule. *ACS Nano* **2019**, *13*, 4322–4334.

(25) Bochicchio, D.; Pavan, G. M. From Cooperative Self-Assembly to Water-Soluble Supramolecular Polymers Using Coarse-Grained Simulations. *ACS Nano* **2017**, *11*, 1000–1011.

(26) Bartók, A. P.; Kondor, R.; Csányi, G. On Representing Chemical Environments. *Phys. Rev. B* **2013**, *87*, 184115.

(27) De, S.; Bartók, A. A. P.; Csányi, G.; Ceriotti, M. Comparing Molecules and Solids across Structural and Alchemical Space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769.

(28) Willatt, M. J.; Musil, F.; Ceriotti, M. Atom-Density Representations for Machine Learning. *J. Chem. Phys.* **2019**, *150*, 154110.

(29) Gasparotto, P.; Ceriotti, M. Recognizing Molecular Patterns by Machine Learning: an Agnostic Structural Definition of the Hydrogen Bond. *J. Chem. Phys.* **2014**, *141*, 174110.

(30) Gasparotto, P.; Meißner, R. R. H.; Ceriotti, M. Recognizing Local and Global Structural Motifs at the Atomic Scale. *J. Chem. Theory Comput.* **2018**, *14*, 486–498.

(31) Marrink, S. J.; Risselada, H. J.; Yefimov, S.; Tieleman, D. P.; de Vries, A. H. The MARTINI Force Field: Coarse Grained Model for Biomolecular Simulations. *J. Phys. Chem. B* **2007**, *111*, 7812–7824.

(32) Abraham, M. J.; Murtola, T.; Schulz, R.; Páll, S.; Smith, J. C.; Hess, B.; Lindahl, E. GROMACS: High Performance Molecular Simulations through Multi-Level Parallelism from Laptops to Supercomputers. *SoftwareX* **2015**, *1-2*, 19 – 25.

(33) Bussi, G.; Donadio, D.; Parrinello, M. Canonical Sampling through Velocity Rescaling. *J. Chem. Phys.* **2007**, *126*, 014101.

(34) Ceriotti, M.; Tribello, G. A.; Parrinello, M. Demonstrating the Transferability and the Descriptive Power of Sketch-Map. *J. Chem. Theory Comput.* **2013**, *9*, 1521–1532.

(35) Paruzzo, F. M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L. Chemical Shifts in Molecular Solids by Machine Learning. *Nat. Commun.* **2018**, *9*, 4501.

(36) Helfrecht, B. A.; Gasparotto, P.; Giberti, F.; Ceriotti, M. Atomic Motif Recognition in (Bio)Polymers: Benchmarks From the Protein Data Bank. *Front. Mol. Biosci.* **2019**, *6*, 1–14.

(37) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn Res.* **2011**, *12*, 2825–2830.

(38) Ceriotti, M. Unsupervised Machine Learning in Atomistic Simulations, between Predictions and Understanding. *J. Chem. Phys.* **2019**, *150*, 150901.

(39) Lou, X.; Lafleur, R. P.; Leenders, C. M.; Schoenmakers, S. M.; Matsumoto, N. M.; Baker, M. B.; Van Dongen, J. L.; Palmans, A. R.; Meijer, E. Dynamic Diversity of Synthetic Supramolecular Polymers in Water as Revealed by Hydrogen/Deuterium Exchange. *Nat. Commun.* **2017**, *8*, 15420.

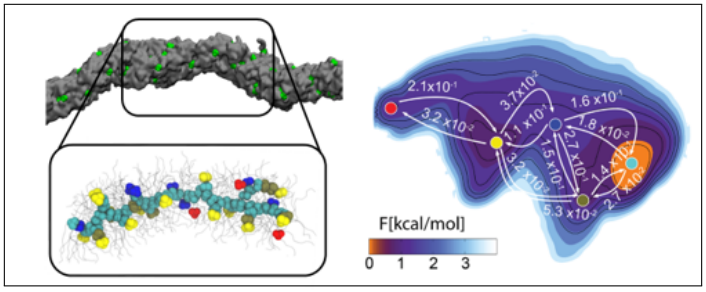
(40) Wang, J.; Olsson, S.; Wehmeyer, C.; Pérez, A.; Charron, N. E.; de Fabritiis, G.; Noè, F.; Clementi, C. Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Cent. Sci.* **2018**, 755–767.

(41) M. Sultan, M.; Pande, V. S. TICA-metadynamics: Accelerating Metadynamics by Using Kinetically Selected Collective Variables. *J. Chem. Theory Comput.* **2017**, *13*, 2440–2447.

(42) Sultan, M. M.; Pande, V. S. Automated Design of Collective Variables Using Supervised Machine Learning. *J. Chem. Phys.* **2018**, *149*, 094106.

(43) Mendels, D.; Piccini, G.; Parrinello, M. Collective Variables from Local Fluctuations. *J. Phys. Chem. Lett.* **2018**, *9*, 2776–2781.

Graphical TOC Entry



Identifying and Tracking Defects in Dynamic Supramolecular Polymers

Piero Gasparotto,[*,†,¶] Davide Bochicchio,[‡] Michele Ceriotti,[*,†] and Giovanni M. Pavan[*,‡,§]

†*Laboratory of Computational Science and Modeling, Institute des Materiaux, Ecole polytechnique fédérale de Lausanne, CH-1015 Lausanne, Switzerland*

‡*Department of Innovative Technologies, University of Applied Sciences and Arts of Southern Switzerland, Galleria 2, Via Cantonale 2c, CH-6928 Manno, Switzerland*

¶*Thomas Young Centre and Department of Physics and Astronomy, University College London, Gower Street London WC1E 6BT, United Kingdom*

§*Department of Applied Science and Technology, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy*

E-mail: p.gasparotto@ucl.ac.uk; michele.ceriotti@epfl.ch; giovanni.pavan@polito.it

Supporting Information Available

Probabilistic Analysis of Molecular Motifs

PAMM is a pattern recognition framework designed to identify molecular motifs based on their meta-stability. The core idea behind the PAMM scheme consists into partitioning the probability distribution function (PDF) of structures sampled from an atomistic simulation or collected from experiments. The input is simply a series of N (preferably high-dimensional) vectors representing local or global environments. This could be any function of the atomic coordinates, such as distances, angles or more sophisticated descriptors. Then one needs to perform a kernel density estimation (KDE) on a grid extracted using farthest point sampling (FPS), which has proven to be well suited for selecting the most widely spread set of landmarks from the initial set. Combing FPS with KDE allows for linear scaling with the number of samples and for the use of very large datasets, which is usually a bottleneck for most clustering algorithms.

The KDE on a grid point $\mathbf{y}_i$, given a set of $\{\mathbf{x}\}$ initial observations, can be written as

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

where w_i is simply a weight, $K_{\mathbf{H}_j}$ an anisotropic multivariate Gaussian kernel, and $\mathbf{H}_j$ is an adaptive bandwidth matrix. Two schemes have been suggested to optimally select the local kernel bandwidth. We here adopted the *fixed localization window* approach, scaling by a factor of 0.7 the automatically determined bandwidth. Having a reliable estimation of the probability density at each grid point, one can perform a density-based clustering to single out the different local maxima in the PDF and fit a multivariate Gaussian to each of the clusters obtained. In this way a GMM (Gaussian Mixture Modelling) can be used to represent the original PDF and define the so-called probabilistic motifs identifiers (PMIs), function of the coordinates giving a score between zero and one that represents the degree of confidence by which a new local structure can be assigned to each of the clusters.

Given a feature vector $\mathbf{x}$, the PMI corresponding to the kth cluster can be written as:

$$\hat{P}_k(\mathbf{x}) = p_k G(\mathbf{x}) / \left(\hat{P}(\mathbf{x})\right) \tag{2}$$

where $\hat{P}(\mathbf{x}) = \sum_{k=1}^{n} p_k G(\mathbf{x})$ is the GMM modelling the PDF, $G(\mathbf{x})$ is the multivariate Gaussian fitted to the kth cluster and p_k is its weight in the mixture. Finally, one can post-process any trajectory and use the predominant PMI to classify each monomer according to its structural state.

Smooth Overlap of Atomic Positions

SOAP is a general state-of-art atom-centered, density-based representation of the atomic environment which has been proven very powerful for both properties prediction and structural classification. A sum of Gaussians centered on each surrounding atom of species α produces a smooth representation of the atomic density around an atom j.

By defining a cutoff function f_c defining the extent of the local environments, the surrounding atomic density can be mapped into a local probability amplitude $\psi^{\alpha}_{\mathscr{X}_j}(\mathbf{r})$:

$$\left\langle \alpha\mathbf{r} \middle| \mathscr{X}_j \right\rangle \equiv \psi^{\alpha}_{\mathscr{X}_j}(\mathbf{r}) = \sum_{i\in\alpha} f_c(\mathbf{r}_{ij}) g(\mathbf{r}-\mathbf{r}_{ij}). \tag{3}$$

One can expand the local density in a basis of orthogonal radial basis functions $R_n(r)$ and spherical harmonics $Y^l_m(\hat{\mathbf{r}})$,

$$\left\langle \alpha nlm \middle| \mathscr{X}_j \right\rangle = \int \mathrm{d}\mathbf{r} R_n(r) Y^l_m(\hat{\mathbf{r}}) \left\langle \alpha\mathbf{r} \middle| \mathscr{X}_j \right\rangle. \tag{4}$$

This amplitude is invariant both to translations and permutations of atoms α, but it is not rotationally invariant. By integrating over all relative rotations $\hat{R}$ tensor products between this atom density evaluated at different points, one can obtain a representation of the atomic environment that is spherical invariant, corresponding to n-body correlations between the atoms. In particular, the second-order (three-body) invariant can be easily computed, on a radial basis, in terms of the

spherical decomposition 4, and corresponds to the SOAP power spectrum

$$\langle \mathscr{X}_j | \alpha n \alpha' n' l \rangle \propto \frac{1}{\sqrt{2l+1}} \sum_m (-1)^m \langle \alpha n l m | \mathscr{X}_j \rangle \langle \alpha' n' l - m | \mathscr{X}_j \rangle . \tag{5}$$

We use $\langle \alpha n \alpha' n' l | \mathscr{X}_k \rangle$ as an agnostic descriptor to fully characterize the local environment surrounding each monomer in the fibers.

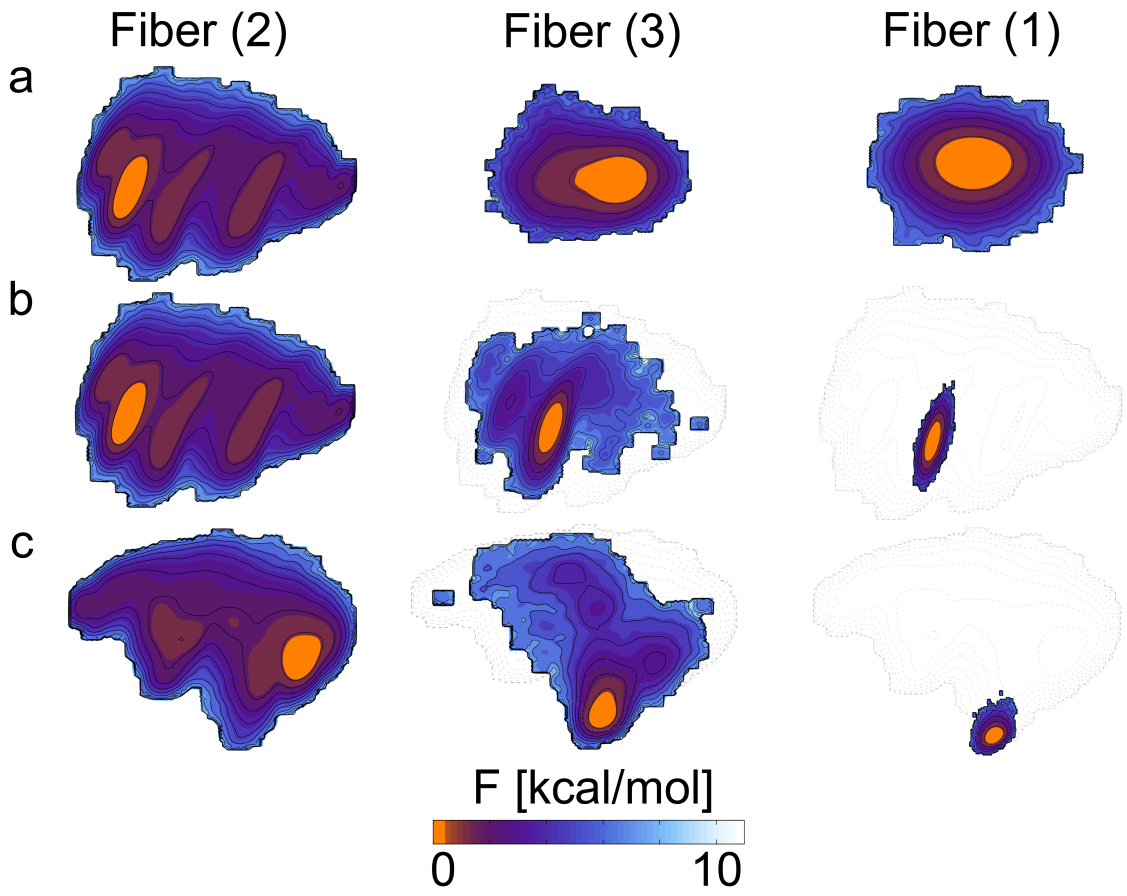


Figure S1: Low-dimensional FES for the different fibers using a PCA model (a) trained separately for the three fibers, (b) trained considering only fiber (2) (i.e. the system with the higher structural variability), and (c) trained considering the full dataset (no labels). Gray dashed lines refer to the contours of the corresponding Fiber (2)'s projection.

Low-dimensional embedding

PCA is useful to extract important features from a high-dimensional dataset and reduce its dimensionality. The algorithm seeks a linear combination of variables such that the maximum variance is extracted from the input features. It then removes this variance and seeks a second linear combination which explains the maximum proportion of the remaining variance, and so on. One can then use the first d principal components to embed a D-dimensional manifold into a d-dimensional one (with $d << D$). To be fully agnostic and as general as possible, we chose to discard any prior knowledge about different fibers and train our PCA on the full dataset containing the SOAP vectors computed from all the three systems. However, it is interesting to check how the PCA trained on a single fiber performs in embedding the high-dimensional features from other fibers. Figure S1, shows the low-dimensional projections for the three fibers training the PCA on the total dataset (Figure S1c), only on fiber (2) (Figure S1b) (which is the fiber presenting the largest structural variability) and on the single fibers separately. One can see how training the model in the system with the higher degree of disorder can be enough to learn how to embed different (possibly more ordered) systems not seen during the training stage.

PCA convergence

One can estimate the minimum number of principal components to be used to sparsify SOAP vectors by considering the amount of variance in the original variables accounted for by each component. Figure S2 shows the percentage of cumulative proportion of variance explained by an increasing amount of principal components. We chose to use 8 components since it is enough to recover more than 90% of the initial variance for the full data set (without distinguishing different fibers).

Choice of structural parameters

SOAP descriptors allow us to adopt a general structural description for local atomic environments avoiding any possible prior human bias (angles, distances, gyration radius or other educated guesses).

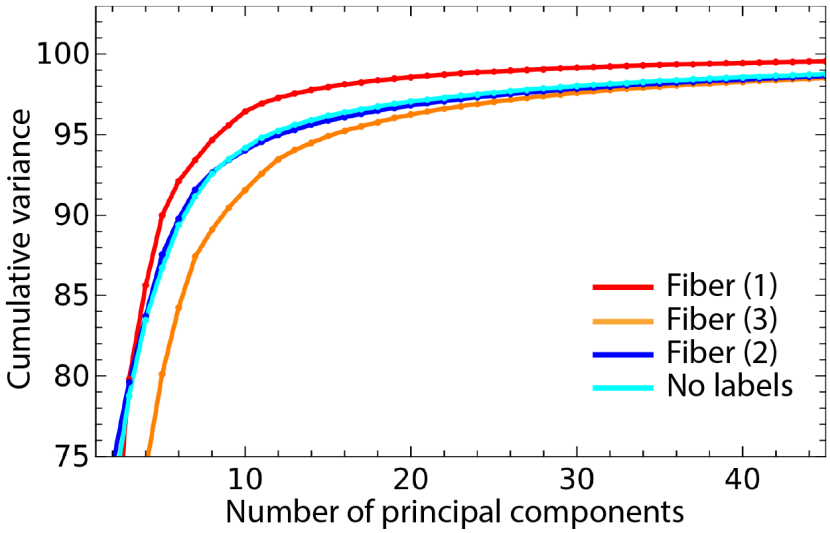


Figure S2: Cumulative variance as function of the number of principal components for the different fibers as well for the full dataset obtained removing the labels of the fibers.

However, few parameters must still be set in order to define the range of structural correlations captured by the SOAP vectors. In the specific, one has to guess a reasonable cut-off distance which has to be large enough to capture all the important physics, but also reasonably small in order to reduce at maximum the computational resources needed to compute and store the SOAP vectors. A good way to rapidly estimate the extent of local structural correlations is to use the Radial Distribution Function (RDF). To compute the RDF we considered only the center of mass of the aromatic ring of each monomer unit. Figure S3 shows that the first neighbour is typically at a distance lower than 6.0 Å.

We used the following parameters: `cutoff` = 8.0 Å, `cutoff_tr_width` = 0.5, `atom_sigma` = 1, `n_max` = 8, `l_max` = 8. All other parameters were set to the default values. Explanations of the parameters are given in the `quippy` library reference, available at `https://libatoms.github.io/QUIP/descriptors.html`.

We tested different cutoff radius in order to find the minimum distance capable of capturing the structural complexity of the fibers, as well as to distinguish between different systems. As before, in this analysis we considered only the center of mass of the aromatic rings for computing the SOAP vectors using different cut-off distances. We then computed the PCA on the full datasets

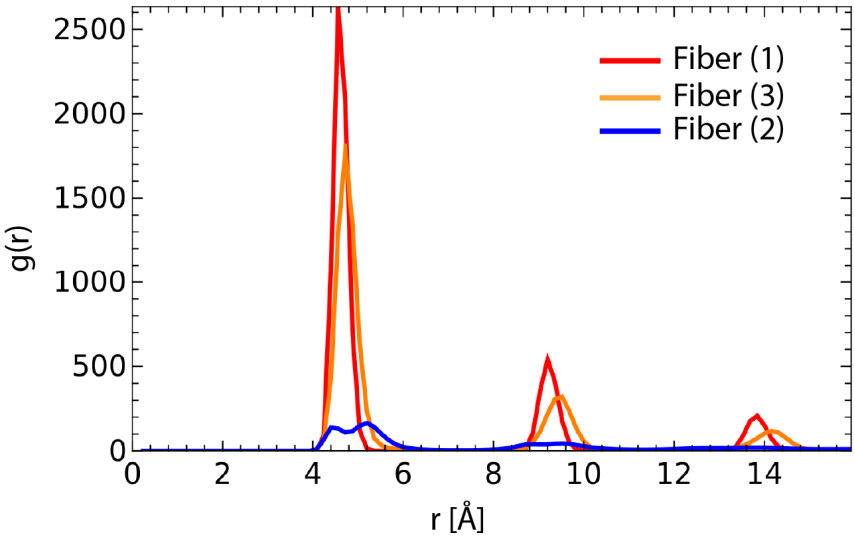


Figure S3: Comparing RDFs for fibers (2), (3) and (1). The RDF is computed considering only the center of mass of each aromatic ring in the monomer.

(removing the labels distinguishing among different fibers) and projected the SOAP features for the three systems onto different maps using the first two principal components. We also checked the robustness of our analysis by comparing the results obtained using cutoffs of 8Å, 12Å or 16Å. In Figure S4, one can see that using `cutoff` = 8.0 is enough to capture the different degree of complexity and population of the states in the three fibers.

Automatic learning of defects

This paper introduces a general method for automatic identification and classification of defects across different classes of supramolecular polymers. In our protocol, in order to make the analysis comparable between different fibers while keeping a detailed view on the stacking of the monomers, we considered the full aromatic ring plus the amide beads of each monomer. Here we show, however, that one could simplify even more the approach and consider only a virtual atom centered at the center of each monomer. Figure S5 shows that using only the information about the centers to build the SOAP vectors could be enough to capture qualitatively all the structural complexity between different fibers. To compute the SOAPs we used the parameters introduced before with `cutoff` = 8.0 Å, while the two-dimensional maps were produced projecting the SOAPs onto the first two

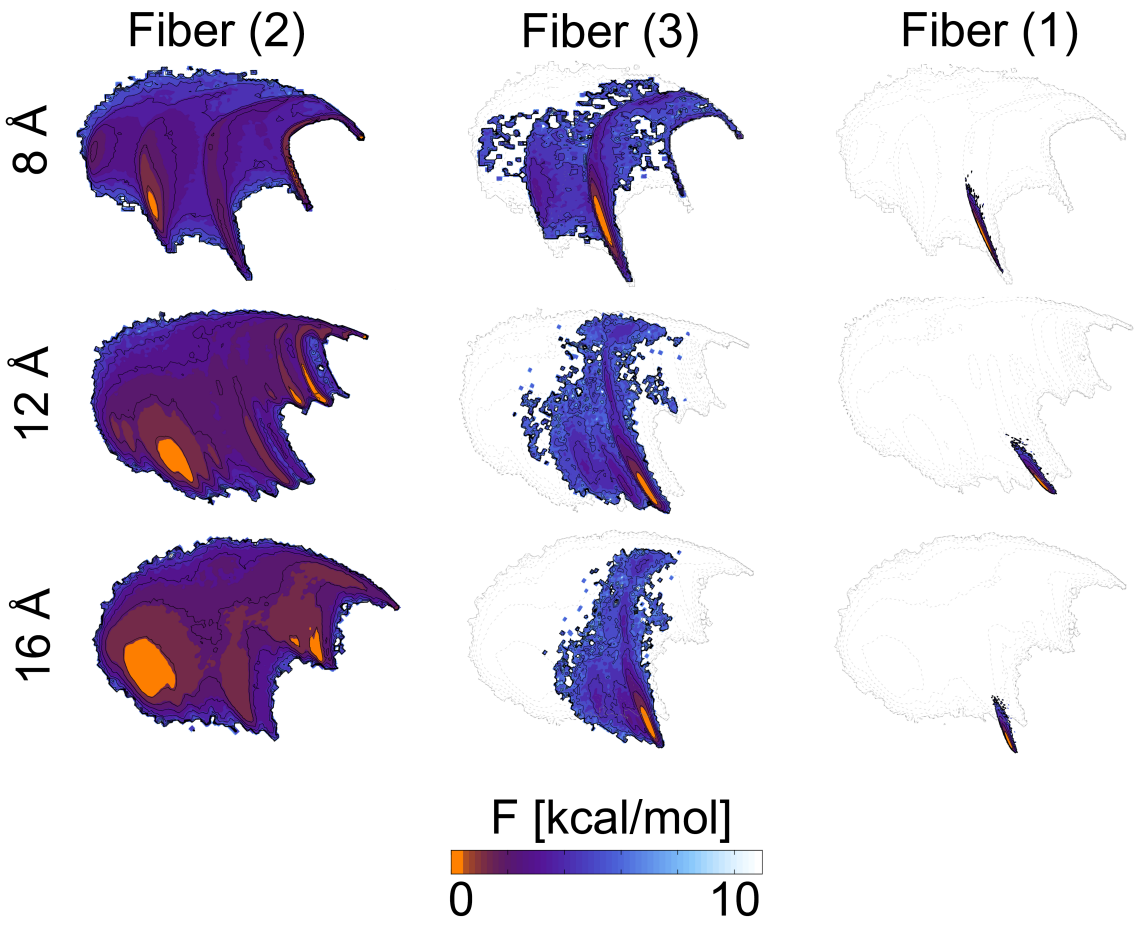


Figure S4: PCA projection of the SOAP vectors for the different systems using `cutoff` = 8.0, `cutoff` = 12.0, `cutoff` = 16.0. The SOAP vectors are computed considering only the center of mass of each aromatic ring in the monomer. In gray are the contours of Fiber (2), which is structurally richer than the other fibers and help the reader to compare the different systems.

principal components obtained from the PCA of the full dataset (without labels for different fibers). This analysis shows the power of SOAP when combined with PAMM for building a completely agnostic, data-driven approach to learn the structural complexity of polymers. Without having any prior knowledge on the systems one could start simply replacing each monomer with a point and study the correlations among those to capture most of the structural and the dynamical nature of the fiber.

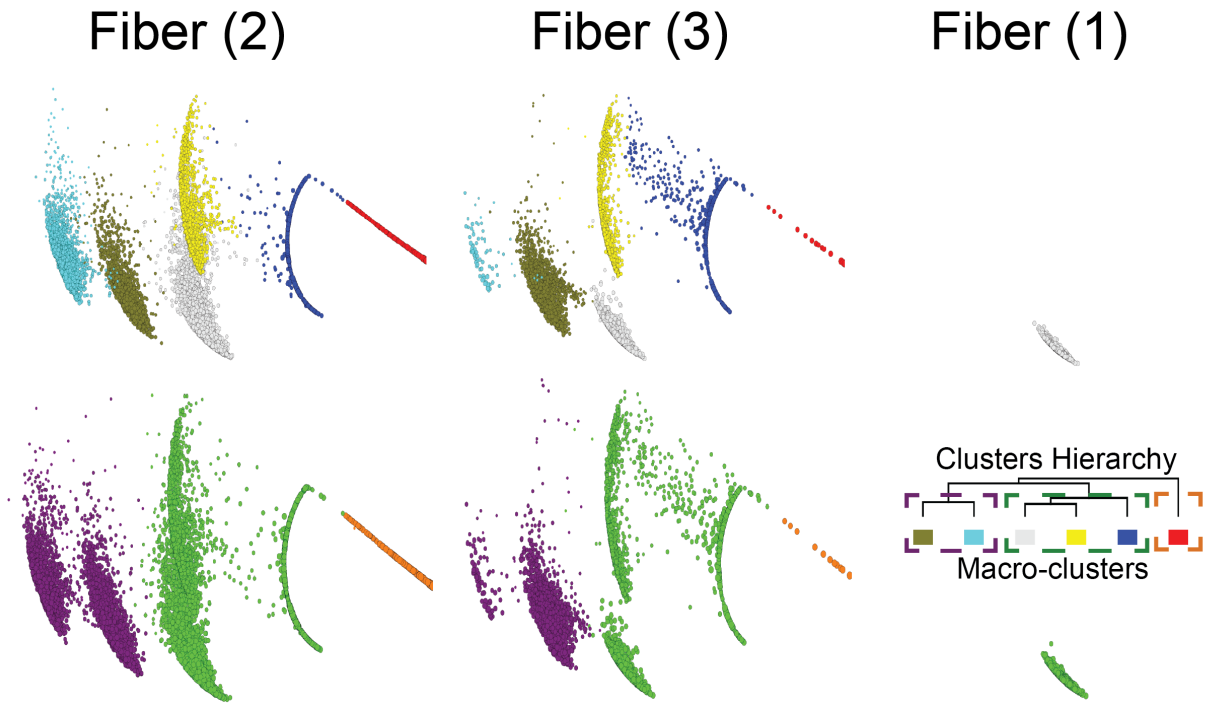


Figure S5: Projection of PAMM clustering of SOAP vectors computed using only the centers of the aromatic rings.