Scuola Superiore di Studi Avanzati - Trieste

Coarse-grained models for self-assembling systems

A thesis submitted for the degree of
Doctor of Philosophy

Candidate:
Guido Polles

Supervisor:
Cristian Micheletti

September 2015

Contents

Chapter 1

Introduction

In the last two decades, an unprecedented upsurge in the research on nano-technologies led to a flourishing of novel applicative avenues in material science and technology and this, in turn, has stimulated theoretical and computational advancements to characterize and control the underlying physical processes. The ability to shape the function of materials at the micro and nano-scale opens astonishing possibilities in a wide range of fields, from electronics to medicine. Not only processors are smaller and faster, also words like nano robots and nano cages made their appearance in the scientific community.

Indeed, together with so-called *top-down* methods as nano-lithography, where micro and nanometric structures are carved out of macroscopic blocks, a large deal of interest has been focused on *bottom-up* methods, where large or complex structures are assembled starting from smaller and simpler elementary blocks. Efficient production of materials using a bottom-up approach, however, faces the issues inherent to the challenging control of the process at such small scales.

In many cases, exceptional results can be obtained by the *unassisted* assembly of suitably engineered building blocks. A proper control of the physico-chemical properties of the construction units can indeed lead, through either a single or multiple steps, to their self-organization in functional target structures. This so-called *self-assembly* process can often be indirectly controlled by tuning the thermodynamic parameters of the system, partly compensating for the lack of direct supervision of the process.

As difficult as it can be to imagine to engineer Lego blocks which assemble in a spacecraft without even opening the box, self-assembly is ubiquitous in nature. Any single cell is, in a sense, a marvellous self-assembling, self-replicating machine, since its generation, growth and functioning is not directly assisted by external guidance.

While current designs of self-assembling processes are still far away from the results of

billion years of evolution, a considerable deal of work has so far been spent to understand and hence harness the physical principles that underpin the general properties of self-assembling systems, from regular crystals, to functionalized surfaces, to nano-metric sized objects.

In particular, theoretical and computational modelling have been extensively used to obtain a detailed description of the actual process. Indeed, on the extremely small scales considered, direct supervision and control often are both expensive and difficult, if not impossible, by experimental means. Computational modelling is, on one hand, an additional tool to describe and understand the properties of such systems. On the other hand, the ability of freely controlling parameters and details of the system provides the opportunity to propose innovative ideas for future physical realizations.

In this thesis we will report on computational work, focusing on two different self-assembling systems and from two distinct perspectives.

In the first part, I will present a computational study of the self-assembly of string-like rigid templates in solution; specifically, I will discuss to what extent one can direct the assembly of the templates into complex three-dimensional structures by suitably designing their shape.

The knotting and linking properties of curves embedded in a three-dimensional space have been a subject of study, both theoretically and, later, computationally, for quite some time. Furthermore, in recent years, chemists were able to synthesize an interesting repertoire of topologically entangled molecules, featuring both knots and interlocked rings[21]. Here, we address the practical problem of assembling a molecule or a colloidal particle with a target entangled topology from a general point of view, adopting purposely simple computational models. Chapter 2 will explore some geometrical properties of the building blocks which can robustly promote the formation of knotted constructs and the ability to target a precise knotted topology. In chapter 3 I will move to a more general type of entanglement, namely the spontaneous linking of ring-like structures. The statistical occurrence of such states is analysed upon varying certain key parameters, such as density, spatial confinement and the shape of the ring-like structures.

In the second part, I will focus on some of the smallest instances of molecular self-assembly in nature, that is viral capsids. The protein envelope of many viruses is indeed assembled inside the host cell from repeated protein units in solution which autonomously organize into a icosahedral shell. However, the dominant pathways of the assembly process is often not known or debated. In most cases, there are no clear indications of whether the assembly follows from a nucleation process, or a hierarchical assembly featuring various intermediate steps. As a step forward clarifying these processes, I used a *top-down* approach: starting from the structure of a assembled capsid,

I explored the presence of a mechanical fingerprint of basic units which compose the full structure. Chapter 4 will next be devoted to the development of a physics-based algorithm to subdivide a capsid in quasi-rigid units, its application to various instances of viral shells, and the exploration of the correspondence between mechanical and functional units. Finally, in chapter 5, I will present a more general method which allows to establish the most significant subdivision in mechanical units of proteins and protein assemblies.

The material presented in chapters 2, 4 and 5 is largely based on the following published works:

- Polles, G., Marenduzzo, D., Orlandini, E. and Micheletti, C. (2015). **Self-assembling knots of controlled topology by designing the geometry of patchy templates.** *Nature Communications, 6, 6423. doi:10.1038/ncomms7423*;

- Polles, G., Indelicato, G., Potestio, R., Cermelli, P., Twarock, R. and Micheletti, C. (2013). **Mechanical and Assembly Units of Viral Capsids Identified via Quasi-Rigid Domain Decomposition.** *PLoS Computational Biology, 9(11).* November 2013 cover article of PLoS Comput. Biol. Selected by F1000, `http://f1000.com/prime/718178685?bd=1`

- Ponzoni, L., Polles, G., Carnevale, V. and Micheletti, C. (2015). **SPECTRUS: A Dimensionality Reduction Approach for Identifying Dynamical Domains in Protein Complexes from Limited Structural Datasets.** *Structure, 23(8), 15161525. doi:10.1016/j.str.2015.05.022.* Cover article for the August 2015 issue of STRUCTURE.

Chapter 3 is based on more recent work, with a manuscript in preparation.

Chapter 2

Self assembly of knotted structures

2.1 Introduction

Self assembling systems are raising huge interest in many fields of material science and nanotechnology. And for a very good reason: the use of self-assembly procedures can open the possibility to build new complex structures or materials with fine-tuned properties in a relatively unassisted way, often delivering high yields and limiting production costs. Self-assembly is also very intriguing from a fundamental point of view: the *spontaneous* emergence of complex structures in the correct thermodynamic conditions is at the same time both striking and fascinating.

Actually, self-assembly is a quite general expression. It has been used in various context and at different length scales, from biologically relevant processes as viral capsids formation, to the production of engineered materials as self assembled monolayers, nanowires or even biomaterials such as DNA origamis.

A substantial body of both experimental and theoretical work has been carried out in order to understand the key principles which can produce target structures or properties from the interaction of simpler building blocks. In this chapter, we address the question whether it is possible to spontaneously assemble a structure with a given topology by merely choosing the correct shape of the building blocks: can we tune geometrical parameters of linear monomers so that they assemble, for example, in a trefoil or a pentafoil knot?

In a material science perspective, knotted structures are interesting candidates for nanocages or nanocarriers, due to their closed and inherently three-dimensional structure. In fact, a particular topology can confer structural stability to an object built up

from simple string-like objects. Some fascinating "molecular knots" have indeed been synthesized[19, 65, 5, 6, 60, 61]. The repertoire of topologically non trivial molecular structures that have been succesfully synthesized is growing but nevertheless still limited. The problem of "topological self-assembly" is an interesting topic also from a fundamental point of view, since the incorporation of entanglement in a self assembly procedure appears to be highly non trivial.

In this chapter I will report on an extensive computational study that I carried out to simulate the assembly of rigid templates which are able to bind in a non-specific manner through small patchy sites. The broader aim is to explore which possible topologies were accessible and the effect of the geometry of the templates on the resulting assemblies.

As a matter of fact, topological self assembly has been previously addressed from at least two different perspectives. Coluzza et al, although in the field of protein folding rather than the one of direct self assembly, explored the possibility of targeting structures with nontrivial topology. Specifically, ref. [16] proposed a method to design a sequence of an heteropolymer to obtain knotted folds. In particular, starting from a viable knotted target structure, a sequence of elements was chosen from a pool of monomers with different interacting sites. The resulting polymer had a minimum of its free energy corresponding to the proposed knotted structure and was able fold with the desired topology.

Additional insight on emergent non trivial topologies in self assembling systems was also obtained by Miller and Wales in ref [50]. In this work, a fixed number of Lennard Jones particles interacting in a dipolar fashion were clustered and allowed to rearrange so to reach their energy minimum. Interestingly, in some of the energy-minimising configurations, the dipoles were aligned along a closed, knotted line. The occurrence of knots as energetically favourable configurations in this setup suggests that it is possible to design stable three-dimensional structures with a non trivial topology.

Here we approach the problem of topological self-assembly from a different, general perspective. Specifically, we focused on the direct assembling of rigid building blocks with a well-defined geometry in solution. In this way, we controlled the shape of the monomers rather than the interactions between them, in order to explore impact of the geometric parameters of building blocks on the topology of assemblies.

2.2 A brief overview on knots

In everyday life we experience knots, for example, when we pick earplugs out of our pocket or Christmas light out of their box. Those physical knots on a long open rope are, however, topologically equivalent to an unknotted, straight rope. In fact, it is easy

to imagine to continuosly translocate any knot along a frictionless rope, until it reaches one end and unties.

An intuitive way to define a knotted curve based on common experience is to think to grab a rope, tie a knot in the middle, and then glue the ends together. In this way, no matter how much we play around with the rope, it will not disentangle unless we cut it open. Also, all the configurations resulting from manipulating or deforming the rope preserve the initial knot type. It comes natural, therefore, to define a knot as the class of equivalence of configurations obtained by these manipulations in the three-dimensional embedding space.

One important point to make is that we can *compose* knots: imagine to tie *two* knots on a rope, instead just of one, and then glue the ends. The resulting rope still represents a single knotted curve, but the knot type is different and more complicated than each of the two single knots we tied. As a matter of fact, it can be also demonstrated that an "anti-knot" does not exists, i.e. you cannot tie two knots that will result in a disentangled closed rope. Indicating the single knots with τ and τ', the resulting knot is called *composite knot* of τ and τ' and it is indicated as $\tau\#\tau'$.

The concept of knot composition is important because it naturally leads to the definition of *prime knots*, i.e. knots that cannot be obtained as the composition of simpler knots.

The classification of prime knots is usually denoted by the minimal *number of crossings* on a bidimensional projection, and an ordinal number to label inequivalent knots with the same number of crossings. In Fig. 2.1 a table of the prime knots up to 8 crossings is shown.

Intuitively, the complexity of knots increases together with the number of crossings. Moreover, the number of inequivalent classes at fixed number of crossings rises very rapidly. There is one knot class counting four crossings, two counting five, three counting six, seven counting seven, twenty-one counting eight, etc.

All unknotted curves are topologically equivalent to the circumference which, of course, counts no crossings, and it is hence denoted as 0_1. The simplest knot type, the so-called *trefoil knot*, has three crossings in its simplest projection, and it is denoted by 3_1. Also, there is only one inequivalent class of knots counting 3 crossings, disregarding chirality. Note, however, that most knots are chiral, meaning that you cannot superimpose them to their mirror image. The achiral knots are quite rare: only 7 out of 35 knots with up to 8 crossings fall in this family.

Knots can be grouped in families also based on other general properties. An important family is the *torus knots* one: some knots can be represented as a single closed curve which does not self-intersect on the surface of a torus (figure 2.2a). 3_1, 5_1, 7_1 are

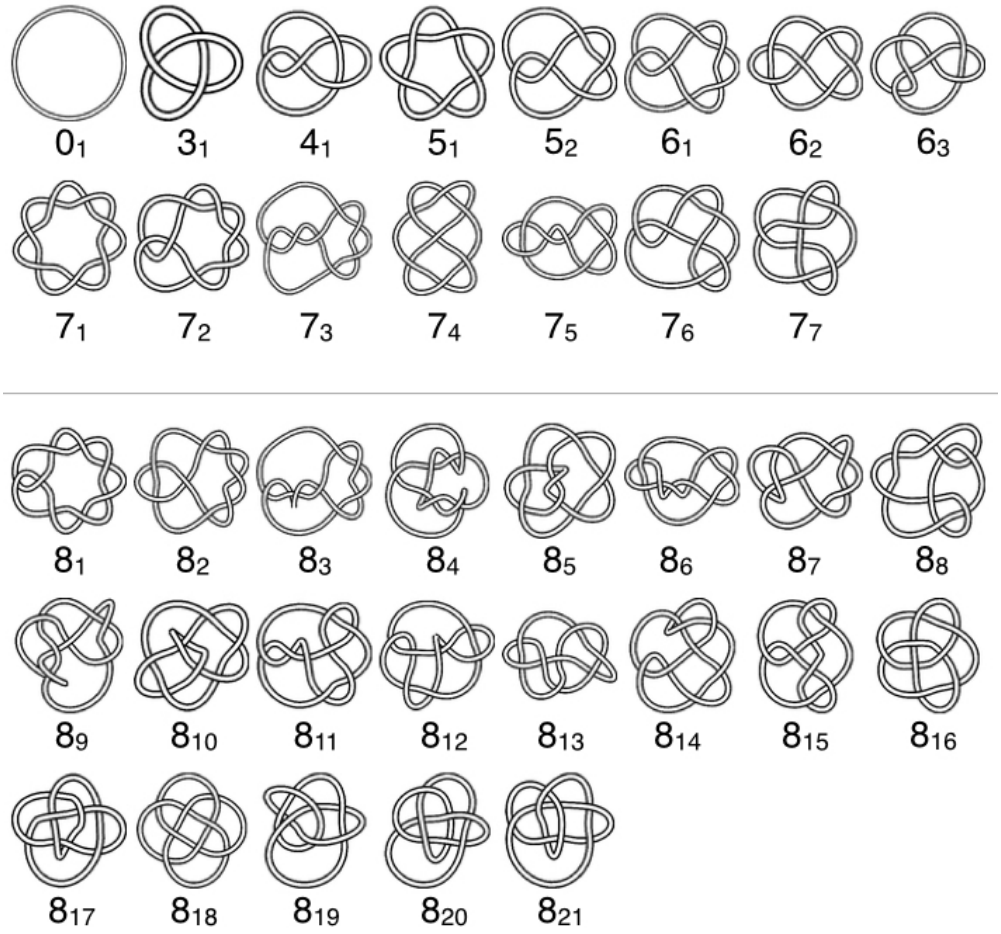

Figure 2.1: Table of prime knots up to 8 crossings.

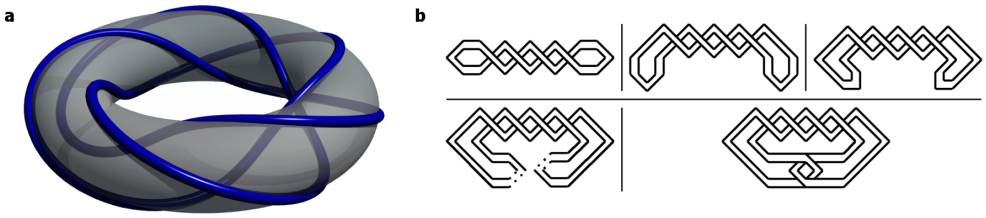

*Figure 2.2: Torus and twist knots. The **a** panel shows an example of 10_{124} torus knot depicted on the surface of a torus (image from Paul Aspinwall, Duke University). In the **b** panel, the creation of a four half-twist stevedore knot.*

the simplest torus knots.

Another family is the *twist knots*. They can be tied by inserting a certain number of half-twists in a unknotted curve and then clasping the ends together (see figure 2.2b). In fact, a twist knot can be untied by inverting just a single crossing, the one at the clasp. 4_1, 5_2 are the simplest twist knots.

Identifying a knot from one of its geometrical representations can be challenging. To this purpose, a number of polynomials have been proposed which are invariant across all the possible knot representations and can be used to identify the type of knot using tables. Here, we mostly used the Alexander determinant[2]. Its calculation from a *knot diagram* (i.e. a bidimensional projection of the curve, which keeps track of over- and under-crossings) has a simple algorithmic formulation and is sufficient to pinpoint knots up to 10 crossings.

2.3 The geometry of building blocks

The aim of generating closed, knotted structures using a self-assembling procedure poses the issue of a suitable choice of the basic building blocks. We started from solutions of a single species of rigid templates.

The basic shape of the templates was chosen to be helical. Their curved and non-planar nature is naturally compatible with the properties of closed, three dimensional structures. Moreover, helical fragments are easily described in terms of only two geometrical parameters: the projected opening angle α and the vertical span h, as described in Figure 2.3. Its Cartesian parametrisation is thus:

$$\begin{cases} x(t) = cos(\alpha t) \\ y(t) = sin(\alpha t) \\ z(t) = ht \end{cases}$$

with $t \in [0, 1]$.

The connection of these parameters with the pitch p and contour length l_c is:

$$\begin{cases} p = \frac{2\pi h}{\alpha} \\ l_c = \sqrt{\alpha^2 + h^2} \end{cases}$$

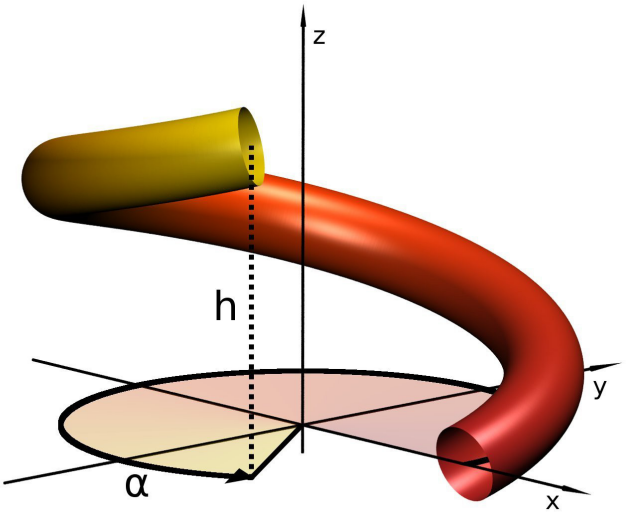

Figure 2.3: Helical parameters α and h describing the shape of templates.

2.4 The assembly model

2.4.1 Building blocks and pairwise interactions

We considered a simple and quite general model of colloidal rigid templates with the ability to attach by the means of two "sticky" patches. The excluded volume interactions of the building blocks in our simulations are obtained by lining "hard" spheres of nominal diameter σ along the helical fragment. The hard spheres of distinct fragments interact mutually with a short-ranged repulsive Weeks-Chandler-Andersen potential (i.e. truncated and shifted 12-6 Lennard-Jones potential):

$$\begin{cases} U_{\mathrm{ij}}^{\mathrm{hs}} = 4\kappa_{\mathrm{hs}}\epsilon\left[\left(\frac{\sigma}{d_{\mathrm{ij}}}\right)^{12} - \left(\frac{\sigma}{d_{\mathrm{ij}}}\right)^{6} + \frac{1}{4}\right] & \text{if } d_{\mathrm{ij}} < 2^{\frac{1}{6}}\sigma \\ 0 & \text{else.} \end{cases} \tag{2.1}$$

where d_{ij} is the distance between the i-th and j-th beads, ϵ is equal to the system thermal energy, $k_{\mathrm{B}}T$, and κ_{hs} is an adimensional parameter quantifying the strength of the interaction. The actual values of the parameters we chose will be discussed later.

The sphere diameter is set to 1/3 of the radius of the helix projection. In order to allow the fragments to self-assemble into higher order structures the building blocks are decorated with two patchy attractive sites on the ends. The attractive interaction has a Gaussian form:

$$U_{\mathrm{ij}}^{\mathrm{patchy}} = -\kappa_{\mathrm{patchy}}\epsilon \exp\left[-\frac{r_{\mathrm{ij}}^2}{2\sigma_{\mathrm{patchy}}^2}\right] \tag{2.2}$$

where κ_{patchy} and σ_{patchy} measure respectively the magnitude and range of the attractive

interaction.

The attractive interaction is kept intentionally simple in order to not impose additional complexity on top of the geometry of the templates.

System setup

In each simulation, 250 copies of the fragment are randomly placed inside a cubic simulation box of volume l_{b}^3 with full periodic boundary conditions. The desired total density of spheres in the simulation, ρ, is fixed by setting $l_{\mathrm{b}} = (n_{\mathrm{hs}}/\rho)^{1/3}$, where n_{hs} is the number of hard spheres in the simulation box and ρ is set to $7.5 \cdot 10^{-3}\sigma^{-3}$, corresponding to a volume fraction of $\sim 0.56\%$.

Because the spherical particles are not individually dispersed in solution but are grouped to line the helical fragments, this volume fraction corresponds to an appreciable crowding of the templates. In fact, since each template typically consists of 15 beads, one has that its specific volume is about 2,000 σ^3. The associated characteristic lengthscale, which measures the separation of neighbouring templates, is therefore $\sim 13\sigma$, which is comparable to the templates' typical size (gyration diameter) which is $\sim 7\sigma$.

The dynamics of the i-th hard-core bead, whose position vector we call $\mathbf{r}_{\mathrm{i}}$, is described by a Langevin equation

$$m\ddot{\mathbf{r}}_{\mathrm{i}} + \gamma\dot{\mathbf{r}}_{\mathrm{i}} + \bigtriangledown_{\mathrm{i}}\left(\sum_{\mathrm{j}\neq\mathrm{i}} U_{\mathrm{ij}}^{\mathrm{hs}}\right) + \boldsymbol{\eta}(t) \tag{2.3}$$

where m and γ are the mass and friction of the $i-$th hard core bead, the index j runs over the other hard core beads, and $\bigtriangledown_{\mathrm{i}}$ is the gradient operator with respect to the i^{th} bead coordinates.

The noise η is uncorrelated across the various beads and, for each of them, satisfies the usual fluctuation-dissipation conditions, $\langle\eta_\alpha(t)\rangle = 0$ and $\langle\eta_\alpha(t)\eta_\beta(t')\rangle = 2k_{\mathrm{B}}T\gamma\delta_{\alpha,\beta}\delta_{\mathrm{t,t'}}$, with α and β running over the three Cartesian components.

Similarly, the evolution of the k-th attractive patch is described by

$$m\ddot{\mathbf{r}}_{\mathrm{k}} + \gamma\dot{\mathbf{r}}_{\mathrm{k}} + \bigtriangledown_{\mathrm{k}}\left(\sum_{\mathrm{l}\neq\mathrm{k}} U_{\mathrm{kl}}^{\mathrm{patchy}}\right) + \boldsymbol{\eta}(t) \tag{2.4}$$

with the index l running over the other patchy centres. Notice that there is no interaction between the hard core beads and the patchy centres, so that their movement is coupled only by the rigid constraint.

The Langevin equations of motion for all beads are integrated numerically with the LAMMPS simulation package [58] with the rigid-body constraint applied to each

construct. The integration time step is $\Delta t = 0.012\tau_{\rm LJ}$, where $\tau_{\rm LJ} = \sigma\sqrt{m/\epsilon}$ and with $m/\gamma = 2\tau_{\rm LJ}$ as in Ref. [35].

With these parameters, the templates' typical Brownian time, that is the time required by an isolated construct to diffuse over a distance equal to its radius of gyration, is $\tau_B \sim 40\tau_{LJ}$, see figure 2.4.

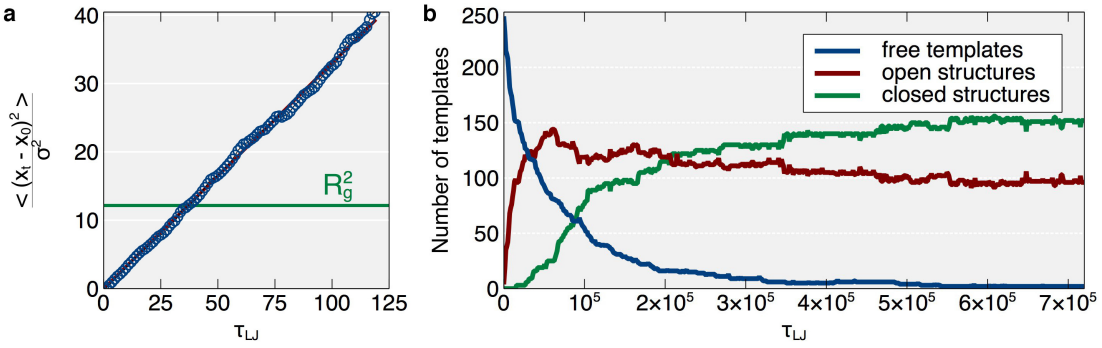

Figure 2.4: Free dynamics and assembly kinetics of helical templates. (a) Time dependence of the mean square displacement of the centre of mass of a single template of geometry $\alpha = 1.7\pi$ and $h = 1.0$. It is seen that the time required to diffuse over a distance equal to its radius of gyration, i.e. the characteristic Brownian time, is about equal to 40 $\tau_{\rm LJ}$. (b) Typical time evolution of the number of templates that are unbound (blue curve) or bound in open (red) or closed (green) oligomers. The time evolution pertains to a single Langevin trajectory where 250 templates, of geometry ($\alpha = 1.7\pi, h = 1.0$) are initially randomly arranged in the simulation box (hence they are all unbound). Stationarity ensues after about $2.5 \cdot 10^5 \tau_{\rm LJ}$.

2.4.2 Interaction parameters

The range for the parameters of the potential is mainly limited by two physical requirements.

On one hand, the bonds between fragments should be sufficiently long-lived in order to observe the production of stable structures. Thus, we wish the typical unbinding time to be much larger than both the microscopic timescale $\tau_{\rm LJ}$ and the typical Brownian time $\tau_{\rm B}$. We chose $\kappa_{\rm patchy} = 25$ in order to obtain a sufficiently long lifetime of the bound state.

On the other hand, an isotropic attractive potential poses the problem of the possible formation of "bundles", in which multiple fragments attach to a single site.

In order to avoid the occurrence of such bundles, we both made the attractive interaction very short-ranged (by lowering $\sigma_{\rm patchy}$) and hardened the Weeks-Chandler-Andersen spheres (by selecting a large value of $\kappa_{\rm hs}$). In this way, the attractive patches are very small compared to the hard spheres they lie on; therefore two bound templates do not leave enough room for additional templates to come in contact with the patchy

zone.

In order to schematize this situation, imagine the patchy ends of three templates coming in close contact. Figure 2.5a,b shows the closest possible contact of three infinitely hard spheres (depicted in grey). In the two panels, two distinct positioning of the patchy centres are shown, with a red dot indicating each of the patchy centre position. In the configuration in panel (a), two patches come in close contact and the third one is substantially non-interacting. In panel (b), all patches are similarly close to each other, forming a "triple contact". To avoid bundles, we thus require the (a) configuration to be more favourable than the (b) one.

In panel (c), the energy difference between the (a) and (b) configurations is shown in the case of infinitely hard spheres and $\kappa_{\text{patchy}} = 1$. Multiple contacts are strongly disfavoured and bundles are substantially suppressed as long σ_{patchy} is not much larger than 0.1.

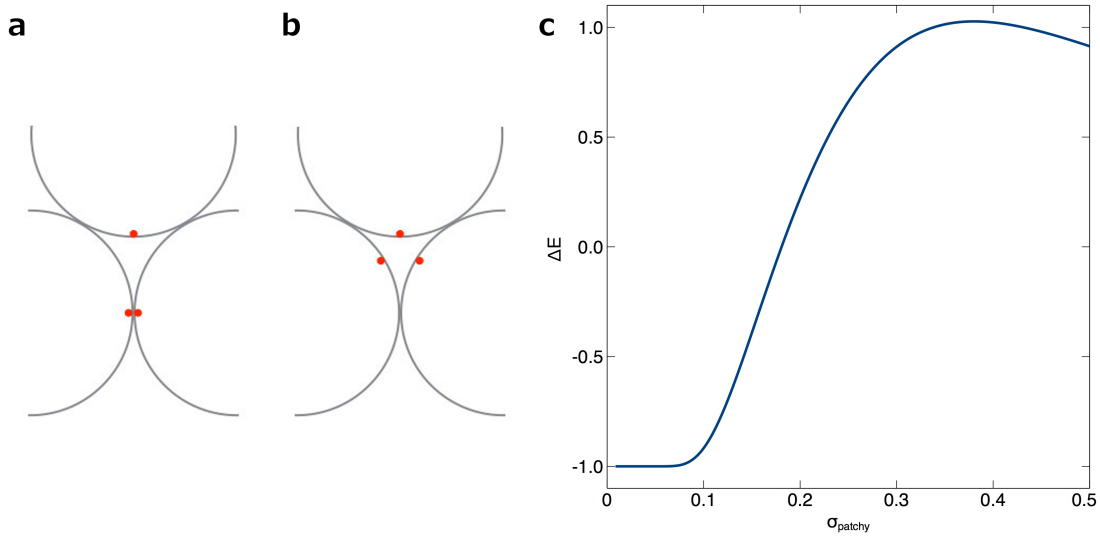

Figure 2.5: Energy difference between triple and double contacts. Panel (a) and (b) show in red the position of the interaction centres for a double and triple contact, respectively. Panel (c) shows the energy difference profile (between the two configurations) as a function of σ_{patchy}.

With these considerations in mind, the values of the parameters of both the patchy potential ($\sigma_{\text{patchy}} = 0.1\sigma$) and for the hard-core interactions ($\kappa_{\text{hs}} = 150$) were thus chosen after several tests in order to avoid the occurrence of multiple contacts at one site and therefore the formation of bundles.

Note that this kind of interaction forces the centrelines of two bound templates to be approximately aligned. Indeed, the rotation of the templates required for the patchy sites to come into contact will align the tangent vectors at the helices ends. Large enough deviations from the aligned conformations cause indeed the detachment of the

patches.

Therefore, two bound templates are kept aligned by the patchy interaction. The effect is more marked for smaller patches and stronger attractive interactions.

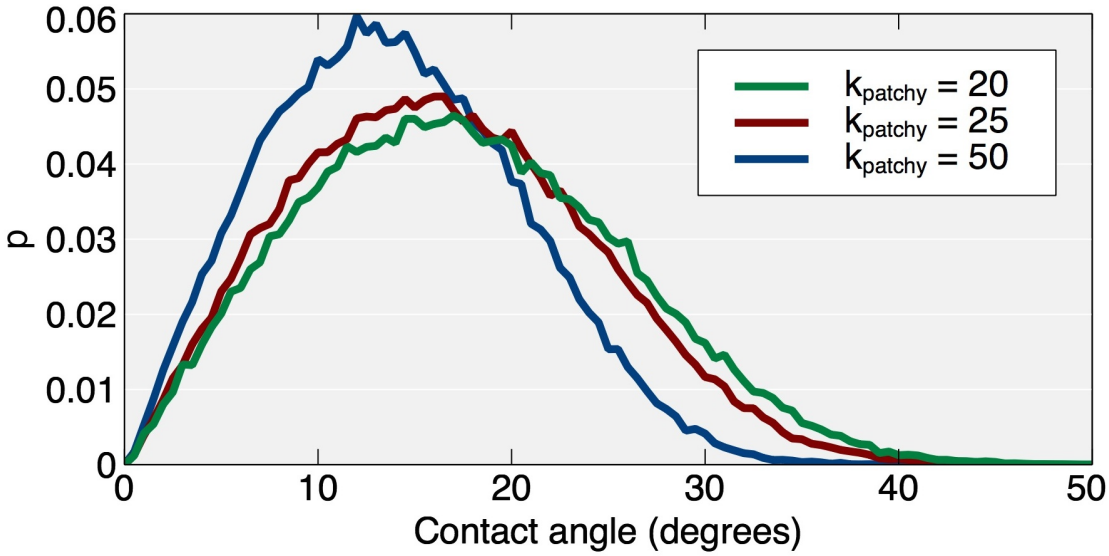

Figure 2.6: Probability distribution of the contact angle formed by two templates at equilibrium for various viable strengths of the attractive potential. The distribution was obtained from simulations of the Langevin dynamics of a system counting only two bound templates (with geometric parameters ($\alpha = 1.7\pi, h = 1.0$)). The typical contacting angle found in the dominant knot type (3_1, $n_t = 3$) is $\sim 20°$. This clearly falls near the maximum of the angle distribution at all potential amplitudes, which suggests that the geometric properties of assemblies and their statistical occurrence should be mostly unaffected by appreciable variations of the binding strength between templates.

Figure 2.6 shows, however, that the distribution of contact angles is not dramatically modified by the strength of the patchy potential. Moreover, the bonding can actually be established for a wide range of contact angles, ensuring some flexibility on the bound configurations.

2.4.3 The parameter space

In order to probe the effects of the templates shape, we performed several simulations varying the templates geometrical parameters.

As previously mentioned, a helical fragment can be described in terms of only two parameters, the projected angle α and its axial rise h (Figure 2.3). We limited the variation of the shapes in the (α, h) parameter space where $1.2\pi \leq \alpha \leq 2.0\pi$, $0.0 \leq h \leq 2.0$.

The exploration of the parameter space was performed by discretizing it in a triangular grid with α and h spacings respectively equal to 0.1π and 0.1 (for a total of 179

points). For each point in the parameter space, we ran 20 assembly simulations using monodisperse solutions of building blocks of the corresponding (α, h) geometry.

2.5 Results

2.5.1 Closure and knotting probability

The first important results regards the ability of templates to assemble in closed structures. We define the *closure probability* as the average fraction of templates which are part of a closed structure. Its dependence on the template shape is conveyed by figure 2.7a, where the closure probability is shown as a function of the geometric parameters α and h. In the figure, darker shades are associated to larger values.

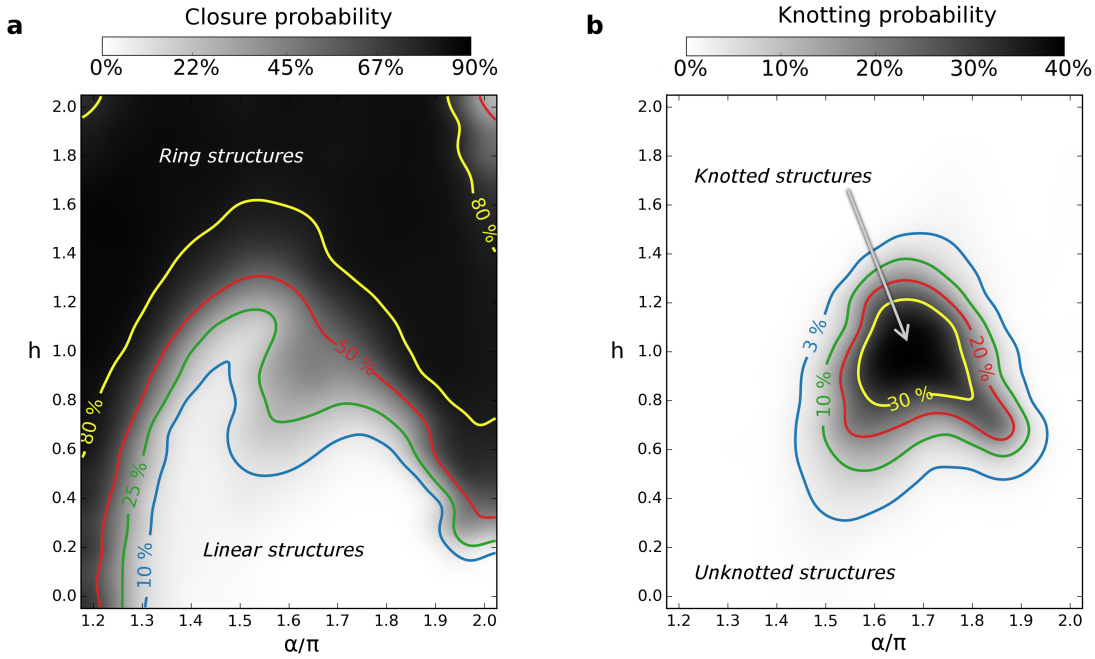

Figure 2.7: Closure and knotting probabilities of the assembled constructs as a function of the templates shape parameters, α and h. The quantities shown in panels (a) and (b) are the closure and knotting probabilities of the self-assembled constructs. These probabilities are more rigorously defined in our framework as the fraction of templates involved in closed (a) and knotted (b) constructs, respectively. The phase diagram is drawn by interpolating data obtained by sampling the two-dimensional parameter space with a triangular grid with α/π and h spacings both equal to 0.1. h is measured in units of the projected helical radius.

One immediately notices that the closure probability is strongly affected by the template geometry. It is easy to imagine that planar templates ($h = 0$) are hardly "fit" to form a closed ring unless they are sufficiently similar to a semicircle ($\alpha = \pi$). In the same way, there are geometrical constraints which allow or disallow non planar templates to interconnect, for example in dimeric structures with the shape of a figure

8 (reminiscent of the slot car tracks popular in the 70's). Where their geometry allows templates to connect at both ends to form dimeric rings, the yield of production of closed structures can be as high as 80%.

Note that, for a diffusion-limited association (i.e. if the typical timescale between productive encounters is much longer than the exploration of relative positioning after a binding event), dimeric products are largely favoured with respect to trimeric or higher order ones. Intuitively, this can be understood by realising that, in low concentrations, the simultaneous encounter of three particles is an event which is way less probable than a normal collision between two particles. In fact, we expect to observe dimeric rings as the most abundant species in the low concentration regime, if a closed dimeric structure is entropically accessible.

The closure, however, is not sufficient to obtain *knotted* constructs; all the dimeric rings, in fact, have trivial 0_1 topology. Together with the constraints to the accessible configurations, there are additional *topological* constraints. Indeed, all ring-like structures may be characterised by their "total curvature", which is simply the integral of the local radius of curvature over the ring length. The self-assembly of closed loops with a nontrivial knot topology, which are of particular relevance for our work, requires that the total curvature of the ring must be equal or exceed 4π, and this leads to further selection in parameter space [51] as it poses a lower bound to the number of templates required for assembly.

The plot in Figure 2.7b shows the *knotting probability*, defined as the average fraction of templates taking part in a knotted structure. Remarkably, one can observe the presence of a zone in the (α, h) space where the production of knotted constructs is definitely noticeable, falling on the boundary between open and closed structures. The high knotting probability in this region, peaking at about 50%, is even more striking giving the very stringent requirements in order to assemble with some efficiency even the simplest of the knots: the geometry of templates shall disfavour closure in dimeric constructs, while concurrently allowing for non-trivial trimeric (or multimeric) closure.

2.5.2 Observed knot repertoire and abundance

We classified the knotted structures obtained during the simulations using the Alexander polynomial. Most of the produced structures where identified as trefoil knots, but the repertoire of accessible topologies turned out to be surprisingly rich, especially given the simplicity of both the model and the geometry of fragments. In Figure 2.8 one can appreciate the most abundant non-trivial topological species (panel a) together with some selected rare occurrences of more complex knots (panel b). In this figure, also the

number of templates which compose each species is indicated as n_t.

Interestingly, the 3_1 knots mostly come in two "flavours". The first type is assembled from three templates and resembling the shape of the ideal knot. The second is assembled using four templates and has a definite, different shape.

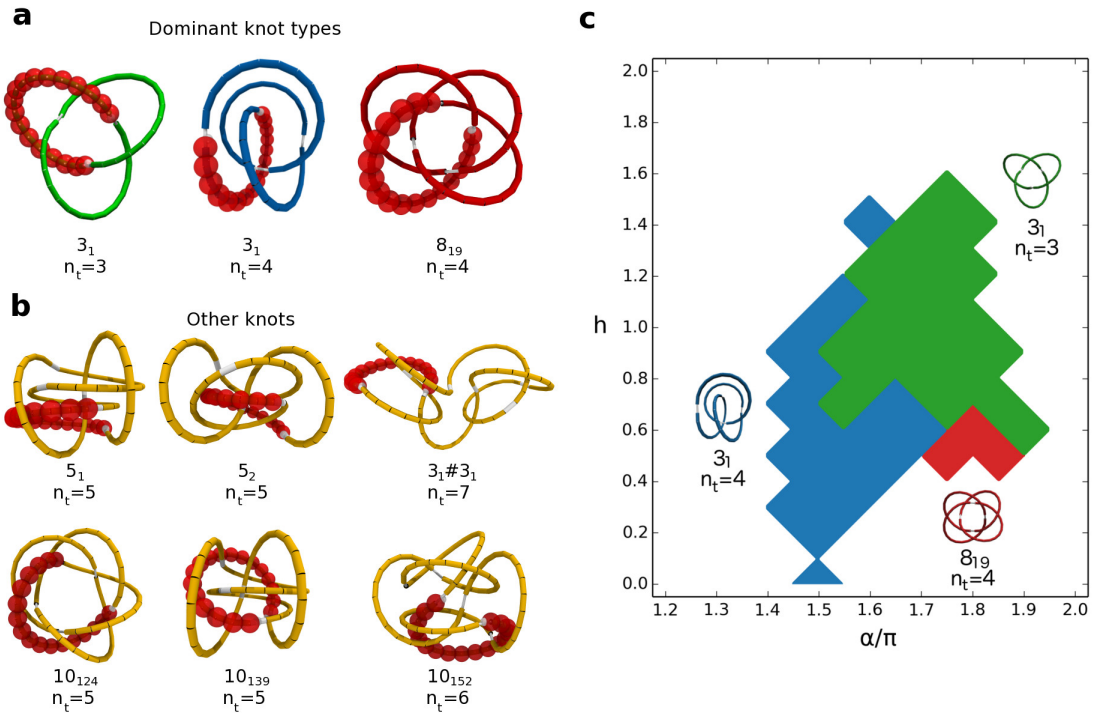

Figure 2.8: Self-assembled knotted structures and topological phase diagram. (a) Representative conformers of the dominant self-assembled knots and (b) of other, rarer knot types. For visual clarity, only one of the n_t constitutive templates is represented explicitly while for the others only the centreline is shown. The templates connecting regions are highlighted with white bonds. (c) Dependence of the dominant topology on the template geometry. In each coloured region the indicated knot is the dominant non-trivial topology one and has a probability of occurrence larger than 1%. Following the left-to-right order shown in panel (a), the peak probability of the tree dominant knot types is about 36%, 12% and 3%, respectively. h is measured in units of the projected helical radius.

Note that all the knotted structures obtained by our self-assembly procedure are chiral knots and, perhaps unsurprisingly, their chirality matches the one of the templates. In fact, the amphichiral Flemish knot 4_1 is not observed at all, although its complexity is relatively low with respect to other constructs. Also, most of the knots are actually torus knots, although some non-torus ones have been observed. The observation of twist knots such as 5_2 clarifies that the assembly of twist knots, although rare, is in fact a possibility.

To answer if template geometry can promote or suppress the products topologies, we analysed the relative abundance of the knotted constructs with respect to the control geometric parameters. In particular, in each (α, h) point we identified what we

called *dominant knot*, i.e. the most abundant non-trivial topology. In order to exclude ephemeral assemblies, we considered only topologies with probabilities (calculated as in the closure probability and knotting probability) over 1%.

In Figure 2.8c, the dominant knots are indicated in different colours in the (α, h) space. This topological phase diagram demonstrates that it is possible to robustly control the incidence of self-assembled constructs having definite topology as the knot type is statistically determined by the template geometry. The first important thing to notice is that, among the wide variety of structures that have been generated, the robust topologies that can be systematically promoted are only three, the trimeric 3_1, the tetrameric 3_1, and the tetrameric 8_{19}. By tuning the geometry of the templates is therefore possible to bolster, for example, tetrameric assemblies with respect to trimeric ones. This is strikingly demonstrated by the switch between the two forms of 3_1 knotted assemblies. Also, note that each type of construct is present in a well-defined geometry: for example, we show in Figure 2.9 the geometrical dispersion of trimeric 3_1 constructs by superposing several structures obtained at the end of one of the simulated trajectories. Here one can note that the geometrical variation inside the class is indeed very small.

Figure 2.9: Monodispersed geometry of knotted templates. The figure presents the structural superposition of 32 distinct trimeric constructs that, at the end of one of the simulated trajectories, are assembled with a trefoil knot topology. It is seen that the geometrical dispersion is very limited. For visual clarity the templates backbone is represented with cylinders rather than with spherical beads. The templates geometry corresponds to $\alpha = 1.7\pi$ and $h = 1.0$.

The most interesting structure, however, is the 8_{19} torus knot. In knot tables, the 8_{19} is marked as the first non-trivial torus knots because, if cut open, it presents three braided strands rather than two as the simpler 3_1, 5_1 and 7_1 torus knots. It is remarkable that such a complicated knot as 8_{19} has a much higher incidence than the 5_1 and 7_1 torus knots which could be expected to be frequently assembled from chiral templates given their nominal simplicity (and which, in fact, are entropically favoured

in fluctuating polymer chains [83, 67]). The appearance of the 8_{19} knot is even more interesting given that the same kind of topology was observed to emerge as an energy minimising configuration of dipolar particles in the work by Miller and Wales [50]. Its spontaneous emergence in two unrelated frameworks suggests that the 8_{19} topology, although being rather complicated (especially in the ideal representation in the table of Figure 2.1), features symmetric and geometrical properties which are optimal in a self-assembly perspective. In this view, this topology is a definitely viable candidate for expanding the current repertoire of "molecular knots".

One important point to make is that while the geometry of the templates promotes different kinds of knots, it is not completely selective, mostly because of the range of possible contact angles between templates. In many points of the (α, h) space we can find a mixture, albeit with different abundances, of different knot types (for example, in the point ($\alpha = 1.8\pi$, $h = 0.8$) a mixture of trimeric 3_1, tetrameric 3_1, 5_1 and 8_{19} is produced).

2.5.3 Mixed geometry and chirality

Another interesting question that naturally arises is whether is it possible to expand the repertoire of accessible structures by mixing different kinds of templates.

A first consideration is that the templates chiral nature poses the question of the effect of mixed chirality, especially in view of the observed absence of achiral structures. We therefore selected a productive geometry ($\alpha = 1.7\pi$, $h = 1.4$) and simulated a racemic mixture of left and right handed templates.

Surprisingly, in this case, the added complexity in the system doesn't enlarge the repertoire of topologies. Indeed, achiral structures remain absent and the resulting assembled knot types are the same obtained from the monodispersed solution. A significant decrease in the closure probability and production of knotted constructs is instead observed (Figure 2.10). This is an indication, at least for this geometry, that templates with mixed chirality do not assemble in newly formed structures and their net effect is instead to hinder the production of closed structures.

We then turned our attention to polydisperse solutions where templates of two different geometries are mixed. In particular, we selected two points in the (α, h) space, characterised by different dominant topologies and mixed them in various ratios, keeping the chirality fixed. A fraction x of the templates in the solution is of the ($\alpha = 1.6\pi$, $h = 0.6$) kind, a geometry which promotes 3_1 assemblies. The remaining fraction of the solute, $1 - x$, is composed by ($\alpha = 1.8\pi$, $h = 0.6$) templates, promoting tetrameric 8_{19} products.

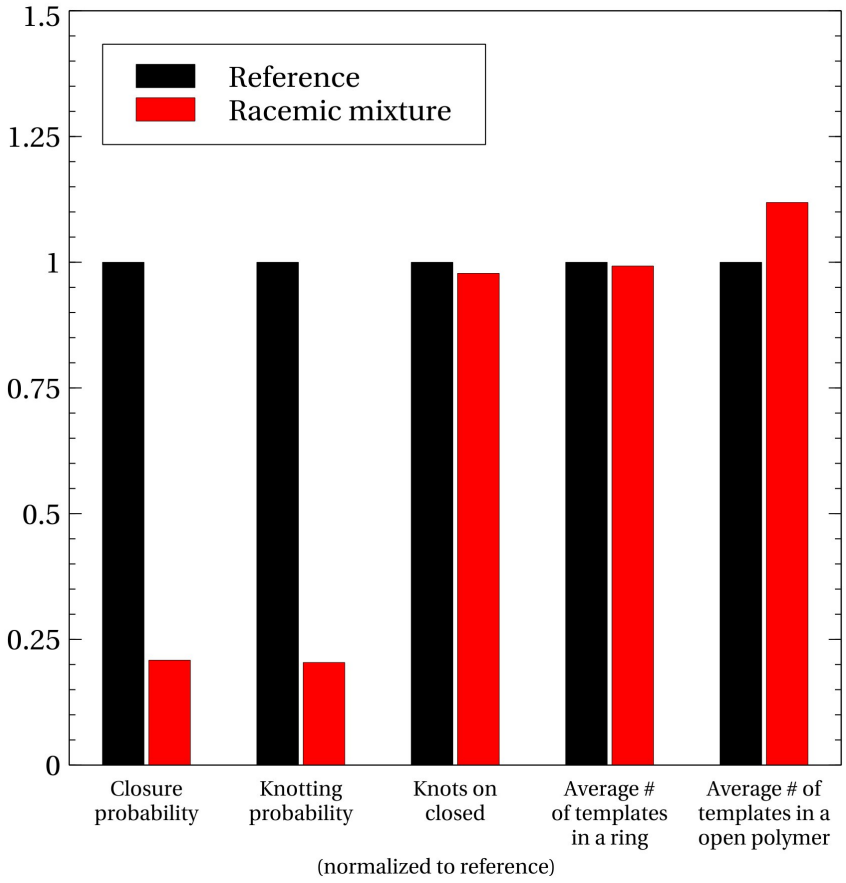

Figure 2.10: Properties of monodispersed versus racemic solutions. The figure illustrates how various properties of monodispersed systems change upon considering solutions where left- and right-handed templates are present in equal proportions. For a straightforward comparison, each data set is normalised to the monodispersed case (black bars, all normalised to unity). The geometry of the considered templates is specified by $\alpha = 1.7\pi$ *and* $h = 1.0$.

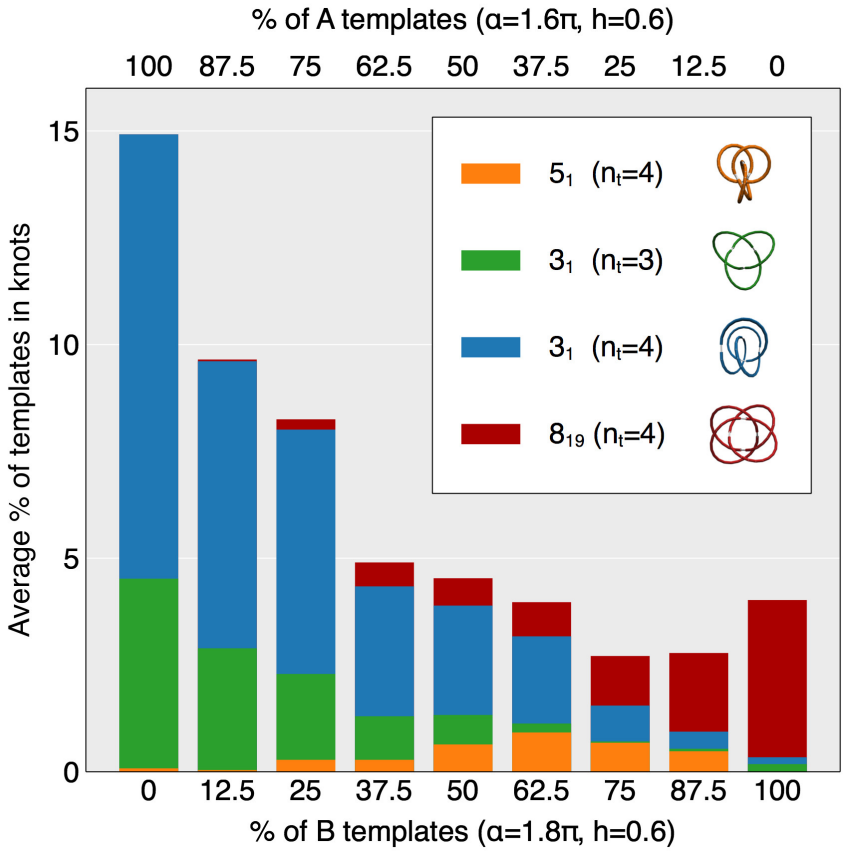

Figure 2.11: Average percentage of templates taking part to various non-trivial knot types as a function of the mixing ratio of two kind of templates.

Figure 2.11 shows the abundance of assembled knots upon the variation of the abundance x of the trefoil promoting templates with respect to the 8_{19} promoting ones. As one can see, along the gradual modification of the relative abundance of the topologies promoted in the monodispersed solutions, an additional 5_1 constructs appear. The occurrence of a new kind of topology is non monotonic with the components ratio in the mixture and is assembled from both kind of templates.

The complexity added by the mixing of different geometries can therefore enrich the repertoire of accessible topologies. The ratio between the components of the mixed solution can thus be considered as a further tunable parameter in promoting topologies not accessible or suppressed in monodispersed solutions.

2.5.4 Density robustness

Arguably, the concentration of the building blocks solution plays an important role in any kind of assembly. We thus explored the robustness of our results among the variation of the crowding in the simulation box.

As previously mentioned, the reference density of hard spheres in the simulation box was set to $7.5 \cdot 10^{-3}\sigma^{-3}$, corresponding to a volume fraction of $\sim 0.56\%$. However small, this corresponds to a appreciable crowding because the hard spheres are not freely moving in the solution, but are lined to form the helical templates. The typical size of the templates is of the same order of the typical distance between them.

We performed simulations on boxes where the density was set as one half, twice and ten times the reference one. The results are summarised in Figure 2.12. As one can immediately notice from the last column, there are no strong qualitative differences in the topological phase diagram across the variation of the density. This indicates that the produced topologies and geometries are robust and depend mostly on the geometry of the templates rather than on other thermodynamic characteristics.

At very high densities (last row of Figure 2.12), there is, however, a noticeable decrease in the closure probability. This is mainly due to the increased incidence of string-like, open structures. Indeed some of these structures percolate through the whole simulation box.

Two effects contribute to the drop in closure probability at high densities. On one hand, the timescale for the encounter and binding of templates is strongly reduced. On the other hand, excluded volume interactions hinder the exploration of the conformational space for dimeric and trimeric structures, strongly increasing its typical timescale. Therefore, intuitively, unless the templates are in the "correct" position, they will hardly rearrange to a closed conformation before another template from the bulk binds to one

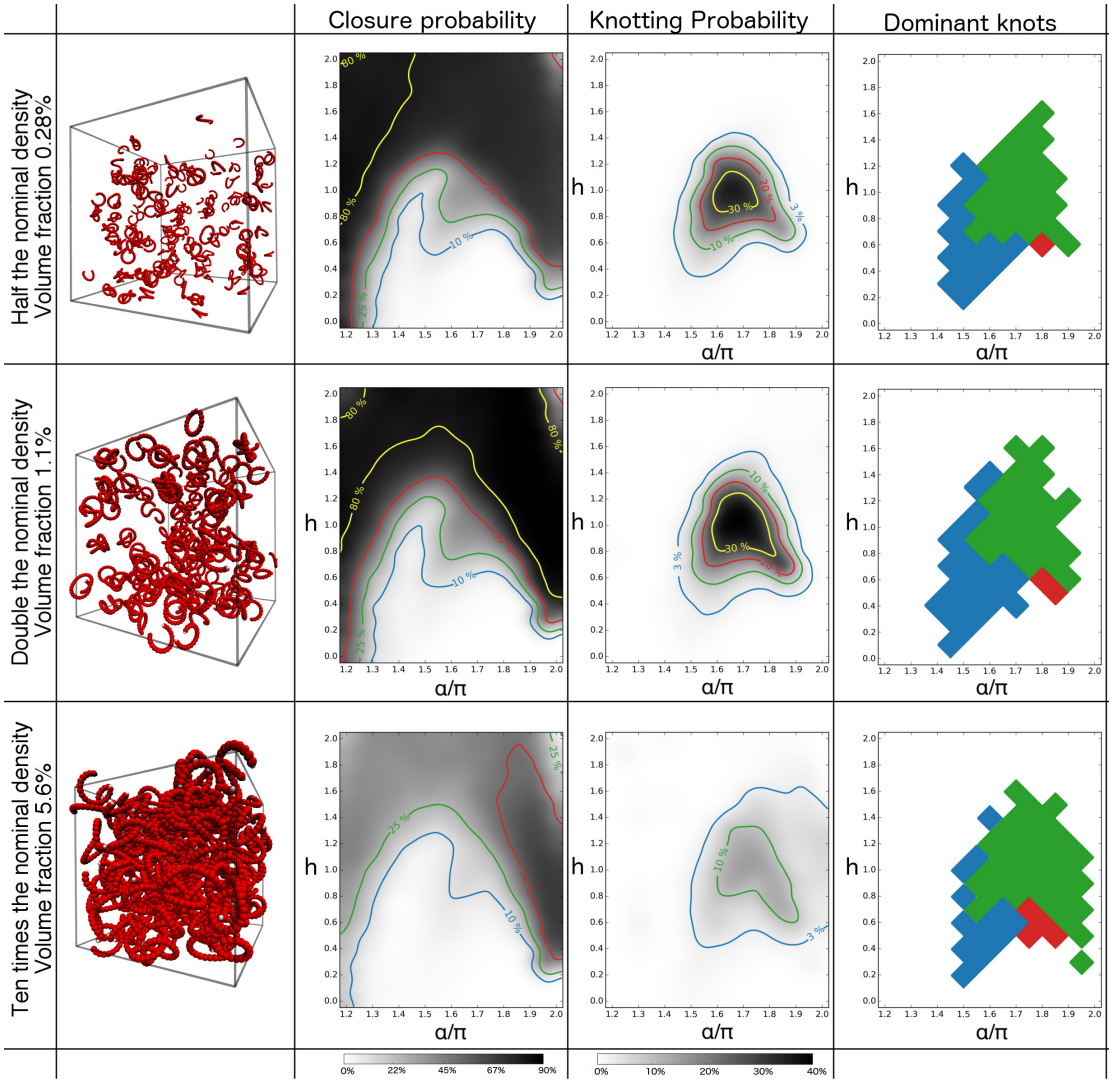

Figure 2.12: Robustness of the closure probability, knotting probability and topological phase diagram on the system density. The columns, from left to right, show snapshots, closure probability, knotting probability and topological phase diagrams. Results are obtained by packing the 250 templates to a monomer density that is half (first row), twice (second row) and 10 times (last row) as large as the reference one considered in the previous sections. The highest density favours the occurrence of constructs that are string-like or which percolate through the whole simulation box. The occurrence of these structures reflects in a lower incidence of ring-like constructs. The high solution density can favour their linking; however in the region of interest where the knotting probability is higher than 5%, one observes that knotted rings take part in generic links in less than 6% of the cases. The topological phase diagram (dominant knots) is visibly robust throughout the wide range of densities studied.

end of the polymer.

Apart from extreme densities, however, the topological phase diagram demonstrates that, by a suitable choice of the templates shape, it is possible to reliably control the statistical, or thermodynamic incidence of self-assembled constructs with definite topology in a manner that is robust upon decreasing or increasing the solution density by a

factor of two.

2.5.5 Geometrical considerations on the phase space accessibility

Although from the model simulation of the assembly naturally emerge the interplay of kinetic and entropic effects, the competition between structures and the spectrum of possible assemblies, it can be interesting to explore more in detail the constraints imposed by the geometry of the building blocks for the simplest structures.

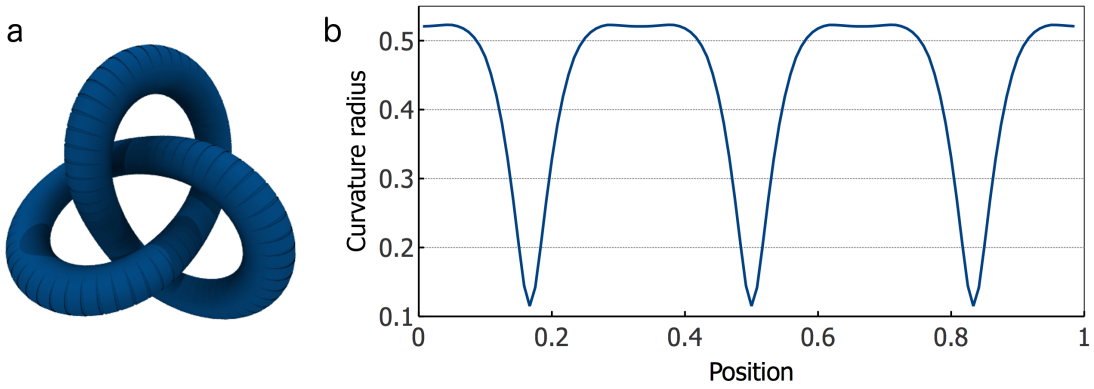

Figure 2.13: Ideal trefoil and its curvature profile along its centreline. (a) Representation of a thick structure following an ideal trefoil centreline. (b) Profile of the radius of curvature along the centreline. The position is indicated as the fraction of centreline covered. The starting and ending point (corresponding to position 0 and 1) is on the centre of one of the curved lobes.

The simplest non-trivial knot, the 3_1 in figure 2.13a has constant curvature along most of its contour length, with the exception of three low curvature points (2.13b). Indeed, one can immediately envisage a way to "break" this structure in three fragments with an approximately helical shape.

Focussing on the formation of 3_1 knots by connecting three helical fragments, we can make some further geometric considerations not resorting to sampling. Indeed, we can analyse the geometrical constraints imposed by the closure requirement and calculate, for example, the angle formed at the contact points in the minimum energy configuration.

This particular case of three fragments is easier to handle because (i) the end-to-end segments of the three fragments form one equilateral triangle and (ii) the lowest energy configuration is three-fold symmetric. This, in turn, makes it feasible to describe the relative position of the fragments in terms of α and h.

First of all, if we consider excluded volume interactions, we can rule out some zones of the parameter space, since a closed, knotted configuration may be incompatible with the self-avoidance condition. In the case of a fragment having a thickness of $\frac{1}{3}$ of the

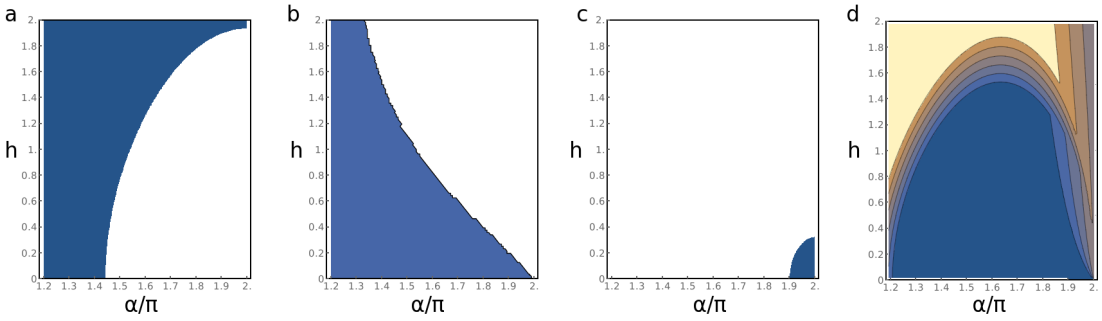

Figure 2.14: Panels (a), (b), and (c): areas in the parameter space not amenable for formation of 3_1 knotted structures by fragments of thickness $\frac{1}{3}$ are depicted in blue. (a) Area excluded by contact at the central zone of the lobes. (b) Area excluded by excessive planarity of the structure. (c) Area excluded by impossibility to bind at the ends. (d) Difference between closed dimers and trefoil configurations in the alignment at the templates' contact points. Trefoil structures are better aligned in the blue zones, dimers in the yellow ones.

projected helical radius, a consistent part of the phase space is not apt to the formation of ideal trefoil knots. In figure 2.14 several excluded zones are depicted. The first panel shows the zone of the phase space where knots cannot form because the central part of the curved lobes and the part with low curvature (see figure 2.13a) would be too close. In panel b, the coloured zone is not amenable to the formation of knots because the resulting knotted structure would result almost planar, again superposing each other if the thickness is finite. Finally, the coloured zone on the lower right corner of panel c is excluded in the case of thick fragments because the ends are so close that binding is impossible.

A second parameter amenable to consideration is the alignment of the templates at the connection points. This analysis is interesting since thick fragments with small connection points at the ends promote aligned configurations with respect to unaligned ones; the resulting values for trimeric structures can then be compared to the alignment for unknotted closed dimers. Figure 2.14d shows the difference in alignment at contact points between knotted 3_1 configurations and 0_1 configurations of two segments. In the yellow zone, dimers are more aligned and thus favoured, while in the blue zone trimeric configurations are preferred. This indeed reflects what is observed from the topological phase diagram in the dynamic model.

Note that these considerations are limited to the simplest 3_1 structures and do not have a straightforward generalisation in the case of more than three helical fragments. Nevertheless, in spite of the simplified assumptions, they can indeed explain some of the salient features of the topological phase diagram. They can thus provide some closer insight on the effect of different specific constraints of these kind of structures.

2.6 Conclusions

In this chapter we explored, by means of an *in silico* model, the effect of the shape of simple helical building blocks on the geometry and topology of the structures resulting from their spontaneous assembly.

We have shown that is indeed possible to control, by a suitable choice of the shape of the templates, the topology of self-assembled constructs.

Indeed, both the closure and the ability to form knotted structures can be controlled by varying the geometric parameters which describe the templates, promoting in turn open, closed or knotted configurations. Moreover, the resulting "topological phase diagram" appears to be robust across a wide range of densities.

The knotted topologies which were obtained by our self-assembly procedure showed interesting features. The most common and easily assembled one is a trimeric 3_1 construct; moreover, a geometrically different, tetrameric form of the same topology can be promoted. Another tetrameric construct can be also sistematically assembled, the rather complex 8_{19} knot, in its symmetric form. To our knowledge, this form has not been considered yet in the field of molecular assemblies and, according to our study, can be a possible, next candidate for synthesis.

Finally, we shown that non-monodisperse solutions can further increase the number of accessible structures, in particular we found that pentafoil knots, practically absent in the monodisperse topological phase diagram, can be observed by the mixing two distinct geometries.

Our topological self-assembly is an example of thermodynamic self-assembly, or "passive" self-assembly in the terminology of Ref. [84]. Normally, we are used to the idea that thermodynamics can only drive the formation of relatively simple structures, whereas further information, or coding, is required to form more complicated assemblies which are often found in the living world. Against this conventional framework, our topological self-assembly provides a notable counterexample: while lacking any active coding and while working in the absence of any external energy input, it can be tailored to drive the formation of knotted structures, with relatively high probability. It is especially remarkable that the selected knots are not in general the simplest ones, as we are able to self-assemble an 8-crossing knot as well as the trefoil and pentafoil knots.

An interesting extension of this work could be to allow to for some flexibility of the templates, as would be relevant for molecular or polymeric building blocks.

Chapter 3

Geometrical factors in link formation

Links are a class of topologies which are closely related to knots. Their definition is based on the same mathematical concepts of knots, involving the mutual entanglement of two or more closed curves, instead of a single one. The mathematical resemblance between the concepts of knots and links motivated us to consider this additional kind of entanglement in our self-assembly framework. Tracing a parallel with the last chapter, a question which naturally arises is whether geometrical constraints, such as the shape of building blocks, allow for controlling and modulating the spontaneous emergence of links.

In the assembly simulations of the previous chapter, some linked structures are actually present (some examples in figure 3.1), but the relative abundance of structures which are both linked and knotted was always less than the 6% of the entangled assemblies. Indeed, the emergence of links in that case was limited in spite of the relatively high concentrations of the templates.

Is it possible to design our system so to increase the incidence of links or, conversely, to discourage their formation? Apart from the fundamental point of view, one can regard the occurrence of links as a desired or undesired property, depending on the specific application. The identification of controllable parameters affecting the emergence of linked structures can thus prove valuable in both situations.

Interestingly, there are also certain instances in living systems where linking is abundant and co-opted for functional purposes. Kinetoplasts are networks of mitochondrial DNA which are found in some species of protozoan parasites. Its striking peculiarity is the arrangement of the genetic material in a network of thousands of DNA rings topologically interlocked, and their linking appears to play an important role in supporting

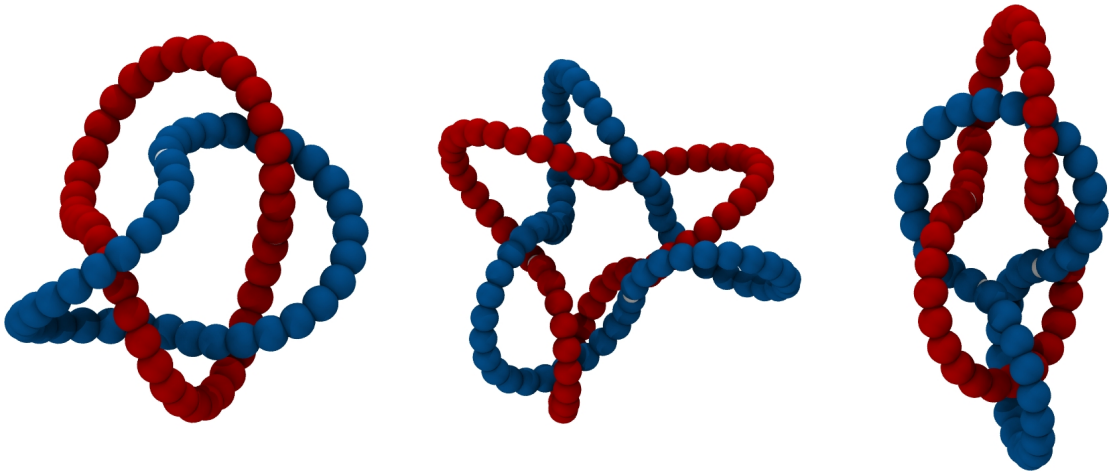

Figure 3.1: Examples of links observed in the simulations in chapter 2.

the full genome replication [30].

From a molecular perspective, the synthesis of so-called "catenanes", i.e. linked rings covalently closed, has been addressed in several chemically-oriented contexts. Supramolecular chemists have been recently able to synthesize a few novel linked structures [21]. As a matter of fact, the synthesis of a link can occur basically in two ways. On one hand, there are *direct methods*, where one allows the covalent closure of elongated, open molecules into rings, inside a concentrated solution of already closed rings. The linking occurs stochastically and specifically when the closure happens to clasp a ring in solution. The yields obtained using this procedure are usually very low. On the other hand, higher production yields can be obtained by stabilising "hooked" configurations of open flexible molecules by the means of metal coordination or hydrogen bonding before initiating the closure reaction which circularizes the molecules, trapping them permanently in a linked state.

In the general spirit of the approach we followed so far, we focused on "unassisted" linking of simple string-like objects. We explore here some geometrical factors, relative to both the templates and the system, which expectedly control the occurrence of self-assembled links.

Intuitively, a first, crucial parameter of the system in this context is the concentration of the components, as we observed in the previous simulations.

Besides looking at the impact of concentration on linking, we shall also consider the effect of the shape of the string-like building blocks, similarly to the previous chapter, including their thickness. The latter ought to play a key role here, because it limits the accessible space for interpenetration of structures.

The last parameter is the extent of spatial confinement. The likely relevance of

this order parameter emerges from a recent study of the effect of shape and dimensionality of spatial confinement on the entanglement of long semi-flexible polymers[57]. Interestingly, confinement has indeed a strong impact on the knotting behaviour of polymers. This study focused on DNA strands modelled as semi-flexible polymeric chains of consecutive cylinders, mimicking the natural persistence length of the nucleic acid. A Monte Carlo sampling of configurations confined in channels and slits revealed that confinement strongly affect the spectrum and incidence of knots along the chain. The effect is strongly dependent on the width of confinement, promoting the entanglement when the size of the slit or channel matches approximately the persistence length.

Although links are mathematically similar to knots, they are, on other aspects, a quite different kind of entanglement. Therefore, a question which naturally arises is whether is it possible to exploit confinement to control linking. And if so, which is its optimal width?

To answer these questions, we extended the simulations analysis framework of the previous chapter to deal with this more challenging type of entanglement. Moreover, we reinforced the assembly simulations with a Monte Carlo study which provide an efficient way to explore the spontaneous emergence of linking at equilibrium.

3.1 Generalities on links

Similarly to the case of knots, introduced in the previous chapter, a practical definition of links can be given by appealing to our intuitive notion: a link is a bundle of two or more circularized ropes which are interlocked and cannot be separated without cutting them open. Links can be classified in classes of equivalence which gather pairs of curves that can be geometrically matched by continuous, non singular transformations that do not involve the breaking of the curves or their strand passages.

Links complexity, like knots complexity, is conveyed by their minimal number of crossings in a two-dimensional projection of the curves. The number of crossings is indeed the first important parameter in the classification. A table of the links up to 8 crossings, is shown in figure 3.2.

Note that in the table there are also links involving three distinct curves. When more than two curves are involved, some interesting cases arise. For example, the link 6_2^3 is the so called *Borromean link*, in which each curve is pairwise unlinked to the other two, but they are inextricably interlocked as a whole.

The algorithmic identification of the link topology, similarly to the knots case, can be performed by computing topological invariants. For example, the Alexander polynomials can be generalised to the case of links.

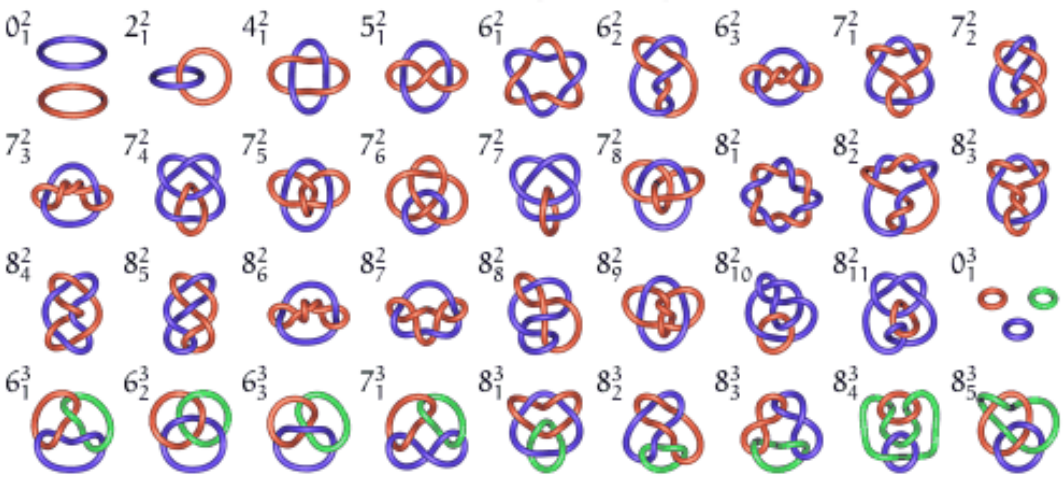

Figure 3.2: Table of links up to 8 crossings.

In this chapter, however, we will use an alternative method to detect linked structures, the *linking number*. Loosely speaking, this invariant is equivalent to count the net number of times a curve winds around the other. It is commonly used because it is straightforward to compute from a two-dimensional projection of the curves which keeps track of over and under-crossings.

The calculation of the linking number requires the curves to be *oriented*, and changing the orientation of a curve introduces a minus sign. For example, the first non-trivial link (2_1^2) of the table in figure 3.2, the so-called *Hopf link*, has a linking number of 1 or -1 depending on the orientations of the curves.

The simplicity of computation of the linking number comes with drawbacks too. Albeit being invariant for a given link type, the linking number does not allow for unambiguously identifying the link type. For example, there are linked curves, such as the Whitehead link (the 5_1^2 link in figure 3.2) that have zero linking number. Moreover, it is not possible to detect, for example, a Borromean link because each pair of the constitutive curves is unlinked.

Because of the above limitations, pairs of curves having a non-zero linking number are said to be *homologically linked* or *algebraically linked*, in contrast to the topologically linked ones. Accordingly, the Hopf link is both topologically and homologically linked, while the Whitehead link, although topologically linked, is not homologically linked.

Despite these limitations, the linking number has the important property that all unlinked curves have zero linking number and curves with a non-zero linking number are topologically linked.

Because we shall focus almost exclusively on simple links, the linking number typically suffices for their identification. We thus adopted it for its simple and transparent

formulation.

3.2 Simulating the formation of linked structures

We investigated the effect of the selected control parameters using two distinct methods and models. On one hand, we adopted the same self-assembly simulation procedure that we used for knots, focussing on simple geometries for our building blocks. Specifically, we shaped our templates as semi-circles or semi-ellipses. On the other hand, we complemented the self-assembly study with a Monte Carlo sampling of a solution of infinitely thin circular or elliptical objects. The sampling aptly complements the assembly study because it allows for an efficient characterization of the linking probability of the system in equilibrium.

Moreover, the comparison between the two combinations of methods and models can provide valuable insight on the role of kinetics and dynamics of excluded volume interactions.

3.2.1 Assembly model

The formation of a link requires the closing of two rings in an interlocked manner. For the sake of simplicity, we focused on very basic shapes which promote the formation of closed rings. Specifically, we considered semi-circles or semi-ellipses as basic building blocks.

Our templates were generated by lining hard spheres on their centreline using the same Weeks-Chandler-Andersen hard-core interaction of the previous chapter. Similarly, patchy particle were placed on the surface of each of the hard spheres at the ends. The radius of the semi-circular templates was set to 8σ, where σ is the nominal diameter of the hard spheres.

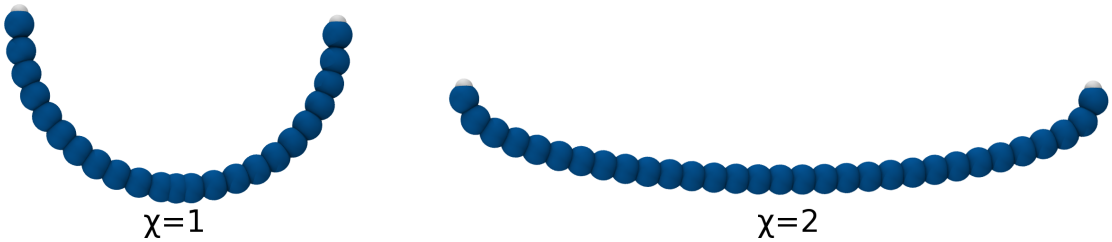

Figure 3.3: A semicircular ($\chi = 1$) and a semielliptical ($\chi = 2$) template. The planar area of the semi-circles and semi-ellipse are the same. Notice the different contour length (number of beads) of the two templates.

The shape of semi-elliptical templates was determined by a positive deformation

parameter, χ. Its Cartesian parametrisation is:

$$\begin{cases} x(t) = 8\sigma\chi\cos t \\ y(t) = \frac{8\sigma}{\chi}\sin t \\ z(t) = 0 \end{cases}$$

with $t \in [0, \pi]$.

Within this definition, the case $\chi = 1$ corresponds to semi-circles. The area of the convex plane portion enveloped by each member of the family of semi-ellipses spanned by varying χ is constant.

The deformation parameter is linked to the eccentricity e by the relation

$$e = \sqrt{1 - \frac{1}{\chi^4}}$$

Examples of templates with $\chi = 1$ and $\chi = 2$ are shown in figure 3.3.

We performed 20 distinct simulations for each point in the parameter space, from which we obtained 20 different ending configurations. The values for the observables reported are the average over these 20 configurations.

3.2.2 System setup

Each assembly simulation counts 500 templates either in bulk or confined in a slab.

For bulk simulations, we used a cubic simulation box with full periodic boundary conditions. The concentration was controlled by setting the side of the box to $\rho^{-\frac{1}{3}}$. Note that, from now on, we will use the diameter of a circular structure (16σ) as unit of length. Hence, the volume fraction is of the order of $\sim 0.01\rho$. Moreover, the values of concentration will be given in terms of *dimers* in unit volume. Hence, one has to bourne in mind that the actual *single template* concentration is two times the values indicated.

Spatially constrained simulations have a limited height h in the z direction and periodic boundary conditions on both x and y. The x and y sides of the box control the concentration of templates ρ, ad were set to $(\rho h)^{-\frac{1}{2}}$.

The confinement is performed by placing two impenetrable walls on $z = 0$ and $z = h$. Any particle is subject to a Weeks-Chandler-Andersen potential

$$\begin{cases} U_{wall} = 4\epsilon\left[\left(\frac{\sigma}{d_z}\right)^{12} - \left(\frac{\sigma}{d_z}\right)^{6} + \frac{1}{4}\right] & \text{if } d_z < 2^{\frac{1}{6}}\sigma \\ 0 & \text{else.} \end{cases} \tag{3.1}$$

where d_z is the particle distance from the wall.

The range of explored concentrations was $0.1 \leq \rho \leq 2$. At lower concentrations, the extremely scarce occurrence of links leads to very poor statistics. At higher concentrations, excessive crowding results in most of the rings being interlocked and the kinetic evolution is extremely slowed.

3.2.3 Monte Carlo conformational sampling of infinitely thin structures

The sampling was performed by generating multiple independent configurations in a box. Each configuration counts 1000 planar structures, composed of a discrete number of points, distanced roughly by 0.25 times the circle radius. This means that a circle is discretized in ~ 25 points.

Concentration, size of the box and periodic boundary conditions are controlled in the same way as in the assembly case.

A random centre of mass position and orientation are generated for each ring. In the case of confined simulations, if the resulting ring configuration crosses the $z = 0$ or the $z = h$ plane, a new position and orientation is generated, until it is accepted.

Typically the number of independent configurations is 10^4, except at low densities ($\rho < 0.1$). For those points, we generated 10^5 configurations, in order to keep the estimated statistical error on the linking probability below 1%.

3.3 Results

3.3.1 Assembly simulations

Solution concentration

The main property we are interested to characterize in our study is the linking probability P_l. We consider the last frames of 20 indipendent simulations; consistently with the definition of the knotting probability given in the previous chapter, the linking probability is defined as the average fraction of templates which form linked structures:

$$P_l = \langle \frac{\#\text{ of templates in linked structures}}{\#\text{ of templates in the simulation box}} \rangle.$$

We started by considering a reference bulk case where no spatial constraint is imposed, and we probed the changes of the linking probability to variations of the template concentration.

Figure 3.4 shows the dependence of P_l on the template concentration ρ. The dependence appears approximately linear at low concentrations, then P_l starts saturating

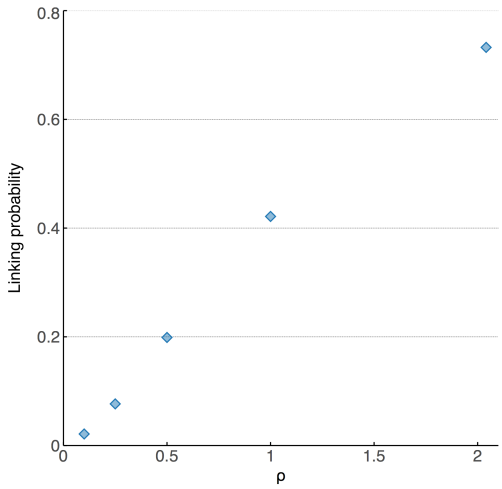

Figure 3.4: Linking probability versus concentration in bulk. Each data point is an average over 20 independent runs.

when ρ is on the order of ~ 1 (i.e. when the average distance between structures is of the order of the structures size). We expect P_l to approach 1 at high densities because of the extreme crowding.

Linking probability under confinement

A second control parameter is the degree of spatial confinement. Apart from being interesting from the fundamental point of view, controlling confinement at a molecular scale is nowadays possible using nanoslits and nanochannels which have opened novel and interesting avenues for molecular manipulation.

Here, we analysed the effect of confining our templates in the z direction, i.e. in a slab of height h.

Figure 3.5 shows the linking probability P_l for semi-circular templates, upon varying the slit height h. The size h is in units of diameters of the assembled rings.

One immediately notices that the behaviour of P_l at fixed concentration is non-monotonic. When h is large the situation is similar to the bulk case. However, when h is reduced down to the template sizes, we observe a maximum in the linking probability. Specifically, it occurs typically around $h{\sim}1.5-2.0$ and the enhancement in P_l is on the order of $\sim$20%.

At even smaller h, the linking probability drops dramatically. A very significant suppression of linked structures can therefore be obtained by carrying out the assembly in a quasi planar environment.

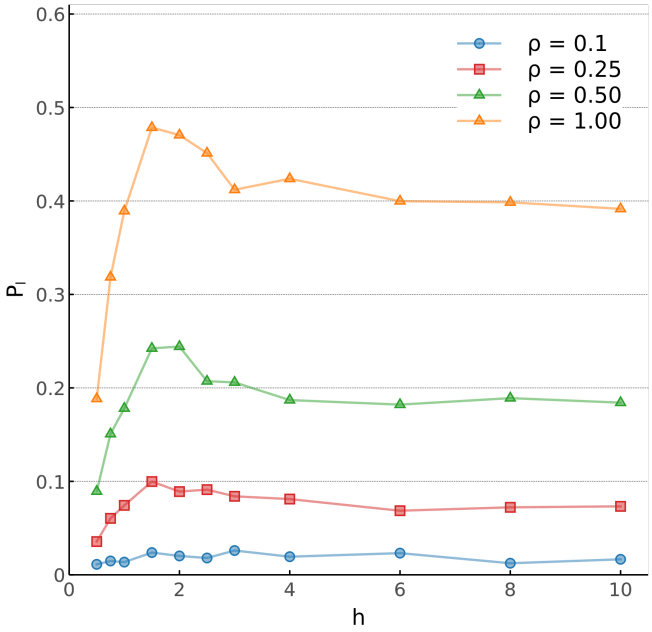

Figure 3.5: Linking probability dependence on the slab height h. The profiles are shown for ρ =0.1, 0.25, 0.5 and 1.

Geometry of templates

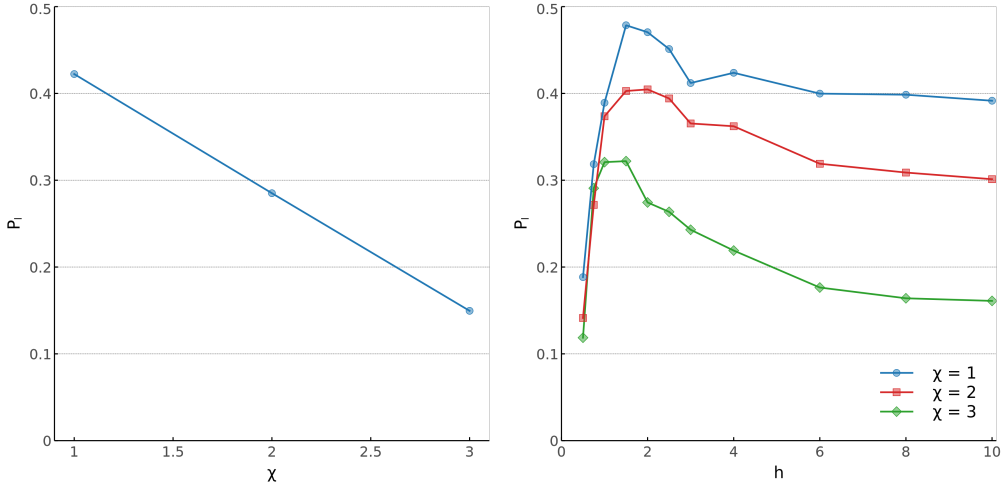

Figure 3.6: Linking probability dependence on the shape of templates ($\chi = 1, 2, 3$) at density $\rho = 1$. (a) P_l values calculated in bulk for the three geometries considered. (b) P_l dependence on h using different template shapes.

The effect of the template geometry on the linking probability is conveyed by figure 3.6, where simulations for semi-circular ($\chi = 1$) and semi-ellipsoidal ($\chi = 2, 3$) templates are summarized. The assembled closed structures taking part in links are closed dimers of templates. Recall that the centreline of dimers covers the same area for each value of

χ, whereas the contour length, and hence the number of beads, is not constant. Eccentric ellipses are less likely to participate in links, the template density being equal, than their circular counterparts. This happens both in bulk (in panel a) and in confinement (panel b). In the latter case, the non-monotonic profile is nonetheless present for both circles and ellipses.

3.3.2 Monte Carlo sampling of closed structures

To gain better insight on the described phenomenology, we then turned to Monte Carlo (MC) sampling of closed structures. This sampling does not reproduce kinetic effects of the assembly process, but has the advantages to be a lot more efficient than dynamics, and provides estimates on what we should expect at thermodynamic equilibrium. Note that the systems considered in the MC sampling and molecular dynamics assembly simulations are related, but not identical. In fact, the fundamental units of MC simulations are the entire rings; no semi-rings or longer multimers are present, unlike the assembly case.

Rings concentration

Figure 3.7a shows the linking probability profile with respect to the ring concentration ρ. The sampling is performed on infinitely thin circular rings.

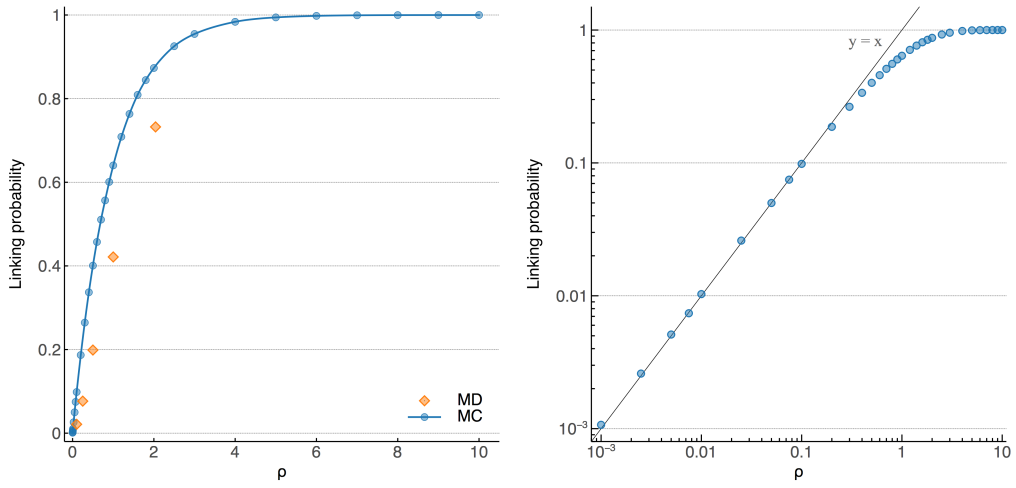

Figure 3.7: Linking probability as measured from Monte Carlo simulations of infinitely thin rings in bulk. In the (a) panel, also data from assembly simulations (MD) is plotted for comparison.

The low concentration regime is shown in the plot of figure 3.7b, together with the $y = x$ line. The sampling shows that the relationship P_l and ρ is indeed linear at

$\rho \ll 1$. As the concentration is increased the linking probability inevitably saturates to 1, similarly to the assembly simulations, due to the significant crowding of the box.

Note that template thickness is expected to hinder linking by reducing the effective space for interpenetration. Because of this effect and the fact that there are no semi-rings, the estimates from this sampling represent *upper limits* on the linking probability values during assembly.

Linking probability under confinement

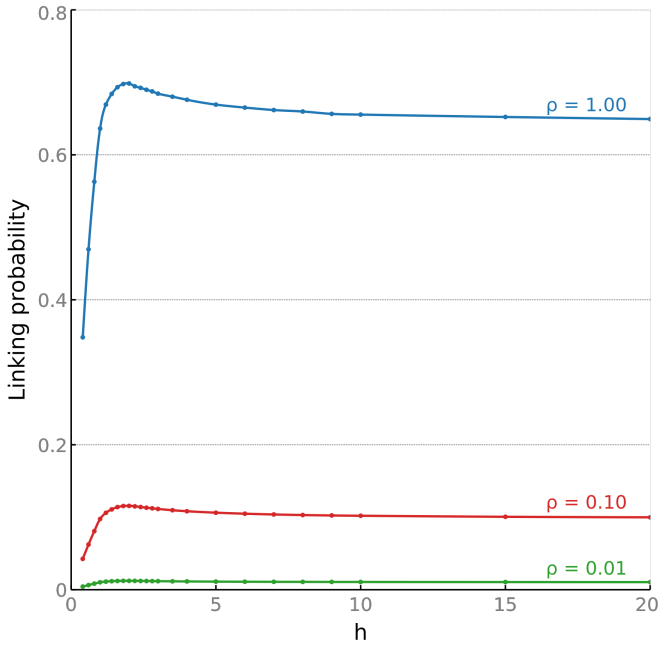

Figure 3.8: Linking probability dependence on the slab height h from Monte Carlo sampling of infinitely thin circular rings.

We repeated the sampling procedure imposing spatial confinement. The resulting curves are shown in figure 3.8. The non-monotonic profile observed in the assembly is recovered also in the case of infinitely thin rings: every curve shows a maximum corresponding to $h = 2$. The corresponding enhancement with respect to the bulk value of P_l is approximately 20%. The fact that the maximum is observable also in the Monte Carlo sampling indicates that the non-monotonicity in the linking probability profile does not originate from aspects that are specific to the self-assembly dynamics, but is a genuine property of the system in thermal equilibrium.

It is apparent from fig. 3.8 that this consideration applies not only to the maximum but also to the presence noticeable drop of the linking probability as h is lowered past the maximum.

Geometry of templates

The last parameter we modified was the geometry of the templates. In figure 3.9 we show the effects of the variation of χ on infinitely thin rings, both in bulk and in confinement.

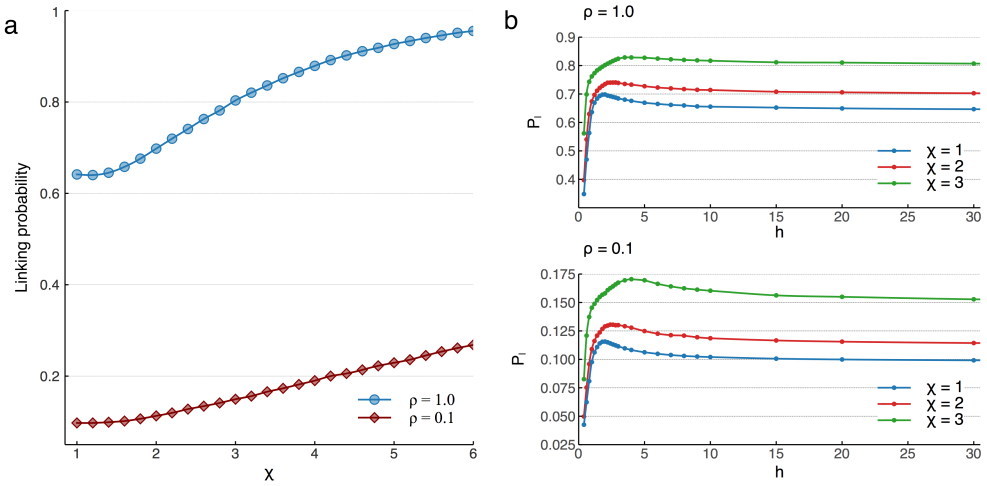

Figure 3.9: Monte Carlo simulation of infinitely thin rings of various geometries for both $\rho = 1.0$ and $\rho = 0.1$. (a) P_l dependence from χ in bulk. (b) P_l versus h profiles for $\chi = 1, 2, 3$.

The profile of P_l on χ at $\rho = 0.1$ shown in figure 3.9a features a "hockey-stick" shape ; specifically it is about constant for $\chi < 2$ and grows linearly for larger values of χ. The profile at $\rho = 1$ also shows an increment of P_l with χ and it asymptotically plateaus to 1. The increase in P_l with χ may be explained by the ability of ellipses to link at larger distances with respect to rings. Figure 3.10 shows the histogram of linked and unlinked configurations with respect to the centre of mass distances for two rings. The added linked configurations at larger distances, indeed, are not compensated by the decrease at short distances caused by the shortening of the minor axis.

The covered area being constant, eccentric rings can "reach" and link to more peers than their circular equivalents. This, in turn, promotes the formation of links between structures.

Another interesting feature of the profiles in figure 3.9b is the widening of the saddle and the translation of the peak value of P_l towards higher values of h, and will be discussed in the next section.

Surprisingly, the plots shown in figure 3.9 feature major qualitative differences with respect to the assembly case. Assembly simulations show a significant decrease of the linking probability when structures become eccentric. Our conformational sampling, however, displays a net *increase* when χ is raised.

We envisage that the observed differences arise for the following reasons.

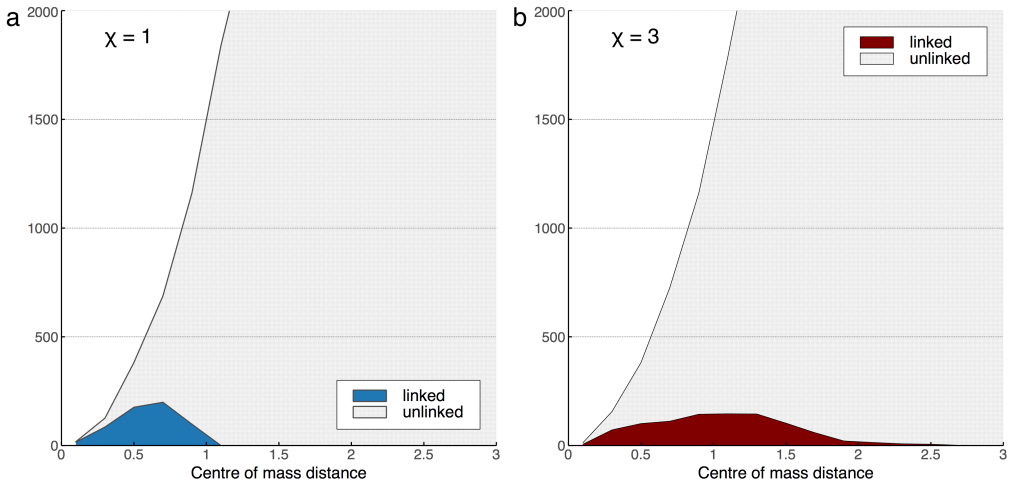

Figure 3.10: Histogram of configurations of infinitely thin rings as a function of the distance of their centres of mass. (a) Count of linked and unlinked configurations for circular rings and for (b) ellipses with $\chi = 3$.

The first reason is the excluded volume interaction of the templates in assembly simulations. The effective accessible area for interpenetration is actually reduced by the substitution of circles with eccentric ellipses. This is perhaps best explained considering the extreme situation where the ellipse is so eccentric that the minor semi-axis is smaller than the hard core spheres diameter. Although the nominal area is kept constant, the cross-sectional area cannot be interpenetrated by another ellipse; the effective space for interpenetration is thus practically zero.

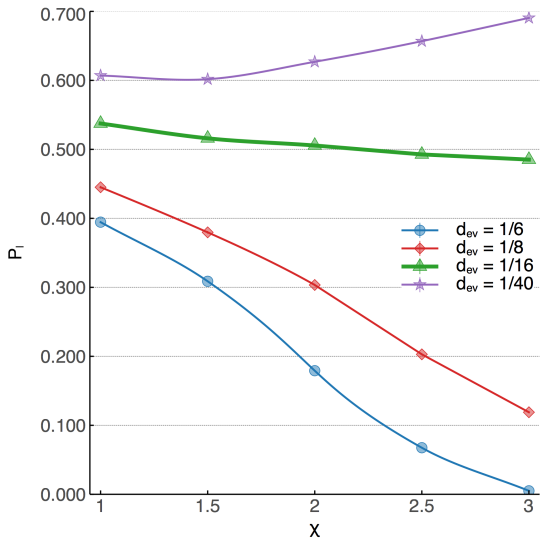

Figure 3.11: P_l dependence on χ for thick rings at $\rho = 1$. Each data point is calculated on 10^5 rings configurations. d_{ev} indicates the thickness of the curve in units of the circle ($\chi = 1$) diameter.

To quantify this effect, we repeated the Monte Carlo sampling in bulk, but replacing the infinitely-thin ellipses with thick ones. Specifically we lined the ellipses contour with circular beads with diameter in the $\frac{1}{40} - \frac{1}{6}$ range. Figure 3.11 shows the P_l dependence on χ, upon the variation of the ellipse thickness. The difference from the infinitely thin case is indeed significant: the thickness of the object counteracts the effect of the eccentricity except for very thin cases. The linking probability shows a monotonically decreasing profile even when the thickness $d_{ev} = \frac{1}{16}$ of the diameter of the ring with $\chi = 1$. The effective thickness of the templates in the assembly simulations is slightly larger of this value because of the shape of the hard-core potential; the Monte Carlo simulations thus provide an upper bound to the linking probability.

The second motivation is possibly related to the kinetic evolution of the system. The duration of the MD assmbly simulations was set a priori to be equal to throughout all considered ring densities and eccentricities. Specifically, this was set equal to $20 \cdot 10^6$ steps, which based on our preliminary runs at $\rho = 0.5$ and $\chi = 1$ amply sufficed to reach the steady state situation. However, this set duration may become insufficient to reach steady state is the system density is too large. Clearly, varying the ring eccentricity does change the *monomer* density at constant template density (see figure 3.3). Accordingly, we cannot rule out that the densest systems (the most eccentric) have not yet reached the steady state. The confined space further worsen the situation, possibly leading to a very significant slowing of the dynamics; the system could thus require very long times to reach a steady state. An indication of this problem is the visible reduction of the number of ring constructs when χ is increased (figure 3.12). Longer simulation times and lower concentrations may be required in this case.

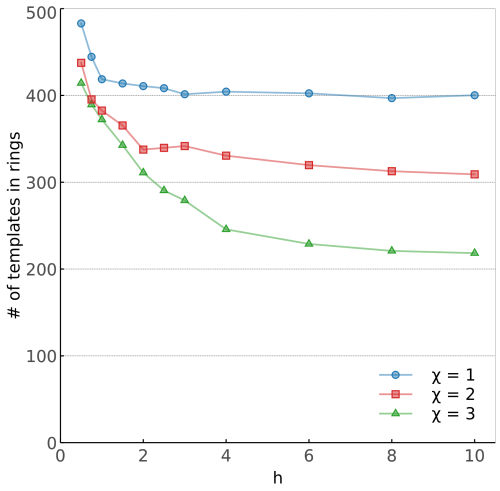

Figure 3.12: Fraction of templates participating in ring structures in the assembly simulations.

3.3.3 Spatial confinement and effective density

The concentration dependence of the linking probability can provide an explanation for the peak in the P_l versus h profile.

Indeed, the presence of a wall limits the rotational freedom of structures at short distance from the wall itself. When the centre of mass of the ring is moved away from the wall, the number of configurations accessible to the ring structures increases, because more and more possible orientations are allowed. When the distance between the centre of mass and the wall is larger than the radius of the ring, all possible orientations are accessible.

In figure 3.13a the transversal density profile is calculated for the beads forming circular rings. Consistently with the above observations, it shows a net decrease in proximity of the hard walls.

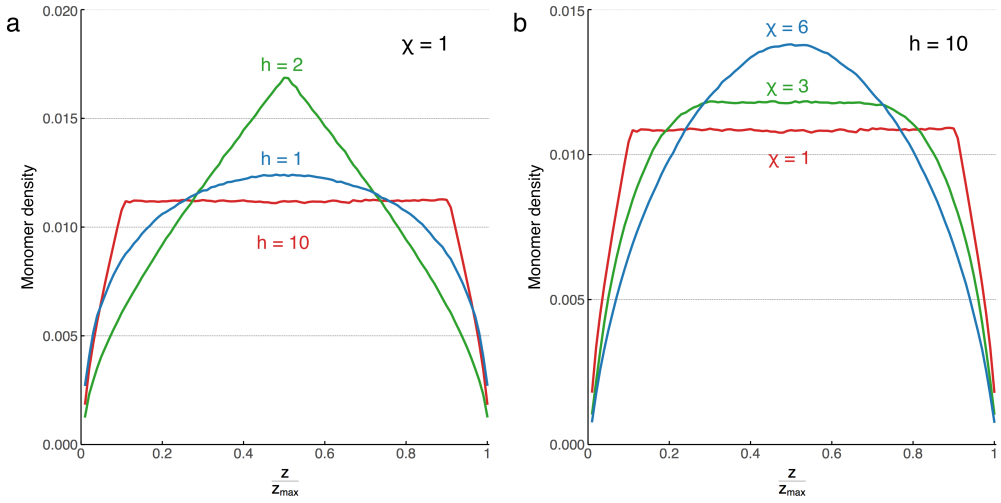

Figure 3.13: Transversal distribution of monomers along the slab. Each profile is the normalized histogram of monomers z positions, subdivided in 100 bins. Each histogram is calculated from sampling 10^6 configurations, for a total of $2.4 \cdot 10^7$ monomers. (a) Profiles calculated for circular rings ($\chi = 1$) at $h = 1, 2$ and 10. (b) Profiles at fixed $h = 10$ and different ring shape, $\chi = 1, 3$ and 6.

The depletion of the "wall layer" results in a larger effective concentration in the rest of the box, which, in turn, can promote the formation of linked structures. The net effect, as one can evince from the comparison of the two curves for different h, is relatively more marked for thin slits. Indeed, as long the wall layer is small compared to the height of the box the effect is negligible. When instead the box height is few times the depletion layer one the effect becomes relevant.

Moreover, this compression effect peaks at two times the circle diameter, which corresponds to the peaks observed in the linking probability. This effect is also completely consistent with the translation of the peak in the P_l profile in the case of ellipsoidal

rings, since the eccentricity of the structures is accompanied by a widening of the wall layer, as shown in figure 3.13b.

3.4 Conclusions

In this chapter I extended the framework introduced in the previous chapter to explore a more complex kind of entanglement, the linking between structures. We focused on planar structures, namely patchy semi-circles and semi-ellipses, and explored the spontaneous emergence of linking during their self-assembly in solution, both in bulk and in presence of spatial confinement. The probability of observing linked structures was probed against variations in the shape of the templates and in the dimensions of the confinement. We complemented the observation for the self-assembling systems with a Monte Carlo sampling of thin, interpenetrable structures which provided valuable insight for the interpretation of the results.

The linking probability profile with respect to the size of the confinement features a non-monotonic behaviour, reaching a maximum when the size of the slab is comparable to the typical size of a template, and then drastically drops when the height is further reduced. Indeed, this demonstrates that it is possible to control to a certain extent the spontaneous emergence of linking by strongly confining the system. The effect can be ascribed to a depletion of the zone in proximity of the walls, resulting in a effective rise in the concentration in the central area.

Moreover, we observed a non-trivial dependence of the linking probability on the shape of the templates. Eccentic ellipses are more apt to linking with respect to circular structures. However, a decrease of the effective space for interpenetration and a drastic slowing in the evolution of the system led, in the case of self-assembling templates, to a net decrease in the formation of linked structures on the timescales considered here. Nevertheless, controlling the shape of basic objects in a solution can be a way to indirectly control the spontaneous emergence of interlocked structures.

Chapter 4

Identification of mechanical domains in viral capsids from quasi-rigid domain subdivision

Challenging most of our definitions of life forms, viruses occupy their own, peculiar spot at the edge between plain macromolecular assemblies and life. Because they rely on the host cell machinery in order to perform their own replication, their genome is usually very short and highly optimized. Indeed, many viruses evolved in the direction of making their genome as compact as possible not only in sequence (thanks to the baffling use of overlapping reading frames and limited incidence of non-coding regions) but in terms of spatial size too [8, 80].

The genomic material of most viruses is encapsidated inside protein shells with diameters in the 20-100 nm range that are organised according to icosahedral symmetry. This is consistent with the principle of genetic parsimony, as shells formed from many copies of a few proteins types can be coded for by relatively short genomic sequences.

The viral capsids formation is indeed an impressive example of self-assembly occurring in nature. The generation of complex macrostructures from the interconnection of simpler protein building blocks is even more striking if one takes into account that most capsids do not act only as mere passive envelopes, but are active players in the cell infection.

The proteins forming the capsid must indeed be able to perform many different tasks essential to the life cycle of the virus. Here we focus on two essential features: the capability of the proteins to self-assemble into full capsids enclosing the genomic material, and that of providing a mechanically stable shell structure. Note however that the shell, in some viruses, can still undergo functional structural changes such as

expansion, with subsequent formation of pores, necessary for the release of the genome into the host cell.

Several experimental and theoretical studies have investigated the kinetics and thermodynamics of capsid assembly, that is the process by which a collection of mono- or poly-dispersed isolated proteins aggregate into a full capsid with correct size and symmetry [20, 87, 10, 17]. For various viruses it has been demonstrated experimentally that individual protein subunits aggregate into one or more types of stable multimers, termed capsomeres, which are the fundamental building blocks out of which the capsid is made.

Furthermore, for those viruses for which structural maturation of the fully assembled capsids are necessary to render these viruses infectious, these usually occur as quasi-rigid body motions of a small number of different functional units [71, 44].

Advanced experimental techniques have been of crucial importance to shed light on specific properties of some viral capsids, and to probe their physico-chemical behaviour. One example are nano indentation experiments, where viral particles are subject to mechanical stress and fatigue by atomic force microscopy; these experiments have singled out the mechanical building blocks of viral capsids and elucidated the mechanisms of genome uncoating[69].

The limitation of many techniques stands, however, in the severe experimental demands or in the difficulties to apply the same methodology to very different viruses.

In this context, our aim was to identify the functional blocks which form a capsid using a *top-down* approach. In fact, both assembly and functional units in capsids can be best rationalized in a mesoscopic perspective. The identification of these "basic units" can help, on one hand, to understand the pathway followed by the process. On the other hand, the correspondence or mismatch between assembly and functional units can reveal the use of different strategies in the production of functional capsid.

Our study is motivated by the observation that by suitably analysing the large-scale internal dynamics of proteins, or protein assemblies, one can decompose them into few dynamical domains, whose quasi-rigid relative motion, meaning motion that can be formulated in terms of rigid translation and rotations, accounts for most of the observed structural dynamics [27, 28, 29, 25, 85, 64, 1, 63]. Based on this observation, supported by successful multiscale simulations where viral capsids were modelled as assemblies of rigid tiles [3], one can therefore expect that these systems can be decomposed into quasi-rigid protein units which, being mechanically stable, could have functionally-important roles.

Therefore, our work focuses on the identification of these functional protein units acting as quasi-rigid blocks. Since experimental information on such units is often

unavailable and difficult to obtain, we have developed a new computational scheme that is designed to complement the experimental efforts by directing the analysis to groups of proteins in the capsid that behave as quasi-rigid units.

The material in this chapter in the methods and results section is largely based on our article "Mechanical and assembly units of viral capsids identified via quasi-rigid domain decomposition" in reference [59].

4.1 Icosahedral viruses

Most viral capsids are arranged in a icosahedral symmetry. The classification of the possible shapes was first proposed by Caspar and Klug[13].

The starting point for their theory was the observation that viruses often assemble identical, repeated building blocks to form an icosahedral capsid. Moreover, also the connections between the proteins are expected to be identical. If both building blocks and contacts are identical, the only possible number of objects which can build up a closed icosahedral shell is 60 [13]. Note that the 60 units need not to be symmetrical themselves, only their arrangement is. However, experimental evidence indicate that capsids could be made of many more constitutive protein units, meaning that they cannot possibly be arranged in a perfectly *equivalent* manner.

The Caspar-Klug theory introduced a successful way to accommodate multiples of 60 asymmetric units on a closed shell, slightly relaxing the requirement of perfect symmetry for the way the units interact.

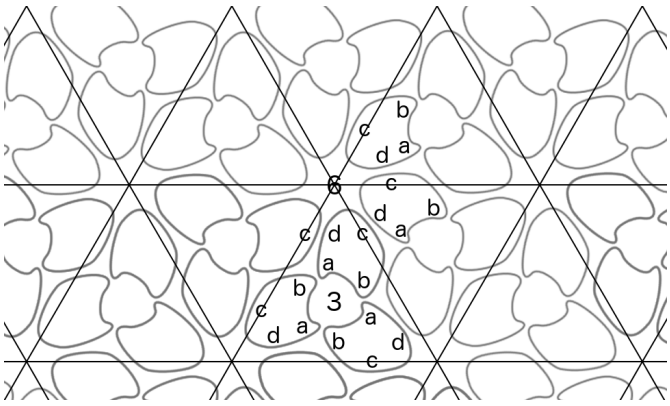

Figure 4.1: Example of arrangement of identical asymmetric proteins on a triangular lattice with 6-fold rotational symmetry. The interactions between protein units are all equivalent: the interaction regions labelled as a,b,c and d are always paired as a-b and c-d.

As shown in figure 4.1, it is possible to place asymmetric units on a plane in a way which is rotational six-fold symmetric, creating a equilateral-triangular net. In this arrangement, all the contacts between the protein units are equivalent.

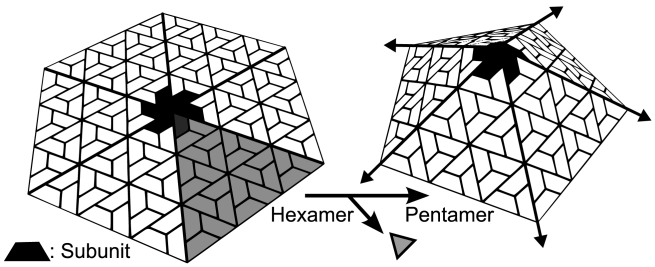

Figure 4.2: Illustration of the quasi equivalence principle. A triangle is removed at a 6-vertex on the triangular net, resulting in a 5-vertex. The image is adapted from ref. [46].

A closed icosahedral surface can be folded from this planar triangular net, while preserving the same contact pattern, by transforming 12 of the vertices with 6-fold symmetry in 5-fold vertices. This is done by removing one of the triangles at the 6-fold vertex and connecting the free edges as shown in figure 4.2. In this way, the capsid can be schematically represented as 12 pentagonal sites separated by a "honeycomb" tessellation of hexagons as in figure 4.3.

From a physical perspective, arranging five or six identical protein units around an axis leads to two distinct situations. The proteins are embedded in a slightly different environment, having respectively 4 and 5 other neighbours. In fact, the constitutive units are not strictly equivalent anymore, but have become *quasi equivalent*, meaning that the environment is approximately, though not exactly, identical. The relaxation of the strict equivalence principle on the contacts between blocks explains how large capsids can accommodate more than 60 equivalent proteins.

The simplest and smallest closed shell (a T=1, see below) is the particular case formed by 60 identical proteins that experience exactly the same environment. In the honeycomb plane representation, this correspond to select only 12 adjacent hexagons and operate the hexagon-pentagon substitution on each one of them. We end up with exactly $12 \cdot 5 = 60$ strictly equivalent positions around the five-fold axes, where we can accommodate the proteins.

If the pentagons are further away in the planar representation the folded icosahedral shells will be larger. Figure 4.3a shows the planar development of an icosahedron from the honeycomb plane. The vertices of the 20 triangles which form the planar development fall, in the honeycomb plane, on the hexagons which have to be substituted by pentagons.

The capsid shells arrangement can be classified according to the *triangulation number*. This is defined as the squared distance of two adjacent five-fold vertices (i.e. of two substituted "pentagons") on the honeycomb plane, measured in units of the triangular lattice spacing. In particular, consider the origin of the axes as the centre of one pen-

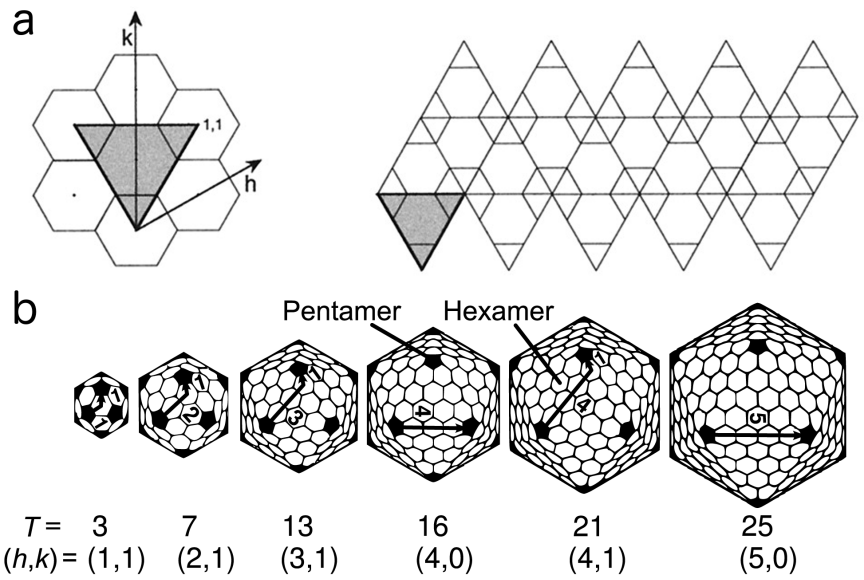

Figure 4.3: Triangulation number is calculated from the numbers h and k of jumps between two pentagons. (a) Connection between the honeycomb plane and planar development of the icosahedral shell. Adapted from ref. [7] (b) Examples of different T geometries. Pentagons are highlighted in black. Adapted from ref. [46].

tagon. Now, consider the natural basis vectors for the 6-fold symmetric tessellation of the plane, i.e. the ones reaching the centres of two neighbouring hexagons. The distance vector from the origin to the centre of the closest pentagon can be expressed in terms of integer numbers h and k of such basis vectors.

The triangulation number is defined as

$$T = h^2 + hk + k^2$$

where h and k are the vertical and diagonal jumps. Therefore, larger T numbers correspond to larger icosahedra (figure 4.3b). The T number is intimately connected to the number of pentagons and proteins in the capsid. In fact, every icosahedron counts 12 pentagons and $10(T-1)$ hexagons. In terms of protein subunits, each capsid requires $12 \cdot 5 + 10(T-1) \cdot 6 = 60T$ of them. For $T > 1$, the proteins are placed in *quasi* equivalent positions, having only a part of them around on the 5-fold symmetry axes.

Many capsids are formed by proteins which are not chemically equivalent. In this case we talk about *pseudo-equivalence* instead of quasi-equivalence, and we indicate it with a lower-case p. For example, L-A virus counts 120 proteins of two different kind. While a $T = 2$ classification would not make sense in the described picture, as no integers h and k exist that yield $T = 2$, here couples of different proteins bind to create 60 larger asymmetric dimers, which then connect in a $T = 1$ fashion. In this case we

talk about a $pT = 2$ capsid.

4.2 Methodology

We apply a top-down approach for identifying capsomeres based on the equilibrium fluctuation dynamics of the fully-assembled capsid. The scheme builds on the notion that the stability of the fundamental functional units ought to reflect in their quasi-rigid character within a thermally-fluctuating capsids.

The starting element is the structure of a fully assembled capsid. From the structure we extract information on the thermal fluctuations using an elastic network model (ENM), which is in turn used to obtain a number of optimal subdivisions in quasi rigid-domains of the full capsid. The last step is the determination of the best subdivision based on objective parameters specifically tailored for viral capsids.

4.2.1 Mechanical characterization

In the last two decades a number of viral structures have been resolved by X-ray crystallography. The sizes span from the $T = 1$ Satellite Tobacco Mosaic Virus capsid, counting 8800 amino acids to the human Adenovirus, counting approximately 1 million amino-acids.

The characterization of the mechanical response of mesoscopic by molecular dynamics simulations of object of this size is infeasible in most cases, and anyway extremely challenging even in the case of the smallest ones.

However, a large body of experimental and numerical evidence has indicated that the principal fluctuations modes, those of lowest energy, have a collective character. This means that the structural deformations associated to these modes entail the concerted displacements of groups of several amino-acids.

As it was first shown by Tirion[79], the collective character of the modes justifies the use of simplified, coarse-grained models (rather that atomistically-accurate ones) for calculating the principal modes of fluctuation of a protein around its reference, native structure.

A commonly used framework for such coarse-grained calculations is provided by elastic network models. The latter rely on a quadratic approximation of the near-native protein free energy,

$$\mathcal{F} = \sum_{\beta,\gamma=1}^{n} \delta\mathbf{r}_\beta \cdot \mathbf{M}_{\beta\gamma} \delta\mathbf{r}_\gamma \tag{4.1}$$

where n is the number of aminoacids, $\delta\mathbf{r}_\beta$ is the vector displacement of the βth amino acid from the native position of a mainchain centroid (typically the C_α atom) and, for

fixed β and γ, $\mathbf{M}_{\beta\gamma}$ is a suitably-defined 3×3 symmetric interaction matrix. We denote by $\mathbf{M}$ the resulting $3n \times 3n$ interaction matrix.

Within the quadratic approximation of Eq. (4.1), the principal modes of structural fluctuations can be calculated exactly with a relatively small computational expenditure, and they correspond to the eigenvectors of $\mathbf{M}$ having the lowest non-zero eigenvalues. In the following we shall indicate with $\lambda_1,\ \lambda_2,\ \lambda_3, \ldots$, the non-zero eigenvalues ranked for increasing magnitude (they are all positive) and with $\mathbf{v}_1,\ \mathbf{v}_2,\ \mathbf{v}_3, \ldots$ the corresponding orthonormal eigenvectors. It can be shown that λ_i is equal to the total mean square structural fluctuation projected on the ith mode, $\lambda_i = \langle \sum_{\beta=1}^{n} |\delta\mathbf{r}_\beta \cdot \mathbf{v}_i|^2 \rangle$, where $\langle \cdot \rangle$ denotes the canonical equilibrium average.

In this study, we shall resort to the beta-Gaussian network model[49] to compute the matrix $\mathbf{M}$ and its eigenvalues and eigenvectors. The model differs from other elastic networks because it allows for capturing effective sidechain-sidechain interactions besides the mainchain ones. The model was previously successfully validated against extensive molecular dynamics simulations of various proteins and protein complexes and differentiates from other elastic network.

The linear size of the $\mathbf{M}$ matrices for the capsids considered here can be as large as $\sim$200000, which makes computationally difficult not only the diagonalization calculation but even the storage in RAM of the matrix itself. We circumvented these difficulties by taking advantage of the sparse character of $\mathbf{M}$.

Indeed, the quadratic interactions in elastic network models are usually activated solely between beads within a certain distance r_{cutoff}. The $\mathbf{M}$ matrix values in our system are set as:

$$\begin{cases} \mathbf{M}_{\beta\gamma} = k & \text{if } \|\mathbf{r}_\beta - \mathbf{r}_\gamma\| < r_{\text{cutoff}} \\ \mathbf{M}_{\beta\gamma} = 0 & \text{else.} \end{cases} \tag{4.2}$$

where r_{cutoff} was set to $10\mathring{A}$, $k = 1$ for non covalently bonded amino acids and $k = 10$ for covalently bonded ones.

The interaction matrix is therefore very sparse, counting approximately $B \times N$ non-zero entries, where N is the number of amino-acids and B is the average number of neighbours ($\sim 6 - 7$).

Another solution which was adopted in order to save computational resources consisted in computing only a small fraction of the whole spectrum of the matrix. Indeed, our work focuses on large scale, collective displacements, which are mostly conveyed by the lowest energy modes of structural fluctuation.

Therefore, we resorted on numerical iterative methods to extract only the lowest

energy eigenvalues and eigenvectors. Specifically, we adopted the Krylov-Shur method [74] implemented in the SLEPc package.

4.2.2 Quasi-rigid domains subdivision

In this work, we used mechanical stability as a proxy for the identification of the functional domains. Specifically, we focused on the identification of suitable subdivisions in quasi-rigid units based on the mechanical characterization from the ENM.

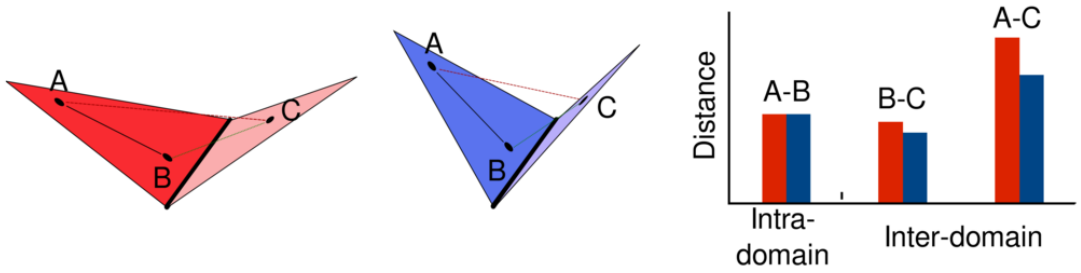

Figure 4.4: Schematic illustration of the relative motion of two rigid domains (triangles) joined by a hinge. Two possible instantaneous configurations are colored in red and blue. The graph on the right illustrates the distances of various pairs of points in the two configurations. The A-B intra-domain distance does not vary while A-C and B-C inter-domain distances do.

The subdivision of viral capsids in quasi-rigid domains is based on the PiSQRD strategy introduced in ref. [64, 1]. The typical starting point of domain decomposition strategies is the observation that for a genuinely rigid body, the distance of any two of its points remains constant as the body moves in space. For biomolecules this property cannot strictly hold because of the incessant structural fluctuations at the atomic or larger scales caused by thermal agitation, interactions with other biomolecules etc. Nevertheless, by suitable analysis of the equilibrium structural fluctuations of proteins or protein complexes it is usually possible to identify connected groups of amino acids whose pairwise distances are practically unperturbed compared to the global structural fluctuations, see figure 4.4. These groups are the sought quasi-rigid domains and their relative motion, consisting of quasi-rigid translations and rotations, usually accounts for most of the molecule's internal dynamics [64, 1, 52]. Interestingly, the quasi-rigid domains do not necessarily coincide with structural subunits of the molecule, thus providing non-trivial information about the relation between the architecture of the protein and its functional dynamics [64].

Accordingly, one can quantify the viability of a tentative capsid subdivision into Q putative quasi-rigid domains by comparing amino acids pairwise distance fluctuations within each domain with those across domains. For good subdivisions, the former should

be much smaller than the latter.

To turn this observation into a quantitative scheme amenable to numerical implementation, we define a cost function, which we call geometric strain, $f_{\beta\gamma}$, for a given pair of amino acids, β and γ as

$$f_{\beta\gamma} = \sum_{i=1}^{N} \left[\frac{1}{\lambda_i} \frac{(\mathbf{v}_{i,\beta} - \mathbf{v}_{i,\gamma}) \cdot \mathbf{d}_{\beta\gamma}}{\|\mathbf{d}_{\beta\gamma}\|} \right]^2 , \tag{4.3}$$

where $\mathbf{d}_{\beta\gamma}$ is the reference, native distance vector of the βth and γth amino acids, N is the number of extracted principal modes and $\mathbf{v}_{i,\beta}$ is the projection of the i-th mode on the three-dimensional subspace corresponding to the βth aminoacid.

N is chosen by retaining all the modes with energy lower than the fifth non-zero mode of a single coat protein, thus ensuring a sufficient level of detail while minimizing computational effort and discarding the mostly irrelevant high-frequency details.

Accordingly, the internal strain of the kth domain $\mathcal{D}_k$, is defined as

$$F_{\mathcal{D}_k} = \sum_{\substack{\beta \neq \gamma \\ \beta,\gamma \in \mathcal{D}_k}} f_{\beta\gamma},$$

where the sum runs over all the pairs belonging to the domain.

The overall strain is the sum of the strain of each of the Q domains

$$F\left[\{\mathcal{D}_k\}\right] = \sum_{k=1}^{Q} F_{\mathcal{D}_k}.$$

Based on previous considerations, the optimal subdivision in Q domains is the amino acid partitioning into Q groups that minimizes the overall strain F.

Notice that the optimization of F needs to be performed separately for all possible values of Q, that is from 2 up to the number of protein units forming the capsid. In fact, the "correct" number of quasi-rigid domains is not known *a priori* and needs to be found based on physical considerations, as will be explained later.

4.2.3 Minimization algorithm

Obtaining an optimal subdivision in quasi rigid domains, at a fixed value of Q, requires to minimize F by a suitable grouping of the amino-acids in Q domains.

The problem of minimizing the strain cost function F is equivalent to finding the

ground state of a Potts model with Q states and energy:

$$E = \sum_{i,j} \delta_{\sigma_i,\sigma j} f_{ij}$$

where σ_i indicates the state of the i-th bead and δ is the Kronecker symbol.

This is a combinatorial problem of size Q^N, thus unapproachable by exhaustive enumeration. We therefore resorted to a greedy algorithm for minimization. Each iteration consists in moving the domain boundaries, i.e. an amino acid close to a domain boundary is randomly selected and assigned to the neighbouring domain. The move is accepted if it leads to a decrease of F and rejected otherwise. The scheme is repeated until the algorithm is unable to further improve the solution, i.e. the count of systematically rejected moves is comparable with the total number of amino-acids.

The boundaries moves are only a small subset of the possible moves. However, comparison with the simplest possible optimization strategy (changing the assignment of a random amino-acid regardless of its position) show no significant differences in the resulting solutions. Domains are, in both cases, spatially compact, even if compact configurations are not explicitly imposed by the basic strategy. This further justifies *a posteriori* restricting the moves to local modifications of the domain boundaries. The convergence to a solution using domain boundaries is up to $\sim$160 times faster in the case of the smaller capsid considered, with respect to the completely random selection. Moreover, the relative efficiency is expected to grow with the capsid size.

To minimize the impact of getting trapped in local minima of F (whose landscape corrugation increases with Q) the greedy-optimization scheme is iterated if the distribution of the domains strain $F_{\mathcal{D}_k}$, $k = 1, \ldots, Q$ is highly heterogeneous (which could be a sign of a suboptimal, very asymmetric solution). Specifically, we first compute the average, μ, and standard deviation, σ, of the domains strain and check if one or more residuals $R_k = |F_{\mathcal{D}_k} - \mu|$ is larger than 3σ. If so, then the two domains with smallest strain are joined while the one with the largest strain is split in two. This amino acid reassignment clearly preserves the total number of domains, Q. The greedy minimization of F is next repeated. The procedure is iterated until convergence to a minimum which features a sufficiently homogeneous energy distribution or when the splitting/joining move is unable to improve the solution.

It is important to note that the capsid here is considered as a whole and we did not enforce any *a priori* knowledge of which amino acids belong to a single protein. Indeed, in principle mechanical domains can cut through proteins, for example when a rather loose loop tightly binds to a different block. The comparison between the mechanical and the proteins boundaries is done a posteriori, providing information on the reliability

of the subdivision itself.

4.2.4 Identification of the best putative building blocks

The algorithm for the subdivision in Q domains was applied on the viral capsid several times, varying Q between 2 and the total number of proteins in the capsid.

The identification of the value(s) of Q giving the subdivision in viable capsomeres is done by monitoring two physical quantities: protein integrity and the number of inequivalent capsomere types. They respectively account for the compatibility of the subdivision with the natural elementary units represented by the single proteins and for the self-similarity of the tiles, which is reflected in a low number of different tiles.

Given a subdivision in domains, an integrity parameter is defined for each protein. For a general subdivision, the amino acids of a protein can be assigned to a number of different domains. However, a good subdivision should preserve the integrity of the protein, i.e. almost the whole protein should belong to a single domain. We thus define the integrity score for a protein as the largest fraction of its amino acids assigned to a single domain. This quantity is then averaged for all the proteins, providing a score for the capsid subdivision.

We also computed the number of similar tiles identified by our subdivision by size inspection. Specifically, we define the size of the i-th domain as the number of amino-acids belonging to the domain itself; we then assigned domains to one tile type if their size is the same within 1% of the average size. The identification of 'good' subdivisions follows from the evidence of a peak in the integrity score. In addition, we expect that the corresponding subdivision is made using no more than a handful of inequivalent types of capsomeres.

4.2.5 Interlocking between capsomeres

To detect possible intertwinings between quasi-rigid units (e.g. due to swapped tails or subdomains of the parent proteins) we computed the *interlocking* parameter. Specifically, we considered separately the two termini of each protein in the capsid, namely the first and last twenty amino acids, and counted the number of amino acids which are assigned to a rigid domain that is different from the dominant one (i.e. the domain to which most of the protein amino acids belong). This calculation returns the number of interlocked amino acids for each terminus of each protein in the capsid. The numbers relative to the N and C terminals are averaged separately, and the largest of the two averages is taken as a measure of the interlocking of the quasi-rigid domains. In other words, if a quasi-rigid domain subdivision has interlocking number equal to 10, it means

that on average one protein has 10 terminal residues assigned to a different domain than its core. Clearly, it also implies that the other terminus has less than 10 interlocked amino acids.

4.3 Results

The scheme is applied to several viruses for which the atomic structure of the fully-assembled capsid is available. The considered set spans several capsids classes and sizes (from the T= 1 Tobacco-Satellite Mosaic Virus and Tobacco-Satellite Necrosis Virus which consist of 60 protein units, to the pT= 7 Murine Polyoma Virus which comprises 360 proteins). We will show that the capsomere identification scheme based on the quasi-rigid capsid decomposition can successfully pinpoint the correct functional units for a group of viruses for which both the assembly process and the structural transitions are known experimentally.

The method is next used to formulate predictions for the capsomeres of several other viruses for which no experimental indication exist yet. This set includes various instances of the T= 3 class for which neither symmetry considerations nor euristic arguments or visual inspection can unambiguously identify candidates for the capsomeres.

We accordingly performed quasi-rigid domain decompositions of several viral capsids for which the atomic structural data is publicly available [12], namely CCMV, MS2, STNV, STMV, as well as L-A, Pariacoto and Polyoma virus. The whole set covers various capsid geometries, namely T=1, pT=2, T=3, and pT=7, and spans a wide range of sizes, from the 60 proteins of STMV (with a total of 8,820 amino acids) to the 360 ones of Polyoma virus (totalling 129,060 amino acids).

4.3.1 Validation Cases

CCMV

We start by considering the cowpea chlorotic mottle virus, which is well suited for validation purposes because it has been extensively studied both experimentally [71, 22, 87, 44, 48] and computationally [76, 86, 77, 15, 47].

CCMV is an icosahedral RNA plant virus whose capsid is constituted of 180 chemically identical protein subunits assembled in the shape of a truncated icosahedron with T=3 geometry. The protein units adopt three different, quasi-equivalent conformations, conventionally denoted as A, B and C [22, 71]. As shown in figure 4.5, the A proteins are organised in groups of five around the five-fold symmetry axes, whereas the B and C proteins cluster alternately in groups of six around the three-fold axes. The pentamers

and hexamers are stabilised by the interactions between the N-terminal arms of the constituent subunits. These intra-capsomeres interactions are complemented by inter-capsomeres ones resulting from the mutual interlocking of the C-terminal arms and the β-barrel of neighbouring protein pairs in different capsomers [71].

According to various experiments these dimers correspond to the capsid assembly blocks for the virion [71, 87]. In the fully-assembled shell the dimeric units involve A/B and C/C pairs in a 2:1 ratio. It should be noted that for A/B and C/C dimers the relative positioning of the subunits (specifically their canting angle) is different. Indeed, the subunits interlocking provides a flexible hinge that, in response to suitable environmental conditions, allows the immature virion to expand[44]. This fact aptly clarifies that the assembly/disassembly units are not necessarily expected to have sufficient rigidity to become the fundamental mechanically-stable units in the assembled capsid [9].

Indeed, for CCMV various studies consensually indicate that these mechanical units correspond to the pentameric and hexameric capsomers [22, 71, 87, 48]. This conclusion can be drawn by considering the details of both the expansion process and the capsid response to nano-indentation. In fact, during the expansion produced by the hinge-motion of the dimers, the pentameric and hexameric capsomers rotate about their axis maintaining an internal quasi-rigid character [47].

In accord with this result, recent coarse-grained simulations of CCMV nanoindentation have demonstrated that mechanical failure occurs along the seams that bridge hexamers and pentamers, which remain largely undeformed by the application of mechanical stress [24, 15].

The above-mentioned phenomenology provides a clear context for benchmarking the proposed strategy for identifying mechanical units in viral capsids. Specifically, for CCMV it ought to return hexamers and pentamers, and not the dimers, as the primary quasi-rigid blocks.

We started by characterising the internal dynamics of CCMV by computing its collective low-energy modes of structural fluctuations and used the data to partition the capsid into a number of putative quasi-rigid units, Q, ranging from 2 up to 180 (the latter corresponding to the number of capsid proteins). The value of Q corresponding to the most plausible subdivision into functional units was found by assessing their compliance with the aforementioned *desiderata*: the preservation of protein structural integrity and the small number of structurally-inequivalent domain types.

To this purpose we computed and analysed the order parameters shown in figure 4.5. We start by discussing box B, which reports the profile of the protein integrity order parameter as a function of the number of imposed quasi-rigid domains, Q. The integrity parameter is evaluated by first computing for each protein the largest percentage of its

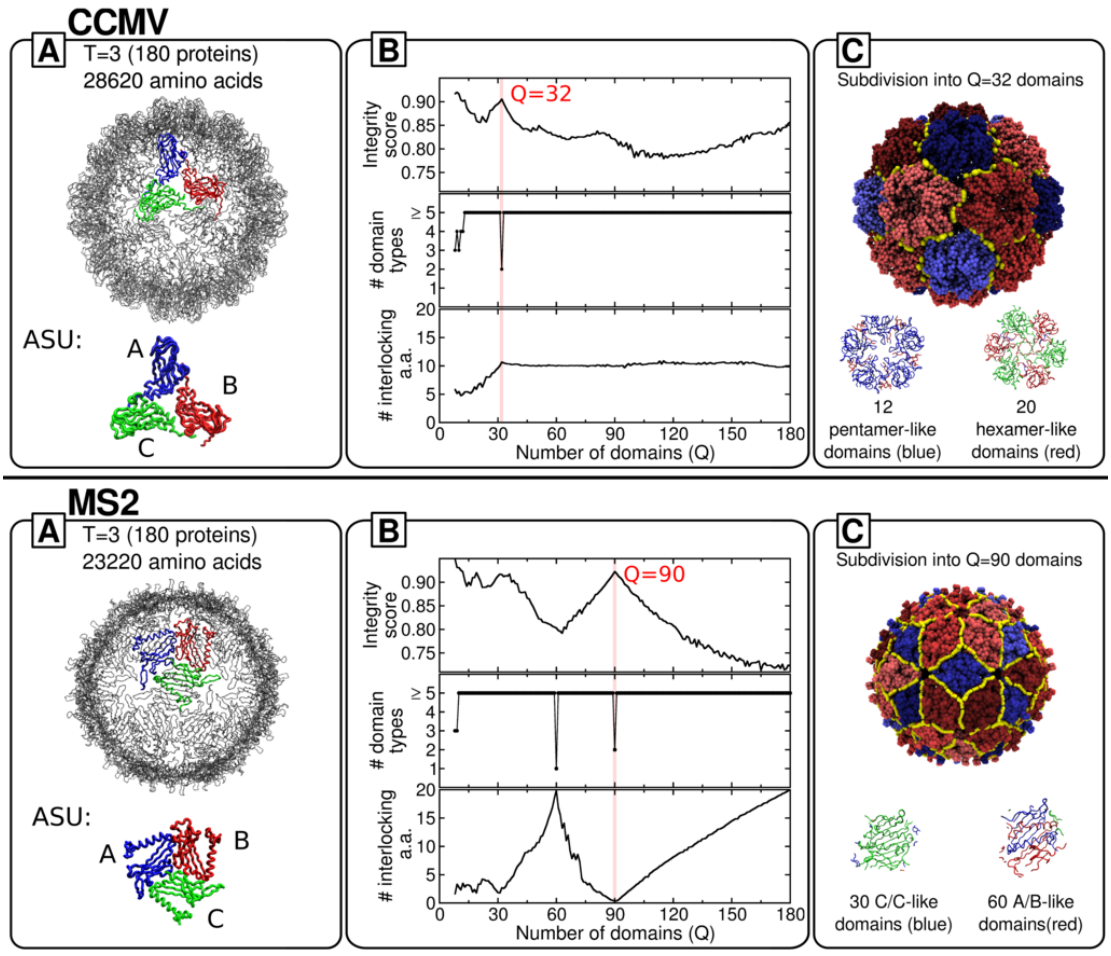

Figure 4.5: Decomposition into basic mechanical units of CCMV and MS2 viral capsids. Each left (A) box shows the capsid structure and its asymmetric structural unit (with distinct quasi-equivalent proteins highlighted in different colors). The middle (B) box shows the order parameters used to identify and characterize the optimal quasi-rigid subdivision. The latter is marked by the red dropline. The corresponding partition in basic mechanical units is represented in the rightmost (C) box. The yellow line marks the boundary between the mechanical units which, for both capsids, come in two different types and are colored in shades of blue and red, respectively. The relationship between the mechanical units and the structurally-inequivalent proteins is illustrated at the bottom of box C.

amino acids that are assigned to the same quasi-rigid block and next averaging this fraction over all proteins. Accordingly, an integrity score of 0.8 implies that, on average, 80% of the amino acids of any protein are in the same quasi-rigid block.

It is seen from figure 4.5 that there exists only one prominent peak of protein integrity (90%) corresponding to a subdivision into $Q = 32$ domains. Furthermore, throughout the considered range of subdivisions, $15 \leq Q \leq 180$, the strain-minimizing partition into $Q = 32$ domains is the only one yielding a limited number of inequivalent domains and can be readily singled out by visual inspection. Specifically, it involves only two distinct domain types, while a minimum of 5 to a maximum of 23 different types is found for all

other values of $Q \geq 15$. Lower values of Q, which correspond to subdivisions into very few macrodomains, are more obviously associated to both high integrity scores and few different domain types, see figure 4.6.

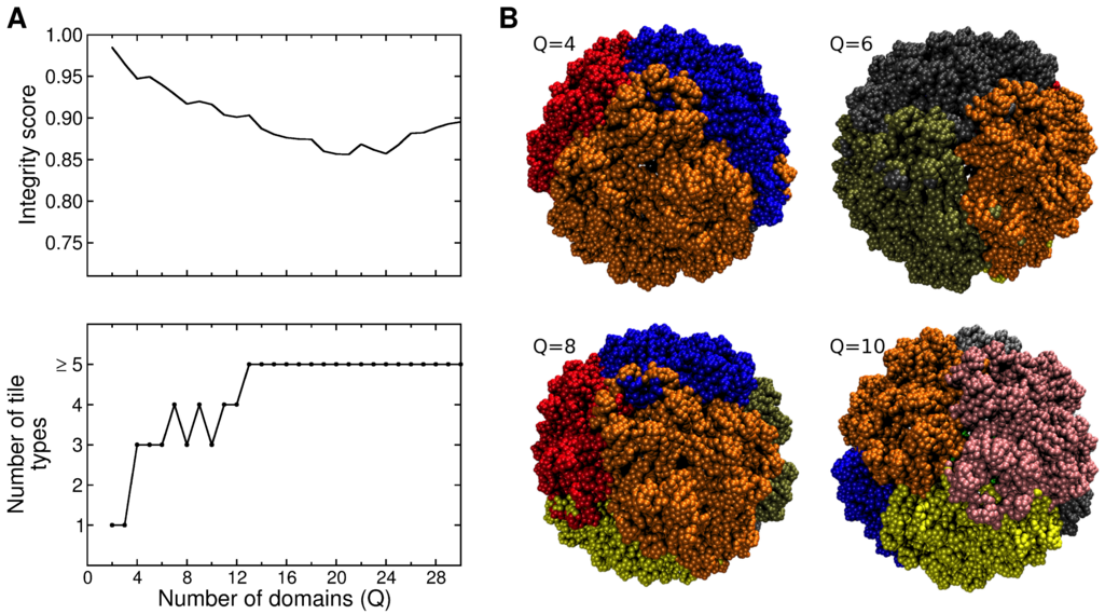

Figure 4.6: Panel A shows a close up of the CCMV profiles for the integrity score and number of tile types for subdivisions from $Q = 2$ up to 30 quasi-rigid domains. Panel B illustrates non-optimal quasi-rigid decompositions of CCMV. The subdivisions correspond to partitions into very few domains as indicated by the Q label. For each of these subdivisions the number of different tiles type is large and ranges from 3 to 4 . For simplicity we therefore used a different color for each domain rather than a different color per domain type.

The combined inspection of the integrity score and the domain types therefore provides a clearcut and non-ambiguous indication of the "innate" character of the CCMV capsid subdivision into 32 quasi-rigid domains which in turn can be grouped into only two structurally inequivalent types. The corresponding subdivision is shown in box C of figure 4.5, with the two domain types colored in shades of blue and red, respectively. The inspection of the subdivision shows that one domain type corresponds to pentameric units and the other to hexameric ones. There are 12 and 20 domains of each type, respectively. By considering the detailed structural representation of the two domain types, shown at the bottom of box C in figure 4.5, it is readily seen that they are, practically, an exact match of the hexameric and pentameric capsomers described before, the only difference being that the interlocked C-terminus is assigned to the "host" dimeric subunit and not to the parent one. The swapping of C-termini across the hexameric of pentameric units yields an integrity score smaller than 100%.

A further relevant parameter to consider for assessing the functional role of the subdivision is the degree of domain interlocking. The corresponding profile is shown in

the bottom graph of box B in figure 4.5 and portrays the average number of a protein's terminal amino acids assigned to a quasi-rigid domain which is not the one containing the protein core.

This parameter is monitored because several viruses, including CCMV, are assembled from protein dimers stabilised by the mutual interlocking of their termini which reach inside the partner protein core. The incidence of such interlockings *across* different quasi-rigid domains provides valuable clues regarding the relationship between the mechanically stable domains and capsid assembly/disassembly. In particular, the absence of cross-domain interlocking ought to be a good indicator that the mechanical domains are viable assembly/disassembly units too. The opposite should hold in case a significant amount of cross-domain interlocking is observed. It should, nevertheless, be borne in mind that cross-domain interlocking can arise after the assembly process.

For the case of CCMV, we observe that the degree of inter-domain interlocking for $Q = 32$ is non-negligible and, indeed, it reflects the above mentioned dimeric swapping of the C-termini between protein subunits. From the previous considerations, this fact indicates that the quasi-rigid hexamers and pentamers do not have the correct level of internal structural independence to be viable candidates as assembly or disassembly blocks. This conclusion is indeed correct given the known role of cross-domain dimers as assembly units.

In conclusion, the emerging quasi-rigid domain subdivision matches correctly the units identified by previous experimental and numerical studies.

Bacteriophage MS2

We next consider the MS2 virus, which is constituted by 180 chemically-identical coat proteins for a total of 23220 amino acids [26, 81]. As for CCMV, the protein units come in three structurally-inequivalent types (conformers), labelled A, B and C in box A of figure 4.5, which form interlocked A/B and C/C dimers and are assembled in a T=3 capsid geometry. However, the arrangement of these units is different: the asymmetric A/B dimer occurs in two groups of 5 around the 6 five-fold axes, and the symmetric C/C dimers are positioned on both ends of the 15 two-fold axes.

The results of the MS2 quasi-rigid domain subdivisions are illustrated in the upper panel of figure 4.5. The protein integrity profile shows one prominent peak corresponding to the subdivision into $Q = 90$ quasi-rigid blocks. These quasi-rigid units come in only two inequivalent types, as illustrated in box B. Detailed inspection of the subdivision reveals that these two types occur precisely in a 2:1 ratio and correspond to the C/C and A/B dimers, which are colored in shades of blue and red, respectively, in box

C. As before, this match of the mechanical domains and structural dimers must be understood with the proviso that protein integrity cannot be fully respected. In fact, amino acids at the boundary of quasi-rigid dimer domain are not necessarily assigned to their sequence-wise nominal dimer. As a result, although the whole A/B and C/C dimers would comprise exactly the same number of amino acids, the two types of quasi-rigid units are structurally diverse enough to be distinguishable by size, see box C in figure 4.5.

It is worth recalling that the MS2 capsid is in the same T=3 class as CCMV. Hence their very different number and types of fundamental quasi-rigid units points to the important role played by specific capsid proteins in shaping the properties and behaviour of viral capsids that are not large enough to be dealt with by continuum approaches [46]. One further major difference between the MS2 and CCMV optimal subdivisions is that the 90 units have a practically negligible degree of interlocking. Indeed, the interlocking profile has a minimum for $Q = 90$. This indicates that the small quasi-rigid units are structurally self-contained dimers. They are therefore viable candidates for being not only the fundamental mechanical blocks of the fully-assembled capsid but can be expected to be structurally-stable even in isolation and hence are also good candidates for being the assembly or disassembly units of the capsid. Indeed, this has been confirmed by isotope pulse-chase experiments [75]. In these experiments, protein subunits of dimers in complex with RNA are labelled differently from those in RNA-free dimers (via different isotopes) and both species are mixed. The fact that no dimers with differently labelled subunits are detected in solution or as part of any of the assembly intermediates suggests that the dimers do not fall apart into individual subunits and that hence the dimer is indeed the unit of assembly.

We emphasize that this *a priori* conclusion has necessarily a tentative character. In fact, because the method is based on the properties of fully-assembled protein shells, it cannot account for the interaction of coat proteins and genomic material during the assembly process. Such interaction can be crucial to aid the fast and correct assembly *in vivo* [56, 75, 34, 36, 18, 33, 45]. However, building on the fact that spontaneous *in vitro* assembly does occur in the absence of the genome, it appears plausible to consider non-interlocked quasi-rigid units as putative assembly units.

These considerations are fully supported by the successful comparison with experimental data for MS2. In fact, it has been established that the capsid is assembled from the A/B and C/C dimeric units [75], and the assembly pathways have been characterized in detail both experimentally and theoretically [17, 53]. In addition, the key role of the dimeric protein-protein interactions for the capsid stability has been indicated by thermal and pressure denaturation experiments [43].

In summary, the MS2 findings reinforce the CCMV indications that the innate functional units identified with the quasi-rigid domain decomposition correspond to those established experimentally.

STNV

The satellite tobacco necrosis virus has been one of the first to be determined at high resolution [42, 32]. With a diameter of only 17 nm, this T=1 RNA plant virus is one of the smallest known. The capsid is composed of 60 chemically and structurally identical coat proteins. Each of these consist of 195 amino acids and their N-terminal arm is positively charged[20, 41], a common feature in many plant viruses. In the fully-assembled, genome-loaded capsids (which are extremely stable [4, 41]) the N-termini interact with RNA loops, achieving charge neutrality. This interaction has been argued to favour an extended and ordered conformation of the N termini, which in turn aids the formation of trimeric capsomere units [20, 37, 11].

The quasi-rigid domain decomposition, whose results are reported in figure 4.7, returns an optimal subdivision for $Q = 20$ mechanical domains. These domains correspond to trimeric units that are monodisperse in size and do not have interlocked termini. This outcome is consistent with the assembly mechanism discussed above, which involves 20 trimers as basic assembly units [20].

A noteworthy implication is that the fundamental units of the assembly process, in which the RNA is known to play a major role, can be correctly identified through the quasi-rigid domain decomposition of the empty capsid. In this regard, it must be borne in mind that elastic network models guarantee by construction the stability of the model capsid for structural fluctuations of the crystal structure. Therefore, ENM approaches can make up for the missing stabilizing interaction of capsid proteins and the packaged nucleic acid. At the same time, the finding indicates that the mechanical-stability of the individual (non interlocked) assembly units is still discernible in the internal dynamics of the fully-assembled capsid.

This remarkable property shows *a posteriori* that even in cases where protein-nucleic acids interplay is important in the assembly process, the quasi-rigid domain analysis of the pure proteic shell can still give valuable clues about the assembly process.

STMV

It is interesting to compare the above analysis with the one for another plant virus, the satellite tobacco mosaic virus (STMV), which presents several similarities with the STNV [4, 41] including the T=1 arrangement of the 60 identical coat proteins (for a

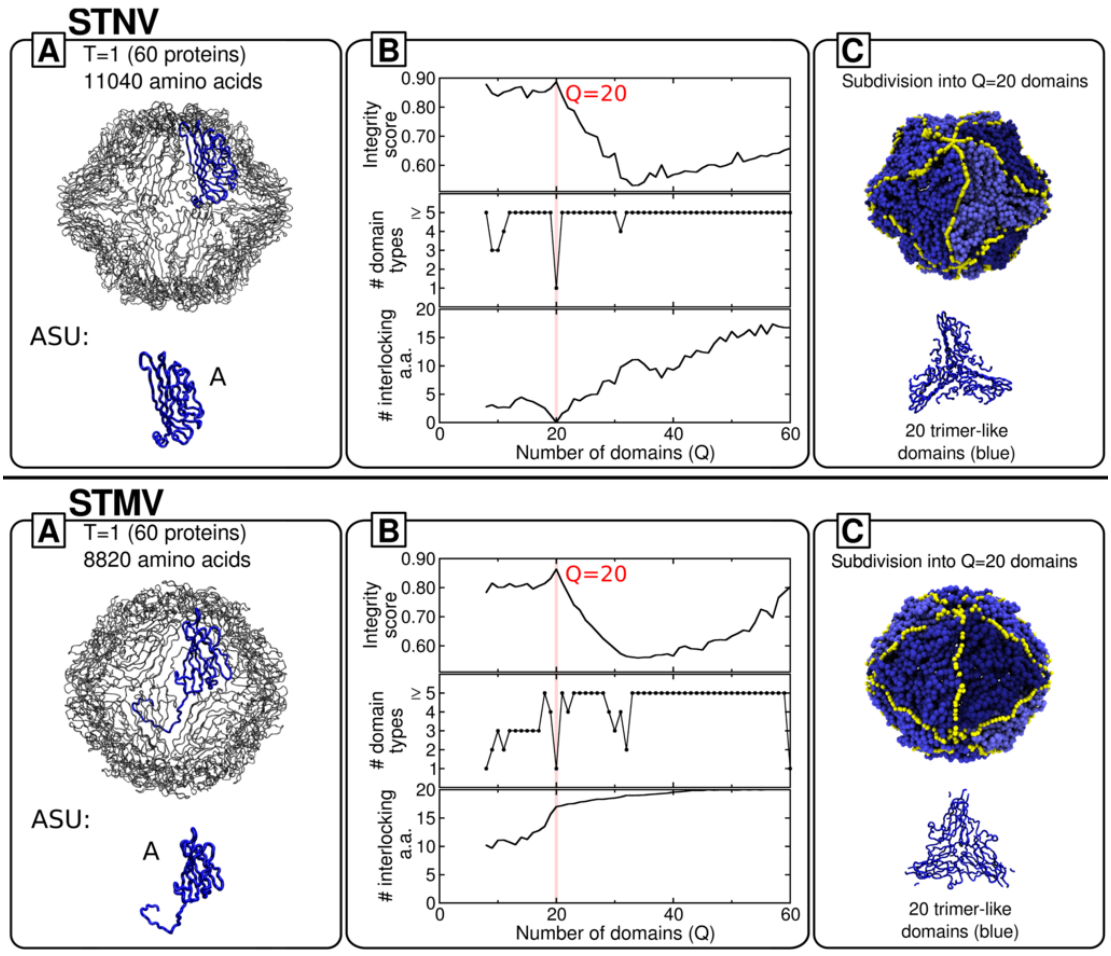

Figure 4.7: Decomposition into basic mechanical units of STNV and STMV capsids. Boxes A, B and C, show respectively the capsid structural organization, the profiles of various order parameters and the optimal decomposition into basic mechanical units. Colors and capsid representations follow the same scheme of figure 4.5.

total of 8820 amino acids) [38].

Because of its relatively small size, STMV has represented an ideal and natural reference for numerical investigations [23, 4]. To the present day, it remains the only virus for which all-atom molecular dynamics simulations have been performed on the fully-assembled capsid, both in the presence and in the absence of the genome[23].

This study as well as coarse-grained simulations [4] provided considerable insight into the internal dynamics of the capsid, its structural stability and resistance to nanoindentation. The consensus indication of these investigations is that the basic mechanical units are trimers of the coat proteins.

While this represents a further point of contact with the STNV, it should be noted that the similarity of their assembly process is still disputed. In fact, it is not yet understood whether the assembly proceeds as a condensation of a protein-RNA complex

[40] or if the collapse of the RNA in a globular state precedes and favours the formation of trimeric and pentameric units [23].

The results of the quasi-rigid domain decomposition of STMV are provided in the bottom panel of figure 4.7. The profiles shown in box B provide a clear indication that the basic rigid units correspond to monodispersed (identical) trimers.

This result is fully consistent with the previously mentioned computational studies of STMV structural stability, and also parallels the results of the related STNV case.

However, at variance with STNV, the analysis of the interlocking profiles shows that, at att values of Q, the trimers present a significant degree of interlocking originating from the interdigitating N-terminal arms of dimers that straddle domain boundaries. This difference to STNV is not surprising, given the lack of amino-acid homology or immunological cross-reactivity between STMV and STNV [39]. As previously discussed for CCMV and unlike STNV, the significant interlocking prevents from concluding that the trimers are plausible building blocks for the assembly of STMV.
As a matter of fact, McPherson *et al.*[39] suggest that the building blocks may be dimers that contact the genomic RNA at the particle 2-fold axes. This open issue could possibly be settled by establishing whether termini interlocking occurs before or after assembly. This information, which is at the heart of the ongoing debate on STMV assembly process, is clearly beyond reach of the present approach which is based only on the fully assembled capsid.

4.3.2 Predictions

We now turn to discuss three viruses for which the basic, mechanically stable functional units are not conclusively known. The following viruses are considered, chosen in order of increasing complexity of the capsid type (T-numbers): the L-A (pT=2), Pariacoto (T=3) and polyoma (pT=7) viruses. We recall that the pT=2 and pT=7 cases refer to non-standard Caspar-Klug geometries.

L-A virus

The L-A virus is a double-stranded RNA (dsRNA) yeast virus whose capsid is composed of 120 chemically-identical coat proteins with a total of 78120 amino acids. The proteins are assigned to two types, A and B, based on their inequivalent positions, see box A in figure 4.8. Similar to several other dsRNA viruses, the A/B dimers are arranged in a T=1 icosahedral capsid. This virus is classified as pT=2 to account for the fact that dimers occupy the positions of monomers in a T=1 structure [14, 54]. Stable empty capsids are observed *in vitro* and it has been suggested that A/B dimers are the

basic assembly building blocks[14]. By inspection one readily recognizes that the A/B asymmetric unit tiling the capsid can be defined in two inequivalent asymmetric ways (see figure 4.8). Because the two alternative pairings have a comparable buried surface area, it is not clear *a priori* which dimer type could be the basic assembly block. As we discuss hereafter, the quasi-rigid domain analysis can provide valuable insights into this open problem.

From the analysis of the graphs in box B of in figure 4.8 it emerges very clearly that the optimal subdivision is attained for $Q = 60$ identical quasi-rigid domains. Because of the high integrity score of this subdivision and the bipartite A/B capsid tiling, it follows that these basic mechanically-stable units necessarily correspond to A/B dimers which, furthermore, are negligibly interlocked.

This result is therefore fully consistent with the experimental indication of A/B dimers being the basic assembly units.

The notable point is that the quasi-rigid domain analysis discriminates very clearly between the two inequivalent asymmetric A/B dimers shown in box A, arguably because of their different networks of intra- and inter-dimer interactions. In fact, upon repeating the quasi-rigid domain partitioning into $Q = 60$ domains, one invariably observes that the strain-minimizing subdivision is the one shown in box C of figure 4.8. Given the robustness of this subdivision we predict that the A/B dimer shown in box C is basic assembly unit of the L-A virus.

Pariacoto virus

The Pariacoto insect virus belongs to the nodaviridae family and has a T=3 capsid [78, 31] constituted by 180 chemically identical coat proteins occupying three inequivalent positions. As shown in figure 4.8 the A units cluster around the five-fold axes while the B and C units are found at the three-fold axes. The capsid consists of 62760 amino acids in total.

While the C-terminal arm of each A protein is located in a channel formed by the A,B,C monomers at the quasi-3-fold axes, the N-terminal arms of the A proteins are involved in an extensive interaction with the encapsidated single-stranded RNA [78].

The inspection of the profiles in box B of figure 4.8 indicates that the optimal subdivision into mechanically-stable units is obtained for $Q = 60$. This partition corresponds to monodispersed, identical trimers, see box C. The other prominent peak for the much smaller number of $Q = 12$ subdivisions corresponds to multiples (pentamers) of these trimeric units. The trimeric units correspond precisely to the A, B, C complexes and their minimal degree of interlocking is suggestive of their role as basic assembly units

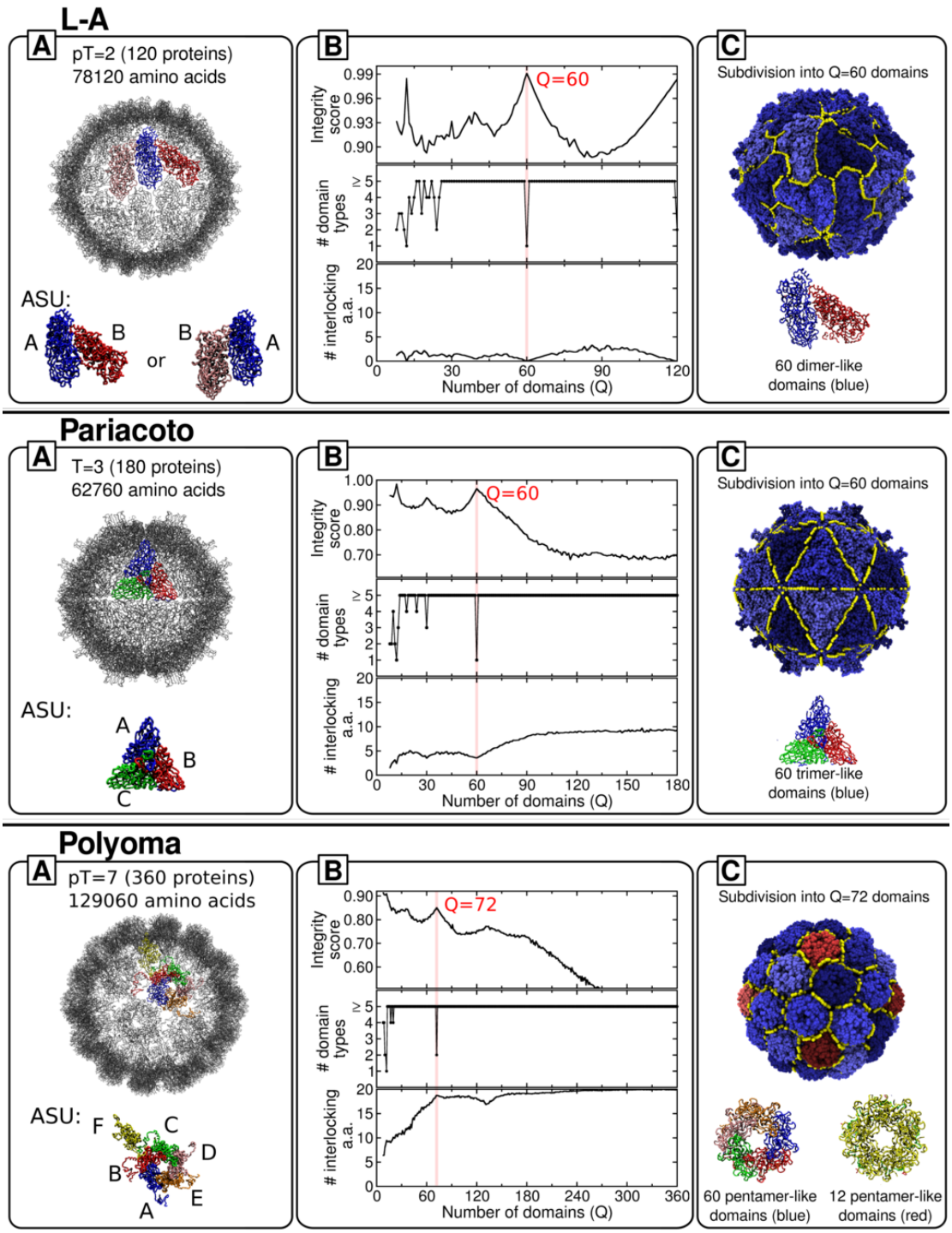

Figure 4.8: Decomposition into basic mechanical units of L-A, Pariacoto and polyoma viral capsids. Panels are organised as in figure 4.5. Boxes A, B and C, show respectively the capsid structural organization, the profiles of various order parameters and the optimal decomposition into basic mechanical units. Colors and capsid representations follow the same scheme of figure 4.5.

for the Pariacoto virus capsid.

The identification of a trimer of proteins as the first stage of assembly is also consistent with the theoretical work by Reddy [66], which is based on calculations of the buried surface area of the coat proteins.

Polyoma virus

We conclude the analysis with the discussion of the murine polyoma virus. This non-enveloped DNA virus has an icosahedral capsid with a pT=7 (non Caspar-Klug) geometry [73]. The shell consists of 360 copies of the main coat protein (VP1) for a total of 129060 amino acids, the largest capsid considered here. The asymmetric structural unit involves six identical coat proteins which are organised into pentameric clusters with structurally inequivalent bonding environments [68], see figure 4.8.

The peak structure of the integrity score profile indicates that the optimal subdivision involves $Q = 72$ rigid domains. As illustrated in box B of figure 4.8, these correspond to pentamers. More precisely two inequivalent types of pentamers are recognized by our approach. The pentameric units shown in box C are therefore expected to be the stable mechanical units for the capsid (though not the assembly ones because of the significant amount of interlocking).

This conclusion is reinforced by the analysis of the suboptimal subdivision into $Q = 12$ domains. These larger domains correspond to five-fold symmetric units made of a central pentamer surrounded by other five pentamers and hence give further support to the capsid's flexibility at pentamer-pentamer boundaries. This prediction could be verified by e.g. using molecular dynamics simulations to analyse the response of the capsid to nano-indentation.

4.3.3 Integrity at the subdomain level

We point out that measuring the integrity score at the level of entire proteins is appropriate for CCMV and the other viruses considered so far, because of the structural compactness of their constitutive proteins. However, when the proteins comprise two or more structural domains, the score can be straightforwardly generalised to capture the integrity of these subdomains.

One such example is given by the subdivision of the Hepatitis E virus-like particle in figure 4.9.

As it is shown in box A, each of the 60 coat proteins features three distinct structural subdomains, named S (a coat domain which composes the envelope for the genetic material), P1 (which forms a protusion around the threefold axis) and P2 (which forms

spikes on the two-fold axis). The optimal subdivision, corresponding to $Q = 50$ domains (coming in two distinct types) is identified by the peak in the integrity score calculated at the protein level and at subdomain level, see the black and blue curves respectively in box B. The fact that the peak of the subdomain integrity is much more prominent than for entire proteins indicates that the basic mechanical domains involve structural subunits from different proteins. This is clearly visible in box C which shows that one domain type corresponds to the spike (formed by the P2 subunits of two neighbouring coat proteins) while the other is a trimer involving the S and P1 subunits of three neighbouring coat proteins.

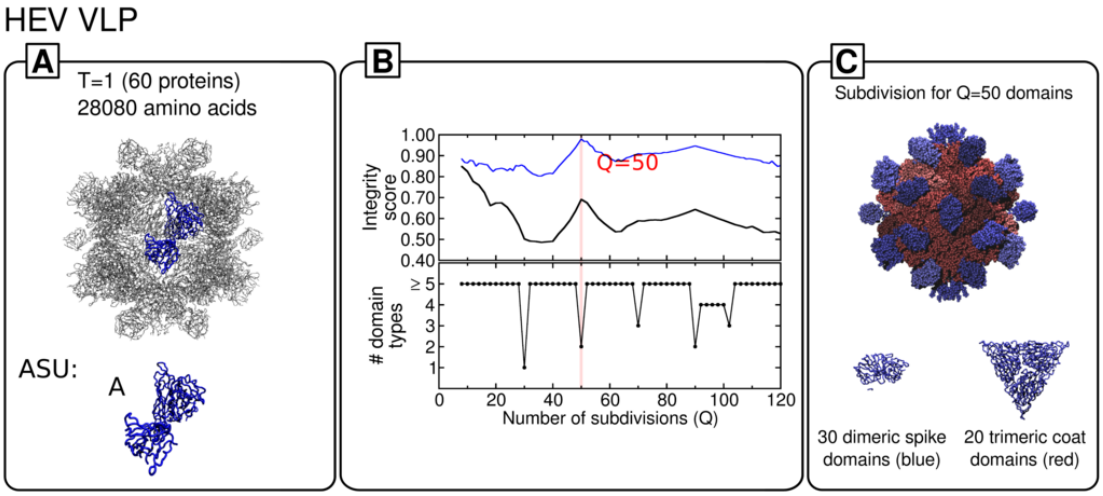

Figure 4.9: Decomposition into basic mechanical units of the HEV virus-like particle.

4.3.4 Strain profile and captured motion

The inspection of the integrity score and the number of domain types provide a immediate and convenient way to identify the correct, innate subdivision in quasi-rigid domains of viral capsids. While the subdivisions are based on the internal dynamics of the capsids, the scores explicitly build on the notion of chain connectivity and on the symmetry of the viral shells. That poses the question whether the optimal subdivisions, selected on the basis of the above considerations, feature some sort of "mechanical fingerprint".

The value of the geometrical strain cost function provide an intuitive and transparent order parameter to explore this issue. As explained in detail in the methods section, it intuitively conveys a measure of the difference between the dynamics described by the ENM model and the dynamics of a coarse grained description where each of domain behaves as a rigid body. The lower its value, the best the rigid body description approximate the "real" dynamics. An immediate property of the geometrical strain is its monotonic decrease when the number of domains Q is increased. Indeed, a description

in $n+1$ rigid bodies is more fine-grained, hence closer to the real model, than one in n rigid bodies. The geometrical strain is identically zero when Q is equal to the number of amino acids, i.e. when each amino acid is assigned to its own distinct domain.

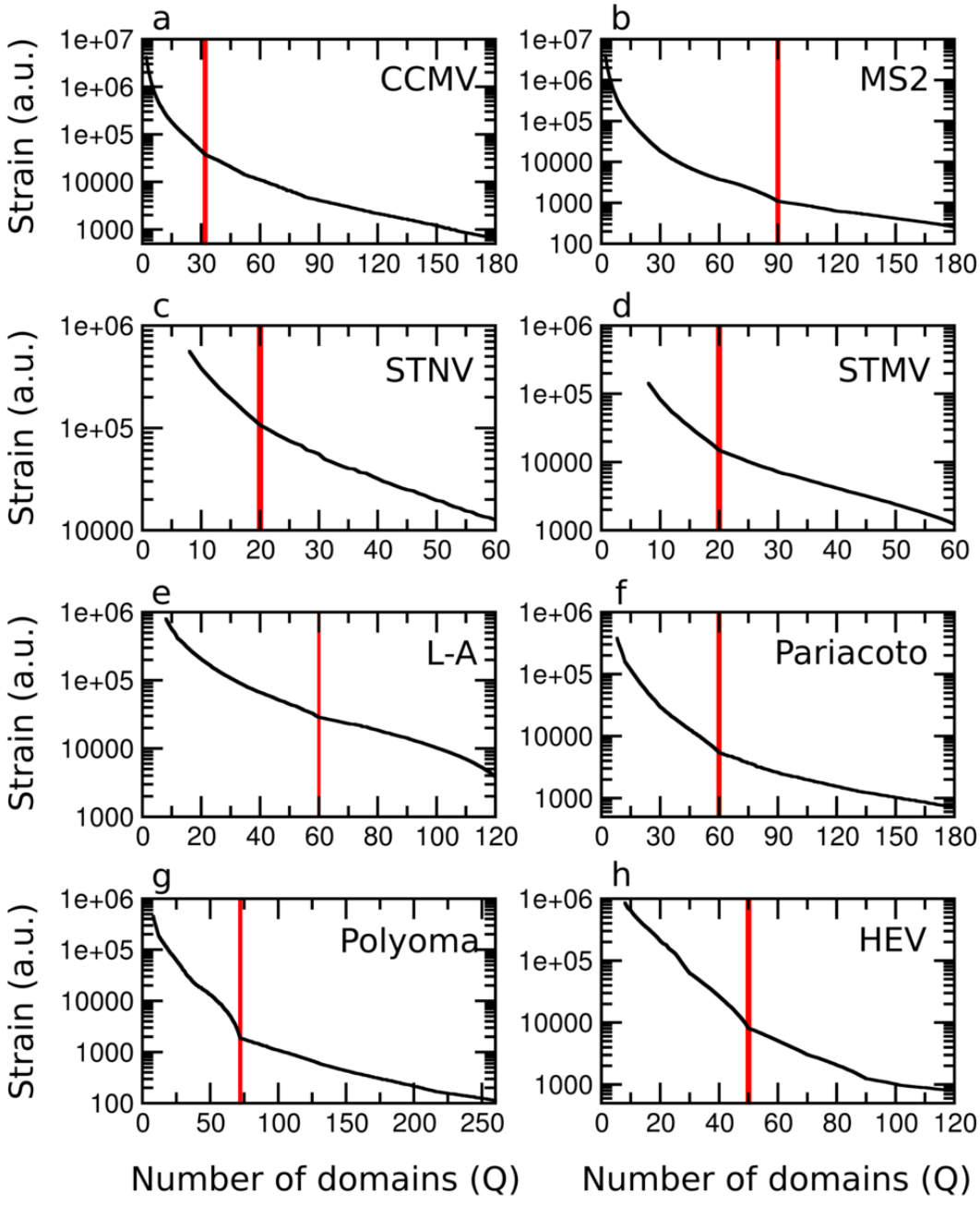

Figure 4.10: Q-dependence of the miminized geometric strain. Panels a-h refer respectively to: CCMV, MS2, STNV, STMV, L-A, Pariacoto, polyoma, HEV. Notice that at the value of Q corresponding to the optimal subdivision (highlighted by the red band) there is usually a kink. The latter signals the change of the slope of the strain curves when the "innate" number of subdivisions is crossed.

The profile of the strain for the considered capsids is depicted in figure 4.10, where the optimal subdivisions are indicated by a red line. It is pleasing that innate subdivisions are associated to visible kinks in the strain profile. The extent to which kinks are pronounced varies from case to case, conveying differences in the aptitude to a quasi-rigid description.

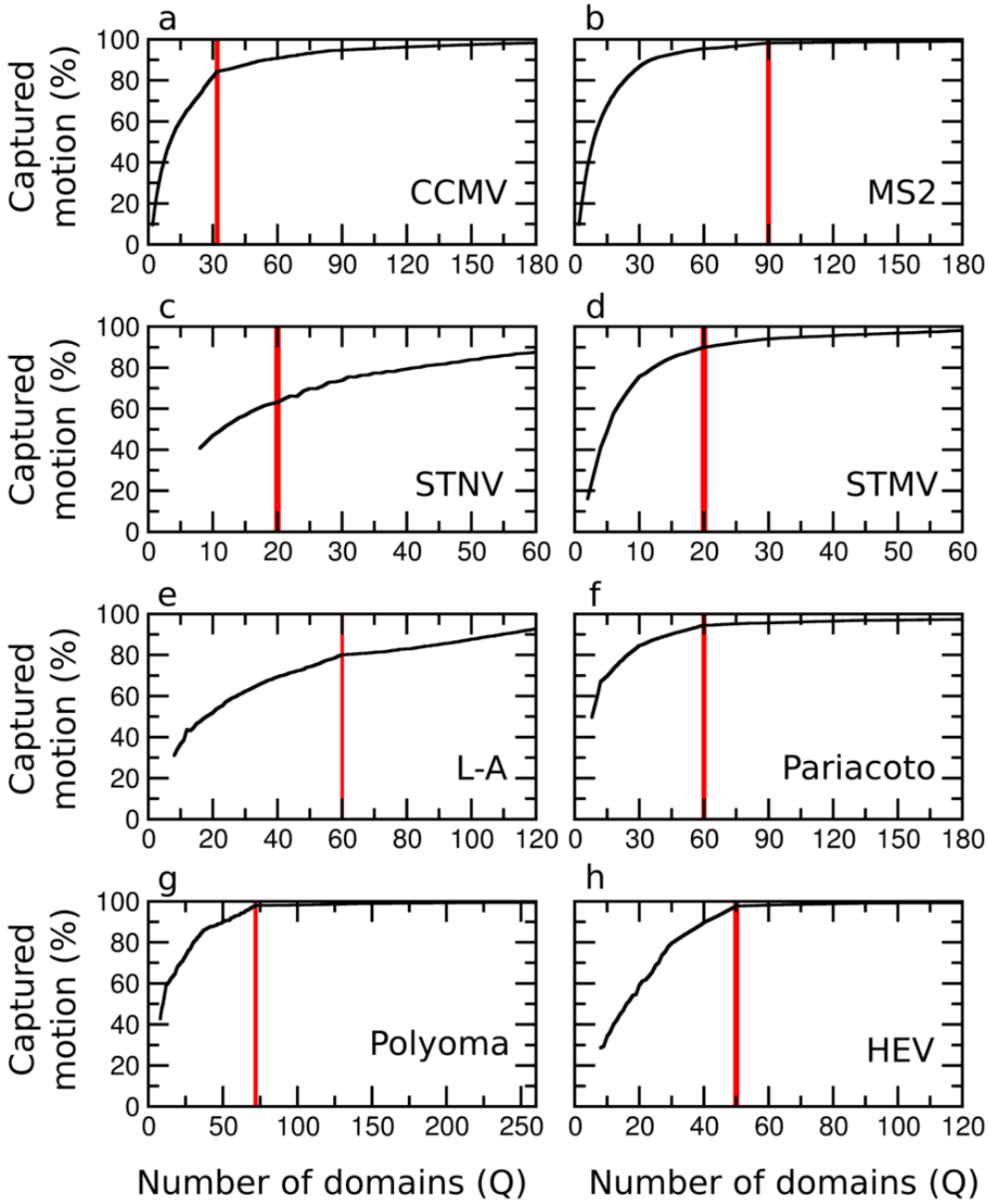

Figure 4.11: Fraction of overall capsid motion (mean square structural fluctuations) that can be ascribed to the pure rigid-like movements of the Q quasi-rigid domains. For each value of Q we considered the domain subdivision which minimizes the geometric strain.

A second, related, quantity we monitored is the fraction of motion captured by the rigid description. Indeed, it is possible to compute a best fit of the rigid domains motion on the fluctuations from ENM, and measure to what extent the rigid body description is a faithful one (details of the calculation are in the supplementary information of ref. [64]). The profiles of the captured motions for the considered capsids are shown in figure 4.11. The first thing to note is that the rigid like motion of the highlighted subdivisions account for a large fraction (typically higher than 80% and never below 60%) of the whole capsid structural fluctuations, confirming their genuine quasi-rigid character. Moreover, consistently with the strain profile, innate subdivisions are associated to kinks of the curves.

The change in the trends of the above measures indicates the presence of a mechanical mark of an innate subdivision. Whereas its identifications from these profiles is not unambiguous or strongly marked, this observation opens the possibility to select an innate partitioning by the analysis of the mechanical character of the returned domains, instead of resorting to specific parameters. In that case, symmetry and connectivity of the structure can be disregarded, leading to the applicability to other classes of protein assemblies. This kind of approach will indeed be the topic of the next chapter.

4.4 Conclusions

We introduced and applied a novel computational strategy that, to our knowledge, represents the first attempt to develop a general and efficient method for identifying the basic, mechanically stable protein units starting from the sole input of the fully-assembled protein capsid. The method relies on the characterization of the internal dynamics of the capsid by means of elastic network models and uses it to optimally decompose the protein shell into blocks that have the characteristics expected for genuine capsid functional units, such as mechanical stability (quasi-rigidity), structural integrity of the constitutive proteins and small numbers of inequivalent block types.

The viability of the scheme was first assessed and validated by considering a set of four viruses (CCMV, MS2, STNV, STMV) for which the fundamental functional units are known. In all cases, the results of the optimal decomposition scheme were fully consistent with available experimental or numerical results for the known mechanical and/or assembly protein units. We next turned to a further set of three viruses, namely polyoma, Pariacoto, L-A viruses, whose functional units are debated or not known, and for which we formulate verifiable predictions.

The positive validation of the method and its affordable computational cost (the first hundred ENM modes of the internal dynamics of capsids of about 60000 amino acids can

be obtained in $\sim$ 2 hours on a single Intel Xeon 2.40GHz core) demonstrates that simple structure-based strategies can provide considerable information on the basic functional units. In particular, they not only aid the understanding of various viral processes, such as assembly and structural modifications, but can also guide the development of their multiscale modelling.

Chapter 5

Spectral-based rigid units subdivision

5.1 Introduction

In the previous chapter I presented a computational method to identify mechanical units in viral capsids. It identifies optimal subdivisions in quasi-rigid blocks, based on large-scale structural fluctuations, as captured by ENMs. An important element of this approach is the supervised inspection of a set of order parameters purposely tailored for icosahedral viral particles, namely: the integrity score and the number of tile types.

The method is very effective in using these auxiliary parameters to pinpoint "innate" subdivisions in quasi-rigid domains with a limited computational burden. The idea that the mechanical properties can be used as proxies to identify functional units, however, is not limited to viral capsids. This paradigm has indeed been proved valuable in many cases, ranging from small proteins to protein assemblies [27, 28, 29, 25, 70, 85, 64, 1]. The development of a method which generalize our algorithm to include other classes of protein assemblies, for which the capsid-oriented order parameters are not applicable, naturally emerge as an interesting challenge.

To achieve this goal, a parameter with more general scope is needed. The key question is: is there a way to quantitatively pinpoint an optimal, *innate* subdivision of a macromolecule or of an assembly which is not characterized by a strong symmetry and a large number of repeated protein units?

The parameters introduced in chapter 4 serve to pinpoint an innate subdivision, i.e. a coarse grained description which retains the essential mechanical features of the protein by using the smallest possible number of degrees of freedom. Those parameters allow for overcoming the difficulty to identify a optimal subdivision based solely on

Structure *August 2015 issue cover.*

the geometrical strain profile (which measures the extent to which the real motion is accurately described by blocks moving relative to each other in a quasi rigid fashion, i.e. rotations and translations). Indeed, the strain profile is monotonically decreasing with the number of blocks: the best description, in that sense, is the trivial subdivision in one block per atom.

This poses the problem of obtaining a quantitative and general measure of the "goodness" of the mechanical description which is not monotonic with respect to the level of coarse graining, thus conveying the information on the mechanical properties of the specific protein assembly.

A second desirable improvement over the capsid strategy is to widen the kind and nature of mechanical information used to describe the internal dynamics of proteins. A coarse grained ENM description is a very convenient procedure when a single, very large crystallographic structure is available. For many systems, however, much more detailed mechanical descriptions can be obtained from all-atoms simulations, or from multiple structures which are well separated in the conformational space, overcoming the ENM limits of small fluctuations. Whereas the strain cost function can be, in principle, generalized to this different kind of information, this possibility was not implemented in our algorithm given the general lack of simulations or multiple structures for viruses.

In this chapter, I will present some results from a novel method, based on the spectral analysis of the matrix of distance fluctuations, that is able to overcome the above mentioned problems. Moreover, I will show that the method works as well for viral capsids, yielding the same innate subdivisions of the previous method. The material presented in this chapter is part of a collaboration with Enzo Carnevale (from Temple University) and Luca Ponzoni, a colleague student in SISSA. My specific contribution, that I will describe hereafter, was in shaping the strategy and specifically the algorithmic design, implementation of parts of the code and the creation of a web interface. We named the algorithm *SPECTRUS*, which stands for *Spectral-based rigid units subdivision.*

The study was featured as the cover article for the Structure issue of August 2015.

5.2 Methods

The quasi-rigid domain subdivision approach stands on the same premises we built on in the previous chapter, namely that, for a genuinely rigid body, the distances between any two of their constitutive points remain unaltered as the object moves in space. For protein domains this can hold only in an approximate way. Nevertheless, in most cases rigid displacements and rotations are actually very good approximations of the real motion of portions of the protein.

Key quantities for dynamical domain decompositions are pairwise distance fluctuations, which are defined as

$$f_{ij} = \sqrt{\langle (d_{ij} - \langle d_{ij} \rangle)^2 \rangle} = \sqrt{\langle d_{ij}^2 \rangle - \langle d_{ij} \rangle^2}.$$

The $\langle\rangle$ brackets denote averages computed over an ensemble of conformations. Here we shall limit considerations to the distance fluctuations of the C_α's, which are customarily taken as representatives of the amino acid.

The conformations in the trajectory can be snapshots of molecular dynamics trajectories or different conformers, or even configurations generated from an ENM. Therefore, the method can be applied in a more general way with respect to the previous one, which was based solely on ENM.

Large pairwise distance fluctuations are not compatible with a rigid motion. Accordingly, the spirit of the approach is to cluster together amino acids with small distance fluctuations and separate groups with large ones. The pairwise distance fluctuations can indeed be considered a *dissimilarity* measure between amino acids.

Based on the dissimilarity measure, we can introduce also a *similarity* matrix, σ, defined as

$$\sigma_{a,b} = e^{-f_{a,b}^2/2\bar{f}^2}$$

where $\bar{f}$ in the Gaussian weighting part is a typical measure for intra-cluster dissimilarities[82].

Here, $\bar{f}$ is computed as the average distance fluctuation of C_α pairs that are closer than 10Å. The introduction of a spatial cut-off avoids similarities between amino acids which are spatially distant. Therefore, the clustering is based on a substantially local similarity measure.

Note that, because of the Gaussian weighting and spatial cut-off, the similarity matrix is sparse, allowing the use of sparse routines in its manipulation.

In our framework, the subdivision in domains is made by performing a clustering of the amino acids based on this metric, assisted by a data preconditioning step.

5.2.1 Spectral clustering

The algorithm is based on the so-called *spectral clustering* method[82, 55].

The latter is one of the most popular modern clustering algorithms and outperforms, in most cases, classic methods as k-medoids [82].

This procedure is best explained in terms of graphs. Consider the data to be clustered as vertices of a graph, and the similarity measure between any two points as the

weight of the edge connecting them. In our case, the vertices correspond to the amino acids, while the similarity matrix defined in the previous paragraph provide the weights of the edges.

The data clustering problem is then recast as the problem of finding a partition of the graph in a way that distinct groups are connected by edges which have very low weights (meaning that amino acids in different groups are not very similar) while edges within a group have large weights (thus amino acids in a group show a strong similarity).

An alternative way to visualize it is a random walk performed on the graph, where the weights between two vertices convey the probability to jump from one to the other. In this case, we look for a partition where the walker will bounce around on a cluster for a long time, and very rarely jumps to a distinct cluster.

An efficient way to find a solution of the clustering problem in Q clusters is to compute the normalized Laplacian matrix of the graph and find its Q lowest eigenvectors (i.e. the eigenvectors corresponding to the smallest non-zero eigenvalues).

Starting from the similarity matrix for the N amino acids, we first compute the symmetric $N \times N$ Laplacian matrix L:

$$L = I - D^{-1/2}\,\sigma\,D^{-1/2}$$

where I is the identity matrix and D is a diagonal matrix with elements equal to $D_{a,a} = \sum_b \sigma_{a,b}$.

From the Laplacian matrix we then compute its Q lowest eigenvectors (i.e. those associated to the Q smallest eigenvalues), which we denote by $\vec{v}^1, \vec{v}^2, \ldots, \vec{v}^Q$.

The heuristic motivation for this procedure lies in the fact that these eigenvectors are indeed connected to the diffusion of the probability density on the nodes of a graph and, in particular, the lowest eigenvectors convey information on the slowest modes of relaxation on the graph.

Each of the N vertex (or amino-acid) is then associated with a point in a space of low dimensionality, namely a Q-hemispherical surface, constructed the Q eigenvectors. This projection operation naturally exposes the clustering properties of the graph as a geometrical segregation on the Q-hemisphere.

The construction of the points is performed constructing an auxiliary $N \times Q$ matrix X, whose columns are the Q lowest eigenvectors of L. Accordingly, the matrix entries are defined as $X_{i,j} = v_i^j$. Each of the N rows of X represent a vector which, when normalized, represents the coordinates of a point on the Q-dimensional unit sphere. These points provide the projection of the original N elements in the lower-dimensional spectral space. In particular, the n-th row is associated with the n-th vertex.

The spirit of this *preconditioning step* is easier to grasp considering an extreme case featuring Q disconnected sub-graphs. In this situation, each disconnected part will evolve independently. Accordingly, the supports of the lowest eigenvectors will be associated to different sub-graphs, as schematically shown in figure 5.1. The points on the Q-sphere corresponding to the i-th sub-graph will therefore have only one non-zero component, the i-th one, collapsing on a single point.

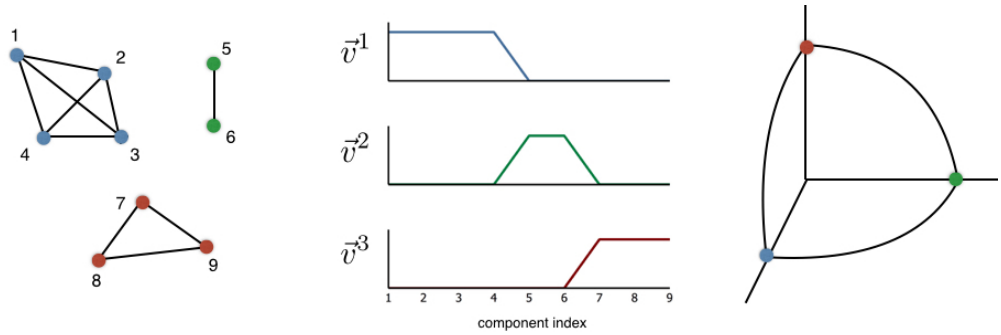

Figure 5.1: A schematic representation of the spectral clustering in three domains for disconnected sub-graphs. The point in the three-sphere surface corresponding to the i-th vertex has coordinates given from the i-th component of the eigenvectors. In this case of three disconnected sub-graphs, points from each sub-graph collapse on the axes.

If the graph is fully connected, we nevertheless expect to observe that the norm of the lowest eigenvectors will be concentrated on the sub-graphs with the least interconnections between them, geometrically exposing the clustering on the Q-sphere.

The last step is, therefore, to cluster the points on the Q-sphere according to their spatial distance. However, the choice of the clustering method is not crucial after the spectral preconditioning.

The first advantage of the spectral projection is, in fact, to make the subdivisions robust across various clustering schemes, due to the spatial confinement of "different" points in the Q-sphere. SPECTRUS uses a k-medoid algorithm because of its algorithmic simplicity and transparent formulation.

The k-medoid method selects randomly Q points to be the "medoids", the representatives of each cluster. Then associate each point to the closest medoid, and calculates a cost which is the sum of the distances of each point from its medoid. Then repeatedly tries to randomly change a medoid and accept the move if the cost decrease, reject it otherwise, until it reaches a solution. The algorithm is repeated with several starting conditions and selects the best result.

Since each point in the Q sphere correspond directly to one amino-acid, the mapping of the clusters of points to the quasi-rigid domains is immediate.

5.2.2 Subdivision evaluation

The above paragraphs explained how to obtain a subdivision in a fixed number Q of domains. Recall, however, that the most challenging problem is identifying the correct, "innate" number Q, which best describes the internal dynamics without introducing unnecessarily fine details.

As in the previous chapter, the selection of an innate subdivision is made in two distinct levels. In the first tier, an optimal subdivision in Q domains is performed for each possible value of Q. As a second level, a particular subdivision is selected from the optimal gropings obtained in the previous step. At this level, the quality of each Q-subdivision is quantitatively assessed using a single parameter, which we called *quality score*. It measures the ratio of the compactness of a clustering performed on the real data with respect to a clustering performed on random points on the Q-sphere.

The sharpness of a partitioning can be intuitively assessed by comparing the inter-cluster and intra-cluster distances. Well separated clusters will have very close points in groups positioned at a relatively large distance. If the points are spread in a diffuse cloud, instead, the distance of a point from its cluster centre can be comparable with the distance from another cluster.

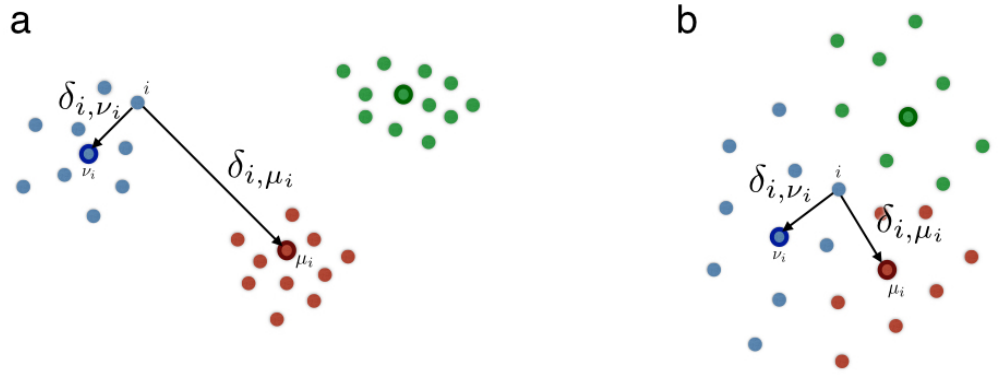

Figure 5.2: A schematic portrait of the inter-cluster δ_{i,μ_i} and intra-cluster distances δ_{i,ν_i} in equation 5.1, used to define the clustering quality score. The colour of the points represent the cluster assignment, and each cluster's medoid is highlighted by a solid border. Note that this is a purposely simple schema. Points used for a clustering in Q domains actually belong to the Q-dimensional unit sphere surface.

We measured the compactness and separation of the Q clusters as

$$\rho(Q) = \text{median}_{i=1,\dots,N}(\delta_{i,\nu_i}/\delta_{i,\mu_i}). \tag{5.1}$$

where i is the index of one element, μ_i is its representative medoid, i.e. the nearest of the medoids, ν_i is the second nearest medoid and $\delta_{i,j}$ indicates the distance on the sphere surface. Figure 5.2 give a schematic representation (on a flat surface, for simplicity) of

the terms appearing in equation 5.1 for a single element. In the panel **a**, a situation where the points are well separated shows that intra-cluster distances δ_{i,ν_i} are typically much smaller than the inter-cluster distances δ_{i,μ_i}. In this situation, $\rho(3)$ is significantly larger than 1. In the panel **b**, instead, the points are dispersed in a single cloud, and distances can be comparable.

In this way, we obtain a measure which conveys simultaneously the compactness of the clusters and their separation, probing the quality of the results from the clustering procedure. The use of the median in place of the average value confer robustness against the presence of outliers.

The *quality score* assigned to the subdivision in Q domains is then straightforwardly defined as this $\rho(Q)$ value normalized to the same measure calculated on a random cloud of points on the Q-sphere.

Representative subdivisions are expected to yield quality scores significantly bigger than 1. The observation of the quality score trend with respect to the number of domains Q is non-monotonic and informative on "innate" subdivisions of the protein, or protein assembly.

As a final assessment on the applicability of the quasi-rigid description, the quality-score profile is recalculated using the 40th and 60th percentile cut instead of the median in equation 5.1. The comparison of the three curves provides an indicative measure of the robustness of the maximum on the quality-score profile.

5.3 Results on capsids

In this section, I will summarize the results obtained by the application of the algorithm to two distinct icosahedral viruses, namely the Satellite Tobacco Mosaic Virus (STMV) and the Triatoma Virus (TrV).

Here, differently from the previous chapter, we do not use any ad-hoc parameter to assign a score to the subdivisions. The algorithm is fed solely with the matrix of distance fluctuations between atoms and do not rely on any notion of symmetry nor chain connectivity. In particular, the quality score presented in the previous chapter is a measure of the quality of the clustering and its calculation does not depend on the structure and spatial organization of the assembly under study.

In both cases, the distance fluctuation matrix was generated using an ENM. This choice was motivated by the presence of a single crystallographic structure and the impracticability of resorting to molecular dynamics for objects of this size.

STMV

The STMV was already considered in the previous chapter and serves here as a validation case. Our spectral decomposition method was applied to cover the range between $Q = 2$ and $Q = 80$ quasi-rigid domains. In this way, we considered a wide number of possible subdivisions, from coarser descriptions featuring just a couple of blocks to very fine ones, where every quasi-rigid unit is smaller than a single protein.

The quality score profile is depicted in figure 5.3b. The profile shows two peaks. The one at $Q = 60, 61$ roughly correspond the trivial subdivision in single proteins. The one at $Q = 20$ (depicted in figure 5.3c) is more informative, and correspond to the same subdivision in trimers observed in the previous chapter (figure 4.7).

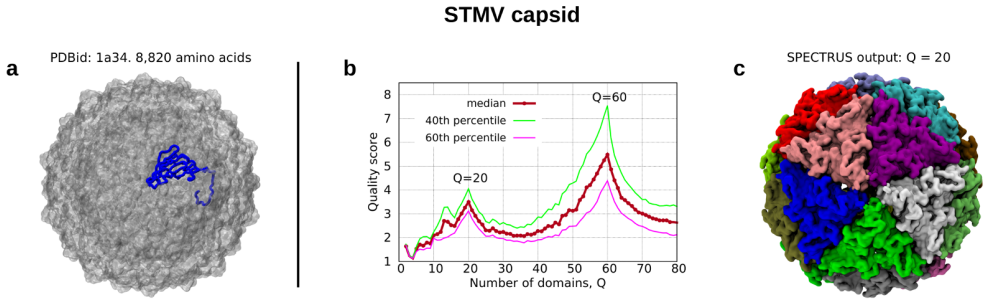

Figure 5.3: SPECTRUS results for STMV virus capsid. The non trivial optimal solution correspond to the peak at $Q = 20$.

This close correspondence between the subdivisions highlighted by the quality score profile and the ad-hoc parameters for the viral capsids demonstrate, on one hand, that the method is general enough to be applicable to macromolecular assemblies of this size. On the other hand, it shows that the distinction between viable and sub-optimal subdivisions can be based solely on the mechanical properties conveyed by the fluctuation matrix, without any additional symmetry or structural considerations.

TrV

The second virus we considered here was the Triatoma virus. Its $pT = 3$ capsid (in figure 5.4a) is composed by 180 proteins which come in three structurally-inequivalent types. The full structure counts a total of 47,220 amino acids[72]. The quality score profile, depicted in figure 5.4b, shows a single peak at $Q = 12$. The corresponding solution in figure 5.4c involves 12 pentagonal units, each composed of 15 proteins. Moreover, the noticeable prominence of this peak indicates a strong innate character of the subdivision.

As far as we know, the only experimental evidence on its basic building blocks comes from the observation of the debris produced by the rupture of viral particles caused

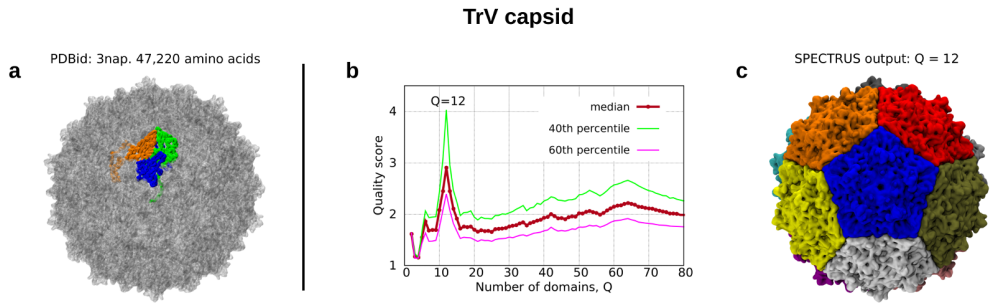

Figure 5.4: SPECTRUS results for Triatoma virus capsid. A definite peak shows an optimal solution corresponding to $Q = 12$ pentagonal tiles.

by AFM nano-indentation[69]. These observations concluded that the most plausible mechanical blocks are, indeed, pentagonal units.

This consensus, on one hand, reinforces from a distinct and independent perspective the results of Snijder *et al.* On the other hand, add confidence in the applicability of the method on large macromolecular assemblies for the identification of the innate mechanical building blocks.

The results presented here show that the method can successfully pinpoint the mechanical domains of viral capsid, overcoming the necessity of the ad-hoc parameters described in the previous chapter. The method thus present itself as a general scheme to identify mechanical subdivision of molecular assemblies in a unsupervised way.

Indeed, in addition to the viral capsid examples presented here, the method was successfully validated and applied to other contexts. As mentioned, these results have been studied by other collaborators, in particular dr. Ponzoni. For this reason, we only mention here the interesting results by the application to Adenylate kinase, GLIC and NavAB membrane channels, that an interested reader can find detailed in ref. [62].

5.4 SPECTRUS Web-server

In order to keep to a minimum the barrier towards using and adopting the SPECTRUS approach, improve the user experience and perform small test runs, I have designed and implemented a web server at the address http://spectrus.sissa.it.

Figure 5.5 shows the main page, where the user can either insert a list of PDB codes or upload a trajectory file for his protein.

If a single PDB code is provided, the software will use the beta Gaussian code to compute the normal modes and will utilize that information to generate the similarity matrix.

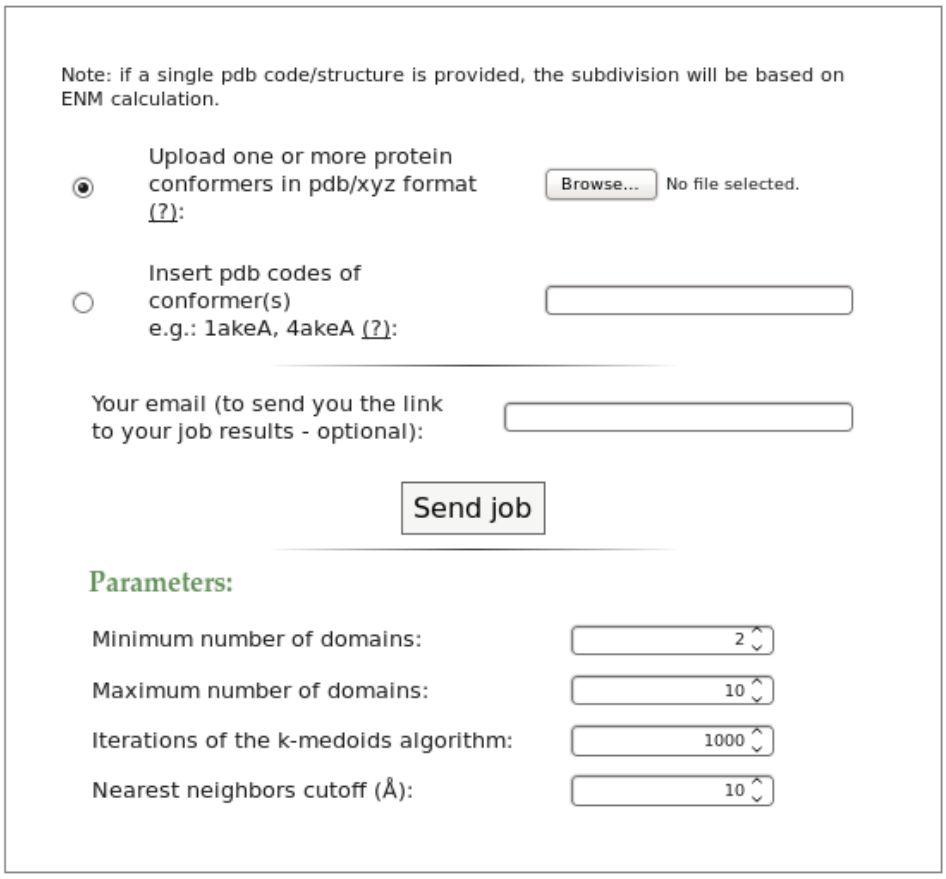

Figure 5.5: SPECTRUS Web Server interface. The user is required to provide the input configurations and the job parameters.

If more than a single PDB code is provided, the software will download the corresponding PDBs and perform a basic check on the number of C^{α} atoms. From each PDB only a single model will be read; hence, each PDB will provide a single configuration. Note that it is also possible to specify a chain identifier after the PDB code (for instance 1akeA). In this case only the specified chain will be used.

If the specified configurations do not have the same number of amino acids, the software will return an error. Note that the web-server will not perform any check on the amino acid type or correspondence. On one hand, this allows for the comparison of homologous proteins. On the other hand, the user should pay attention to provide already aligned PDBs.

The conformational data can be provided also by uploading a file. The user can choose to provide a zip archive containing PDB files or a raw trajectory in xyz format. The xyz format used here is a concatenation of frames in ascii format. Each frame consist of:

- a single line containing the number of atoms N;
- a comment line;
- N lines, one per C_{α}, containing the name of the atom (which will be ignored and assumed to be a C_{α}) and its three x, y, z Cartesian coordinates separated by spaces or tabs.

The xyz file can be easily generated from other trajectories using the VMD software. Note that xyz files do not retain bounding box informations, thus special care has to be taken to unwrap the proteins across the eventual periodic box boundaries.

Because of the current limits on computational resources, as today the web-server is able to process only small files and trajectories (up to XXX atoms and up to 50M). In order to process larger files there is a link to download the full version to be compiled and run locally.

In the same page , it is possible to regulate the most important parameters for the run. The user can set the minimum and maximum number of domains, the number of k-medoids iterations and the cut-off for the research of the nearest neighbours.

Finally, it is possible to be alerted when the job finishes by providing an email address.

Submitting the job will start the server-side computation and the user will be automatically taken to the summary page where it is possible to visualize the job status, the logs and, in case of a successfully completed job, the resulting clusterizations along with the associated quality score profile.

The results page provides a javascript interface from which it is possible to visualize the molecule and the subdivisions. The quasi rigid domains are highlighted with distinct colours in the visualization app.

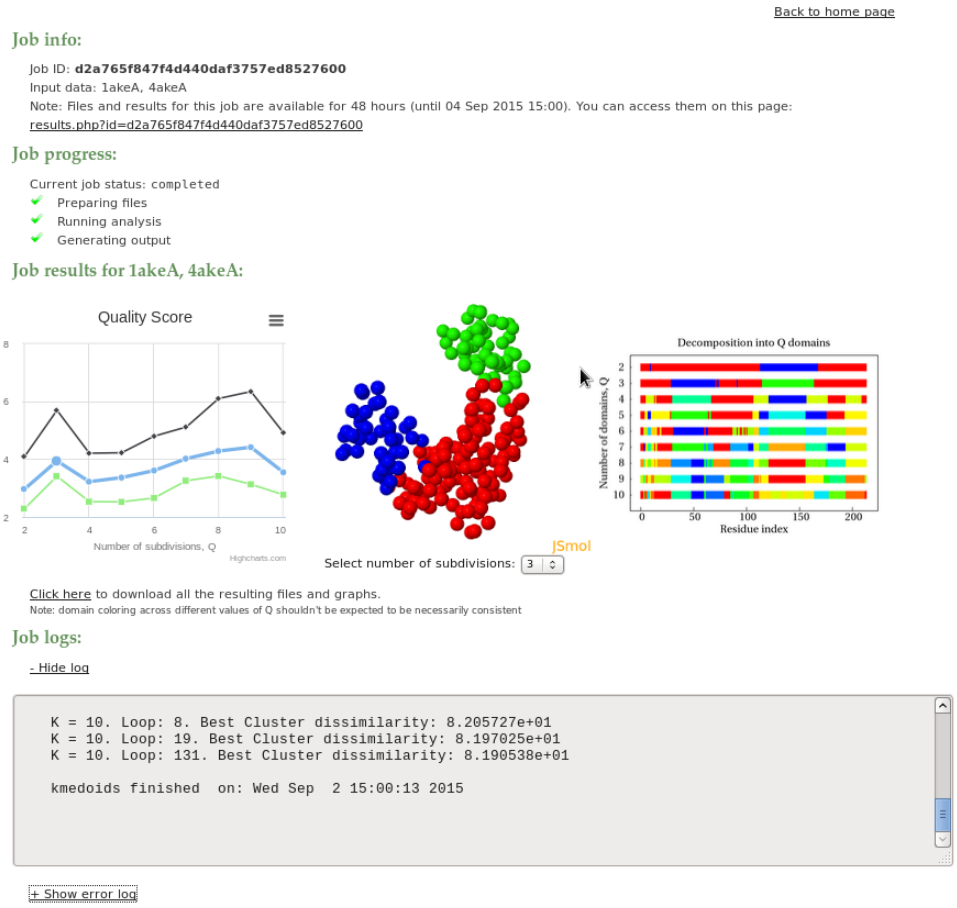

Figure 5.6: SPECTRUS Web Server results page

5.5 Conclusions

In conclusion, a novel general method was introduced, allowing the identification of innate subdivisions of proteins or protein assemblies in quasi-rigid domains.

The SPECTRUS algorithm utilizes the matrix of pairwise distance fluctuations of amino acids to generate a similarity measure between them. The algorithm then subdivides the full object in quasi rigid domains by clustering the amino acids based on this similarity measure.

A key step in the clustering procedure is the optimal dimensional reduction adopted from the spectral clustering method, in which the similarity matrix is projected in a space of lower dimensionality. The importance of this step is two-fold. On one hand,

it increase the robustness of the clustering, making it substantially independent on the clustering algorithm of choice. On the other hand, it allows for the comparison of the quality of the clustering with respect to a reference case, where the similarity is completely random. This comparison is based on a measure which takes into account both the typical distance of points within a cluster and the distance between closest clusters. The ratio between the subdivision and the random case led to the definition of a mono-dimensional, non-monotonic order parameter, named quality score, which conveys to which extent the returned subdivision is "innate" to the considered protein.

The observation of the quality score profile provides a clear indication on the best, non-trivial decomposition of a protein or protein assembly in quasi-rigid domains, while, at the same time, measuring the confidence on its applicability.

SPECTRUS is made available either as downloadable source code or as a web-server. The web-server, available for the subdivision of small proteins, present a user-friendly interface to both submission and the visualization of the results.

Aknowledgements

I would like to express my deepest gratitude to my advisor, prof. Cristian Micheletti, for his mentorship and continuous support. This work would not have been possible without his patient guidance.

I want also to thank the people I collaborated with: Enzo Orlandini, Davide Marenduzzo, Reidun Twarock, Giuliana Indelicato, Paolo Cermelli, Vincenzo Carnevale, Gianpaolo Gobbo and, last but not least, Raffaello Potestio. Not only the collaborations were scientifically fruitful, I also found kind support and enriching interactions. A particular mention goes to Luca Ponzoni for all the interesting discussions and fruitful cooperation.

My gratitude to all the people in the SBP sector in SISSA for the outstanding and friendly environment. Many thanks to Alessandro Laio and Giovanni Bussi for their teaching and helpfulness.

Many thanks also to all the people who supported me during these years. In particular, my deepest and warmest thanks go to my family, which always supported and encouraged me. Really, no place is like home.

Bibliography

[1] T. Aleksiev, R. Potestio, F. Pontiggia, S. Cozzini, and C. Micheletti. PiSQRD: a web server for decomposing proteins into quasi-rigid dynamical domains. *Bioinformatics*, 25:2743–2744, 2009.

[2] J. W. Alexander. Topological invariants of knots and links. *Transactions of the American Mathematical Society*, 30(2):275–275, 1928.

[3] A Arkhipov, P L Freddolino, and K Schulten. Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure*, 14(12):1767–1777, Dec 2006.

[4] Anton Arkhipov, Peter L Freddolino, and Klaus Schulten. Stability and dynamics of virus capsids described by coarse-grained modeling. *Structure*, 14(12):1767–77, December 2006.

[5] Jean François Ayme, Jonathon E. Beves, Christopher J. Campbell, and David a. Leigh. The self-sorting behavior of circular helicates and molecular knots and links. *Angewandte Chemie - International Edition*, 53(30):7823–7827, 2014.

[6] Jean-François Ayme, Jonathon E Beves, David A Leigh, Roy T McBurney, Kari Rissanen, and David Schultz. A synthetic molecular pentafoil knot. *Nat Chem*, 4(1):15–20, January 2012.

[7] T. S. Baker, N. H. Olson, and S. D. Fuller. Adding the third dimension to virus life cycles: Three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. *Microbiology and Molecular Biology Reviews*, 63(4):862–922, 1999.

[8] A. Ben-Shaul and W.M. Gelbart. Viral ssrnas are indeed compact. *Biophysical Journal*, 108(1):14 – 16, 2015.

[9] Tristan Bereau, Christoph Globisch, Markus Deserno, and Christine Peter. Coarse-grained and atomistic simulations of the salt-stable cowpea chlorotic mottle virus (SS-CCMV) subunit 26-49: β-barrel stability of the hexamer and pentamer geometries. *J. Chem. Theory Comp.*, 8(10):3750–3758, 2012.

[10] A. Borodavka, R. Tuma, and P.G. Stockley. Evidence that viral RNAs have evolved for efficient, two-stage packaging. *Proc. Natl. Acad. Sci. USA*, 109:15769 – 15774, 2012.

[11] Alexander Borodavka, Roman Tuma, and Peter G Stockley. Evidence that viral RNAs have evolved for efficient, two-stage packaging. *Proc. Natl. Acad. Sci. USA*, 109(39):15769–74, September 2012.

[12] M Carrillo-Tripp, CM Shepherd, IA Borelli, S Venkataraman, G Lander, P Natarajan, JE Johnson, CL Brooks, and VS Reddy. VIPERdb2: an enhanced and web API enabled relational database for structural virology. *Nucleic Acids Res*, pages 436–442, 2009.

[13] D. L. Caspar and a. Klug. Physical principles in the construction of regular viruses., 1962.

[14] J. R. Castón, B. L. Trus, F. P. Booy, R. B. Wickner, J. S. Wall, and A. C. Steven. Structure of L-A virus: A specialized compartment for the transcription and replication of double-stranded RNA. *J. Cell Biol.*, 138:975 – 985, 1997.

[15] M Cieplak and M O Robbins. Nanoindentation of virus capsids in a molecular model. *J. Chem. Phys.*, 132(1):015101–015101, Jan 2010.

[16] Ivan Coluzza, Peter D. J. van Oostrum, Barbara Capone, Erik Reimhult, and Christoph Dellago. Sequence controlled self-knotting colloidal patchy polymers. *Phys. Rev. Lett.*, 110:075501, Feb 2013.

[17] E. C. Dykeman, N. E. Grayson, K. Toropova, N. A. Ranson, P. G. Stockley, and R. Twarock. Simple rules for efficient assembly predict the layout of a packaged viral RNA. *J. Mol. Biol.*, 408:399 – 407, 2011.

[18] K M Elsawy, L S Caves, and R Twarock. The impact of viral RNA on the association rates of capsid protein assembly: bacteriophage MS2 as a case study. *J. Mol. Biol.*, 400(4):935–947, July 2010.

[19] Edward E. Fenlon. Open problems in chemical topology. *European Journal of Organic Chemistry*, pages 5023–5035, 2008.

[20] R.J. Ford, A.M. Barker, S.E. Bakker, R.H. Coutts, N.A. Ranson, S.E.V. Phillips, A.R. Pearson, and P.G. Stockley. Sequence-specific, RNA-protein interactions overcome electrostatic barriers preventing assembly of satellite tobacco necrosis virus coat protein. *J. Mol. Biol.*, 425(6):1050 – 1064, 2013.

[21] Ross S. Forgan, Jean Pierre Sauvage, and J. Fraser Stoddart. Chemical topology: Complex molecular knots, links, and entanglements. *Chemical Reviews*, 111(9):5434–5464, 2011.

[22] JM Fox, Guoji Wang, JA Speir, and NH Olson. Comparison of the native CCMV virion with in vitro assembled CCMV virions by cryoelectron microscopy and inage reconstruction. *Virology*, 218:212–218, 1998.

[23] P. L. Freddolino, A. S. Arkhipov, S. B. Larson, A. McPherson, and K. Schulten. Molecular dynamics simulations of the complete satellite tobacco mosaic virus. *Structure*, 14(3):437–449, Mar 2006.

[24] M M Gibbons and W S Klug. Influence of nonuniform geometry on nanoindentation of viral capsids. *Biophys. J.*, 95(8):3640–3649, Oct 2008.

[25] H. Golhlke and M. F. Thorpe. A natural coarse graining for simulating large biomolecular motion. *Biophys. J.*, 91:2115–2120, 2006.

[26] R. Golmohammadi, K. Valegård, K. Fridborg, and L. Liljas. The refined structure of bacteriophage MS2 at 2.8 Å resolution. *J. Mol. Biol.*, 234:620 – 639, 1993.

[27] S. Hayward, A. Kitao, and H. J. C. Berendsen. Model-free methods of analyzing domain motions in proteins from simulation: A comparison of normal mode analysis and molecular dynamics simulation of lysozyme. *Proteins: Structure, Function, and Genetics*, 27:425–437, 1997.

[28] K. Hinsen. Analysis of domain motions by approximate normal mode calculations. *Proteins*, 33:417–429, 1998.

[29] K. Hinsen, A. Thomas, and M. J. Field. Analysis of domain motion in large proteins. *Proteins*, 34:369–382, 1999.

[30] Robert E. Jensen and Paul T. Englund. Network News: The Replication of Kinetoplast DNA. *Annual Review of Microbiology*, 66:473–491, 2012.

[31] K.N. Johnson, L. Tang, J.E. Johnson, and L.A. Ball. Heterologous RNA encapsidated in Pariacoto virus-like particles forms a dodecahedral cage similar to genomic RNA in wild-type virions. *J. Virol.*, 78:11371–11378, 2004.

[32] T.Alwyn Jones and Lars Liljas. Structure of satellite tobacco necrosis virus after crystallographic refinement at 2.5 Å resolution. *J. Mol. Biol.*, 177(4):735 – 767, 1984.

[33] A Kivenson and M F Hagan. Mechanisms of capsid assembly around a polymer. *Biophys J*, 99(2):619–628, Jul 2010.

[34] M Kozak and D Nathans. Fate of maturation protein during infection by coliphage MS2. *Nature, nature new biology*, 234:209–211, 1971.

[35] K. Kremer and G. Grest. Dynamics of entangled linear polymer melts: A molecular dynamics simulation. *J. Chem. Phys.*, 92:5057–5086, 1990.

[36] Deborah A Kuzmanovic, Ilya Elashvili, Charles Wick, Catherine O'Connell, and Susan Krueger. The MS2 coat protein shell is likely assembled under tension: a novel role for the MS2 bacteriophage A protein as revealed by small-angle neutron scattering. *J. Mol. Biol.*, 355(5):1095–111, February 2006.

[37] Stephen W. Lane, Caitriona A. Dennis, Claire L. Lane, Chi H. Trinh, Pierre J. Rizkallah, Peter G. Stockley, and Simon E.V. Phillips. Construction and crystal structure of recombinant STNV capsids. *J. Mol. Biol.*, 413(1):41 – 50, 2011.

[38] S B Larson, J Day, A Greenwood, and A McPherson. Refined structure of satellite tobacco mosaic virus at 1.8 Å resolution. *J. Mol. Biol.*, 277(1):37–59, 1998.

[39] S. B. Larson, S. Koszelak, J. Day, A. Greenwood, J. A. Dodds, and A. McPherson. Double-helical RNA in satellite tobacco mosaic virus. *Nature*, 361:179–182, 1993.

[40] S. B. Larson and A. McPherson. Satellite tobacco mosaic virus RNA: structure and implications for assembly. *Curr. Opin. Struct. Biol.*, 11:59–65, 2001.

[41] Daniel S. D. Larsson and David van der Spoel. Screening for the Location of RNA using the Chloride Ion Distribution in Simulations of Virus Capsids. *J. Chem. Theory Comp.*, 8(7):2474–2483, July 2012.

[42] Lars Liljas, Torsten Unge, T.Alwyn Jones, Kerstin Fridborg, Seved Lövgren, Ulf Skoglund, and Bror Strandberg. Structure of satellite tobacco necrosis virus at 3.0 Å resolution. *J. Mol. Biol.*, 159(1):93 – 108, 1982.

[43] Sheila M B Lima, Ana Carolina Q Vaz, Theo L F Souza, David S Peabody, Jerson L Silva, and Andréa C Oliveira. Dissecting the role of protein-protein and protein-nucleic acid interactions in MS2 bacteriophage stability. *The FEBS Journal*, 273(7):1463–75, April 2006.

[44] H. Liu, C. Qu, J. E. Johnson, and D. A. Case. Pseudo-atomic models of swollen CCMV from cryo-electron microscopy data. *J. Struct. Biol.*, 142:356 – 363, 2003.

[45] J. P. Mahalik and M. Muthukumar. Langevin dynamics simulation of polymer-assisted virus-like assembly. *J. Chem. Phy.*, 136(13):135101, 2012.

[46] Ranjan V. Mannige and Charles L. Brooks, III. Periodic Table of Virus Capsids: Implications for Natural Selection and Design. *PLoS ONE*, 5(3):e9423, March 2010.

[47] Yinglong Miao, John E Johnson, and Peter J Ortoleva. All-atom multiscale simulation of cowpea chlorotic mottle virus capsid swelling. *J. Phys. Chem. B*, 114(34):11181–95, September 2010.

[48] JP Michel and IL Ivanovska. Nanoindentation studies of full and empty viral capsids and the effects of capsid protein mutations on elasticity and strength. *Proc. Natl. Acad. Sci. USA*, 103:6184–6189, 2006.

[49] C Micheletti, P Carloni, and A Maritan. Accurate and efficient description of protein vibrational dynamics: comparing molecular dynamics and Gaussian models. *Proteins*, 55(3):635–645, May 2004.

[50] Mark A. Miller, , and David J. Wales. Novel structural motifs in clusters of dipolar spheres: knots, links, and coils. *The Journal of Physical Chemistry B*, 109(49):23109–23112, 2005. PMID: 16375267.

[51] J. W. Milnor. On the total curvature of knots. *Annals of Mathematics*, 52:248 – 257, 1950.

[52] G. Morra, R. Potestio, C. Micheletti, and G. Colombo. Corresponding functional dynamics across the hsp90 chaperone family: insights from a multiscale analysis of md simulations. *PLoS Comput Biol*, 8, 2012.

[53] V. L. Morton, E. C. Dykeman, N. J. Stonehouse, A. E. Ashcroft, R. Twarock, and P. G. Stockley. The impact of viral RNA on assembly pathway selection. *J. Mol. Biol.*, 401:298 – 308, 2010.

[54] H Naitow, J Tang, M Canady, R B Wickner, and J E Johnson. L-A virus at 3.4 Å resolution reveals particle architecture and mRNA decapping mechanism. *Nat Struct Biol*, 9(10):725–728, Oct 2002.

[55] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[56] C Z Ni, Rashid Syed, Ramadurgam Kodandapani, John Wickersham, David S Peabody, and Kathryn R Ely. Crystal structure of the MS2 coat protein dimer: implications for RNA binding and virus assembly. *Structure*, 3(3):255 – 263, 1995.

[57] Enzo Orlandini and Cristian Micheletti. Knotting of linear DNA in nano-slits and nano-channels: A numerical study. *Journal of Biological Physics*, 39(2):267–275, 2013.

[58] S. J. Plimpton. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys.*, 117:1–19, 1995.

[59] Guido Polles, Giuliana Indelicato, Raffaello Potestio, Paolo Cermelli, Reidun Twarock, and Cristian Micheletti. Mechanical and Assembly Units of Viral Capsids Identified via Quasi-Rigid Domain Decomposition. *PLoS Computational Biology*, 9(11), 2013.

[60] Nandhini Ponnuswamy, Fabien B L Cougnon, Jessica M Clough, Dan Pantos, and Jeremy K M Sanders. Discovery of an organic trefoil knot. *Science*, 2012.

[61] Nandhini Ponnuswamy, Fabien B L Cougnon, G. Dan Panto, and Jeremy K M Sanders. Homochiral and meso figure eight knots and a solomon link. *Journal of the American Chemical Society*, 136:8243–8251, 2014.

[62] Luca Ponzoni, Guido Polles, Vincenzo Carnevale, and Cristian Micheletti. SPECTRUS: A Dimensionality Reduction Approach for Identifying Dynamical Domains in Protein Complexes from Limited Structural Datasets. *Structure*, 23(8):1516–1525, 2015.

[63] R. Potestio, T. Aleksiev, F. Pontiggia, S. Cozzini, and C. Micheletti. Aladyn: a web server for aligning proteins by matching their large-scale motion. *Nucleic Acids Res*, 38:41–45, 2010.

[64] R. Potestio, F. Pontiggia, and C. Micheletti. Coarse-grained description of protein internal dynamics: an optimal strategy for decomposing proteins in rigid subunits. *Biophys. J.*, 96:4993–5002, 2009.

[65] Thirumurugan Prakasam, Matteo Lusi, Mourad Elhabiri, Carlos Platas-Iglesias, John Carl Olsen, Zouhair Asfari, Sarah Cianférani-Sanglier, François Debaene, Loïc J. Charbonnière, and Ali Trabolsi. Simultaneous self-assembly of a [2]catenane, a trefoil knot, and a solomon link from a simple pair of ligands. *Angewandte Chemie - International Edition*, 52(38):9956–9960, 2013.

[66] V.S. Reddy and J.E. Johnson. Structure-derived insights into viral assembly. *Adv. Virus Res.*, 64:45 – 68, 2005.

[67] S. Y. Shaw, J. C. Wang. Knotting of a dna chain during ring closure. *Science*, 260:533–536, Apr 1993.

[68] D.M. Salunke, D.L. Caspar, and R.L. Garcea. Self-assembly of purified polyomavirus capsid protein VP1. *Cell*, 46:895–904, 1986.

[69] J. Snijder, C. Uetrecht, R. Rose, R. Sanchez-Eugenia, G. Marti, J. Agirre, D. Guérin, G. Wuite, A. Heck, and W. Roos. Probing the biophysical interplay between a viral genome and its capsid. *Nature chemistry*, 5(6):502–509, 2013.

[70] G. Song and R. L. Jernigan. An enhanced elastic network model to represent the motions of domain-swapped proteins. *Proteins*, 63:197–209, 2006.

[71] J. A. Speir, S. Munshi, G. Wang, T. S. Baker, and J. E. Johnson. Structures of the native and swollen forms of cowpea chlorotic mottle virus determined by X-ray crystallography and cryo-electron microscopy. *Structure*, 3:63–78, 1995.

[72] G. Squires, J. Pous, J. Agirre, G. S. Rozas-Dennis, M. D. Costabel, G. A. Marti, J. Navaza, S. Bressanelli, D. M. A. Guérin, and F. A. Rey. Structure of the *Triatoma virus* capsid. *Acta Crystallographica Section D*, 69(6):1026–1037, Jun 2013.

[73] T. Stehle and S.C. Harrison. Crystal structures of murine polyomavirus in complex with straight-chain and branched-chain syaliloligosaccharide receptor fragments. *Structure*, 4:183–194, 1992.

[74] G. W. Stewart. A krylov–schur algorithm for large eigenproblems. *SIAM J. Matrix Anal. Appl.*, 23(3):601–614, March 2001.

[75] P. G. Stockley, O. Rolfsson, G. S. Thompson, G. Basnak, S. Francese, N. J. Stonehouse, S. W. Homans, and A. E. Ashcroft. A simple, RNA-mediated allosteric switch controls the pathway to formation of a T=3 viral capsid. *J. Mol. Biol.*, 369:541 – 552, 2007.

[76] Florence Tama and Charles L. Brooks, III. The mechanism and pathway of pH induced swelling in cowpea chlorotic mottle virus. *J. Mol. Biol.*, 318(3):733–47, May 2002.

[77] Florence Tama and Charles L. Brooks, III. Diversity and identity of mechanical properties of icosahedral viral capsids studied with elastic network normal mode analysis. *J. Mol. Biol.*, 345(2):299–314, January 2005.

[78] J. Tang, M. N. James, I. N. Hsu, J. A. Jenkins, and T. L. Blundell. Structural evidence for gene duplication in the evolution of the acid proteases. *Nature*, 271:618–621, 1978.

[79] M M Tirion. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys. Rev. Lett.*, 77(9):1905–1908, Aug 1996.

[80] Luca Tubiana, Anže Lošdorfer Božič, Cristian Micheletti, and Rudolf Podgornik. Synonymous mutations reduce genome compactness in icosahedral ssrna viruses. *Biophysical Journal*, 108(1):194 – 202, 2015.

[81] K. Valegård, L. Liljas, K. Fridborg, and T. Unge. The three-dimensional structure of the bacterial virus MS2. *Nature*, 345:36 – 41, 1990.

[82] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[83] V.V. Rybenkov, N. R. Cozzarelli, A. V. Vologodskii. Probability of dna knotting and the effective diameter of the dna double helix. *Proc. Natl. Acad. Sci. USA*, 90:5307–5311, Jun 1993.

[84] G. M. Whiteside and M. Boncheca. Beyond molecules: Self-assembly of mesoscopic and macroscopic components. *Proceedings of the National Academy of Sciences of the United States of America*, 99(8):4769–4774, 2002.

[85] A. Zen, V. Carnevale, A. M. Lesk, and C. Micheletti. Correspondences between low-energy modes in enzymes: dynamics-based alignment of enzymatic functional families. *Protein Sci*, 17:918–929, 2008.

[86] D Zhang, R Konecny, NA Baker, and JA McCammon. Electrostatic interaction between RNA and protein capsid in CCMV simulated by a coarse-grain RNA model and a Monte Carlo Approach. *Biopolymers*, 75(4):325–337, 2004.

[87] A. Zlotnick, R. Aldrich, J. M. Johnson, P. Ceres, and M. J. Young. Mechanism of capsid assembly for an icosahedral plant virus. *Virology*, 277:450 – 456, 2000.