universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

„Theory and simulations of designable modular bionic proteins“

verfasst von / submitted by

Dott. Chiara Cardelli

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Doktorin der Naturwissenschaften
Doktorin der Natural Sciences (Dr. rer. nat.)

Wien, 2018 / Vienna 2018

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 796605411
Dissertationsgebiet lt. Studienblatt / field of study as it appears on the student record sheet:	Physik
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Christoph Dellago

Contents

Abstract

Self-assembly is the process by which a set of building blocks spontaneously reaches a stable pre-determined ground state with well defined structure. The large number of potential applications, such as active response to the environment [1] or catalysis, makes self-assembling a most researched feature in material science.

Heteropolymers are important examples of self-assembling systems [2,3], where the structure is determined by the sequence along the chain of an alphabet of building blocks. The technology for the synthesis and manipulation of heteropolymers is already advanced [4]. However, so far it has not been possible to design systems that both form precise structures and have a high variability of these structures, comparable to what proteins (natural heteropolymers) can achieve.

In this doctoral thesis, I demonstrate through computational studies over a broad spectrum of polymer geometry and alphabets that directional interactions allow to design heteropolymers with precision and variability comparable to proteins.

I introduce a computational method to measure the number of accessible compact configurations of a polymer to verify, following the predictions of the mean field theories for protein freezing [5], that the role of the directional interactions is to reduce such number, that is connected to the minimal

alphabet necessary to design the polymer.

Moreover, I introduce a simple and experimentally accessible criterion based on the radial distribution function [6], to discriminate a priori designable from not-designable polymer backbones. I show that this criterion is fulfilled both by for natural proteins than for our designable heteropolymers.

The latter two tools can guide the engineering of new types of self-assembling heteropolymers that will open new applications for polymer-based materials science.

Finally, I present a computational study of designed heteropolymer conformations with complex knotted topologies. The unprecedented self-assembled knotted structures can then be externally locked simply by controlling the interaction between the end monomers, paving the way to applications in the design and synthesis of active materials and novel carriers for drug delivery.

Zusammenfassung

Self-assembly ist der Prozess, bei dem ein Satz von kleinen Bausteinen sich spontan zu einer vorbestimmten, stabilen Struktur zusammensetzen. Die zahlreichen potentiellen Anwendungen, wie Katalyse oder aktive Reaktion auf die Umwelt [1], machen self-assembly zu einem der aktivsten Forschungsgebiete in den Materialwissenschaften.

Heteropolymere sind wichtige Beispiele für solche self-assembling Systeme [2, 3]. Sie sind Ketten, die aus kleineren Bausteinen zusammen gesetzt sind, die einem sogenannten Alphabet entnommen werden. Die räumliche Struktur in der sich ein Heteropolymer anordnet, ist durch die Reihenfolge der Bausteine entlang der Kette bestimmt.

Die Technologie für die Synthese und Manipulation von Heteropolymeren ist bereits fortgeschritten [4]. Es ist jedoch noch nicht möglich Systeme zu designen die viele unterschiedliche Strukturen präzise formen können, wie das Proteine (natürliche Heteropolymere) tun.

In dieser Dissertation demonstriere ich durch computergestützte Studien eines breiten Spektrum von Polymergeometrien und Alphabeten, dass direktionale Wechselwirkungen es erlauben, Heteropolymere mit Präzision und Variabilität vergleichbar derer von Proteinen zu designen.

Mean Field Theorien für Protein freezing [5] zeigen, dass es, um Heteropolymere zu designen, notwendig ist, die Anzahl der zugänglichen Konfig-

urationen zu reduzieren, die mit dem minimalen Alphabet zugänglich sind. Ich führe eine Methode ein, die zeigt, dass direktionalen Wechselwirkungen diese Rolle erüllen.

Weiters, führe ich ein einfaches und experimentell zugängliches Kriterium basierend auf der *radial distribution function* [6] ein, das designbare von nicht-designbaren Polymer-backbones unterscheidet. Ich zeige, dass dieses Kriterium sowohl von natürlichen Proteinen als auch von den von uns untersuchten designbaren Heteropolymeren erfüllt wird.

Die beiden letztgenannten Resultate sind Leitlinien für die Entwicklung neuer Arten von self-assembling Heteropolymeren, die neue Anwendungen für die Polymer-basierten Materialwissenschaften eröffnen.

Am Schluss, stelle ich eine Studie von designbaren Heteropolymerkonformationen mit komplexen verknoteten Topologien vor, die den Weg zu Anwendungen in der Konstruktion und Synthese von aktiven Materialien, sowie zu neuen drug-delivery Methoden ebnen können.

Danksagung

I owe my deepest gratitude first of all to my supervisors Ivan Coluzza and Christoph Dellago for their continuous encouragement and support in these years, and for giving me the possibility to participate in an international research community and a stimulating scientific environment. I thank my colleagues Valentino Bianco, Francesca Nerattini and Luca Tubiana, who together with Ivan Coluzza formed a very united and collaborative group, where every idea and doubt where fruitfully discussed as a team. I thank Carina Karner and Francesca Nerattini for the both scientific and human support through all these years. I thank Lorenzo Rovigatti and Francesco Sciortino for the uniquely insightful scientific discussions. Finally, I thank all friends and family and in particular Flavio for the moral and practical support through these years.

Preface

The following research articles are contained in this work.

- **Chapter 1**: Chiara Cardelli, Valentino Bianco, Lorenzo Rovigatti, Francesca Nerattini, Luca Tubiana, Christoph Dellago, and Ivan Coluzza. The role of directional interactions in the designability of generalized heteropolymers. *Scientific Reports* 7, 4986 (2017). I.C. designed the research, C.C. performed the simulations, I.C. V.B. and C.C. performed the data analysis. All the authors wrote the manuscript and discussed the research.

- **Chapter 2**: Chiara Cardelli, Francesca Nerattini, Luca Tubiana, Valentino Bianco, Christoph Dellago, Francesco Sciortino, and Ivan Coluzza. General methodology to identify the minimum alphabet size for heteropolymer design. Submitted to *PNAS* in date 11/05/2018. I.C. and F.S. designed the research, C.C. performed the simulations, C.C. I.C. performed and discussed the data analysis. All the authors wrote the manuscript and discussed the research.

- **Chapter 3**: Chiara Cardelli, Luca Tubiana Valentino Bianco, Francesca Nerattini, Christoph Dellago, and Ivan Coluzza. Heteropolymer design and folding of arbitrary topologies reveals an unexpected role of alphabet size on the knot population. Submitted to *Physical Review X*

(PRX) 23/05/2018. I.C. designed the research, C.C. performed the simulations, C.C. and V.B. adapted the simulation methods, C.C. and L.T. performed and discussed the data analysis. All the authors wrote the manuscript and discussed the research.

Introduction

The topic of this thesis is in general terms the approach of learning from nature, through computer simulations, to create new self-assembling materials. Self-assembly is one of the most researched subjects both in material science and in biophysics. In the process of self-assembly, building blocks spontaneously reach a stable ground state with organised spacial arrangement, driven by weak (non-covalent) interactions. However, the term does not have a unanimous formal definition in literature. Therefore, different processes, from the self-organisation of molecules into supramolecular structures to the growth of semiconductor quantum dots on solid substrates, have been called self-assembly. Here, we follow the definition in Ref. [7], restricting the term to "processes that involve pre-existing components (separate or distinct parts of a disordered structure), are reversible, and can be controlled by proper design of the components".

Self-assembly was studied originally in bio-chemistry, where it plays a fundamental role in many biological processes, since it is the way nature builds most of its complex structures. Examples are the folding of proteins into their native 3D structures, or the assembly of nucleic acids – RNA and DNA – into their functional forms (e.g. t-RNA).

By observing nature, it becomes clear that self-assembly can have great potential also in engineering artificial materials. In fact, by simply designing

an appropriate set of building blocks, a desired 3D structure can self-assemble without external intervention. This feature makes the process extremely versatile and low-cost, and thus promising for large-scale production of materials. Moreover, since the desired structure is the thermodynamically stable ground state of the system, self-assembled systems spontaneously avoid defects and thus the desired structure is obtained with a high precision. Another characteristic of self-assembly is reversibility, a property not generally possessed by materials constructed with other methods, that potentially enables to create materials sensitive to perturbations and able to adapt to changes in the environment. Finally, self-assembly can be hierarchical, as demonstrated by nature through DNA strands that can, in turn, self-assemble into bigger structures up to chromosomes, or different protein chains that assemble into complex quaternary structures (e.g. hemoglobine).

Self-assembly of artificial heteropolymers

Self-assembly can occur with building blocks of different length scales, from the nanoscale, including molecular self-assembly, to the microscale. At the nanoscale, self-assembly has many possible applications ranging from sensing and imaging devices, syntheses of 3D photonic crystals for novel optical computers or photovoltaic cells, to nanomachines and drug delivery [8].

At the mesoscopic and microscopic scale, notable self-assembly examples are realised from colloidal particles, such as colloids with different exotic shapes [9], colloids coated with DNA strands [10] colloidal particles with "patches" (anisotropic points) [11, 12]. In particular, the latter *patchy particles*, assemble both computationally and experimentally into patterns with various complex three-dimensional architectures (see following reviews [13–

16] and the works in Ref. [12, 17–21]), and allow the direct experimental observation of self-assembly in real time using confocal microscopy [22].

However, in the examples mentioned above as well as in most other self-assembling systems, in order to obtain a different final structure, pattern, or material, it is necessary to construct a new set of building blocks. Differently in nature, by only using the same set – also called alphabet– of 20 amino acids, proteins can fold into a huge number of different 3D structures. Similarly, nucleic acids assemble into their different functional forms by using only four different nucleotides.

Both proteins and nucleic acids are heteropolymers: the final 3D self-assembled structure is determined by the sequence of the building blocks along a polymer chain. The identification of the sequence that drives the polymer to fold into a chosen target structure is normally referred to as *design*. In what follows, we define as *designability* the property of a heteropolymer to have at least one sequence that reliably folds into a unique target structure [5]. The 3D structural information is determined by a 1D sequence information, always using the same alphabet of building blocks. This strategy is very convenient for nature, because it limits the number of building blocks from nature to synthesize (or acquire from outside), and for the same reason it would also be a very convenient approach to the fabrication of artificial materials. DNA was already taken as inspiration for novel artificial materials [23], such as DNA-origami [24] or DNA molecular spiders and nanorobots [25, 26]. However, DNA assembles in functional forms that fluctuate among similar degenerate structures more with respect to proteins in their native structures, since it does not need a fixed precise structure for its natural purposes. Instead most proteins, with the exception of a few ones called disordered proteins, fold into their unique 3D structures and maintain

it fixed with very high precision up to 1 Å [27] (as root mean square displacement RMSD from the native structure). Such high precision is fundamental for most protein functions. For instance, many proteins serve as catalyzers (enzymes), and thus must have a precise shape of their active sites to ensure specificity in the bonding of the desired ligand. In addition, the amino acid alphabet is extremely versatile, since with only 20 letters it is possible to construct the 139807 known proteins in the Protein Data Bank [27]. Both this high variability of possible very different structures and the high precision of the self-assembled structure make proteins an ideal source of inspiration for novel materials. However, up to know it has not been possible to construct artificial heteropolymers with such precision and variability.

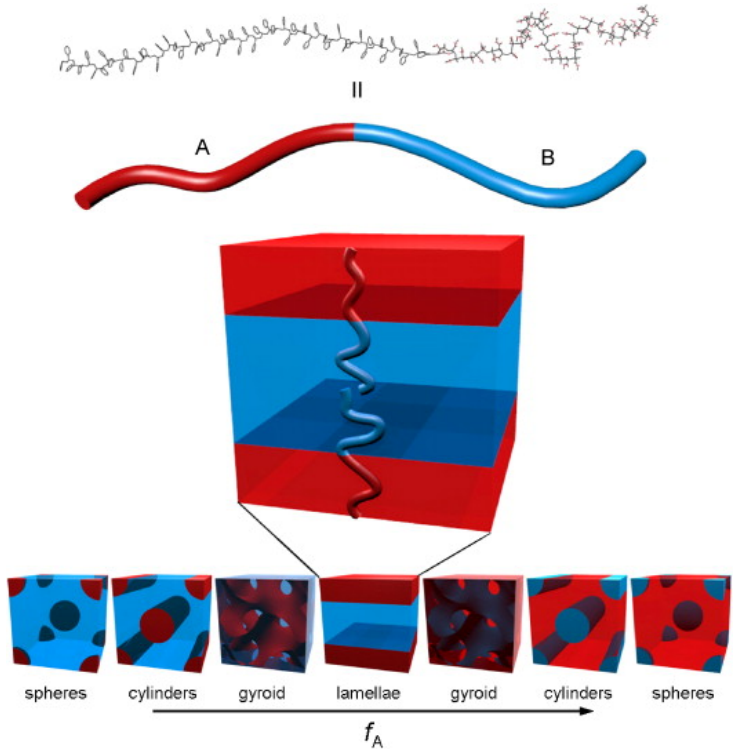

Figure 1: Example of AB block copolymer microphase separations. Figure reproduced from Ref. [28].

Nevertheless, in material science artificial heteropolymer synthesis and manipulation is at an advanced stage. For instance, it is possible to synthesize block copolymers, polymers consisting of blocks of different monomers or, in other words, with sequences of long blocks of repeating monomers.

Commonly block copolymers are formed by two different chemical blocks (AB copolymers) and are synthesised though different techniques that control the length of the different blocks, allowing low polidispersity. Triblocks [29], tetrablocks [30, 31], or star polymers – linear chains connected to a central core – with up to 7 blocks [32] can also be made. In block copolymer melts the blocks of different components tend to segregate, forming microphase separated structure with various morphologies such as layer, cylindrical or globular domains, depending on composition and temperature [4] (see Fig. 1). However, they do not adopt a specific structure of each single chain.

Recently, efforts to overcome this limitation have been made [33–37]. In particular, in the computational work presented in Ref. [2], Khokhlov and collaborators designed AB copolymer sequences to obtain microphase-separated structures with better definition and larger characteristic length scales with respect to simple symmetric AB copolymers. Another contribution in this direction is given by Moreno et al. [3], in which a similar micro-phase separation is obtained by designing specific hydrophilic-hydrophobic patterns. However, contrary to proteins, these two methodologies do not provide control over the detailed shape of the structure of a single chain.

An important contribution in designing heteropolymers with precise structures is represented by the "tube protein model" developed by Maritan and co-workers [38, 39], in which total hydrophobicity and the bending rigidity of the polymer "pre-sculpt" the conformational free energy landscape, making the heterpolymer designable in some specific cases [38]. The authors

can obtain the folding of a heteropolymer with 30 monomers into a precise unique structure, by designing the sequence through an alphabet of 4 building blocks. In this model however, the number of possible target structures is highly reduced to helical like structures, thus the variability of possible target structures is not high. In a more recent work [40], the homopolymer recovers protein-like secondary structures (both helix-like and β-strand like), but a design of the heteropolymer to fold into a specific structure is not performed. In a different version of the model, directional interactions are added [41] to the homopolymer.

It seems extremely difficult to construct designable artificial heteropolymer, especially to make them able to fold with a precision and variability of structures comparable to the ones of proteins. What do artificial heteropolymers lack that grant proteins such precision and versatility? Answering this question, by understanding the features that control the designability of proteins, is the key for the generation of new self-assembling materials. One possible answer is that the self-assembly properties of proteins follow from unique characteristics of their building blocks, i.e., from the unique arrangement of the atoms in amino acids, from which follows the specific geometry of the protein skeleton. A different answer could be, that it is possible to narrow down from the complexity of proteins to a few simple key ingredients that make the design process possible.

Computational methods are a powerful tool to address this question [42]. In fact, studies *in silico* can be both employed as a tool to reproduce phenomena not directly observable experimentally, and also make it possible to artificially isolate different ingredients of a phenomenon in order to understand its nature. The latter can be achieved through *coarse-grained* models, simplified representation with fewer degrees of freedom, that are widely used

to simulate complex biological systems, such as proteins, nucleic acids, lipid membranes or carbohydrates. Coarse-grained models have been introduced to study complex phenomena that by using all-atom models would require simulation times too long or systems too large to simulate for the state of the art of the computational power available. An example of such complex phenomena is protein folding [43]. In fact, the systematic folding of natural or artificial sequences by using all-atom protein models it has been observed, to the best of our knowledge, only in a few cases [44–52]. However, the importance of coarse-grained models is independent from computer engineering and experimental advances. In fact, another fundamental use of coarse-graining is helping in physically understanding a phenomenon, by artificially playing with the different ingredients of a model and thus discriminate the essential ones that make the phenomenon possible.

The simplest protein (and general heteropolymer) coarse-grained models are the ones on lattice [53–58], in which the different monomers of a polymer chain are arranged on a grid (see example in Fig. 2 a). A step towards off-lattice coarse grained protein models is represented by the "Caterpillar" protein model [59, 60], in which the specific interactions between the amino acids are represented by different long range isotropic potentials for each different pair of amino acids, while the backbone hydrogen bonds are reproduced by simple directional interactions (see Fig. 2 b). In his work the author has designed several natural protein structures, and successfully demonstrated that the artificially generated sequences were able to spontaneously reach the target structures with high accuracy. Thus, the Caterpillar model highlights that two key ingredients seem sufficient to reproduce most of the self-assembly properties of proteins: the directional interactions that mimic the hydrogen bonds, and the specific sequence of heterogeneous interactions.

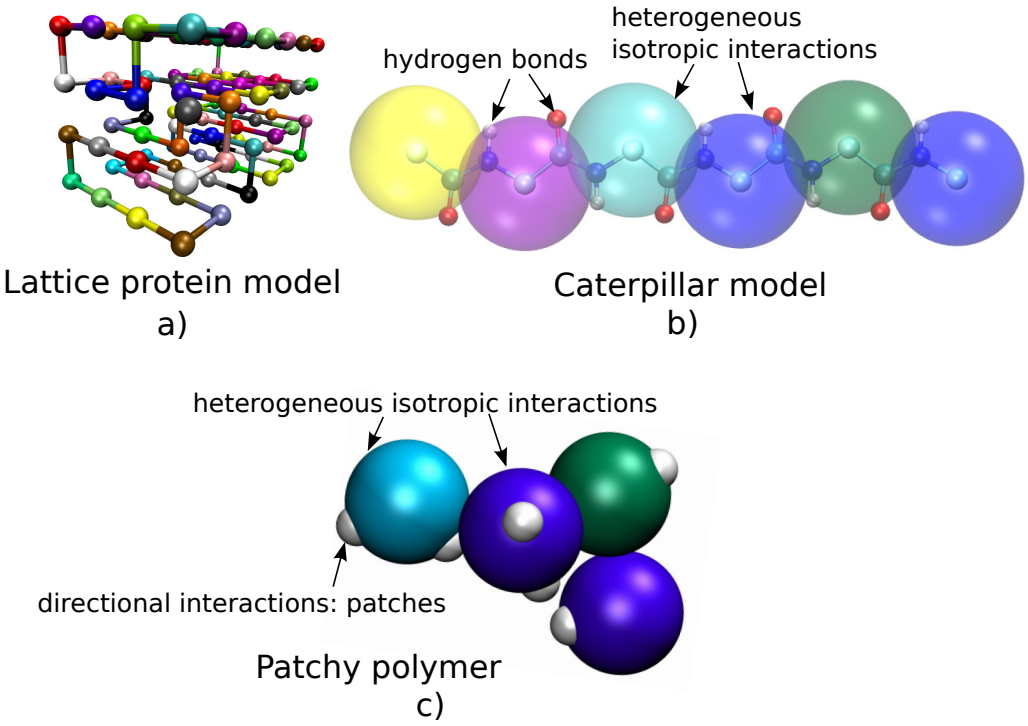

Figure 2: Schematics of different heteropolymer models.

a) An example of lattice protein model. b) The Caterpillar coarse grained protein model [59, 60]. The directional interactions are limited to the homogeneous directional interactions of the backbone, and the characteristic of different amino acids (different colors in the figure) are introduced by long range heterogeneous isotropic interaction. The spheres represent the self-avoiding diameter (4 Å) of the beads. c) The patchy-polymer model. Here the directional interactions are represented by patches, that can change in number and orientation.

The Caterpillar protein model has been used as inspiration for an artificial heteropolymer model, the "patchy polymer" model [61, 62], consisting of a polymer of patchy particles (see Fig. 2 c). In this model, the authors employ the same key ingredients that have been proved to be fundamental by the Caterpillar model and, in particular, the directional interactions are represented by the patches. Since the model does not aim at reproducing natural protein structures, the geometry of the monomers in the patchy polymer is different from the protein backbone geometry and different target structures are obtained. For some specific cases, the patchy polymer has been proven to precisely refold artificial sequences into unique target structures [61, 62].

The patchy-polymer model differs from non-designable traditional heteropolymer models mainly in the addition of the directional interactions to the heterogeneous isotropically interacting monomers.

A question then spontaneously arise: what is the role of the directional interactions in the designability of heteropolymers? In this doctoral thesis in Chapter 1, I study heteropolymers with a broad spectrum of number M of patches, obtaining completely new spaces of different target structures with respect to protein-like structures. For each of these cases, I test the designability of the heteropolymer for different alphabet sizes, by employing the methodology of the Monte Carlo (MC) simulations SEEK, DESIGN and FOLD (SDF) [62]. The SDF method is briefly described in the following.

i) SEEK: In this step, an extensive sampling of the heteropolymer conformations and sequences is performed, by employing both conformational MC moves and sequence mutation moves. The most designable target structure is the global free energy minimum of this space, being the structure that is the thermodynamically most favoured in the space of both the conformations and sequences or, in other words, the structure that most sequences prefer.

ii) DESIGN: optimizes the sequence to fold into the target structure chosen, by performing an extensive sampling of possible sequences while maintaining the structure fixed, and selecting the sequence found at the free energy minimum.

iii) FOLDING: performs only conformational moves, starting from a stretched structure while keeping the designed sequence fixed, in order to verify whether the designed sequence has the global free energy minimum close to the target structure, within a high precision in the range of fractions of the monomer size. If the folding fails with the most favourable target structure, the heteropolymer with a certain number of patches M and alphabet size q is labelled as not designable as a whole.

Through the SDF method, I probe in Chapter 1 the emergence of designability by changing the number M of patches per monomer, starting from a standard heteropolymer (no patches) up to chains of monomers with 10 patches, for different alphabet sizes. I show that there is a sweet spot in the number of directional interactions that allows to design heteropolymers that self-assemble into a unique predetermined structure, with precision comparable to what proteins can achieve. Such precision is obtained even up to small alphabet sizes of $q = 3$, comparable to the state of the art of non-designable block copolymers.

From the new spaces of target structures obtained through the SEEK MC simulation, I show in Chapter 1 the emergence of a simple qualitative criterion to discriminate designable from non-designable polymer backbones. By computing the radial distribution function for each different M averaged over the ensemble of target structures, I identify as designability criterion the appearance of a particular peak that dominates over the peak relative to the random packing of the heteropolymer. I show that the peak is a feature

of designable heteropolymers, as it is also dominating the radial distribution function of natural proteins. The criterion can probe the designability *a priori* – without having to design and fold the heteropolymer – by only looking at properties of the ensemble of possible target structures. It is also experimentally accessible, by measuring the structure factor $S(q)$ through x-ray or neutron scattering of a population of polymers. Therefore, this criterion can be used as a guideline to engineer new artificial heteropolymers that can reach the full potential of biopolymers, to go beyond the current protein- and DNA-based materials [23, 33, 63].

The peak can be interpreted as a qualitative measure of how much the introduction of the patches confines the system to populate only conformations that fulfill the directional interactions, such as in proteins. In natural proteins, in fact, it is known that the directionality of the hydrogen bonds and the geometry of the backbone 'pre-sculpt' the conformational space of proteins to the sub-space of currently known proteins structures with specific secondary structures [41, 64, 65], as it is shown for instance in recent works by Giacometti and co-workers [66,67] where protein-like secondary structures were obtained via a coarse-grained model where side chains introduced steric interactions. This 'pre-sculpting' seem to allow for the design of sequences capable of selecting a unique target structure [60].

Measuring quantitatively the role of directional interactions

The designability criterion in Chapter 1 suggests that the role of the directional interactions is in general that of pre-sculpting the accessible conformational space of a heteropolymer, by making accessible only those confor-

mations that fulfill the directional interactions, such as in proteins. This suggests that it could be possible to find several other designable heteropolymer skeletons, different from the protein one, that can have the same role.

Indeed the mean field theory for heteropolymer freezing, the Random Energy Model (REM) [5, 53, 54, 68], predicts that it should be possible to construct designable artificial heteropolymers by following the above principle. The REM introduced the concept of protein freezing transition [53], namely the transition between two phases of the polymer chain, of which one is dominated by very few conformations with low discrete energies, while the other is a continuum spectrum of exponentially many of them. Later versions of the theory [53,54,68], demonstrated that for a general heteropolymer is possible to reach a "frozen" thermodynamic state dominated by a single conformation, i.e., it is possible to design a heteropolymer to fold into a unique target structure. In order for the system to be designable, it has to be possible for the sequence to minimize the energy below the continuum spectrum.

The REM introduced the concept of protein freezing transition [53], namely the transition between two phases of the polymer chain, of which one is dominated by very few conformations with low discrete energies, while the other is a continuum energy spectrum of exponentially many of them. Later versions of the theory [53, 54, 68], demonstrated that for a general heteropolymer is possible to reach a "frozen" thermodynamic state dominated by a single conformation, i.e., it is possible to design a heteropolymer to fold into a unique target structure. To identify heteropolymer with a single frozen conformations, the REM predicts that ones should look among the sequences with the lowest energy in the target structure. Typically this procedure is carried out with a Monte Carlo simulation with only sequence mutations and performed

at low temperature. The lower the energy of the designed heteropolymer, the higher the heteropolymer stability. The REM dictates that for the system to be designable, it has to be possible for the sequence to minimise the energy below the continuum spectrum.

For this to be realisable, the number of conformations accessible by the system needs to be reduced. In fact, a clear review of the REM can be found in the work of Pande *et al.* [5], shows that for a system to be designable it needs to respect the inequality :

$$q > e^{\omega} \tag{1}$$

where q is the number of different constituents (i.e., the alphabet size) and ω is the conformational entropy per monomer of the system, defined as the logarithm of the total number of accessible *compact* conformations per monomer [5, 69]. The inequality 1 states that the designability of a heteropolymer increases with the alphabet size q and with decreasing the conformational entropy per monomer ω. It follows that the designability can be enhanced either by increasing the alphabet size as in Go-models [70], in which each amino acid is constructed to interact only with a specific subset of residues, or by decreasing e^{ω}, meaning reducing the number accessible compact conformations. In other words, the inequality 1 identifies the minimum alphabet necessary to design a heteropolymer, that is the smallest integer q_{min} greater than e^{ω}. Thus, by controlling ω it is possible to reduce the alphabet size of components employed in the polymer synthesis.

It is important to stress that the compact polymer conformations are a subset of total possible ones. For a lattice polymer – that is the test case for the REM theories – the definition of compact conformation is unequivocal, meaning a conformation with the maximum number of contacts possible on a lattice. Lattice heteropolymers are also a straightforward way to reduce

e^{ω} and obtain designable heteropolymers [53–58]. In fact, here the number of accessible compact conformations is reduced, with respect to an off-lattice polymer, by forcing the polymer to assume only conformations on the lattice. For an off-lattice polymer, the results in Chapter 1 suggest that the role of the directional interactions in terms of the REM might be that of decreasing ω, by reducing the number of accessible compact conformations. However, for off-lattice polymers the REM does not give an operative definition of compactness, making it difficult to establish a general methodology to estimate ω and in turn the designability of a heteropolymer.

In Chapter 2 we verify that the directional interactions indeed reduce ω, by introducing a methodology to calculate ω for off-lattice polymers. The methodology is based on estimating the total conformational entropy s, defined as the logarithm of the total number of conformations (including non-compact ones), for different numbers of patches. From s, we count firstly as compact structures the ones selected as target structure in the designability study in Chapter 1, that is the most designable target structure found as global free energy minimum of the SEEK plus the the structures within its thermal fluctuations (1 k_BT from the minimum). We calculate e^{ω}, and thus the computed minimum alphabet necessary to design those structures q_{min}, and compare it with the verified minimum alphabet obtained by the designability characterization in Chapter 1, for different numbers of patches.

Then, different compactness definitions are tested and linked to the variability of the different possible target structures. In fact, if the target structure is chosen further apart from the free energy minimum, it is expected to be less designable. By enlarging the ensemble of compact structures, we calculate how much q_{min} needs to be increased in order to design less favoured structures. Thus, we show how much variability of different structures can

be ensured for different numbers of patches, in order to identify the best candidates with the highest variability of different structures.

The experimentally accessible criterion based on the radial distribution function presented in Chapter 1 and the computational method in Chapter 2 can guide the engineering of new types of self-assembling heteropolymers that can reach both the precision and versatility that proteins can achieve, and will therefore open new applications for polymer-based materials. An example of possible application is developed in Chapter 3, where I treat a particularly challenging problem: the design of heteropolymers that spontaneously self-assemble in complex pre-determined knotted topologies.

An application of self-assembling heteropolymers: design of knotted polymers

As we are used to observe at the macroscopic level with ropes, knots form spontaneously in polymer chains of sufficient length and flexibility and their complexity increases as the polymer becomes longer [71–74]. Several studies for different artificial polymer models have been performed [75–78]. The computational work in Ref. [79] shows that knots or different complexity are formed by clusters of Stockmayer particles – interacting via a Lennard-Jones potential plus a point dipole – by changing the strength of the dipole. At the molecular scale, knots are found both in DNA [80–83] and in about 800 native protein structures [84–88], as shown by a recent exhaustive analysis of the PDB [89] (for an example of knotted protein see Fig. 3.4). As most proteins, as well as our patchy polymers, consist of a backbone with two ends, they cannot form knots in the mathematical definition that requires a closed loop. However, imaginarily connecting the ends through different

strategies [87,90,91] allows to analyse their knot topology. It has been shown that knotted proteins are able to fold without the aid of chaperones, therefore the folding path of the knot must be uniquely encoded in their amino acid sequence [92–94]. On the contrary in DNA, its knotting and unknotting is mediated by topoisomerase enzymes [81, 95, 96]. Clear recent reviews of natural and artificial knots at different scales are found in Ref. [97–99].

As our everyday experience teaches us, knots have different applications depending on their topology: the "bowline" permits to create a ring of fixed size that does not shrink if the two lengths are pulled, the "figure-eight" knot stops ropes from running out of a retaining device, the "clove hitch" is best for tying a rope around an object. There is no reason not to speculate that several applications for knots could be found also at the molecular or microscopic scale. In fact, studies show that knots enhance mechanical and thermodynamical stability in knotted proteins [84, 86, 100–103] or polymers [61]. This stabilisation of the polymer structure could have applications in engineering cargo-carrier constructs, for instance for gene or drug delivery [104], as the stability of the structure of the cargo is essential to ensure the success of the delivery [101, 105, 106]. Moreover, most known knotted proteins are natural catalyzers (enzymes) and the knot is usually located in the active site, therefore knots can have an important effect on catalytic activity both in nature [107, 108] and in synthetic polymer knots [109].

However, to this date the synthesis of a chosen knot topology is still a challenging task. Most of the progress in this direction is represented by molecular knots engineered via topological synthetic chemistry, a discipline that started with the pioneering works of Sauvage and collaborators 30 years ago [110]. Since then, different synthetic strategies have been tested for the fabrication of the simplest knot – the trefoil knot – based on a combinations

of covalent interactions and metal coordination using ligands with helical shape [99] (for an example see Fig. 3.1). Only recently, a few examples of more complex knots were synthesized: the pentafoil knot [111, 112], the figure-eight (or Savoy knot) [113], and the 8_{19} [114]. In the design of knotted artificial polymers, the state of the art is still limited to design of sequences for lattice polymers that enhance the total knotting probability, without being able to design a specific knot [115]. Of the billions of types of knots existing, only the four ones mentioned above have been selectively synthesized using small building blocks, making the synthesis of specific knots still an open problem.

In Chapter 3, we present a significant advancement to the solution of this problem, showing that patchy polymers can be designed to fold into specific chosen knot types with a precise structure, that can be locked externally by closing the ends of the polymer. Through a Monte Carlo method that enhances the correct sampling of different knots, we categorise the space of possible knot types of patchy polymers. With a chain of only 50 monomers, we observe a broad variety of knots of complexity up to 20 crossings, and we estimate the occurrence probability for each knot type. We tested and succeeded with high precision in the folding of all knot types experimentally constructed up to now [110–114], plus also the unprecedented twist knot 5_2 and the complex pretzel knot with 10 crossings, the 10_{124} (see Fig. 3.3 and 3.6). Thus, our design strategy can be used as a guideline for experimental construction of more complex knots.

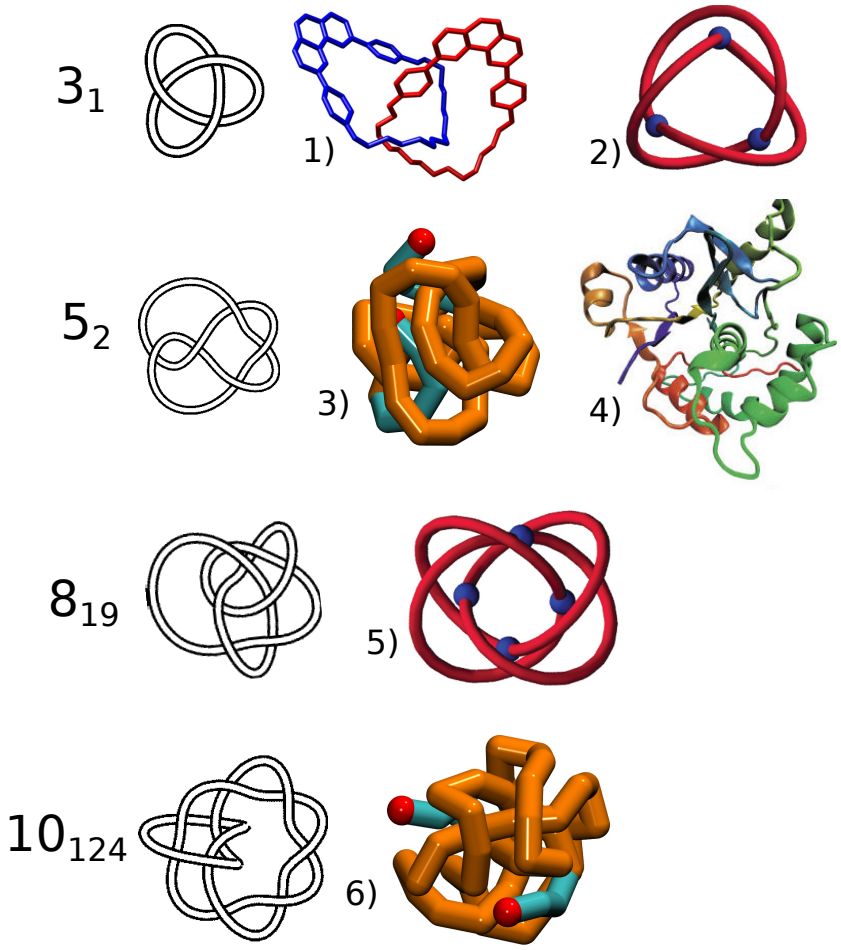

Figure 3: Examples of experimentally and computationally obtained knots. On the first two columns the name and symmetrical representation (from [116]) is reported. 1) Example of 3_1 "catenane" synthesised structure from Ref. [117]. 2) Example of computationally obtained 3_1 and 5) 8_19 structures from helical templates from Ref. [118]. 4) 5_2 natural protein [84]. 3) 5_2 and 6) 10_124 folded structures obtained by designing patchy polymers in the present thesis, see Chapter 3.

Outline of the thesis

The following three chapters consist each of self-contained scientific paper, published or submitted to a peer-reviewed scientific journal. Thus, each chapter is an independent work with its own abstract, introduction, results and conclusions. Therefore, I ask the reader to excuse redundancies between the chapters. In Chapter 1, lengths are in units of the radius of the monomer bead to compare with previous works on the patchy polymers, in Chapter 2 and Chapter 3 the length unit has been changed to the diameter of the monomer bead to conform to the customary standards used in the field.

The order of the chapters coincides with the order of the different subtopics lined out in the introduction. Firstly, the reader can explore the open problem of self-assembly of artificial heteropolymers in Chapter 1, that shows a designability characterization of a broad spectrum of different alphabet sizes and different numbers of directional interactions. Here, we show that the latter allow to design heteropolymers with high precision, comparable to proteins. Moreover, we introduce a qualitative measure, based on the radial distribution function, of the role of the directional interactions in the designability of different classes of heteropolymers. This results in an experimentally accessible criterion to discriminate a priori designable from not-designable polymer backbones.

In Chapter 2, we develop a computational method to determine the minimum alphabet q_{min} necessary to design a heteropolymer, estimating quantitatively the role of the directional interactions in reducing q_{min} to accessible values. The method also allows to measure the variability of different possible target structures, as function of the number of directional interactions (the patches). In this way, it is possible to estimate the best candidates of number of patches that ensure the highest variability of designable target

structures.

In Chapter 3, the reached high precision and versatility find an application in the design of knotted polymers with chosen topology and precise structures, that can have several applications as active materials, catalysers and novel carriers for drug delivery.

Chapter 1

The role of directional interactions in the designability of generalized heteropolymers

1.1 Introduction

Gaining control of self-assembly processes is key for the generation of smart materials, with applications ranging from energetics [119], to photonic crystals [120] and biomimetic scaffolding [121]. Among the most promising and versatile classes of systems that undergo self-assembly are polymers composed of monomers of different species, also known as heteropolymers [4, 30, 122, 123]. Remarkable examples are natural self-assembling systems such as DNA [24], RNA [124] and, in particular, proteins [125, 126]. Their extremely rich self-assembling behaviour is mainly controlled by the *sequence* of chemically different building blocks, or monomers, that compose the polymer. The collection of all possible monomer types is the *alphabet* of the system. For proteins, different sequences of the same alphabet of 20 amino acids lead

to the huge number of proteins expressed in nature, making the alphabet extremely versatile. Indeed, it is the specific sequence that drives a heteropolymer to uniquely collapse (fold) into a target conformation, providing precise control over the resulting structure.

Artificial heteropolymers synthesis and manipulation is at an advanced stage, as demonstrated by the numerous materials based on block copolymers [4]. Typically, block copolymers are synthesised from two monomer types (AB copolymers), but there are examples of linear chains synthesis with an alphabet size of 4 (ABCD block copolymers) [30, 123] or star polymers with up to 7 types [32]. The different types are organized in patterns of blocks of the same repeated monomer, from two up to five blocks [4], and by controlling the relative length of the blocks it is possible to realize rich phase diagrams [4, 122, 127–133]).

However, it is extremely difficult to drive the folding of a single chain towards a very specific structure, especially with the large variety of different structures and the high accuracy, fractions of inter-monomer distances, obtained by biopolymers [59, 63, 126, 134, 135] that is necessary for many applications (*e.g.* catalysts). Recently, efforts to overcome this limitation have been made [33–37], in particular by Khokhlov *et al.* [2], where copolymer chains have been iteratively "painted" to undergo a hydrophobic-hydrophilic micro-phase separation and by Moreno *et al.* [3], who have engineered specific hydrophilic-hydrophilic patterns to produce a similar micro-phase separation. However, contrary to biopolymers, these two methodologies do not provide control over the detailed shape of the target structure.

According to mean field theories (MFT) [5, 54, 68] it should be possible to construct artificial heteropolymers that, similarly to proteins, drive the

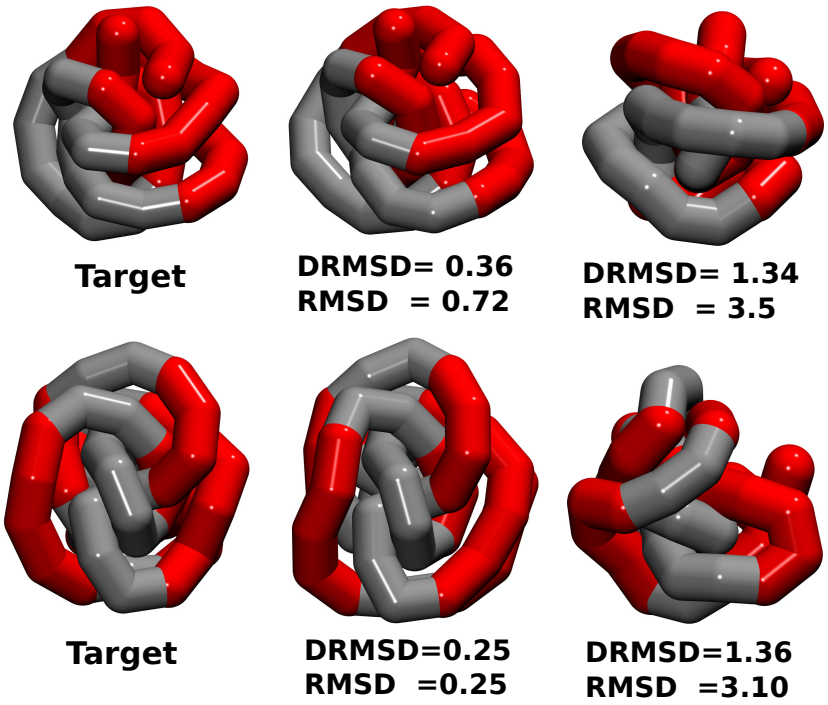

Figure 1.1: Functionalized patchy polymer structures. Simulation snapshots showing the target structures with the desired functionalizations (left), the final folded structures obtained using our design procedure (centre) and examples of misfolded structures (right), the last two aligned with respect to the correspondent target structure. The values of $DRMSD$ (distance root mean square displacement, see Eq. 3.6) and $RMSD$ (root mean square displacement see Eq. 3.11) with respect to the target structure are in units of the bead radius, that for proteins has been seen to correspond to approximately 2 Å, so in protein units the two folded structures would have a $RMSD$ with respect to the target structure in the range $0.5 - 1.4$ Å and the two molten globule structures in the range $6 - 7$Å. Such functionalized polymers can, in turn, be used as Janus (one-patch) particles (top) or triblock Janus (two patches) particles (bottom).

collapse of specific sequences into given target structures. The identification of such sequences is normally referred to as *design* and, in what follows, we define *designability* as the property of a heteropolymer to have at least one heterogeneous sequence that reliably folds into a given stable target structure [5]. A clear review of the MFT mentioned above can be found in the seminal works of Pande *et al.* [5], who showed that the designability of a heteropolymer increases with the alphabet size and decreases with the conformational entropy per bead. Thus, the designability of a heteropolymer can be enhanced following two strategies: i) by increasing the alphabet size as in Go-models [70], where each amino acid can specifically interact only with a subset of residues and, ii) by decreasing the configurational entropy per particle, i.e. reducing the number of possible configurations.

A straightforward way to adopt the second strategy is represented by designing lattice heteropolymers [53–58], in which the number of possible configurations (thus the configurational entropy) is reduced by the topology of the lattice.

In order to adopt the second strategy for off-lattice polymers a possibly effective way is to introduce directional interactions, *i.e.* interactions that depend on the mutual orientations of the monomers, in addition to the heterogeneous isotropic interactions. In this way, the conformational entropy per bead decreases because of constraints on the mutual orientations. In fact, the directionality of the interactions introduces a frustration in the system that "pre-sculpts" the configurational space of the system, which will contain just a fraction of the total compact configurations. Another effective way to reduce the conformational entropy is to increase the chain stiffness. An important contribution in this direction is represented by the "tube protein model", developed by Maritan and co-workers [38, 39, 41, 136, 137], in

which control of the conformational free-energy landscape is achieved by tuning the total hydrophobicity and the chain stiffness of short polymers, which can be made designable in some specific cases [38, 41]. Natural proteins take advantage of both routes by exploiting the directionality of the hydrogen bonds and the geometry of the backbone [39, 41, 65, 137]. From an experimental standpoint, in fact, decreasing the conformational entropy via these two routes is more feasible then increasing the alphabet like in Go-Models, which might be the reason why directional interactions (such as hydrogen bonds) are so important in biology.

In this work, we show that the presence of directional interactions in a generalised heteropolymer allows for the realisation of compact structures with functionalized regions on their surfaces with an accuracy comparable to natural catalysts, *i.e.* of the order of a fraction of amino acids distances. To give an idea of the huge potential of such an approach, in Fig. 1.1 we present simulation snapshots of heteropolymers that self-assemble with high-accuracy into predetermined structures: two spherical objects with patterned (functionalized) surfaces. The sequences we design are able to make the two heteropolymers reliably fold into a Janus-like (top) and a triblock-Janus-like (bottom) object with $DRMSD$ (distance root mean square displacement see Eq. 3.6) of 0.36 and 0.25 with respect to the target structure which correspond to an $RMSD$ (root mean square displacement see Eq. 3.11) of 0.72 and 0.25 with respect to the target structure, respectively. Depending on the chemical nature of the functionalization, such folded structures could exhibit catalytic functions or be used as building blocks in a hierarchical self-assembling process [138]. Of course, these applications would also require a precise control of inter-chain interactions in order to avoid amorphous aggregations [139].

However, in order to exploit the folding of artificial heteropolymers for

applications, we ought to first address the following question: "Can we predict *a priori* whether an engineered heteropolymer is designable?" In other words, if one engineers heteropolymer architectures that differ in terms of number, geometry or nature of the directional interaction, is it possible to predict a priori which ones are designable (i.e., for which ones one can find at least one specific sequence that drives the system to fold into a given target structure)?

Here we show that the answer to this question is represented by a distinct peak in the radial distribution function dominating over the random packing of the heteropolymers. A minimal set of directional interactions is an effective way to induce such a peak (Fig. 1.2), which guarantees the designability (Fig. 1.3).

1.2 Results and discussion

Our working hypothesis is that the addition of directional interactions will result in designable chains as previously suggested by results obtained for the "patchy polymer" model [61,62] and the "caterpillar" protein model [59,60]. In the following we will use two heteropolymer models with different chain stiffness (see Methods): the freely-rotating chain (FRC), marked by open locks in the figures, and the freely-jointed chain (FJC), marked by closed locks in the figures. Briefly, the "chemical character" of each monomer is given by a isotropic interaction potential, represented as a simple square-well like shape, with a different pre-factor (more or less attractive or repulsive) for every different pair of monomer types. The directional interaction between the patches is homogeneous along the chain and is the potential derived by Irbäck et al. [140], commonly used to model hydrogen bonds (see Methods).

By changing the number of patches per monomer we will probe the emergence of designability starting from a standard heteropolymer (no patches) up to chains of monomers with 10 patches. The number of patches unambiguously identifies their arrangement on their surface, which is chosen so as to yield the most symmetrical geometry (on the poles for two patches, equispaced on the equator for three patches, on a tetrahedron for four patches, *etc*). The bottom of Fig. 1.3 shows all the patch arrangements employed.

Strictly speaking, proving that a heteropolymer with a certain number of patches is not designable would require testing all possible sequences for the presence of a unique collapsed equilibrium structure. Here we rely on the statistical definition of designability, where the methodology of the Monte Carlo simulations SEEK, DESIGN and FOLD (SDF) was proven to be able to discriminate between designable and not-designable structures [62]. Briefly, the SDF method: i) seeks the most designable target structure through an extensive sampling of the heteropolymer conformations and sequences (see Fig. A.1 in Appendix A.3); ii) designs the sequence that should optimally fold into the target structure; iii) tests whether the designed sequence correctly folds into the target structure. If the folding fails with the most favourable target structure, the heteropolymer with a certain number of patches is labelled as non-designable. More details can be found in Sec. Methods 3.2.1. Here we apply the SDF method to chains made of 50 monomers. From now on, all the distances will be in units of the bead radius R_{bead} and the energies in units of $k_B T_{Ref}$, where T_{Ref} is a reference temperature that sets the scale of interactions.

In Figure 1.3 we show the results for the designability for all cases studied. The systems without directional interactions (patches) are not designable for any alphabet. By adding directional interactions, the heteropolymer becomes

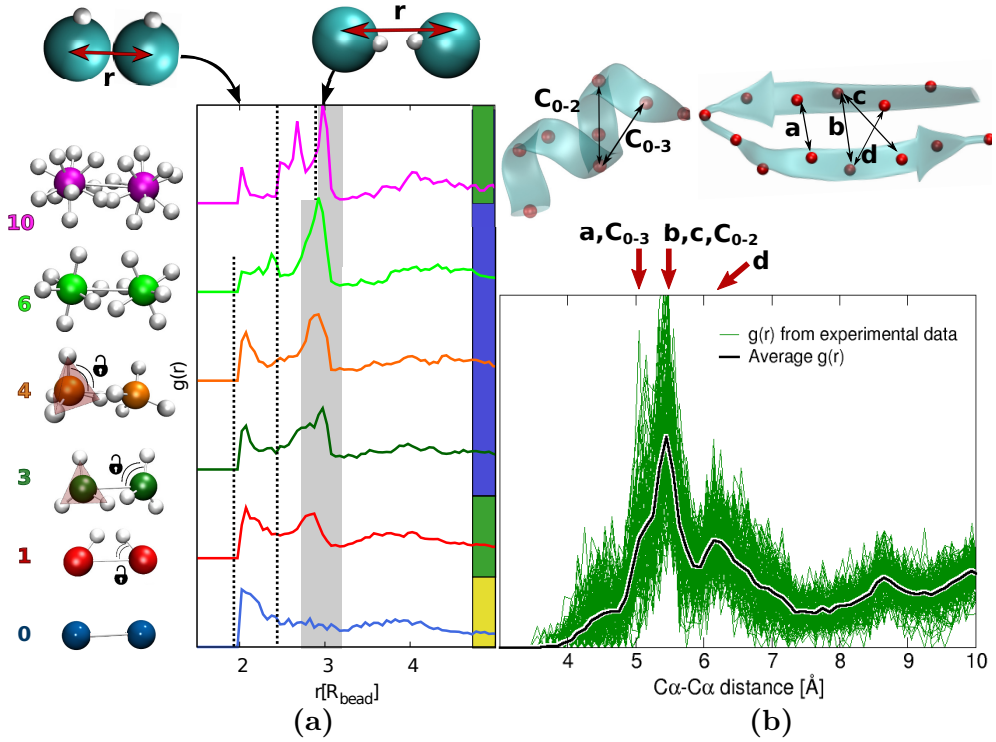

Figure 1.2: a) Radial distribution functions $g(r)$ for patchy polymers. We show one $g(r)$ curve for each patch arrangements because we observed that it does not change significantly for different alphabet sizes. The $g(r)$ has been averaged on the 40 most probable conformations for different patches arrangements (y-axis): white spheres indicate the patches located at the surface of the monomer. The $g(r)$ are y-shifted for the sake of visualisation. The random packing peak is delimited by two dashed lines, while the grey band highlights the range of the directional peak. Note that for 10 patches the random packing peak splits and moves towards larger distances (Fig. A.4 in Appendix A.3). The ratios between the random packing peak and the directional peak areas are: 0.75 for 1 patch, 0.23 for 3 patches, 0.27 for 4 patches, 0.22 for 6 patches, and 1.33 for 10 patches. b) $g(r)$ calculated over the $C_\alpha - C_\alpha$ distances for 145 high resolution protein structures from the Protein Data Bank (green lines). The thick black line is the average. The first peaks dominating the $g(r)$ correspond to the characteristic distances of the backbone hydrogen bonds, which determine the secondary structure [141] (α-helix and β-sheet shown on top). All $g(r)$ have been calculated by neglecting the contribution of the first neighbour along the chain.

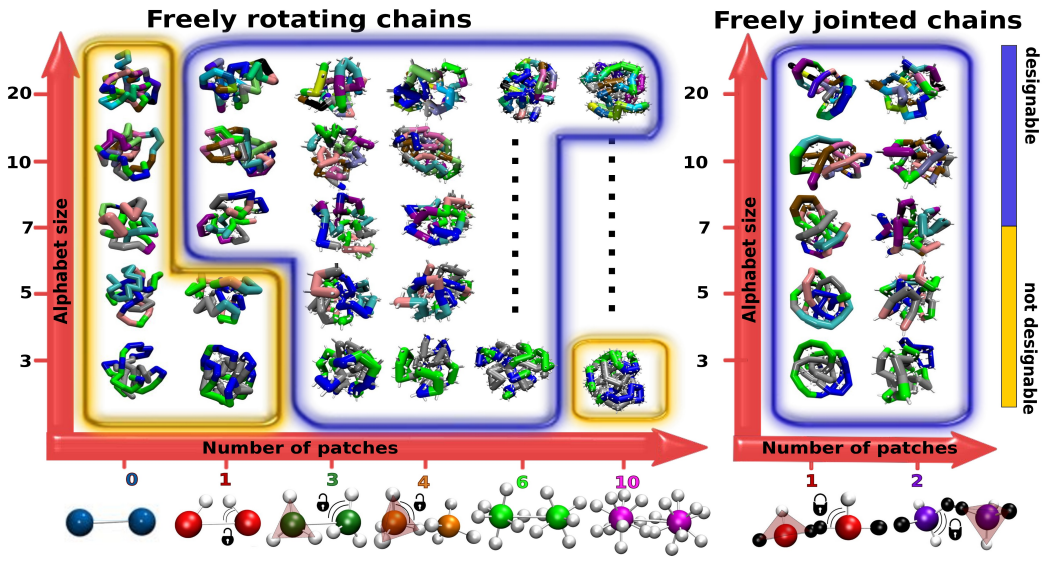

Figure 1.3: Designability diagram for different patch numbers and alphabet sizes. Within the yellow regions, the heteropolymer is not designable, while the blue regions correspond to designable systems. For each case, the chosen target structure is shown. Bottom panel: a schematic illustration of patchy polymers. White spheres indicate the patches located at the surface of the central monomer. For FRC (left) we consider from 0 (simple heteropolymer case) to 10 patches and 1 and 2 patches for FJC (right). With the exception of the one patch geometry, in all the other cases the patches are fixed on the vertices of a platonic solid or in the most symmetric way (see Appendix A.2). The two patch FJC are placed on a tetrahedron completed by the anchoring points (in black). For both models, we consider alphabet sizes q ranging form $q = 3$ to $q = 20$ (see Methods 3.2.1 for the definition of the interactions).

designable for a wide range of patch numbers and alphabet sizes, both for the FRC and FJC models. The emergence of designability coincides with the appearance of a peak in the pair radial distribution function—highlighted with the grey band in Fig. 1.2a)—located at the distance $r \simeq 3$, at which the patch-patch interaction is most favourable. The presence of such an isolated intermediate peak between the first ($r \simeq 2$) and second random close packing neighbours ($r \simeq 4$) indicates that the directional interactions are inducing a geometrical frustration in the system. The frustration strongly biases towards a subset of compact conformations. In fact, in the non-designable case without patches, the directional interaction peak is not present. For one patch, the directional interaction peak is present but is lower than the peak corresponding to the close packing. These chains with one patch are not always designable (borderline), and their designability depends on the alphabet size q (Fig. 1.3). Increasing the number of patches the directional interaction peak becomes dominating over the random packing peak, making the chains designable in a broad interval of q (3 to 20). Increasing further the number of patches up to 10, the random packing peak increases again, splits and moves towards higher distances (see Fig. A.4 in Appendix A.3) due to the self-avoiding of the many patches and also the system loses directionality: this corresponds in turn to a loss of designability (Fig. 1.3). It is important to stress that the strength of the directional interaction is such that even the chains with one patch are always maximally bonded, implying that arbitrarily increasing the relative strength of the directional interactions will not suppress the first peak of the $g(r)$ (see Fig. A.1 in Appendix A.3). Thus, when the patch-patch peak dominates over the random close packing peak, the system becomes robustly designable in a broad interval of alphabet sizes q, both for the FRC and FJC models.

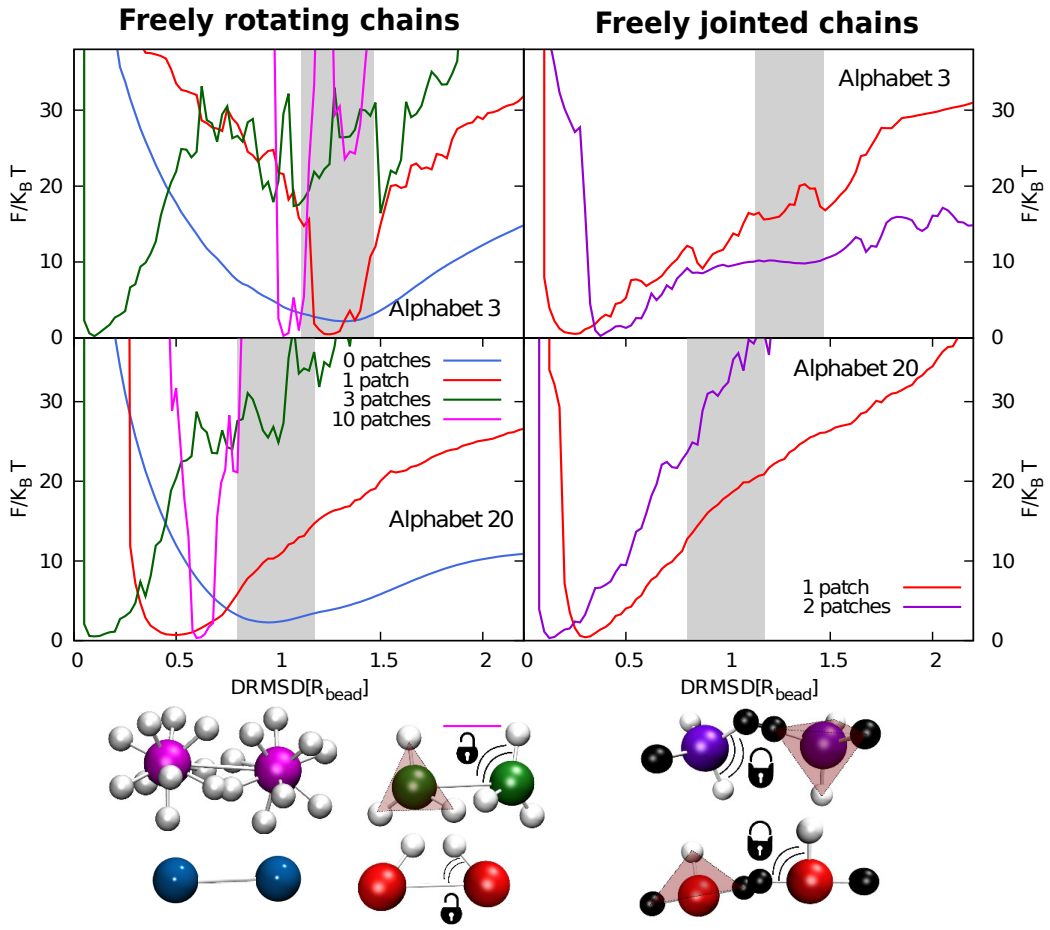

Figure 1.4: FOLDING free energy landscapes for some significant cases. The free energy is plotted as a function of the distance root mean square displacement ($DRMSD$) for FRC (left panels) and FJC (right panels), for different patch numbers and alphabet sizes q at temperature 0.4. $DRMSD = 0$ corresponds to the target structure. We define as molten globule region for each alphabet size (grey area) the $DRMSD$ range spanning $1k_BT$ from the position of the global minimum of the free energy of the corresponding bare polymer without patches. All global minima with $DRMSD$ values below the grey area are considered correctly folded (see Methods for details), see Fig. 1.1 for examples of folded and molten globule structures. The geometry of the patches is reported on the bottom of the figure, with colours corresponding to the free energy curves. The global minimum of the system with 10 patches and $q = 3$ is at the border with the grey area and it has been categorised as not designable.

According to the MFT [5], given an alphabet of size q, a system is designable when the sequence entropy per particle $\ln(q)$ exceeds the conformational entropy per particle ω. Hence, the designable–to–not-designable transition, where $\omega \sim \ln(q)$, allows us to estimate ω. Following the diagram in Fig. 1.3, we estimate ω to be $> \ln(20)$ for 0 patches, $\sim \ln(5)$ for one patch in the freely rotating chain, and $< \ln(3)$ for up to 6 patches. At 10 patches ω increases again to a value between $\ln(3)$ and $\ln(20)$. Since e^{ω} is related to the number of compact conformations, the latter is reduced by the patches by approximately an order of magnitude. A more precise evaluation of ω requires the study of intermediate values of $\ln(q)$ which cannot be achieved simply changing the alphabet.

The presence of the directional interaction peak is a fundamental fingerprint of designability. Indeed, we find it to be a general feature also of natural proteins. In Fig. 1.2b we show the radial distribution function for some characteristic examples out of 145 analysed high resolution protein structures from the Protein Data Bank. Here, the conformational space is shaped by the directionality of the hydrogen bonds, which forces the carbon C_α of the amino acids to be at the typical distances for the different types of secondary structure. Hence, the peaks analysed in Fig. 1.2b are equivalent to the directional interaction peak highlighted in grey in the patchy polymers. Since proteins are precisely designable because of the peculiar hydrogen-bond geometry of the backbone, we expect the latter to be an ideal template for general heteropolymers and patchy-polymers.

These results are in agreement with MFT [5, 54, 68], which predicts that the decrease of conformational entropy allows for better sequence design. Here, we introduce the presence of a peak in the $g(R)$ not related to the random packing as an estimate of the reduction of the conformational en-

tropy, in order to have a simple tool to anticipate the designability of different polymer architectures. Thus, we propose the presence of such a peak as a general criterion to engineer designable polymers able to fold into unique target structures, with the same accuracy and versatility of natural proteins. The geometry of the protein skeleton is a particular choice, and our results suggest that, by following our criterion, several others can be found.

Another important result is the characterization of the transition from a not-designable to a designable number of patches (Fig. 1.3). We start by noting that the freely rotating chain is sensitive to the choice of the alphabet: a minimum alphabet size $q = 7$ is required to guarantee designability for any number of patches (Fig. 1.3). To characterise the designable–to–not-designable transition, we have performed the SDF trial for each point in the designability diagram. The last step of the SDF is the calculation of the free energy difference between structures, grouped together according to the distance root mean square displacement ($DRMSD$) from the target conformation at $DRMSD = 0$ (see Methods). The accuracy of the refolding varies considerably for each scenario and depends on the alphabet size, the number of patches and the local environment of the monomers. In Figure 1.4 we show the FOLDING free energy profiles for some significant points in the designability diagram for $q = 3$ and $q = 20$ (see Fig. A.3 in Appendix A.3). At low temperature ($T = 0.4$) all curves show a global free energy minimum located at different $DRMSD$, however, they are considered folded only if the global minimum corresponds to the folded structure and not to a disordered molten globule structure. The different nature of the global minimum can be discriminated by increasing T and pushing the system to unfold ((see Fig. A.2 in Appendix A.3). What is striking in the figure is the high refolding accuracy of the FJC model, with smooth folding profiles and global minima very close

to the target structure. We ascribe such a high accuracy to the stronger constraint experienced by monomers in the FJC compared to the FRC model, in agreement with MFT results [5, 54, 68]. In fact, the patches in the FJC have less rotational freedom, thus the conformational entropy ω is decreased further. Finally, we note that such high accuracy is particularly interesting for single-patch chains, which, according to the designability criterion based on the $g(r)$, is a borderline case (Fig. 1.2a).

1.3 Conclusions

In conclusion, we extend the concept of *designability* to heteropolymers, bridging the gap between naturally foldable biopolymers and artificial polymers. We demonstrate that there is a minimal set of ingredients that makes it possible to exert a precise control over the conformation of the folded structure. Indeed, we show that starting from a traditional heteropolymer model, we can attain designability by introducing a few directional interactions (patches). The patches make the system designable even with alphabet sizes comparable to the case of block-copolymers (3 and 5), which are synthesizable and manipulable with the current advanced technology [4, 30, 123]. From an experimental standpoint, in fact, small alphabet sizes are more feasible. We demonstrate that the directional interaction peak in the $g(R)$ is a universal fingerprint for designability and can be used as a criterion to engineer new heteropolymers. This criterion will allow artificial heteropolymer to self-assemble into an enormous variety of highly exotic structures with high control over the detailed shape. This will expand the possibilities of current copolymer technology to the full potential of biopolymers, providing a way of going beyond the current protein- and DNA-based materials [63, 142, 143].

Heteropolymer design could find several applications depending on the scale of the building units. At the micrometre scale, chains could be realised using colloidal particles with applications in the design of materials with new mechanical, electronic or photonic properties tailored in 3D. Colloidal particles are widely available, their interactions can be tuned to be either attractive or repulsive and can range from a few nm to several micrometres [14]. Flexible strings of polymeric colloids were made by inducing electric dipoles to line the particle up before thermal fusion [144]. Such strings were observed to collapse into a random compact configuration under influence of an induced depletion attraction. Control over the sequence can theoretically be achieved in a similar fashion as for protein synthesis where the chains are grown from a solid surface [145]. Each monomer is loaded in the sample alternating washing cycles to remove the previously unbound monomers. The limiting factor of this technique is in the bonding probability that must be extremely high in order for the growth to continue until chain lengths of 20-50 residues. Currently, we are working on strategies to overcome such limitations in close collaboration with experimental groups. At the atomic and nanometre scales, control of the sequence is in reach of current polymer synthesis technology [30, 32, 123]. As we show, artificial chains could be designed to place chemical groups with a precision similar to what protein can achieve [146, 147], but also to work in different environments than the ones suitable for proteins or DNA.

1.4 Methods

The patchy polymer model we employ has already been proven to be able to refold artificial sequences into unique target structures [61, 62]. Here we con-

sider two standard heteropolymer models where heterogeneous isotropically-interacting monomers are bonded along the chain *via* a harmonic potential with two different anchoring geometries: i) in the freely rotating chain (FRC) model the spring bonds the centers of the monomers; ii) in the freely jointed chains (FJC) model the spring connects two anchoring points on the monomer surface and opposite to each other (see bottom of Fig. 1.3 for a sketch). An experimental realisation of the FJC model could be represented by covalently bonded chemical units or surface grafted colloidal particles. While possible experimental examples of patchy particles that could serve as monomers for the FRC model, i.e. where the patches can rotate with respect to the bead, are the lock and key colloids as in Ref. [148], the solid colloids with surface-mobile DNA linkers in Ref. [149], DNA coated emulsion droplets with mobile DNA patches [150], colloidal particles with an induced electric dipole [144]. The isotropic and bonded interactions are complemented by non-specific additional attractions provided by patches arranged on the monomer surface. The resulting directional interaction is anisotropic in nature as it depends not only on the distance between two monomers but also on their relative orientations. The key parameters of the models are the alphabet size of the isotropic interactions, the number of patches (which is the same for all the monomers in a chain) and their the geometrical arrangement. The isotropic interaction energy $E_{AB}(r)$ between two different sub-units of types A and B is represented as a simple-square-well like shape (Figure 3.1)

$$E_{AB}(r) = \begin{cases} \epsilon_{AB}\left[1 - \dfrac{1}{1.0 + \mathrm{e}^{2.5(r_{max}-r)}}\right] & \text{if } r > R_{bead} \\ \infty, = & \text{if } r < R_{bead} \end{cases} \tag{1.1}$$

where r is the distance between the centres of the beads and R_{bead} is the hard core radius, which is the same for each bead. ϵ_{AB} is a different pre-factor for every different pair of monomers. The cut-off distance $r_{max} = 6R_{bead}$ is

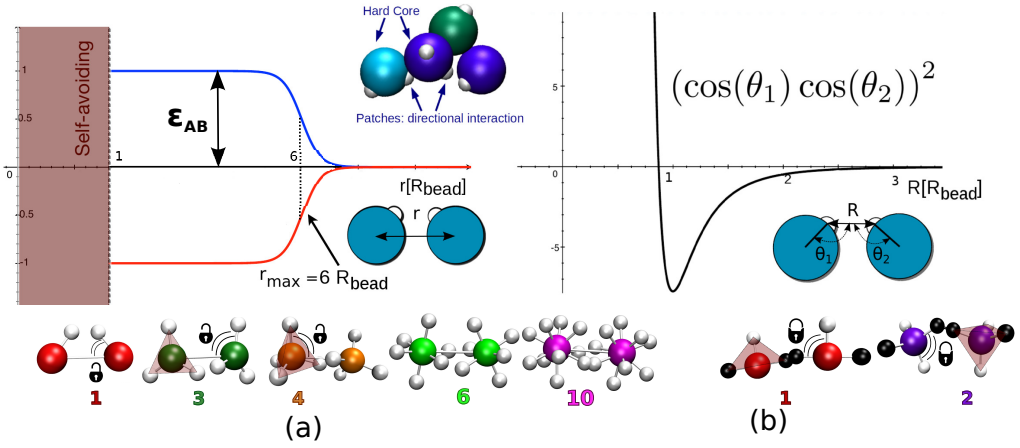

Figure 1.5: Interaction potentials. a) Isotropic square-well like interaction as a function of the distance of the centers of the spherical monomers r (see left inset). The well depth is controlled by a pre-factor ϵ_{AB}, which is different for each different pair of monomers. The pre-factors are grouped in an interaction matrix, the size of which depends on the number of monomer types in the range 3 to 20 (find the matrices used in Appendix C.1). For an alphabet of size N, the set of $N(N+1)/2$ heterogeneous interactions are Gaussian random numbers with zero average and standard deviation of $0.33 k_B T_{\text{Ref}}$, similarly to protein coarse-grained protein models [59, 151, 152], where T_{Ref} is a reference temperature that sets the scale of interactions. Inset: schematic illustration of patchy polymers. Small white spheres indicate the patches. b) Directional potential. The radial Lennard-Jones contribution in the plot is multiplied by the directional contribution $(\cos\theta_1 \cos\theta_2)^2$, where θ_1 and θ_2 are the angles between the patch vector and the distance vector (right inset).

the distance at which $E_{AB} = \epsilon_{AB}/2$ and was derived with a trial and error approach on coarse-grained proteins in the caterpillar protein model [59,60].

As directional interaction between the patches we employ the potential derived by Irbäck et al. [140], commonly used to model hydrogen bonds. It is represented by a 10-12 Lennard-Jones type potential multiplied by a factor containing the angles between the patches and the bead radius (Fig. 3.1), so that the energy is minimum if the patches face each other (when they are opposite to each other the radial part of the potential is ~ 0)

$$E_p = s\,\epsilon_p\,(\cos\theta_1 \cos\theta_2)^\nu \left[5\left(\frac{\sigma}{R}\right)^{12} - 6\left(\frac{\sigma}{R}\right)^{10}\right]. \tag{1.2}$$

Here R is the distance of the patches as in Fig. 3.1 (right inset), $\epsilon_p = 3.1\ k_B T$ and $\nu = 2$ [140] while we set $\sigma = R_{bead}$. The scaling factor s is chosen to not over favour the patch contribution over the isotropic one. If its value is too large all sequences form regular structures that depend solely on the symmetries of the patch arrangements on the beads. On the other hand, if it is too small all sequences fail to self-assemble and collapse into random glassy three-dimensional structures. Using a trial and error approach, we found 4 to be good number [62]. The neighbour beads along the chain are bonded via a harmonic spring potential.

In order to find whether the polymer is designable or not, we identify for each different number of patches and alphabet size at least one pattern (sequence) that has a global free energy minimum into a given structure. To increase the chances to find such a pattern, we first perform a SEEK MC simulation, in order to find potentially designable target structures that are not known from nature, unlike for proteins. In the SEEK, we explore at the same time different structures and sequences, and we extract a target structure from the global minimum of the free energy landscape obtained

(as in A.1 in Appendix A.3). The global minimum corresponds to the structure with the highest number of sequences that fold into it, thus the most designable one [153]. Other structures further from the global minimum might be good candidates as well, so the solution is not necessarily unique. On the contrary, if this structure will be not designable, it is highly unlikely and thus unfeasible for practical purposes to find another structure that will be designable.

The target structure is then redesigned in the DESIGN, where we explore only the different sequences maintaining the structure frozen. Here we choose the optimised sequence in the global minimum of the free energy, which in our method corresponds to a low potential energy and a high heterogeneity of the sequence.

Starting from a fully stretched structure we then perform a FOLDING Monte Carlo simulation, in order to study the self-assembling properties of this pattern. Here we explore the conformational space keeping the pattern fixed in the designed sequence.

We project the FOLDING free energy onto an order parameter, namely the root mean square displacement of the inter-particle distance ($DRMSD$) between the target structure and each sampled structure:

$$DRMSD = \frac{1}{N}\sqrt{\sum_{ij}(|\Delta\vec{r_{ij}}| - |\Delta\vec{r_{ij}}^{T}|)^2} \tag{1.3}$$

where $\Delta\vec{r_{ij}}$ is the distance between the sphere i and j while $\Delta\vec{r_{ij}}^{T}$ is the same distance calculated over the target structure, and N is the chain length (50 in our case). Most studies dedicated to proteins adopt another order parameter, the $RMSD$. This differ from the $DRMSD$ in that its definition contains the positions of the atoms instead of their distances:

$$RMSD = \frac{1}{N}\sqrt{\sum_{ij}(|\vec{r_{ij}}| - |\vec{r_{ij}}^{T}|)^2} \tag{1.4}$$

The $DRMSD$ has already been shown to be a proper order parameter to study the folding process [59]. $DRMSD = 0$ corresponds uniquely to the target structure. The closer the global minimum is to $DRMSD = 0$, the smaller is the corresponding ensemble of structures. Thus, if the free energy landscape has a clear global minimum close to $DRMSD = 0$, we can identify at least one pattern that drives the system to fold into a unique target structure: the polymer is designable. Based on this qualitative definition we already clearly separate two groups: one with global minimum with position $DRMSD < 0.6$ and one with global minimum in the range $DRMSD \in [0.8, 1.4]$.

However, in order to define more precisely a threshold we increase the temperature and push the system to unfold. Upon increasing the temperature the minimum significantly shifts towards higher $DRMSD$ (as in A.2 in Appendix A.3), consistently with a temperature induced folding-unfolding transition. In some cases, our temperature resolution was high enough to observe two simultaneous minima. The configurations with $DRMSD$ values corresponding to the position of the second minimum are the molten globule structures. Interestingly, for the same alphabet size the position of the molten globule minimum is conserved for different number of patches (see A.2 in Appendix A.3), and for 0 patches is the only global minimum observed at all temperatures (as in A.2 in Appendix A.3). Hence, the 0 patch chain was never designable. Thus, we choose as definition for the threshold between folded and not folded for each alphabet size (grey area in Figure 1.4) the $DRMSD$ range spanning $1K_BT$ from the position of the global minimum of the free energy of the case with 0 patches. Systems with global minimum at any temperature below this area are labeled as designable, the others as non-designable.

We observed from the simulations that the position of the minimum corresponding to the molten globule does not change significantly with the number of patches while it varies with the alphabet size. The latter is because the bare heteropolymer without directional interactions, although it never folded, starts to feel the influence of the larger alphabets on its designability. Even when the system does not reach the folded state for the target structure that we identified via the SEEK, and hence the SDF trials fail, it might still be possible to find a handful of structures that are designable. However, since a heteropolymer with few and hard-to-find designable structures is not a good candidate for potential applications, we label it as not-designable.

In all Monte Carlo simulations we enhance the sampling with the Virtual Move Parallel Tempering algorithm [154], performing each simulation at 16 different temperatures in the set [3, 2.5, 2.0, 1.6, 1.4, 1.2, 1.0, 0.9, 0.8, 0.75, 0.7, 0.65, 0.6, 0.55, 0.5, 0.4]. The SEEK, DESIGN and FOLDING steps are each composed by 10 independent simulations, run until we observe that the free energy landscape does not vary anymore and that all the 10 independent simulations give the same results, i.e. the free energy surfaces for each of the 10 independent simulations overlap within the statistical error. So each time we categorise a minimum in the FOLDING free energy landscape as folded or non-folded, the minimum is found always at the same position over 10 independent simulations.

For the radial distribution functions of proteins, the normalisation has been performed on the same ideal gas with an average density, to make the $g(r)$ of proteins with different lengths comparable. All the $g(r)$ have been calculated by neglecting the contribution of the beads (or amino acids) directly connected to each other along the chain, in order to ignore their trivial contribution to the first neighbour's peak.

Chapter 2

General methodology to identify the minimum alphabet size for heteropolymer design

2.1 Introduction

Self-assembling artificial [119–121, 155, 156] and natural materials [24, 51, 124, 125, 157–162] posses the ability to organise themselves into complex and heterogeneous structures. This ability is mainly driven by reversible interactions (e.g. van der Waals interactions, hydrogen bonds, dipolar interactions, depletion interactions) [163].

Considerable effort has been spent to understand how self-assembling behaviour arises and how it can be controlled through careful design of the conformational and physicochemical properties of the primary components of the material [14, 164–166]. A particularly successful strategy to control self-assembly is to join the components into hetero-chains, e.g., proteins or small RNAs, following a specific sequence of the components along the chain. The

sequence can be optimised to drive the system to fold into a unique target structure of the whole chain. The search for such optimised sequences for a specific target structure is commonly referred to as design [43, 68, 160, 167], and a system for which it is possible to find such a sequence for at least one target structure is called designable. The mean-field theory of the random energy model (REM) [5,68,69,168] predicts that for a system to be designable it needs to respect the inequality $q > e^{\omega}$, where q is the number of different constituents (i.e., the alphabet size) and ω is the conformational entropy per monomer of the system, defined as the logarithm of the total number of accessible *compact* conformations per monomer [5,69]. This inequality can be used to identify the minimum alphabet necessary to design a heteropolymer, that is the smallest integer q_{min} greater than e^{ω}, $q_{min} \equiv \lceil e^{\omega} \rceil$. From this, it follows that reducing ω also reduces the alphabet size q_{min}. A holy grail of the field of artificial polymer-based materials is the ability to control ω such that the alphabet of components employed in the polymer synthesis is reduced to a minimum, all while keeping a large conformational space of possible target structures. It is important to stress that the *compact* polymer conformations are smaller in number than the total possible ones, hence $\omega < s$ where s is the polymer entropy. An operative definition of *compact* for off-lattice polymers is not given in the REM, making it difficult to establish a general methodology to estimate ω and in turn the designability of a heteropolymer.

Recently, Cardelli *et al.* have studied the self-assembling properties of a general heteropolymer consisting of patchy particles monomers (patchy-polymer) across the transition from $q < e^{\omega}$ to $q > e^{\omega}$ [6]. Such a study has shown that the directional interactions are the key for designability, inducing the transition to designable polymers at $q > q_{min} = \lceil e^{\omega} \rceil$. This suggests that the role of the directional interactions is that of decreasing ω by reducing the

number of accessible conformations.

In this work, we verify this intuition by introducing a general methodology to estimate ω for any polymer model and apply it to patchy polymers with different numbers of patches. Since the total number of compact conformations (and thus ω) does not depend on the specific sequence [1], we estimate ω employing a homopolymer version of the patchy polymer. Following a thermodynamic path from a reference system with known entropy, we compute the total number of conformations of the homopolymer employing advanced Monte Carlo (MC) methods [154, 169, 170]. From the total number we then select different subsets of compact conformations, according to different operative definitions of compactness. The definitions we introduce here can be easily quantified for off-lattice polymer systems.

The methodology to estimate ω allows us to test the hypothesis and prediction of the REM on the ensemble of most designable conformations [6]. Having verified the validity of the theory, we can use it to predict the size of the alphabet necessary to enlarge the ensemble of target conformations. Finally, we observe that for 4 and 6 patches the necessary alphabet grows very slowly with the space of the possible designable target conformations, making them the most versatile geometry for heteropolymer design.

2.2 Results

In this work we use the freely-rotating-chain (FRC) patchy polymer model, introduced in Ref [6]. Each bead has a diameter σ and presents M patches (inset in Fig. 2.5), whose arrangement on the bead surface is chosen according

[1] For each sequence some conformations are more likely than others, but this does not prevent the polymer to explore them.

to the most symmetrical geometry (equidistant on the equator for $M = 3$, on a tetrahedron for $M = 4$, on an octahedron for M=6 and see Supplementary Informations for $M = 10$).

The FRC is a general model for a designable heteropolymer, for which it has been demonstrated that the designability can be controlled by the number of patches $M = 0, 1, 3, 4, 6, 10$ [6]. However, since ω does not depend on the sequence, from now on we employ a homopolymer version of the patchy polymer (see Eq. 2.8). We consider the same patch arrangements of Ref. [6], shown at the bottom of Fig. 2.5. For these cases, we aim at computing the conformational entropy Ω—that, according to the REM [5] is defined as the logarithm of the number of *compact* conformations $\mathscr{N}_c$ of a heteropolymer composed of N monomers—as a function of the number of directional contacts between the patches, P, and of the number of isotropic contacts between the beads, Q (see Methods). Hence, we have to compute the total conformational entropy $s_{pp}(P, Q) = \ln(\mathscr{N}_{pp}(P, Q))$ of the patchy polymer, that is the logarithm of the total number of polymer conformations $\mathscr{N}_{pp}(P, Q)$, and from the total number select the subset of the compact conformations.

The total conformational entropy $s_{pp}(P, Q)$ computed with Monte Carlo simulations is not an absolute entropy, it is defined up to an arbitrary constant. For this reason, as explained in the following and represented in Fig. 2.1, at first we need to compute the absolute entropy of a self avoiding polymer $s_{saw} = \ln(\mathscr{N}_{saw})$, where $\mathscr{N}_{saw}$ is the number of conformations of a self-avoiding chain. It is important to stress that the self-avoiding polymer model is equivalent to the FRC model when the bead-bead and patch-patch interactions are switched off (see Eq. 2.9).

To properly compute s_{saw}, we need to know the number of conformations

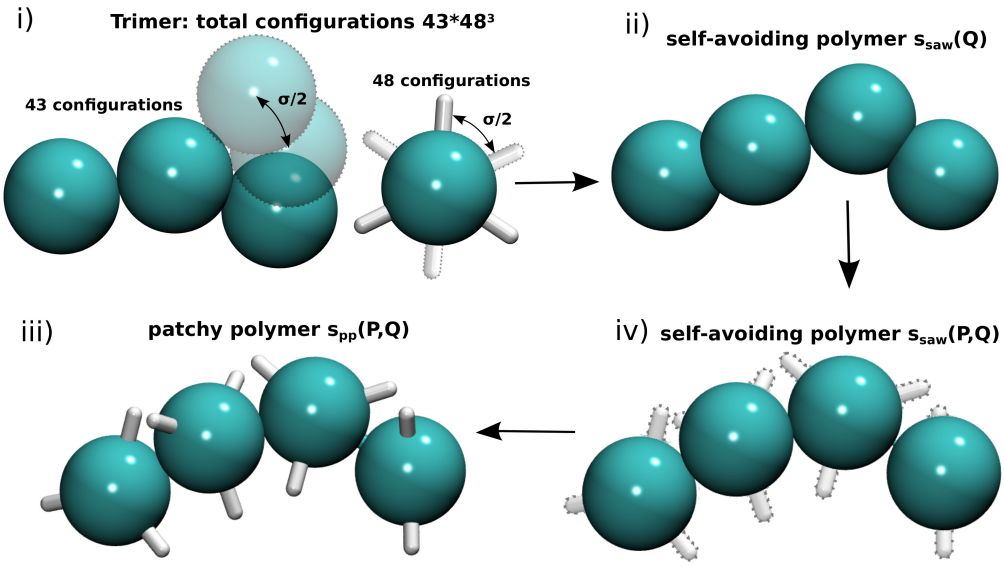

Figure 2.1: Scheme of the path used to calculate the total conformation entropy of the patchy homopolymer starting from an analytically known system (trimer). i) The absolute total entropy $s_{saw}|_{N=3}$ of a self-avoiding trimer is calculated analytically. Accordingly, the absolute total entropy for the self-avoiding polymer with $N = 50$ $s_{saw}|_{N=50}$ is calculated, employing the gran canonical method in Ref. [169, 170]. ii) With the same method, we calculate $s_{saw}(Q)$ for $N = 50$ as a function of the number of isotropic contacts Q, and shift it so that $\ln \int \mathscr{N}_{saw}(Q)dQ = s_{saw}|_{N=50}$ defined in Eq. 2.2. iii) For $N = 50$ we calculate via the enhanced MC method in Ref. [154] $s_{saw}(P,Q) = \ln \mathscr{N}_{saw}(P,Q)$ with the additional variable P, naming the number of patch-patch contacts. We shift $s_{saw}(P,Q)$ so as $\ln \int \mathscr{N}_{saw}(P,Q)dP$ overlaps with $s_{saw}(Q)$ (as in Fig. S1 in the Supporting Information). iv) For $N = 50$ we calculate via the enhanced MC method in Ref. [154] for the patchy homopolymer $s_{pp}(P,Q)$ and we shift it upon $s_{saw}(P,Q)$.

of a reference state. The chosen reference state is a trimer of self-avoiding bonded beads, whose entropy $s_{saw}|_{N=3}$ can be calculated analytically, provided a characteristic length scale is introduced to discretise the continuum

space. The REM [5] introduces $a \equiv \sigma/2$ as the radius of the tube following one chain. All the other conformations that fall within it are considered equivalent. Following this definition we consider $a \equiv \sigma/2$ as the characteristic length. Since the bond spring is strong, we assume that each bead can only be placed on the surface of the other one. Hence, the number of conformations of each bead is simply the number of ways we can place a sphere of diameter σ on the surface of an identical sphere. The close packing of hard spheres would give 12 different arrangements, where each conformation is distant by σ (or more) from the other, along with the surface. By partitioning the total surface into sections of side length a, we obtain $12 \times 4 = 48$ different conformations. In fact, starting from the close packing conformations, each bead can be shifted of a in four directions along with the surface getting a different conformation. In the trimer, we fix the position of the first bead to avoid over-counting configurations related by translations of the entire chain. The position of the second bead with respect to the first one is also fixed, to avoid over-counting configurations related by rotations of the entire chain. From the total 48 conformations that we should consider for the third bead, we have to subtract 5 because the third bead is not allowed to overlap (completely or partially) with the first one. On top of this, we have to count also the rotational degrees of freedom of all the beads because the patches break the rotational invariance symmetry. Following the same criteria, each bead can be rotated in 48 different conformations, leading to 48^3 rotational states for the trimer. Hence, the total trimer's conformational entropy $s_{saw}|_{N=3,a=\frac{\sigma}{2}}$ is:

$$s_{saw}|_{N=3} = \ln(43) + 3\ln(48) \tag{2.1}$$

Starting from the trimer as the reference system, following the gran canonical method proposed in Ref. [169, 170], we can calculate the absolute value

of the total entropy for a self avoiding polymer of length $N = 50$. The value calculated from the simulation $s_{saw}^{simul}|_{N=50}$ needs to be shifted considering the absolute entropy of the trimer, thus by the difference between the absolute value in Eq. 2.1 and the value calculated by the simulation $s_{saw}^{simul}|_{N=3} + 3\ln(48)$. Note that the term $s_{saw}^{simul}|_{N=3}$ does not take into account the rotational degrees of freedom, because they are not included in the simulation method used. For the same reason, we also have to sum $50\ln(48)$, the rotational degrees of freedom for $N = 50$. Thus the total absolute entropy for $N = 50$ results as following:

$$\begin{aligned} s_{saw}|_{N=50} = s_{saw}|_{N=3} - \left(s_{saw}^{simul}|_{N=3} + 3\ln(48)\right) + \\ + s_{saw}^{simul}|_{N=50} + 50\ln(48) \end{aligned} \tag{2.2}$$

Then, we compute with the same method $s_{saw}(Q)$: it is important to stress that such a method does not include the sampling over the patch-patch bond collective variable P and thus samples conformational entropy $s_{saw}(Q)|_{N=50}$ only as a function of the number of isotropic contacts Q. We shift such surface $s_{saw}(Q)$ in such a way, that the total entropy is corresponding to the absolute value of $s_{saw}|_{N=50}$ in Eq. 2.2:

$$\ln \int \mathscr{N}_{saw}(Q)dQ = s_{saw}|_{N=50} \tag{2.3}$$

Then, we compute via Monte Carlo simulations enhanced via the Virtual Move Parallel Tempering (VMPT) algorithm [154] both $s_{saw}(P,Q)$ with the additional collective variable P and $s_{pp}(P,Q)$ for the patchy homopolymer (see Eq. 2.8). By firstly aligning the surfaces $s_{pp}(P,Q)$ and $s_{saw}(P,Q)$, and then aligning the curves $\ln \int \mathscr{N}_{saw}(P,Q)dP$ and $s_{saw}(Q)$, we get the absolute surface $s_{pp}(P,Q)$.

We observe a good agreement between the absolute conformational entropy of the self-avoiding polymer $s_{saw}(Q)$ calculated via the gran canoni-

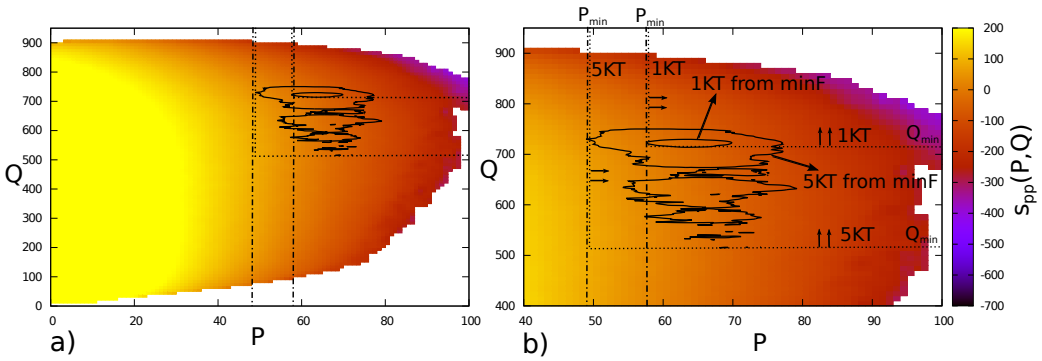

Figure 2.2: Shifted $s_{pp}(P,Q)$ for 3 patches, with indicated the different areas of $s_{pp}(P,Q)$ defined as compact conformations to calculate ω. a) total surface. b) zoom of the data shown in panel a). The areas arise from the different definition of compact. The solid curves enclose the compact conformations within 1 k_bT (inner curve) and 5 k_bT (rough outer curve) from the free energy minimum $minF$. The vertical and horizontal lines represent respectively P_{min} and Q_{min} calculated as the lowest P and Q within 1 k_bT and 5 k_bT from $minF$. The dotted lines enclose the corresponding areas at their right and above, as the compactness has been defined as $P > P_{min}$ and $Q > Q_{min}$. For the definition of compactness based on the criterion $P > P_{min}$ the whole areas to the right of the dashed-dotted lines have been considered.

cal method and $\ln \int \mathcal{N}_{saw}(P,Q)dP$, calculated via the VMPT MC simulation [154], in the range of Q that they both sample (see Fig B.1). The VMPT method samples a much wider range of Q, thanks to its sampling enhancement. Fig. 2.2 panel b) shows the conformational entropy for the patchy-polymer $s_{pp}(P,Q)$, shifted as described in Methods. We observe that $s_{pp}(P,Q)$ has its maximum for low P and Q, while the free energy minimum was found in the area with higher P and Q enclosed in the ellipse labelled in panel b) as "1 k_bT from minF". This observation is the first indication that the number of populated conformations at low temperatures are reduced by

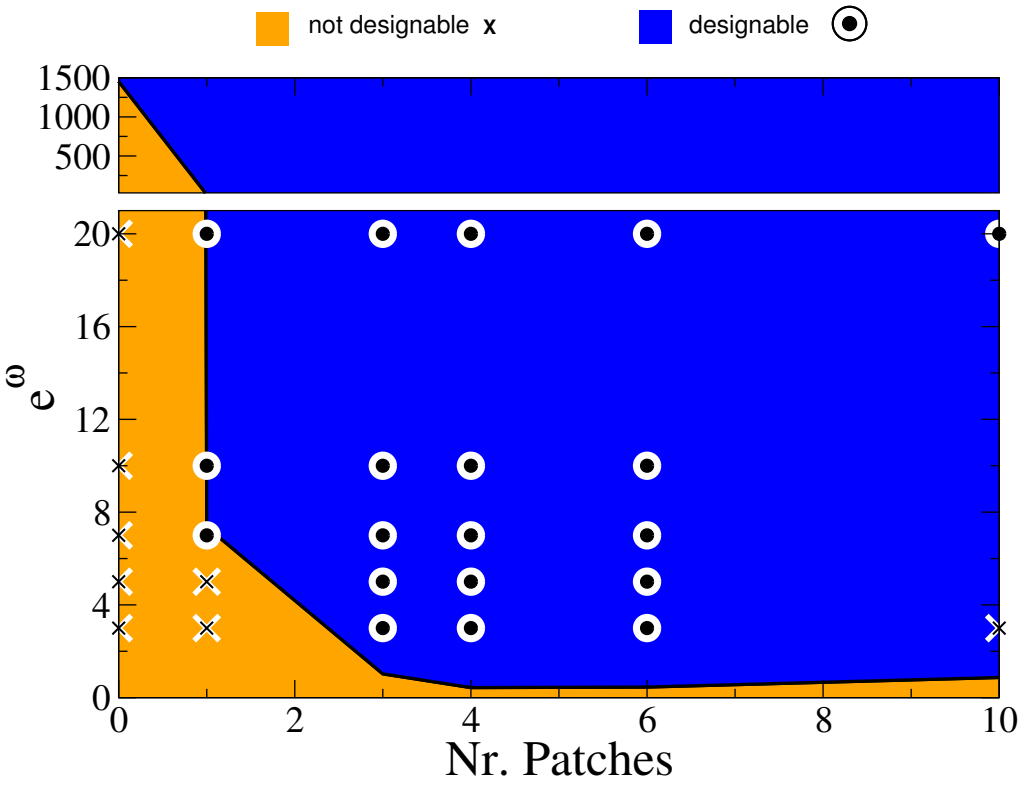

Figure 2.3: The broken line is $\exp(\omega)$ calculated on the ensemble of compact conformations within 1 k_bT from the free energy minimum $minF$ (enclosed in the ellipse of 1 k_bT form $minF$ in Fig. 2.2), for patches M=0,1,3,4,6,10. $\exp(\omega)$ in the REM represents the alphabet q_{min} at which the transition between not designable and designable occurs. Accordingly two areas are defined: yellow area (not designable) and blue area (designable). The points (circles and crosses) represent the alphabets verified in [6]: the circles are the cases verified as designable while the crosses the ones verified as not designable.

the presence of the directional interactions between the patches.

From the absolute value of $s_{pp}(P,Q)$ we now can calculate $q_{min} = \lceil e^{\omega} \rceil$ by counting the number of compact conformations ω. According, to the REM, the compact off-lattice conformations are the ones with "small temporal and

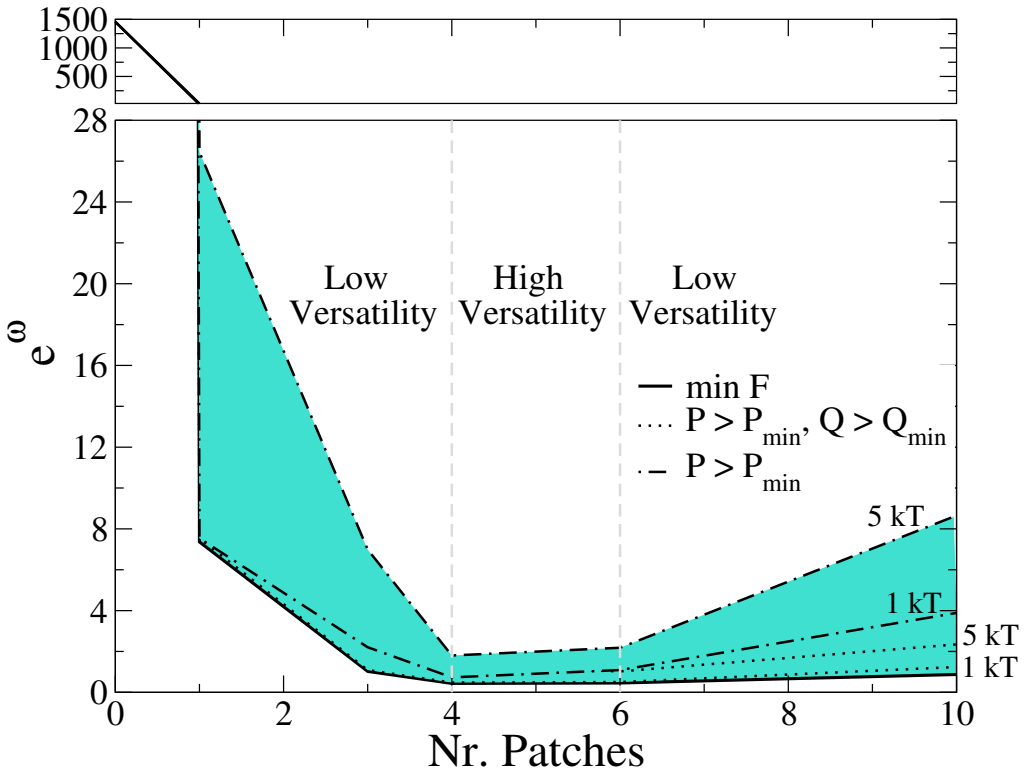

Figure 2.4: $e^{\omega} = q_{min}$ alphabet at which the transition between not designable and designable occurs vs the number of patches. The curves represent the calculated alphabet $\exp(\omega)$ with different definitions of the ensemble of compact structures, as defined in Fig. 2.2. The lowest solid line is two overlapping lines: e^{ω} calculated on the ensemble enclosed in the curve of 1 k_bT form $minF$ and 5 k_bT form $minF$ in Fig. 2.2. The dashed and dotted-dashed lines correspond to e^{ω} calculated within the ensembles defined by the corresponding vertical and horizontal lines in Fig 2.2. We observe that for the systems with 1 and 10 patches the minimum alphabet at which the system is designable increases considerably by increasing the ensemble of possible target conformations, while for the systems with 4 and 6 patches it remains low.

spatial density fluctuations" [5], but no operative definition is given to identify them. To overcome this problem, here we calculate e^{ω} as the sum of the

number of conformations found in the basin within 1 k_BT from min F, as indicated by the inner ellipse in Fig 2.2. Such a definition is based on the operative definition of the designability given in the work of Cardelli et al. [6], where the design is performed for the structures found within 1 k_BT of the global free energy minimum of the $P - Q$ space. Hence, our definition of compactness includes a measure of the effect of the directionality on the conformational entropy ω.

In Fig. 2.3, we show the number of compact conformations per monomer (e^{ω}) for different patch numbers. e^{ω}, following the REM, predicts the minimum alphabet size q_{min} for the system to be designable, i.e. it traces the transition line between designable and not designable alphabets for different patch numbers. In Fig. 2.3, we compare this transition line with the designability categorisation given in [6]. We observe considerable accordance between the two methods: the designable points verified in [6] are within the designable area predicted by REM, with the exception of the 10 patches case.

In Fig. 2.4, we extend the predictions of the REM by testing different definitions of compactness that correspond to larger basins of possible target conformations, to predict how much would the minimum alphabet grows with a bigger variability of target conformations. The dotted lines in Fig. 2.4 represent e^{ω} calculated summing all compact conformations with $P > P_{min}$ and $Q > Q_{min}$ as delimited by the dotted lines in Fig 2.2, where P_{min} and Q_{min} are the lowest P and Q values found within 1 k_bT (or 5 k_bT) $minF$. The dashed-dotted lines in Fig. 2.4 represent e^{ω} calculated summing all compact conformations with $P > P_{min}$ for 1 k_bT (or 5 k_bT), as delimited by the corresponding dotted lines in Fig. 2.2. We observe that by enlarging the basin of possible target conformations e^{ω} increases, meaning that the system

requires a larger minimum alphabet in order to be designable. However, in the systems with 4 and 6 patches q_{min} does not increase significantly by increasing the basin of possible target structures, therefore with these number of patches one can choose different target structures that maintain designability with low alphabet sizes, thus allowing a higher variability of structures. Finally, the system with 10 patches loses more easily designability by increasing the variability of target structures, this is explained by the loss in directionality that makes the monomer with M=10 act as a bigger self-avoiding sphere, as shown in [6].

2.3 Conclusions

In this manuscript, we present the first accurate calculations of the REM conformational entropy ω for a patchy polymer as a function of the number of patches. The values of ω are absolute as they were obtained with a thermodynamic path starting from the analytically solvable model of a self-avoiding trimer all the way to the full patchy polymer. The calculation of the path required extensive simulations for each number of patches investigated.

The main conclusion that can be drawn from our work is that the inequality $q > e^{\omega}$, defining the minimal alphabet for heteropolymer design, is valid and accurately reproduces the designability phase diagram previously obtained by Cardelli et al. [6]. Moreover, the inequality allows the prediction the minimal alphabet necessary to expand the space of target structures for the design. This knowledge combined with our methodology for the calculation of ω provides a powerful tool for heteropolymer engineering. Finally, we could identify in the 4 and 6 patches geometries the optimal ones for heteropolymer design since the minimal alphabet grows very slowly with

the space of different conformations included in the calculations of ω. Our methodology can be directly transferred to any heteropolymer model to establish the minimal alphabet making the system designable. In particular, we think it will be particularly effective in cases where the conformational space is restricted by directional interactions such as hydrogen bonding (Proteins, RNA, the building blocks in DNA origami) or dipolar interactions (dipolar or magnetic colloids). Above all, the method can be applied to compare the designability of different protein models and estimate the minimum alphabet necessary to design them, that seems to be already of 4 letters for the Caterpillar coarse-grained model [59], as suggested by an upcoming work of some of the present authors.

2.4 Methods

FRC model details

The full Hamiltonian of the FRC patchy polymer is given by:

$$H_{FRC} = \sum_{i<j}^{N} U(R_{ij}) + 4 \sum_{i+2<j}^{N} \sum_{l,k=1}^{M} U_{pp}\left(\mathbf{R}_i\,;\,\mathbf{R}_j\,;\,\mathbf{r}_{l,i}\,;\,\mathbf{r}_{k,j}\right) + \\ + \sum_{i=1}^{N-1} U_B(R_{i\,i+1}) + \sum_{i+2<j}^{N} U_{bb}(R_{ij}) \tag{2.4}$$

where N is the total number of beads, and R_{ij} is the distance between the beads i and j. The indexes i, j run over the N beads, while the indexes l and k run on the M patches of the beads i and j, respectively.

The first term represents a self-avoiding potential acting between all the beads and is defined as $U(R_{ij}) \equiv \infty$ if $R_{ij} \leq \sigma$ and $U(R_{ij}) = 0$ otherwise.

The second term represents the directional interaction and is given by the potential derived by Irbäck et al. [140], commonly used to model hydrogen

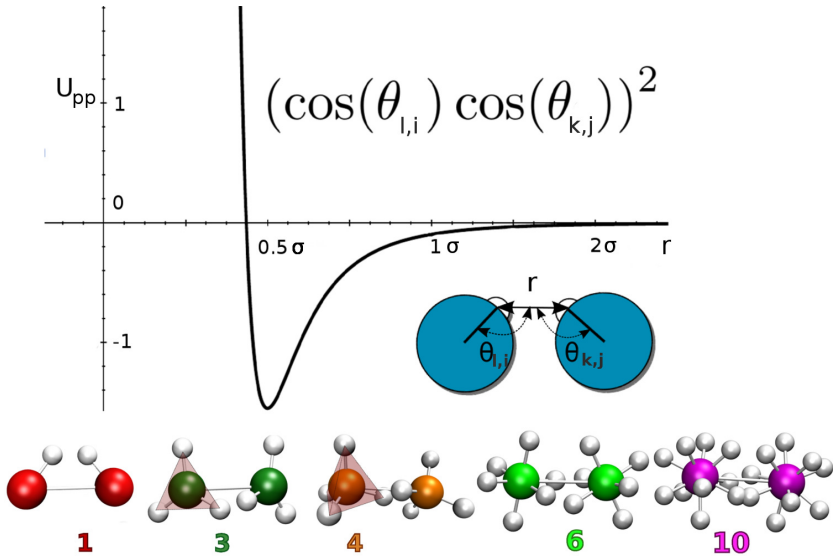

Figure 2.5: Patchy polymer model. Top: Directional potential of interaction between the patches: the radial Lennard-Jones contribution in the plot is multiplied by the directional contribution $(\cos\theta_{l,i}\cos\theta_{k,j})^2$, where $\theta_{l,i}$ and $\theta_{k,j}$ are the angles between the vector that connect each patch to the respective bead center and the vector that connects the two patches (right inset). Bottom: schematic illustration of two consecutive beads of the patchy polymers. The central bead is represented smaller than its size for the sake of visualisation. The small white spheres indicate the patches that are placed on the surface of the bead.

bonds

$$\begin{aligned} & U_{pp}\left(\mathbf{R}_i\,;\,\mathbf{R}_j\,;\,\mathbf{r}_{l,i}\,;\,\mathbf{r}_{k,j}\right) \equiv \\ & \equiv \epsilon_p \left(\cos\theta_{l,i}\cos\theta_{k,j}\right)^{\nu} \left[5\left(\frac{\sigma}{2r_{lk}}\right)^{12} - 6\left(\frac{\sigma}{2r_{lk}}\right)^{10}\right], \end{aligned} \tag{2.5}$$

where the $\theta_{l,i}$ and $\theta_{k,j}$ are the angles between the vector $\mathbf{r}_{lk} \equiv \mathbf{r}_{l,i} - \mathbf{r}_{k,j}$ and the vectors $\mathbf{r}_{l,i} - \mathbf{R}_i$ and $\mathbf{r}_{k,j} - \mathbf{R}_j$, connecting the patches l and k to the center of their bead, respectively (Fig. 2.5). Accordingly, we have $r_{lk} \equiv |\mathbf{r}_{lk}|$. The potential U_{pp} tends to align the vectors $\mathbf{r}_{lk}$, $\mathbf{r}_{l,i} - \mathbf{R}_i$ and $\mathbf{r}_{k,j} - \mathbf{R}_j$ with minimum when the patches are at distance $r_{kl} = \sigma/4$, representing a

conformation where the patches face each other, while it vanishes when $\mathbf{r}_{lk}$ is orthogonal to one of the vectors $\mathbf{r}_{l,i} - \mathbf{R}_i$ or $\mathbf{r}_{k,j} - \mathbf{R}_j$. When $\theta_{l,i} < \pi/2$ or $\theta_{k,j} < \pi/2$ we fix $U_{pp} = 0$ to avoid anti-parallel arrangements of the patches. U_{pp} is cut off at $R_{i,j} = 1.5\sigma$. For the values of the prefactor and exponent we take $\epsilon_p = 3.1\ k_BT$ and $\nu = 2$ [140]. Note that the term U_{pp} excludes the interactions between beads that are first and second neighbours along the backbone.

The third term in the Hamiltonian is a harmonic bonding potential, connecting consecutive beads along the chain

$$U_B\,(i,\, i+1) = K(R_{i\ i+1} - \sigma)^2 \tag{2.6}$$

where $K = 10\ k_BT$. The last term represents the isotropic bead-bead interaction and is modelled with the potential

$$U_{bb}(R_{ij}) = \epsilon_{ij}\left[\frac{1}{1.0 + \mathrm{e}^{-2.5(3\sigma - R_{ij})}}\right] \tag{2.7}$$

where ϵ_{ij} is a sequence-dependent prefactor depending on the types of bead i and j. U_{bb} is essentially a sigmoidal function with constant value $U_{bb} \sim \epsilon_{ij}$ for $R_{ij} \leq 2.5\sigma$, decaying to $U_{bb} \sim 0$ for $R_{ij} \geq 3.5\sigma$, with the inflection point $U_{bb}(3\sigma) = \epsilon_{ij}/2$. The factor 4 in front of the term U_{pp} is fixed to balance the directional contributions U_{pp} with respect to the isotropic one U_{bb}.

The FRC homopolymer Hamiltonian is obtained by setting this last term $U_{bb} = 0$ in Eq. 3.1:

$$\begin{aligned} H_{homo} = &\sum_{i<j}^{N} U(R_{ij}) + s\sum_{i+2<j}^{N}\sum_{l,k=1}^{M} U_{pp}\,(\,\mathbf{R}_i\,;\,\mathbf{R}_j\,;\,\mathbf{r}_{l,i}\,;\,\mathbf{r}_{k,j}\,) + \\ &+ \sum_{i=1}^{N-1} U_B(R_{i\ i+1}), \end{aligned} \tag{2.8}$$

while the self-avoiding chain Hamiltonian is obtained by setting $U_{bb} = 0$ and

$U_{pp} = 0$ in Eq. 3.1:

$$H_{saw} = \sum_{i<j}^{N} U(R_{ij}) + \sum_{i=1}^{N-1} U_B(R_{i\ i+1}), \tag{2.9}$$

Entropy Sampling

We employ Monte Carlo simulations to compute the total conformational entropy i. e., the logarithm of the total number of polymer conformations, s_{saw} and s_{pp} of the self-avoiding polymer (see Eq. 2.9) and a patchy homopolymer (see Eq. 2.8) respectively. Both are homopolymer models, since s_{saw} and s_{pp} (and Ω) do not depend on the specific sequence. The entropy of the patchy polymer s_{pp} is computed via canonical Monte Carlo (MC) simulations where only the polymer conformations are sampled via a set of conformational moves. The conformational entropy is projected onto both the number of contacts Q and the number of patch-patch contacts P.

The number of bead-bead isotropic contacts Q is associated to the U_{bb} term:

$$Q = \sum_{i+2<j}^{N} \left[\frac{1}{1.0 + \mathrm{e}^{-2.5(3\sigma - R_{ij})}} \right]; \tag{2.10}$$

and the number of directional contacts P is connected instead to the U_{pp} term. P is given by the number of $i-j$ pairs of patches that satisfy the conditions: $r_{ij} < 0.625\ \sigma$, $\theta_{l,i} > 0.8\ \pi$ and $\theta_{k,j} > 0.8\ \pi$ (as shown in Fig. 2.5).

We enhance the sampling with the Virtual Move Parallel Tempering (VMPT) algorithm [154], iteratively accumulating the biasing potential over Q and P, and performing each simulation at 32 different temperatures in the set [3.0 2.75 2.5 2.0 1.9 1.7 1.6 1.5 1.45 1.4 1.35 1.3 1.25 1.2 1.15 1.05 1.025 1.0 0.95 0.925 0.9 0.8 0.75 0.7 0.65 0.6 0.590 0.575 0.55 0.5 0.45 0.4]. For each patch geometry, we perform 10 independent simulations run in subsequent simulation blocks of 10^9 MC steps. We repeat the iterations until

we observe that the conformational entropy landscape does not show an appreciable difference from the previous run and among the 10 independent simulations.

The entropy of the self-avoiding chain s_{saw} is computed both with the VMPT scheme described above ($s_{saw}(P,Q)$) and with the gran canonical scheme introduced in Ref. [169, 170] ($s_{saw}(Q)$). The latter method allows to compute the ratio between the partition functions of polymers with N and $N-1$ monomers via a gran canonical Monte Carlo simulation, where the chain grows by adding one bead at a time and sampling all its possible conformation. Hence, starting from a trimer it is possible to compute the entropy difference between $N = 50$ and $N = 3$, which is analytically known from Eq. 2.1.

To pass from the curve $s_{saw}(Q)$ to the surface $s_{saw}(P,Q)$ for a self avoiding polymer we perform new Monte Carlo simulations of the FRC self-avoiding patchy-polymer (defined in Eq. 2.9) for each value of M. We sample its conformational entropy $s_{saw}(P,Q)$ with the additional collective variable P. By integrating over the order parameter P, $\ln \int \mathcal{N}_{saw}(P,Q)dP$, we obtain the same curve $s_{saw}(Q)$ up to an additive constant. Thus, by aligning $\ln \int \mathcal{N}_{saw}(P,Q)dP$ with the absolute entropy $s_{saw}(Q)$ we compute the proper shift for the surface $s_{saw}(P,Q)$. A key difference between the two simulations we employed to compute $s_{saw}(Q)$ and $s_{saw}(P,Q)$ is that in the latter also single particle rotational moves are performed, allowing us to create a path between the bare self-avoiding polymer and the full patchy-polymer. Moreover, it is worth noticing that we could not compute directly the absolute value of $s_{saw}(P,Q)$ with the gran canonical method of Ref. [169, 170] because it does not sample properly the polymer conformations corresponding to large values of P.

The last step consists of a Monte Carlo simulation of the patchy-polymer with the patches interaction enabled, allowing us to sample the conformational entropy $s_{pp}(P, Q)$ for different number of patches. The computed surface $s_{pp}(P, Q)$ is then aligned with $s_{saw}(P, Q)$ to get the absolute entropy for the patchy-polymer.

All the shifts are obtained by minimising the sum of square distances of the points of the conformational entropy profiles.

Chapter 3

Heteropolymer design and folding of knots

3.1 Introduction

Knots have been used since ancient times to perform a variety of tasks: tying objects or animals, securing ropes, sewing, etc. Advancements in mathematics, physics, biology and chemistry [171] in the past half-century have made clear that knots are not only macroscopic object, but can arise on polymers and molecules as well, either as the result of spontaneous entanglement [71–74] or through the action of enzymes in DNA [81, 95, 96], or as specific folds in proteins [84–88]. Knots have been reported even in clusters of dipolar spheres [79] and in the disclination lines formed by colloidal dispersions in cholesteric liquid crystals [172–174]. Molecular knots, much like their macroscopic counterparts, change the physical properties of the chains on which they are tied, and can serve specific functions, for example enhancing mechanical and thermodynamical stability in knotted proteins [84, 86, 100–103] or polymers [61].

There is, in fact, growing interest in synthesizing specific artificial knots, which could be used for example as reusable catalysts for different chemical reactions due to their stability [109] or to construct drug delivery carriers [104]. In the past 30 years, topological synthetic chemistry has taken considerable steps toward producing knotted molecules using metal coordination and helically shaped ligands [99, 175]. Nonetheless, only 4 type of knots can be produced as of today: the trefoil [110], Savoy or figure-eight knot [113], pentafoil [111, 112] and 8_{19} knots [114]. Computational investigations [118, 176, 177] show that self-assembly of helical templates, mimicking the helical ligands used in topological synthetic chemistry, reproduce a mixture of different knots including more complex ones than the ones mentioned above, although only in small traces.

However, to this date, the synthesis of a specific knot with controllable topology and precise structure is still a challenging task. A possible strategy to progress in this direction is to imitate protein folding, employing heteropolymers. This can be achieved in principle by introducing so called-patchy polymers [6, 61, 62], chains composed by different monomers characterised by the presence of patches on their surface, which guarantee a directional interaction akin to hydrogen bonding in proteins. Monomer pairs interact not only via this short range directional interaction, but also through an alphabet of different isotropic attractions/repulsions. The key parameters of the models are the alphabet size of the isotropic interactions q and the number of patches M (which is the same for all the monomers in a chain). Computational studies have shown that patchy polymers can fold precisely into a variety of three-dimensional structures which are dictated by the heteropolymer sequence along the chain [6, 61, 62]. Furthermore, the presence of knotted folds has been reported in a previous study [61], where the authors

show that a knotted structure externally locked by controlling the interaction between the end monomers is stabilised for temperature unfolding.

Here we consider two standard heteropolymer models where heterogeneous isotropically-interacting monomers are bonded along the chain *via* a harmonic potential with two different anchoring geometries: i) in the freely rotating chain (FRC) model the spring binds the centres of the monomers; ii) in the freely jointed chains (FJC) model the spring connects two anchoring points on the monomer surface and opposite to each other (see bottom of Fig. 3.1 for a sketch). Possible experimental realisations of the FJC chains could be surface grafted colloidal particles or covalently bonded chemical units. Instead, constructing a FRC chain would need patchy particles where the patches are free to rotate with respect to the central bead. Examples of such are the solid colloids with surface-mobile DNA linkers in Ref. [149], the lock and key colloids as in Ref. [148], or emulsion droplets with mobile DNA patches [150].

We study the space of possible knots in chains of 50 monomers of both the FJC model with $M = 1, 2$ patches per bead and the FRC model with $M = 1, 3$, each with two alphabet sizes $q = 3$ and $q = 20$. We show that the knot spectrum is controlled by the number of patches and the alphabet size and presents a broad variety of knots of complexity up to 20 crossings. Interestingly, increasing the alphabet size tends to simplify the knot spectrum. We design the sequence to obtain precisely folded specific knot topologies, including the twist knot 5_2 and complex pretzel knot 10_{124}, both of which have not been synthesised.

3.2 Methods

3.2.1 Patchy polymer models

The full Hamiltonian of the patchy polymer is given by:

$$H = \sum_{i<j}^{N} U(R_{ij}) + \sum_{i+2<j}^{N} U_{bb}(R_{ij}) + s \sum_{i+2<j}^{N} \sum_{l,k=1}^{M} U_{pp}\left(\mathbf{R}_i\,;\,\mathbf{R}_j\,;\,\mathbf{r}_{l,i}\,;\,\mathbf{r}_{k,j}\right) + \sum_{i=1}^{N-1} U_B(R_{i\,i+1}) \tag{3.1}$$

where N is the total number of beads, and R_{ij} is the distance between the beads i and j. The indices i, j run over the N beads, while the indices l and k run on the M patches of the beads i and j, respectively.

The first term represents a self-avoiding potential acting between all the beads and is defined as $U(R_{ij}) \equiv \infty$ if $R_{ij} \leq 1\sigma$ and $U(R_{ij}) = 0$ otherwise. The subsequent three energy terms in the Hamiltonian account for the isotropic interaction between two beads, U_{bb}; the directional interaction between patches, U_{pp}; and the bonding energy between consecutive beads, U_B. The parameter $s = 4$ rescales the strength of the directional interaction with respect to the isotropic interaction [62].

The second term represents the isotropic bead-bead interaction and is modelled with the potential

$$U_{bb}(R_{ij}) = \epsilon_{ij}\left[\frac{1}{1.0 + \mathrm{e}^{-10(3-R_{ij})}}\right] \tag{3.2}$$

where ϵ_{ij} is a sequence-dependent prefactor depending on the types of bead i and j. The possible monomer types are defined by an alphabet of size q, giving a $q \times q$ symmetric matrix of possible values for ϵ_{ij}. The interaction matrices are generated according to a Gaussian distribution with null average and standard deviation equal to the ones typically used to model real amino acids [178–183]. In this work we consider two different alphabets with size $q = 3$ and $q = 20$ respectively; the interaction matrices are reported in the

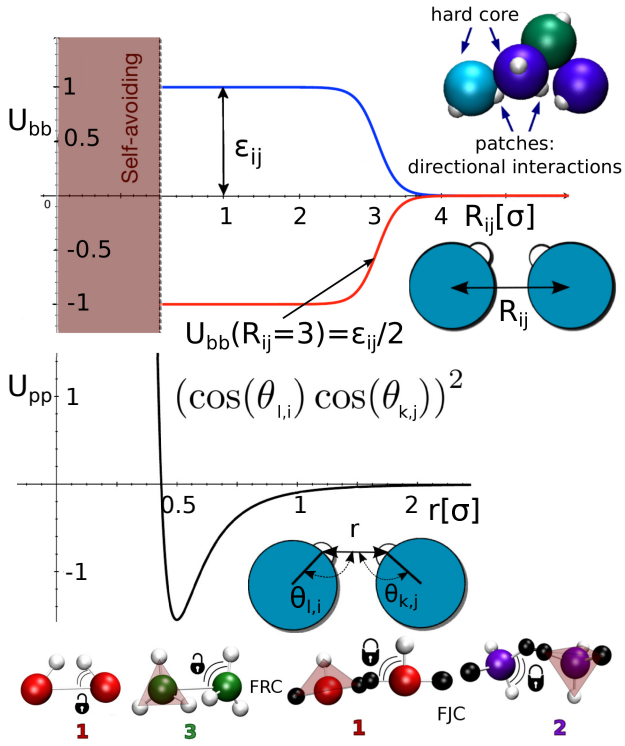

Figure 3.1: Interaction potentials. a) Isotropic square-well like interaction as a function of the distance between the centres of the spherical monomers r_{ij} (see left inset). The well depth is controlled by a pre-factor ϵ_{ij}, which is different for each different pair of monomer types. The pre-factors are grouped in an interaction matrix, the size of which depends on the number of monomer types in the range 3 to 20 (the matrices used are given in the Supplemental Material). For an alphabet of size q, the set of $q(q+1)/2$ heterogeneous interactions are Gaussian random numbers with zero average and standard deviation of 0.33 k_BT. Inset: schematic illustration of patchy polymers. Small white spheres indicate the patches, placed on the surface of the central sphere. b) Directional potential. The radial Lennard-Jones contribution in the plot is multiplied by the directional contribution $(\cos\theta_{l,i}\cos\theta_{k,j})^2$, where $\theta_{l,i}$ and $\theta_{k,j}$ are the angles between the vector that connects each patch to the respective bead centre and the vector that connects the two patches (right inset). Bottom: schematic illustration of the patch geometries. The central bead is represented smaller than its size for the sake of visualisation.

Supplemental Material. We considered in particular two matrices with a balanced number of self-attractive and self-repulsive types, since these are the most representative. In fact, with Gaussian distributed matrices, the probability of having only self-interactions of one kind is low ($\frac{1}{2}^q$).

U_{bb} is essentially a sigmoidal function with constant value $U_{bb} \sim \epsilon_{ij}$ for $R_{ij} \leq 2.5\ \sigma$, decaying to $U_{bb} \sim 0$ for $R_{ij} \geq 3.5\ \sigma$, with the inflection point $U_{bb} = \epsilon_{ij}/2$ at $3\ \sigma$. For computational reasons, the interaction is truncated when $r_{ij} > 4.25\ \sigma$.

The third term represents the directional interaction and is given by the potential derived by Irbäck et al. [140], commonly used to model hydrogen bonds

$$U_{pp}\left(\mathbf{R}_i\,;\,\mathbf{R}_j\,;\,\mathbf{r}_{l,i}\,;\,\mathbf{r}_{k,j}\right) \equiv \epsilon_p \left(\cos\theta_{l,i}\cos\theta_{k,j}\right)^{\nu} \left[5\left(\frac{1}{2r_{lk}}\right)^{12} - 6\left(\frac{1}{2r_{lk}}\right)^{10}\right], \tag{3.3}$$

where the $\theta_{l,i}$ and $\theta_{k,j}$ are the angles between the vector $\mathbf{r}_{lk} \equiv \mathbf{r}_{l,i} - \mathbf{r}_{k,j}$ and the vectors $\mathbf{r}_{l,i} - \mathbf{R}_i$ and $\mathbf{r}_{k,j} - \mathbf{R}_j$, connecting the patches l and k to the centre of their bead, respectively (Fig. 3.1). Accordingly, we have $r_{lk} \equiv |\mathbf{r}_{lk}|$. The potential U_{pp} tends to align the vectors $\mathbf{r}_{lk}$, $\mathbf{r}_{l,i} - \mathbf{R}_i$ and $\mathbf{r}_{k,j} - \mathbf{R}_j$, and has its minimum when the patches are at distance $r_{kl} = 1/4\ \sigma$, representing a conformation where the patches face each other, while it vanishes when $\mathbf{r}_{lk}$ is orthogonal to one of the vectors $\mathbf{r}_{l,i} - \mathbf{R}_i$ or $\mathbf{r}_{k,j} - \mathbf{R}_j$. When $\theta_{l,i} < \pi/2$ or $\theta_{k,j} < \pi/2$ we fix $U_{pp} = 0$ to avoid anti-parallel arrangements of the patches. U_{pp} is cut off at $R_{i,j} = 1.5\ \sigma$. For the values of the pre-factor and exponent we take $\epsilon_p = 3.1\ k_BT$ and $\nu = 2$ [140]. Note that both the terms U_{bb} and U_{pp} exclude the interactions between beads that are first and second neighbours along the backbone.

The last term, U_B, is a harmonic potential joining consecutive beads. U_B distinguishes the two classes of polymer models we consider:

$$U_B^{FRC}(i,\, i+1) = K_B(R_{i\ i+1} - 1)^2 \tag{3.4}$$

$$U_B^{FJC}(i,\, i+1) = \begin{cases} K_B(r_{v(i)\ v(i+1)} - 1/4)^2 & \text{if } r_{v(i)\ v(i+1)} > 1/4 \\ 0 & \text{if } r_{v(i)\ v(i+1)} < 1/4 \end{cases} \tag{3.5}$$

In the FRC Freely Rotating Chain model, the bond is applied directly between the centres of consecutive beads along the chain and the rest distance is equal to σ (see Fig. 3.1 a), bottom panel). In the FJC Freely Jointed Chain model, each bead has two special patches which act as anchoring points between subsequent beads along the chain (see Fig. 3.1 b), bottom panel). Given two consecutive beads i and $i+1$, in the above formula we indicate the anchoring points with $v(i)$ and $v(i+1)$. The spring rest distance in this case corresponds to the minimum of the patches potential $1/4\sigma$. For both models, $K_B = 160\ k_BT/\sigma^2$.

Except for the FJC with one patch geometry, in all the other cases the patches are fixed on the vertices of a platonic solid on the surfaces of the bead i.e., at distance $\sigma/2$ from the bead centre. From now on we will indicate with M the number of free patches, excluding the anchoring points. In the FJC with $M = 1$ the patch is placed at 90 degrees with respect to the two anchoring points (in black). In the FJC with $M = 2$ the two patches are placed on a tetrahedron completed by the anchoring points. In the FRC $M = 3$ they are placed on the vertices of an equilateral triangle. For each case, we consider two alphabet sizes $q = 3$ and $q = 20$ (see Methods for the definition of the interactions). In the FRC model, the patches can rotate solidly on the bead. In the FJC model, only the free patches can rotate

along the axis defined by the two patches involved in the bonding interaction between subsequent beads.

Monte Carlo simulations

We employ the methodology of the Monte Carlo simulations SEEK, DESIGN and FOLD (SDF), that was proven to be able to discriminate between designable and not-designable structures [6, 62]. Briefly, the SDF method: i) performs an extensive sampling of the heteropolymer conformations and sequences to chose a target structure; ii) designs the sequence that should optimally fold into the target structure; iii) tests whether the designed sequence correctly folds into the target structure.

The sampling capability of each of these steps is enhanced by using the Virtual Move Parallel Tempering with multiple replicas at different temperatures, coupled to a bias potential to escape local free energy minima [154]. The bias potential helps the simulations to correctly sample the whole space and is then removed. For each patch geometry, we perform 10 independent simulations, run in subsequent simulation blocks of 10^9 MC steps. We repeat the iterations until we observe that the conformational entropy landscape does not show an appreciable difference from the previous run and among the 10 independent simulations.

In a DESIGN simulation, we consider a fixed target structure and explore its sequence space through point mutations and monomer type swaps, with the aim to find an optimal sequence which minimises the free energy while maintaining a high heterogeneity (see ref. [61] SI). This exploration is performed using a VMPT scheme with 5 parallel simulations, each composed of 16 replicas running at temperatures [0.1, 0.3, 0.5, 0.7, 0.9, 1.2, 1.5, 2, 2.5, 3, 4, 6, 7, 8, 9, 10]. All simulations are performed with a bias potential on the

energy of a given sequence and its heterogeneity, defined as in ref. [61].

FOLD simulations are complementary to DESIGN ones. In this case, we fix the polymer sequence and explore the conformational space of the chain through a combination of crankshaft, pivot and monomer displacement moves. We consider 10 parallel simulations each running on 32 different replicas at temperatures 3.0, 2.5, 2.2, 2.0, 1.9, 1.7, 1.6, 1.55, 1.5, 1.45, 1.4, 1.35, 1.3, 1.2, 1.15, 1.1, 1.05, 1.025, 1.0, 0.99, 0.975, 0.95, 0.925, 0.9, 0.875, 0.85, 0.825, 0.8, 0.75, 0.7, 0.65 and 0.6. The free energy (see Fig. S.7 in the Supplemental Material) and bias potential are both projected on the total number of contacts between the patches P (distance below 0.625 σ and angles θ_{li} and $\theta_{kj} > 0.8\ \pi$), and the DRMSD from the target structure, defined as

$$DRMSD = \frac{1}{N}\sqrt{\sum_{ij}(R_{ij} - R_{ij}^{T})^2} \tag{3.6}$$

where R_{ij} is again the distance between beads i and j while R_{ij}^{T} is the same distance calculated over the target structure. For an example of a FOLD free energy landscape see Fig. S.7 in the Supplemental Material.

Finally, SEEK simulations combine the exploration of both conformational and sequence space at the same time. For each polymer model, we perform 10 independent simulations, each of which running a VMPT scheme with the same 32 temperatures used for FOLD.

From SEEK, we obtain free energy profiles mapped on two different pairs of global variables (See Fig. S.8 and Fig. S.9 in the Supplemental Material). The first pair is given by the total number Q of contacts between the spheres calculated as the total isotropic energy $\sum_{i,j} U_{bb}(R_{ij})$ with all $\epsilon_{ij} = 1$ and the total number P of contacts between the patches. For this pair of global variables we save a histogram of 850 Q by 100 P bins and for each bin we store the minimum energy conformation of each replica. The second pair is

the Average Crossing Number (ACN) and the end-to-end distance R_{ee} of the polymer, for which we consider a histogram of size 100 by 200 bins. The average crossing number is defined as

$$\frac{1}{2\pi}\sum_{i=1}^{N-1}\sum_{j=1}^{N-1}\frac{(\mathbf{r}_{i+1}-\mathbf{r}_i)\times(\mathbf{r}_{j+1}-\mathbf{r}_j)\cdot(\mathbf{r}_i-\mathbf{r}_j)}{|\mathbf{r}_i-\mathbf{r}_j|^3}, \tag{3.7}$$

where the $\mathbf{r}_{i+1}-\mathbf{r}_i$ approximates the vector tangent to the polymer backbone at the position of the monomer i. The ACN is a continuous variable – as needed for the bias potential – that corresponds to the average number of intersections of the "polymer shadow", meaning the average of the number of intersections over all projections of the polymer structure onto all possible planes, i.e., the planes orthogonal to all the directions passing from the centre of the polymer. The ACN correlates with the number of crossings of the knot, with higher values of ACN corresponding to more complex knots [184]. SEEK simulations are performed with a bias potential $W(ACN, P)$.

SEEK simulations are intended to provide possible targets for DESIGN and FOLDING simulations. These target conformations are given by the minimum energy conformations obtained for each bin in the histograms. We perform 10 independent simulations each with 32 replicas that exchange temperatures (not conformations), thus the conformation with minimum energy will be a low temperature conformation for each replica. Hence, we have up to 320 independent conformations at low temperature per bin.

Through SEEK we can obtain the ensemble averages of several observables. In order to do so, for a given variable a, we have to weight its value for each conformation by the unbiased probability P of said conformation within its bin in the histogram (ACN, R_{ee}):

$$P_k = \frac{e^{-\beta E_k}}{\sum_{i=1}^{n_b} e^{-\beta E_i}} \tag{3.8}$$

where n_b is the number of conformations for each bin b (up to 320). Thus, for a variable a, its ensemble average is:

$$\langle a \rangle = \frac{\sum_{b=1}^{n} (\sum_{i=1}^{n_b} a_i P_i) e^{-\beta F_b(ACN, R_{ee})}}{\sum_{b=1}^{n} e^{-\beta F_b(ACN, R_{ee})}} \tag{3.9}$$

where n is the total number of bins and $F_b(ACN, R_{ee})$ is the unbiased canonical free energy of the bin b.

Knot analysis

To assign a topological state to the conformations of linear patchy polymers, we must first of all circularise them. We do so by using the Minimally Interfering Closure [91]. The topology of each circularised chains is then identified by calculating the Dowker code of its projection and comparing it against a table including all knots up to 16 crossings. To do so, we used the KNOTFIND routine included in KNOTSCAPE [185].

To calculate the probability of a given knot topology, for instance the 3_1, we perform a SEEK ensemble average as described in eq. 3.8, taking the variable a to be:

$$a_i = \delta_{\tau_i, 3_1} \equiv \begin{cases} 1 & \text{if topology is } 3_1 \\ 0 & \text{otherwise} \end{cases} \tag{3.10}$$

where the Kronecker delta is evaluated over the topology τ of conformation i.

Results

The two models of patchy polymers considered here, Freely Jointed Chains (FJC) and Freely Rotating Chains (FRC), idealise several possible experimental realizations of patchy systems, at a scale ranging from molecular to

colloidal polymers. The demonstrated ability of patchy polymers to fold into designed three-dimensional structures [6] opens the possibility to use them to experimentally obtain a variety of different knots in a reliable way.

We consider different possible realizations of FRC and FJC. In particular, we study FRC with $M = 1$ and $M = 3$ patches per bead and FJC with $M = 1$ and $M = 2$. For both FJC and FRC we consider one alphabet of size $q = 3$ and one of size $q = 20$, giving us 4 different cases per model.

To characterise what knots can be obtained with each system, we start by analyzing the knot spectrum of FJC and FRC polymers with $N = 50$ beads, correlating the probability to obtain each knot with two different geometrical characteristics of the polymer conformations: the Average Crossing Number (ACN) and the end-to-end distance (R_{ee}). To do so, we perform a series of SEEK simulations (see methods) sampling both the conformational and sequence space at the same time. For each simulation, we project the free energy on the $ACN \times R_{ee}$ plane storing its values in a two-dimensional histogram. To each bin, we associate a population of up to 320 low energy conformations each of which is sampled independently by one of the parallel replicas in the simulation. This strategy allows us to identify the most probable knot type for each pair of ACN, R_{ee} values, guiding the design of knots with controlled geometrical properties.

In Fig. 3.2 we show the results for the FJC model with $M = 1$ since it presents the most varied knot spectrum, the corresponding results for FRC are qualitatively equivalent and are reported in the Supplemental Material. Here and in the following, we denote knot types using the Alexander-Briggs notation [186], commonly used for knots with up to 10 crossings. In this notation, each knot type is indicated in the form X_Y, where X stands for the number of crossings and Y distinguishes the knot from others with the

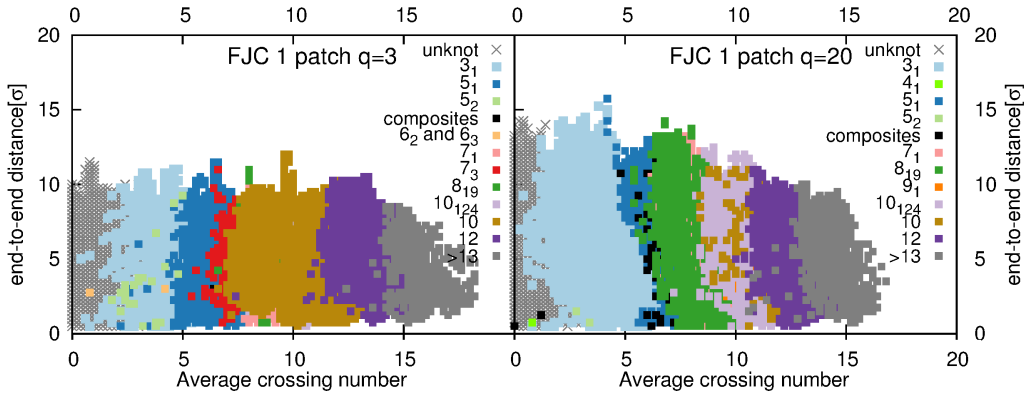

Figure 3.2: Topological 'phase diagram' of the most probable knot type for each couple of ACN, R_{ee} for the FJC model with 1 patch and with $q = 3$ and $q = 20$.

same number of crossings.

The knot spectra in Fig. 3.2 shows that the knot complexity correlates with the Average Crossing Number, both for alphabet 3 and alphabet 20, leading to the formation of several *knot domains*. In both cases, the dominating knot for a certain value of ACN has a comparable number of essential crossings, so trefoil knots dominate the spectrum for $ACN \simeq 3$, 5_1 knots for $ACN \simeq 5$, etc. Nonetheless, some knots show a great spread in ACN, indicating a vaster set of accessible conformations corresponding to them, while others are confined to a narrow set of ACN values. It is important to notice that since Fig. 3.2 only reports the most probable topology for each bin, the actual spread of each knot is greater with respect to the reported domains. Interestingly, there is also a vertical spread with some knots, like the 5_2 for FJC $M = 1, q = 3$ being more likely found when the two ends of the chain are close to each other.

Starting from the knot spectrum of each bin of the $ACN \times R_{ee}$ histogram, we obtain the total knot spectrum of each model by weighting the probability of each knot in a bin by the canonical probability of the bin, as explained in

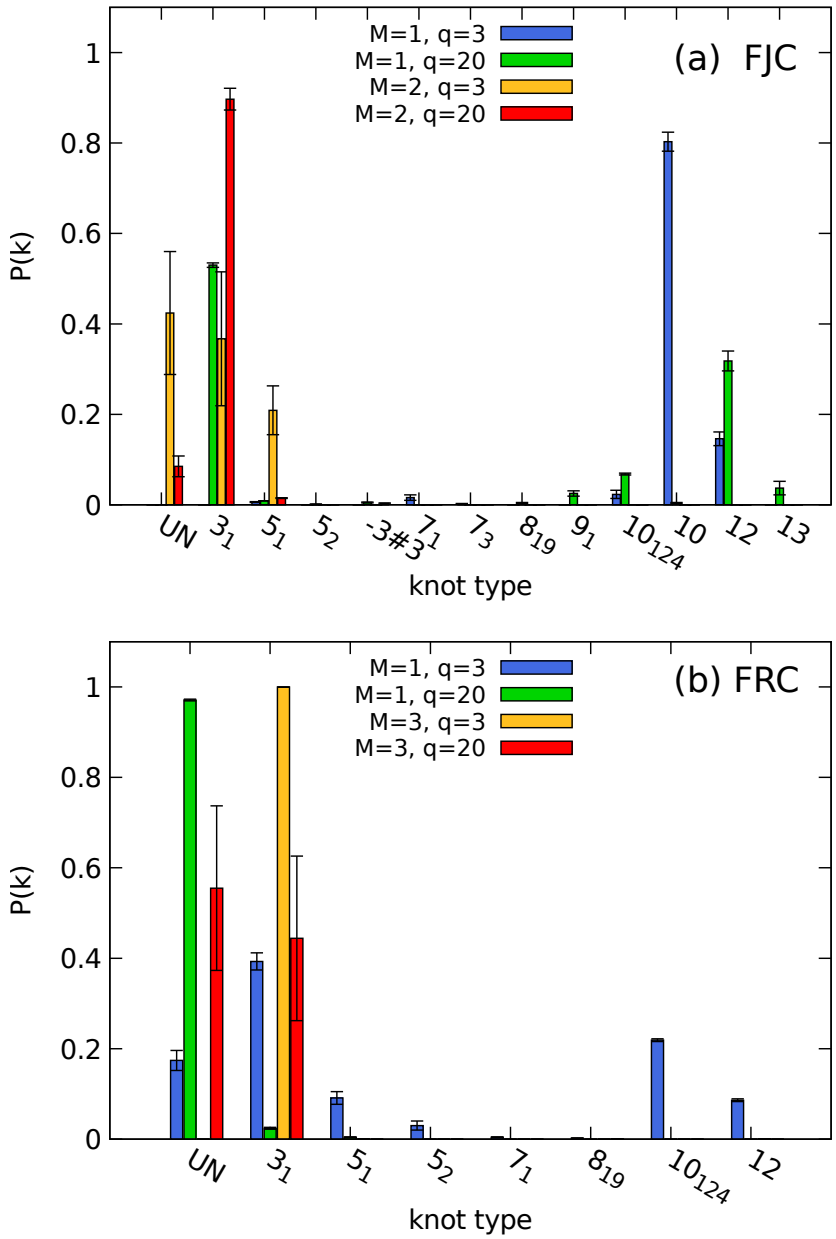

Figure 3.3: Normalised total probabilities of each knot type a) for the FJC model with $M = 1, 2$ patches and $q = 3, 20$ b) for the FRC model with $M = 1, 3$ patches and $q = 3, 20$. The error bars are represented as $\pm$ the standard deviation, calculated selecting 10 different random sets of half of the conformations in each bin.

details in Methods. The knot probability histograms so obtained are reported in Fig. 3.3 a) and b) for the FJC and FRC model respectively. These knot spectra allow us to characterise the effect of the backbone (FJC vs FRC), number of patches, and alphabet size on the probability of observing different knots.

First of all, we note that in all the cases considered here torus knots are the most abundant. In particular, 3_1, 5_1, 8_{19}, and 10_{124} are dominating topologies. Interestingly the same topologies dominate the spectrum of knots which can be formed by self-assembly of helicates [99, 111, 114, 118, 176, 177] and by low temperature phases of bead-and-stick attractive homopolymers [187]. However, patchy polymers can be used to tie twist knots as well, like the 5_2 knot that is present in all cases in traces ($< \%1$) and is relatively abundant for FRC $M = 1$, where it reaches a probability $P_{5_2} \simeq 0.03$, equal to 0.3 that of the 5_1 knot. Furthermore, as shown in the next subsection, twist knots can still be designed and folded even in the FJC model, despite their rarity.

An important result of our simulations is that both the FRC and FJC models show knot spectra that differ significantly from the one observed for self-avoiding polymers [188]. Moreover, the histograms are strongly affected by the model used, the geometry of the patches and the size of the alphabet. In particular, we notice that the models with one patch have richer knot spectra, and that increasing the alphabet from $q = 3$ to $q = 20$ favors the formation of simple or trivial knots.

In order to better understand these phenomena, we analyse the patch-patch contact distribution as a function of the arc separation s along the backbone, reported in Fig. 3.4 for the FJC model and Fig. 3.5 for the FRC model. From these we find that the systems FJC 2 patches $q = 3, 20$ in Fig. 3.4, and FRC 1 patch $q = 20$ and 3 patches $q = 3, 20$ in Fig. 3.5, have a

larger number of contacts at separations smaller than 10 beads, particularly for $s = 3$, corresponding to local kinks. As can be seen in Fig. 3.3, the same systems have a spectrum of knots skewed towards simpler topologies. The connection between the knotting probability and the distribution of local contacts was observed for homopolymers [188]. The study shows that the crumpled local structures significantly lower the probability of forming loops large enough to be pierced by other portions of the chain, and therefore lower the probability of forming knots. Hence the patches, while sculpting the conformational space allowing for designability, have a strong influence on the knot spectrum.

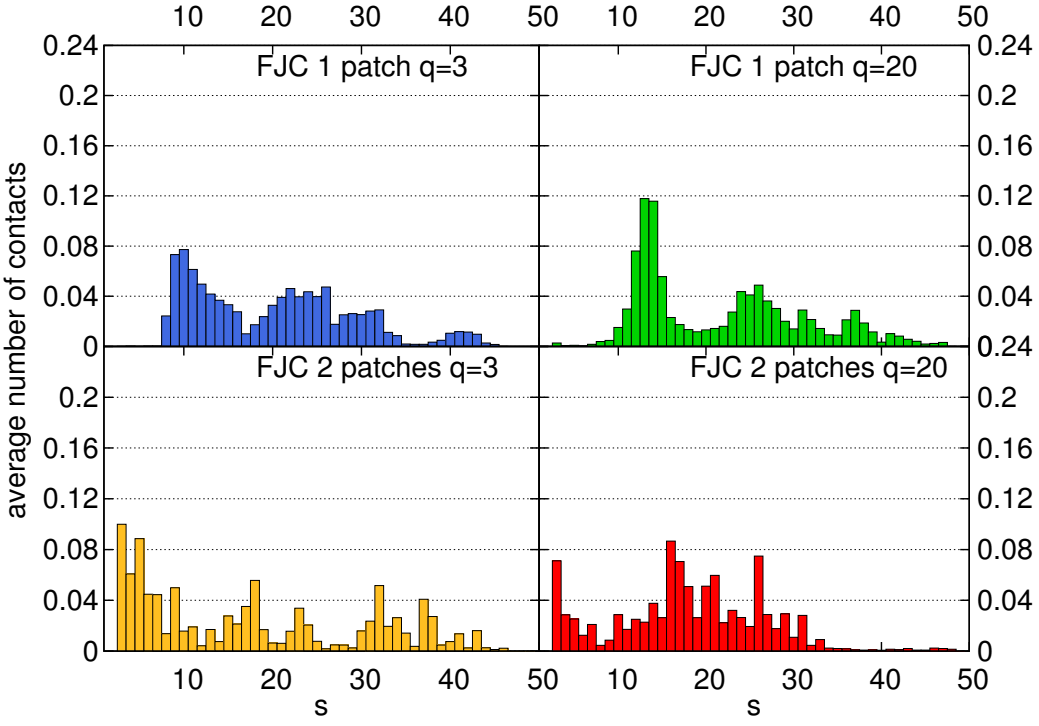

Figure 3.4: Normalised distribution of the patch contacts for FJC as a function of their chain separation s in monomers units.

Following the rationale of the influence of local contacts when comparing

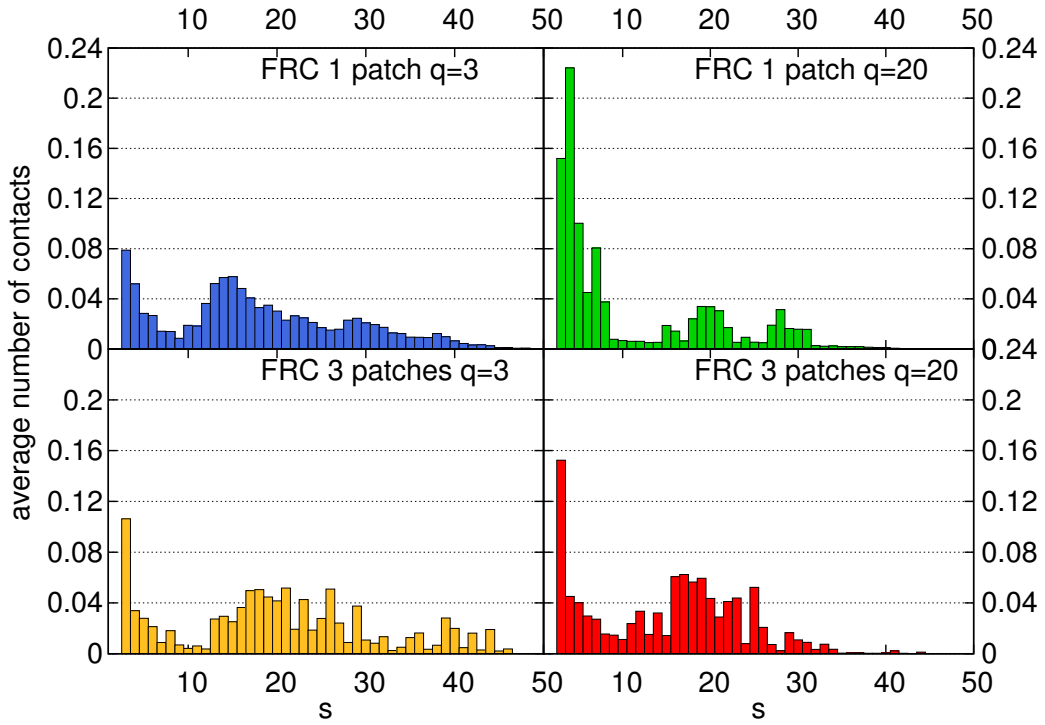

Figure 3.5: Normalised distribution of the patch contacts for FRC as a function of their chain separation s in monomers units.

the FJC with the FRC model, we further note that for the FJC model the probability of having directional contacts at scales below 10 beads (Fig. 3.4) is considerably lower than for the FRC model, and consequently the knot spectrum is skewed toward highly complex knots. Conversely, the higher probability for FRC of having local contacts (Fig. 3.5) translates into a higher density of peaks in the low knot-complexity region of the spectrum (Fig. 3.3).

Finally, the most remarkable effect made apparent by Fig. 3.3 is that incrementing the alphabet size to 20 letters, as in proteins, leads to a dramatic simplification of the knot spectra. Furthermore going from $q = 3$ to $q = 20$ has a slightly different effect in FJC and FRC. While in the first case it systematically enhances the 3_1 knots, in the latter it favours unknots, in one case (FRC with 1 patch) bringing the probability of unknots from 0.18 to 0.98.

The increase in the likelihood of unknots and trefoil knots is at first surprising. Indeed, based on the fact that increasing the alphabet size improves the designability of a model [6], one might expect that a larger alphabet would bring a richer knot spectrum. However, the number of possible conformations with a given topology decreases rapidly with knot complexity, due to the increase in the number of essential crossings [189]. Since a larger alphabet has a higher power to optimise different structures, it will be able to optimise both the structures with high topological complexity and the ones with low topological complexity, which are more abundant in number and therefore will dominate. In fact, we measured the knot spectrum for a plain version of the FJC and FRC models, in which the isotropic attractions are switched off ($U_{bb} = 0$ in Eq.3.2), but the patch-patch interaction are still present. The results show that for FRC $M = 1$ we observe an equal partitioning between the unknot (48%) and 3_1 (52%) topologies, while for FJC

$M = 1$ the 3_1 (91%) dominates the spectrum.

Our hypothesis that increasing the alphabet makes accessible more configurations compatible with simple knots, is further confirmed by looking at the sequence energy (second term of Eq. 3.1) of the sequence designed to fold into the knotted structures reported in Sec "Designing and folding specific knots". We find that for FRC $M = 3$ and $q = 20$, within the same model knots with different complexity (unknot, 3_1, 10_{124} and 8_{19}) have similar sequence energies, which differ between each other in the range of the thermal fluctuations from the free energy minimum. If on the other hand the shift in the peaks for $q = 20$ was of enthalpic origin we should have expected energy differences well above the thermal fluctuations. We thus propose that this trend is of entropic nature, with larger alphabets being able to optimise the larger conformational basins of simple knots better.

This result also implies that at small alphabet size, complex knots might dominate due to the difficulty to optimise the large number of existing unknotted or simply knotted conformations with a reduced set of letters. In fact, as reported in Fig. S.5 in the Supplemental Material, increasing the size of the alphabet to $q = 20$ significantly reduces the length and frequency of homopolymeric stretches found by the SEEK simulation compared to $q = 3$. While for $q = 3$ we observe homopolymeric stretches of up to 10 monomers, the probability of having repetitions with the size larger than 3 is negligible when $q = 20$. It is important to stress that the interaction matrices are Normally distributed (see Methods), and with the self-interactions distributed between attractive and repulsive. Hence, stretches of different letters correspond also to having alternatively self-attractive/self-repulsive stretches.

The presence of alternating long self-attractive/self-repulsive homopolymeric stretches forces the chain to form non-local contacts to reduce the

chain energy and thus makes it difficult to optimise conformations rich in crumpled regions, which correspond to unknots or simple knots like the 3_1. As a consequence, only a small subset of a large number of conformations with simple topologies will be optimised, and thus these topologies will not dominate the spectrum[1]. Again, this line of reasoning is confirmed by the distributions of patch contacts. In fact, as can be seen from Figs. 3.4 and 3.5, for the FJC model going from $q = 3$ to $q = 20$ strongly enhances the first peak when $M = 1$ and completely cuts off the tail of the distribution for $M = 2$. In the case of FRC, which allows local contacts already at $q = 3$, increasing the alphabet size leads both to a strong increase in the height of the peaks at small s and to the suppression of contacts at $s > 35$.

Another observation supporting our results comes from a previous study by Wüst et al. [190], in which the authors showed that the sequence of a two letters HP model could be used to design lattice heteropolymers to adopt either prevalently knotted or prevalently unknotted conformations. In particular, they found that sequences containing long contiguous sequences of monomers of the same type folded into highly knotted structures, since they forced the chain to form large loops. On the other hand, sequences which contained only short repeating stretches were almost always unknotted [190]. An interesting conclusion of the work of Wüst et al. was that the low number of knotted natural protein structures could be produced by an active evolutionary pressure over the sequence of the protein to avoid long repeti-

[1] Also for $q = 3$, the sequence energies for the designed knots – optimised for each knot topology for the structure with the highest probability of having that knot topology – are similar for different knot complexities. This does not imply that with $q = 3$ it will be possible to optimise the large number of existing unknotted or simply knotted conformations, as many of them will be not optimised due to the presence of the homopolymeric stretches.

tions that favours knots. From our results instead, we can draw a different conclusion: the increase in the alphabet size alone spontaneously drives the system to reduce the complexity of the knots or eliminate them altogether. Hence, the 20 amino acids alphabet of proteins has the additional property of reducing the spontaneous occurrence of knots without losing the possibility of optimizing complex topologies, a hypothesis that has not been proposed before.

We notice here that while several knotted protein folds have been reported, no knots has been found in RNAs uploaded to the PDB [191]. This might at first seem to be at odd with our claim, since RNA has a four letter alphabet. We note though that not only the reason for the absence of knots in natural RNAs might be evolutionary, several studies have pointed out that RNA tends to fold hierarchically [192, 193], forming very stable local branches before long range contacts can be established. As a result, RNA folds in the PDB appear not to be fully equilibrated [194, 195].

Finally, the difference between FJC, where enlarging the alphabet favours the trefoil knot, and FRC, where it favours trivial knots, originates from the bending rigidity of the polymer. In fact, recent studies on the knotting probability of homopolymers have shown that while completely flexible chains favours unknotted conformations, a slight increment of the persistence length strongly enhances the probability of obtaining simple knots like the trefoil [77, 187, 196]. To confirm that this is the case also in our setup, we computed the persistence length of pure self-avoiding freely rotating chains and freely jointed chains, in the absence of any attractions between the beads. The results, reported in the Supplemental Material Fig. S.6 show that while the FRC model is completely flexible, the FJC has a small persistence length of about 3 monomers. This also further explains the negligible probability of

having patch-patch contacts at low values of s for FJC $M = 1$.

Overall, patchy polymers combine the advantages of directional interactions to shape the conformational space towards regions rich in complex knots, with the control offered by heteropolymers on the specific target structure.

3.2.2 Designing and folding specific knots

Having analysed the most abundant knot types in each model and the effect of the number of patches and alphabet sizes on their probability, we now move on to demonstrate that different topologies can be designed and folded reliably both in FJC and FRC. Furthermore, we demonstrate that this applies even to rare knots such as the 5_2 knot in the FJC model. To do so, we select a set of representative topologies from our various models. These are 3_1, 5_1 and 5_2 for FJC M=1; 5_1 for FJC M=2; 3_1, 8_{19}, 10_{124} for FRC M=3. Whenever possible, we design conformations in which the ends are close to each other, so that they can be externally connected in a second step to trap the knot.

Our design and folding scheme works as follows. For a given topology, we identify the corresponding global free energy minimum in the ACN,R_{ee} projection. This is given by the structure with the highest number of sequences that fold into it, making it the most designable one [153]. This solution is not necessarily unique, as other structures further from the global minimum might still provide designable candidates. Among the conformations in the vicinity of the free energy minimum in the ACN, R_{ee} plane, we identify the one with the lowest free energy that also has values of P and Q in the vicinity of the free energy minimum of the corresponding P, Q SEEK spectrum.

Following these criteria, we choose the best conformation for each knot type and use it as input for a DESIGN [62] simulation (see Methods), in

which we explore the space of sequences while maintaining the structure frozen.

DESIGN simulations provide us with an optimal chain sequence for a given target structure. To test the validity of the designed sequence, we verify with the FOLDING simulation that the target conformation is the global energy minimum in the conformational space of the polymer with the fixed designed sequence. If the global free energy minimum is close to the target structure the polymer is correctly folded, following the criterion in [6] (see Fig. S.7 in Supplemental Material).

Following this scheme, we can design and fold all target topologies, with a remarkable precision in the range of fractions of the monomer size, comparable to the folding precision that proteins can reach. In Fig. 3.6 we show the folded structure in comparison with the target structure, together with the root mean square displacement from the target structure ($RMSD$) – often used to categorised protein folds – defined as:

$$RMSD = \sqrt{\frac{1}{N}\sum_i \delta_i^2} \tag{3.11}$$

Here, δ_i is the distance between the bead i of the folded structure and the corresponding one of the target structure.

We note in particular that we are able to fold knots up to 5 crossings in conformations with close ends which could be easily locked externally. A similar result should be achievable for more complex knots like the 8_{19} and 10_{124} by simply using longer chains.

Finally, our procedure demonstrates that the FJC model can be used to design and fold highly localised knots, as shown by the 3_1 knot in Fig. 3.6, which takes up only 14 beads, less than one third the length of the chain.

3.3 Conclusions

In this work, we studied the topological properties of two different models of hetero patchy polymers, namely Freely Rotating Chains and Freely Jointed Chains. Both models offer several viable experimental realizations and might prove highly valuables to produce knotted supra-molecular constructs. In this perspective, we characterised the knot spectrum of both chain models, for different patch numbers and two different alphabet sizes of the heterogeneous interactions.

We show that increasing the number of patches suppresses knots with high complexity, by enhancing the probability of local contacts on the chain. Interestingly, increasing the alphabet simplifies the knot spectrum, driving it to populate mainly the trefoil and the unknot.

The latter finding has intriguing implications in the field of knotted proteins where the low percentage of knotted structures recorded so far suggested that it might be the result of explicit selection pressure [93, 100, 103, 171, 198–200] . Here, instead, we propose that simply the alphabet of 20 letters is a strong deterrent for the formation of knots.

Finally, we take several knots, both toroidal and twist, some of which have never been experimentally tied, like the 5_2 and 10_{124} and we design them by optimizing the sequence of heterogeneous interactions. As we demonstrate in the results section, the designed sequences were able to fold back into the target knotted structure, suggesting that our approach might indeed open the pathway for the experimental realization of complex supra-molecular knots. We are able to select and fold also structures with close ends, which could be easily locked externally to guarantee the thermal stability of the structure [61]. Moreover, the structures fold back with a precision of fractions of the bead size, comparable to what proteins can achieve and necessary to

open possible applications as catalysers or drug delivery carriers.

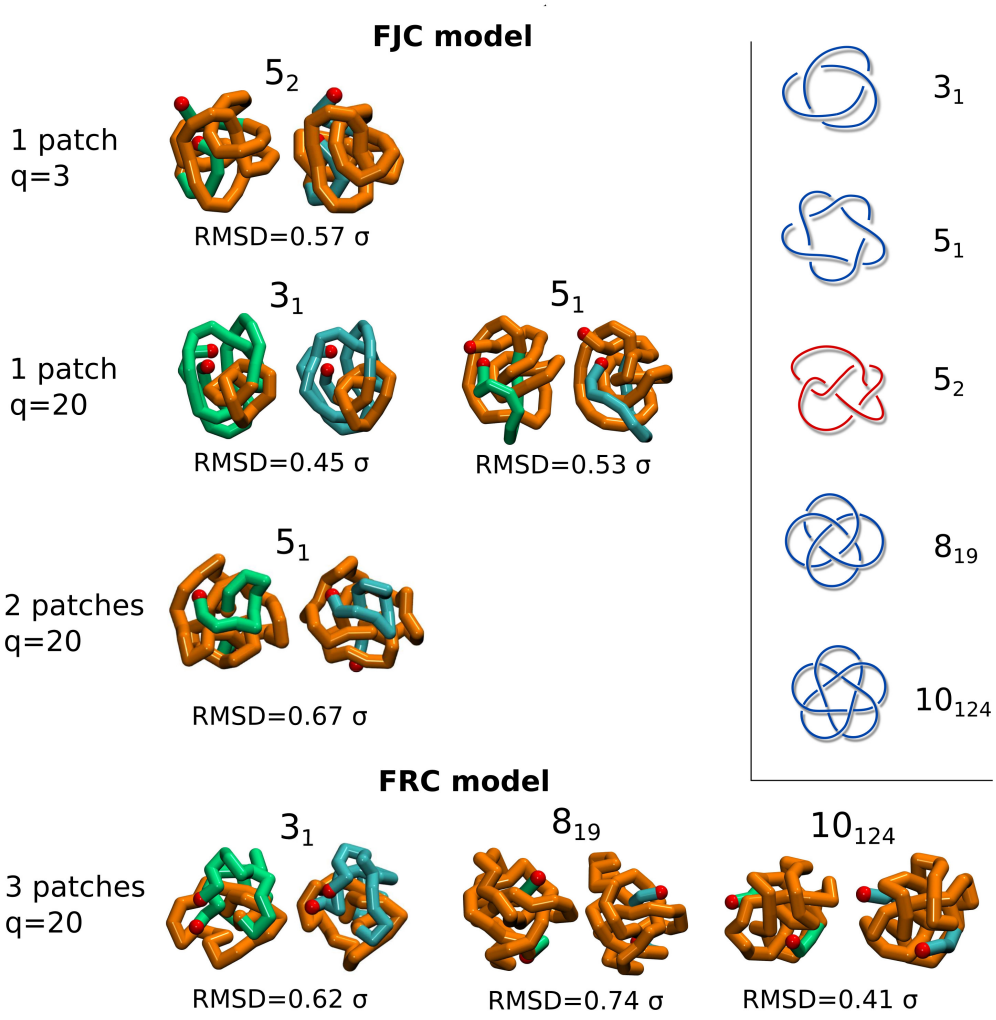

Figure 3.6: Designed knotted conformations. Each pair of structures represents the target (left) and folded (right) conformations. The part of the polymer involved in the knot is represented in orange, the end monomers in red, and the rest in green (for the target) and cyan (for the folded). For each pair, the value of RMSD (root mean square displacement between the folded and the target) is reported. The knotted portions have been identified using Kymoknot [197]. The diagrammatic representation of all designed knots are reported on the right. The 5_2 (colored in red) is a twist knot, while the other are torus knots.

Appendices

Appendix A

The role of directional interactions in the designability of generalized heteropolymers - supplementary material

A.1 Supplementary theory

A guiding theory for heteropolymer designability is given by the Random Energy Model (REM) [5, 54, 68]. A clear review can be found in the seminal works of Pande *et al.* [5], where it is shown how the designability of a heteropolymer increases with the total number of possible bonds for each bead (valence) and decreases with the conformational entropy per bead. Hence, it is reasonable to assume that directionality (the patches) combined with isotropic interactions would increase designability, because the valence (i.e. the total number of possible bonds) remains constant, while the conforma-

tional entropy per bead decreases. In fact, the introduction of the patches decreases the entropy by favouring the system to populate more specific structures with the patches along particular directions.

On the other hand, if the number of patches increases to much, the interactions become again close to isotropic and the designability decreases again. Hence, a model is needed that explicitly brings about the designability from a basic heteropolymer model by controlling the alphabet size and the conformational entropy per particle.

A.2 Symmetry of patches on the surface

Considering n as the number of patches on the surface, for freely jointed chains (FJC) with $n = 3$ the anchoring points are places perpendicular to the patch, while for $n = 4$ they are placed on a tetrahedron completed by the anchoring points. The freely rotating chains (FRC) are fixed on the vertices of a platonic solid or in the most symmetric way: equispaced on the equator and placed on the vertices of a equilateral triangle, of a tetrahedron and of a octahedron for the $n = 3$ $n = 4$ $n = 6$ cases, respectively. For $n = 10$ there is no platonic solid, so the patches are placed on the surface in the most symmetric way by using the following numerical procedure:

1. n patches are randomly placed on a sphere, their positions given by the set of vectors $\{\vec{r_1}, \ldots, \vec{r_n}\}$

2. we assign a fictitious energy to the system, defined as $U = \frac{1}{2} \sum_{i \neq j} |\vec{r_i} - \vec{r_j}|^{-1}$

3. we minimise U by attempting to move a randomly chosen patch, accepting the move if the total energy of the system consequently decreases.

Formally, this can be regarded as a Monte Carlo (MC) simulation performed at temperature zero.

4. We iterate this procedure until convergence of U.

For completeness, we note that the above method produces a patch distribution which is independent of the definition of distance between two patches, being it the Euclidean distance or a spherical distance, for all the values of n considered here. In addition, the patch distribution makes sure that two particles cannot be involved in more than one bond.

A.3 Supplementary results figures

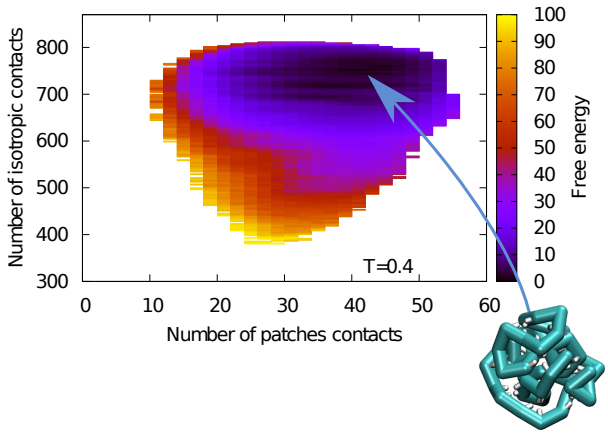

Figure A.1: Free energy landscape sampled by SEEK for one patch and alphabet 3 in the freely rotating chain model. The free energy is in function of the total number of contacts between the spheres (distance below 6 R_{bead}) and the total number of contacts between the patches (distance below 1.25 R_{bead} and angles θ_1 and $\theta_2 > 0.8\ \pi$). The target structure is chosen in the global minimum of this landscape. Following the above definition of patches contacts, in the target structure the 80% of the patches are maximally oriented between each others. However, we observe by looking closer at the structure, that all the patches are interacting. Nevertheless, the close packing peak in the radial distribution function dominates on the directional interaction peak. Hence, the close packing is not suppressed even when all the directional interactions are fulfilled. Thus, increasing the relative strength of the directional interactions will not make the first peak disappear.

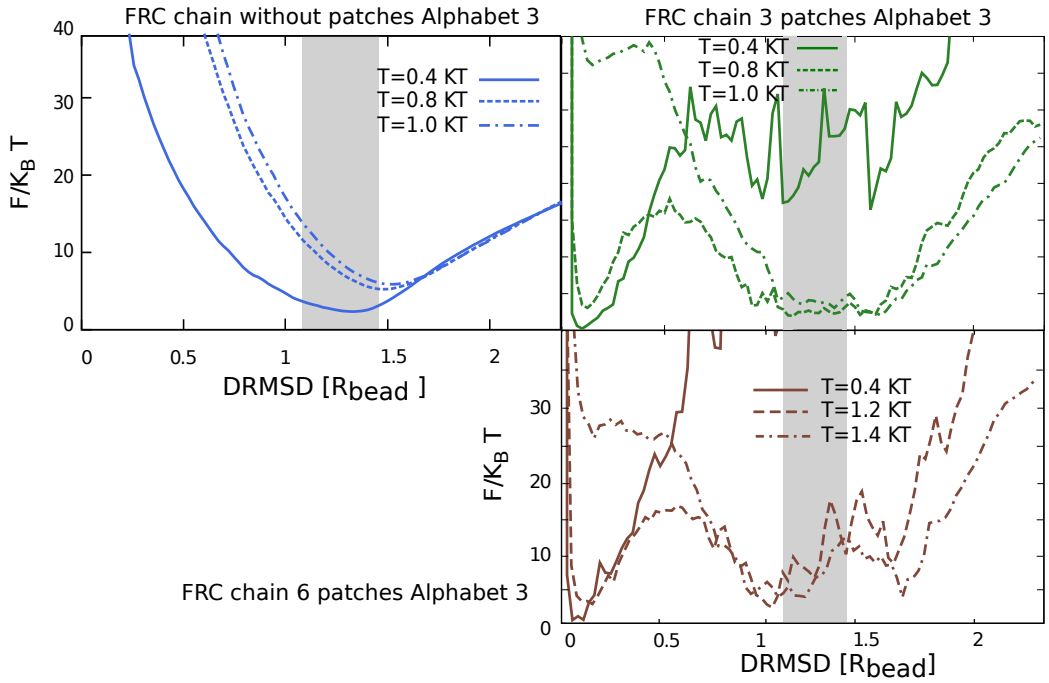

Figure A.2: FOLDING free energy landscape as a function of the distance root mean square displacement ($DRMSD$) for an example of a non-designable system (left) and two designable ones (right). In the cases on the right, our temperature resolution was high enough to observe two simultaneous minima. The configurations with $DRMSD$ values corresponding to the position of the second minimum are the molten globule structures. Interestingly, for the same alphabet size the position of the molten globule minimum is conserved for different number of patches, and for 0 patches is the only global minimum observed at all temperatures. Hence, the 0 patch chain is never designable.

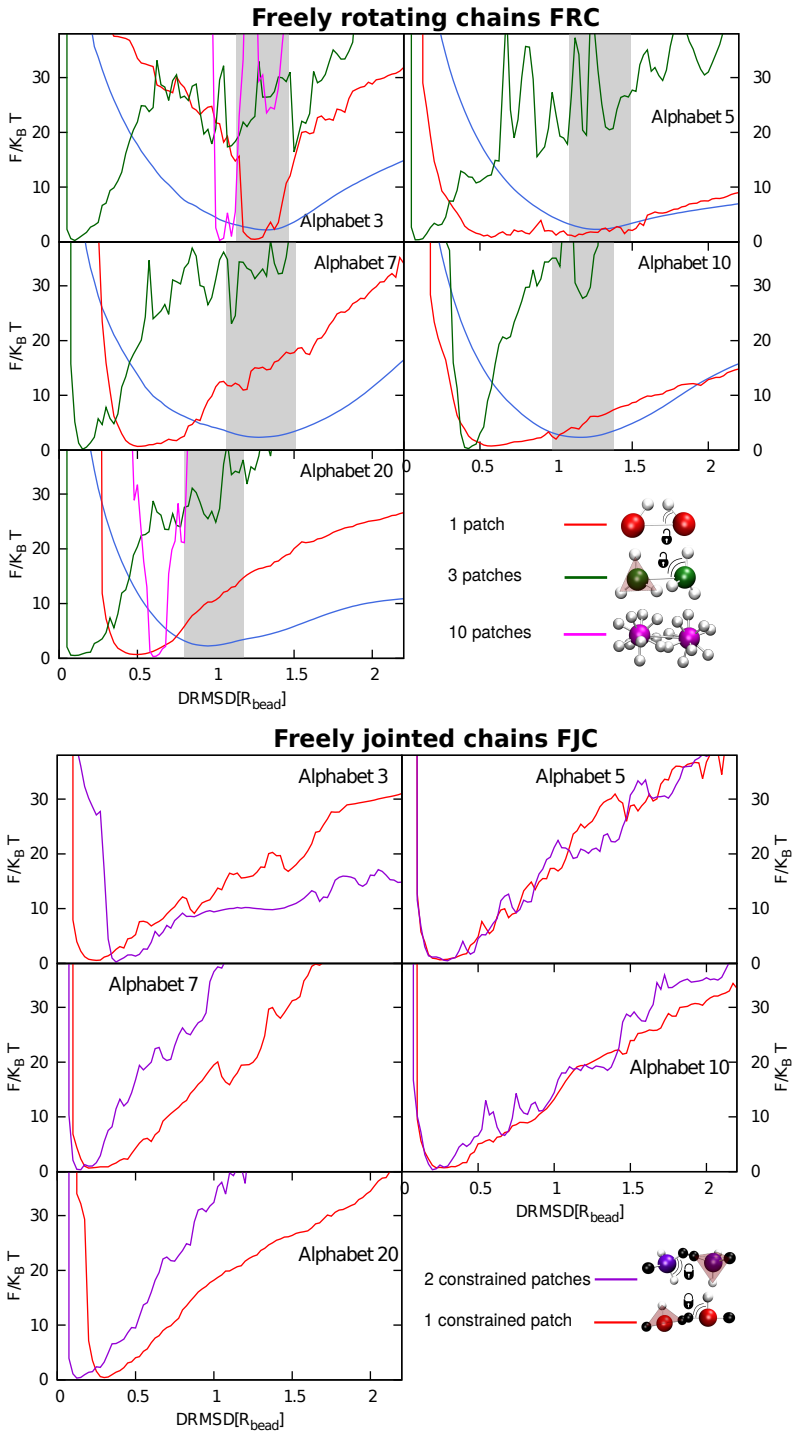

Figure A.3: FOLDING free energy landscapes. The free energy is plotted as a function of the distance root mean square displacement ($DRMSD$) for freely rotating chain (left) and freely jointed chain (right), for different patches numbers and alphabet sizes at temperature 0.4.

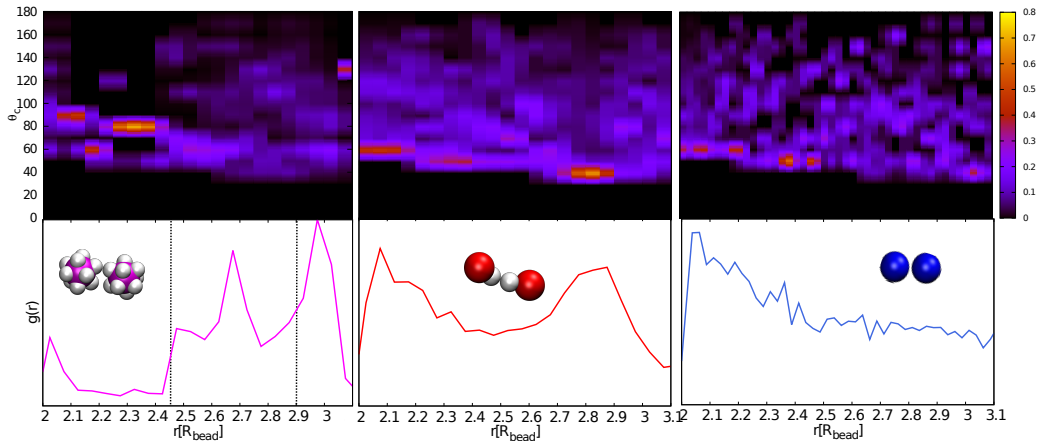

Figure A.4: Probability that one bead forms an angle θ_c with two other beads located at distance r from the central one. The probability is averaged on the 40 most probable conformations and we exclude the angles formed by three consecutive beads along the chain. a) without patches. b) 10 patches.

For the polymer without patches we observe in Figure S.4 that the prevalent angle is 60° corresponding to a random packing. For 10 patches we observe that the high number of patches decreases the random packing at distances close to $2R_{bead}$. In fact, at small distances $(2.00 - 2.45R_{bead})$ there is a component of directionality that cannot be attributed to the patch-patch bond, but it is instead due to the inter-penetration of the corona formed by the 10 patches, which can occur only with discrete angles. While at intermediate distances $(2.45 - 2.90R_{bead})$ be observe that the peak at 60° reappears. In fact, this correspond to the new random packing, which is shifted due to the hindrance of the corona formed by the many patches. In the last part we observe only at distances $3R_{bead}$ the reappearing of some directionality due to the patch-patch bond.

A.4 Isotropic heterogeneous interactions matrices

Interaction matrix for alphabet size 3

	1	2	3
1	0.405427	0.399805	-0.408428
2		0.083519	-0.185762
3			-0.294699

Interaction matrix for alphabet sizes 5

	1	2	3	4	5
1	0.073272	-0.056012	-0.176851	0.055725	0.072086
2		0.115497	-0.358047	-0.606938	-0.02605
3			0.327363	0.441482	-0.221434
4				-0.003289	-0.046151
5					0.417604

Interaction matrix for alphabet sizes 7

	1	2	3	4	5	6	7
1	-0.488145	-0.099172	-0.665188	-0.513159	-0.140222	0.027036	-0.023432
2		-0.411612	0.731346	0.494311	0.067639	0.39224	0.06542
3			0.339552	-0.428898	-0.463306	-0.061358	0.325872
4				0.356201	-0.377746	0.60377	-0.715795
5					-0.035632	0.013969	0.282897
6						0.581666	0.402024
7							-0.279583

Interaction matrix for alphabet size 10

	1	2	3	4	5	6
1	0.036338	-0.109579	0.450045	0.211862	-0.144162	-0.156263
2		-0.014035	0.001334	0.219005	0.672716	0.174732
3			-0.422396	0.2583	-0.064613	0.015065
4				0.341895	-0.130257	-0.503189
5					0.320228	-0.077332
6						-0.151698
7						
8						
9						
10						

Interaction matrix for alphabet size 10

7	8	9	10
-0.06087	0.335178	0.383069	0.398842
0.117673	-0.097265	-0.038571	0.267376
0.767522	-0.068141	-0.099905	-0.223041
0.175065	-0.009685	-0.07783	0.319185
0.072921	-1.035992	-0.465169	-0.18564
-0.016638	-0.029539	-0.19378	-0.247656
-0.049106	0.12019	0.448171	-0.742203
	0.023326	-0.221476	-0.065685
		-0.173327	-0.215073
			-0.022223

Interaction matrix for alphabet size 20

	1	2	3	4	5	6
1	0.081553	0.426631	0.568848	-0.502789	-0.12822	-0.435738
2		0.353419	-0.097927	0.389741	0.243083	-0.336347
3			-0.056157	-0.179468	0.111268	-0.135234
4				0.701963	-0.041826	0.388943
5					-0.073145	-0.186174
6						0.553226
7						

Interaction matrix for alphabet size 20

7	8	9	10	11	12	13
-0.348879	0.128948	0.045576	-0.28945	0.048781	-0.354339	0.273338
0.166491	-0.072994	0.017063	-0.324963	0.169753	0.027633	-0.870442
-0.029389	-0.079183	-0.235617	0.197951	-0.698542	0.131127	0.061888
0.534033	0.275594	-0.097178	0.125153	0.246242	-0.430059	-0.530827
0.016316	0.411457	-0.402362	0.49164	-0.031645	0.034439	-0.129541
0.485911	0.142607	0.089138	-0.26214	-0.30248	0.237108	-0.528338
-0.078432	0.291858	0.03107	-0.038642	0.480908	-0.198523	0.357243
	0.392551	-0.059625	0.200538	-0.468885	-0.05802	-0.284558
		-0.117411	0.081203	-0.173082	-0.186838	-0.626013
			0.084789	0.045141	-0.120325	-0.037095
				-0.161365	-0.352376	0.066567
					-0.000727	0.179239
						0.355301

Interaction matrix for alphabet size 20

14	15	16	17	18	19	20
0.900379	-0.305614	0.11782	0.141584	0.194476	-0.021256	-0.111908
0.125427	0.342717	0.099973	-0.244877	-0.357507	0.319781	-0.236908
-0.619739	0.36774	-0.161619	-0.18599	-0.678679	-0.142067	-0.003966
-0.093322	-0.433012	0.29192	0.010121	-0.627608	0.323705	0.174866
0.026183	-0.129373	-0.034842	-0.494044	0.381081	0.323671	0.046249
0.500438	-0.027833	0.357806	-0.29547	-0.524512	-0.240258	-0.504283
0.1966	-0.153644	0.765287	0.442432	-0.286014	0.227902	0.466241
-0.096826	0.311937	0.000396	0.147358	0.395159	0.372012	0.362592
0.220451	0.073642	0.060668	0.383668	-0.593914	-0.303292	0.074573
0.01467	0.217459	-0.630983	-0.432735	-0.09005	-0.205353	0.056339
0.287437	-0.296898	-0.029977	-0.803003	0.255198	-0.546963	-0.122039
-0.169458	0.109036	0.237348	0.381594	0.008841	-0.102676	-0.091471
-0.186524	0.253168	-0.127866	0.090694	0.031939	-0.376384	-0.007257
-0.152785	-0.266961	-0.094049	-0.276912	-0.452899	0.025443	-0.4211
	-0.275975	0.667552	0.354217	-0.136487	0.086701	0.174718
		0.003189	0.086487	0.135358	-0.01824	0.346642
			-0.223676	-0.102268	0.362454	0.215395
				-0.067085	0.173723	0.679878
					0.33719	-0.109486
						0.084862

Appendix B

General methodology to identify the minimum alphabet size for heteropolymer design - supplementary material

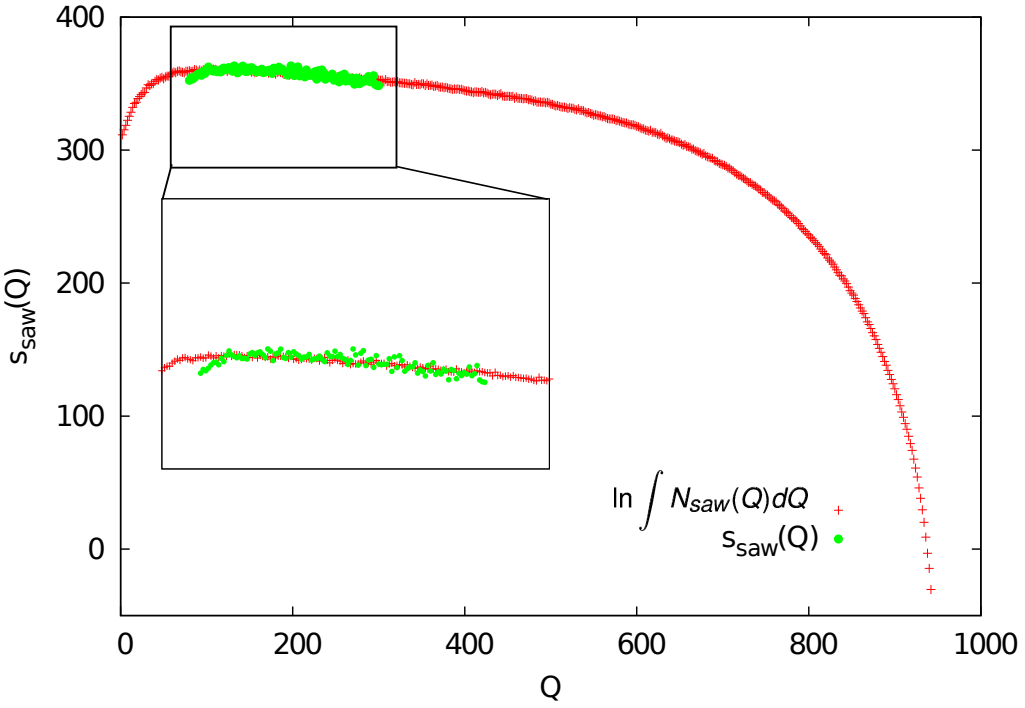

Figure B.1: Comparison of the conformational entropy of the self-avoiding polymer of length N=50 between $\ln \int \mathcal{N}_{saw}(P,Q)dP$, calculated via the VMPT method [154], shifted on $s_{saw}(Q)$, calculated via the gran canonical method [169, 170].

Appendix C

Heteropolymer design and folding of knots - supplemental material

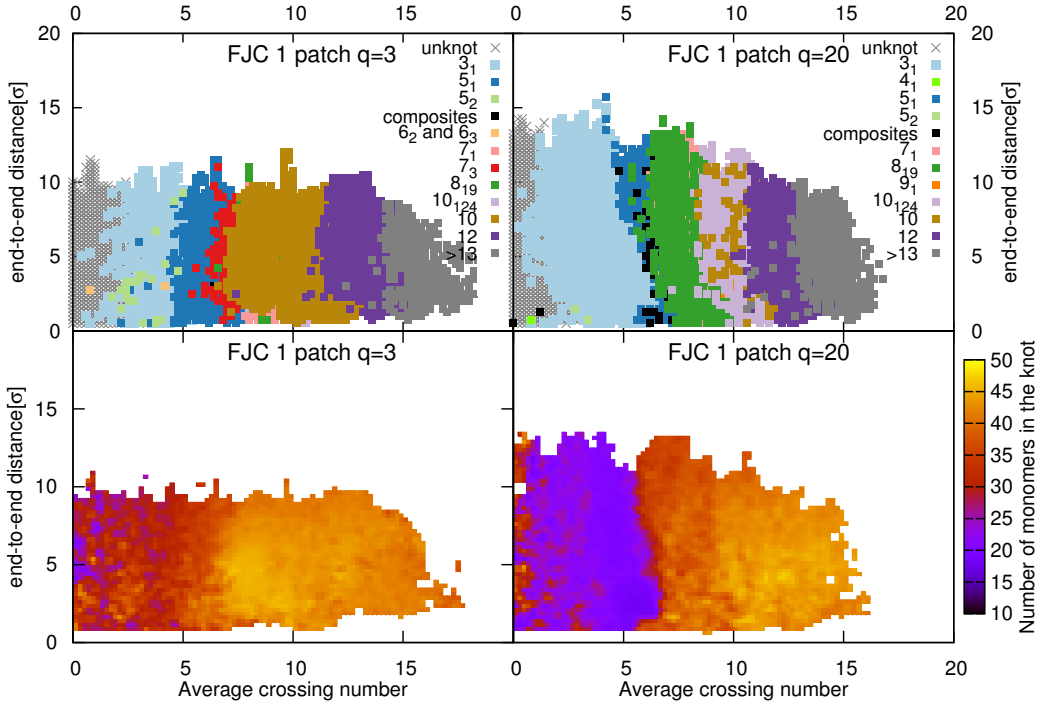

Figure C.1: FJC model with 1 patch and with $q = 3$ and $q = 20$ Top: Topological 'phase diagram' of the most probable knot type for each pair of ACN, R_{ee}. Bottom: spectra of the average number of monomers involved in the knot for each pair ACN, R_{ee}.

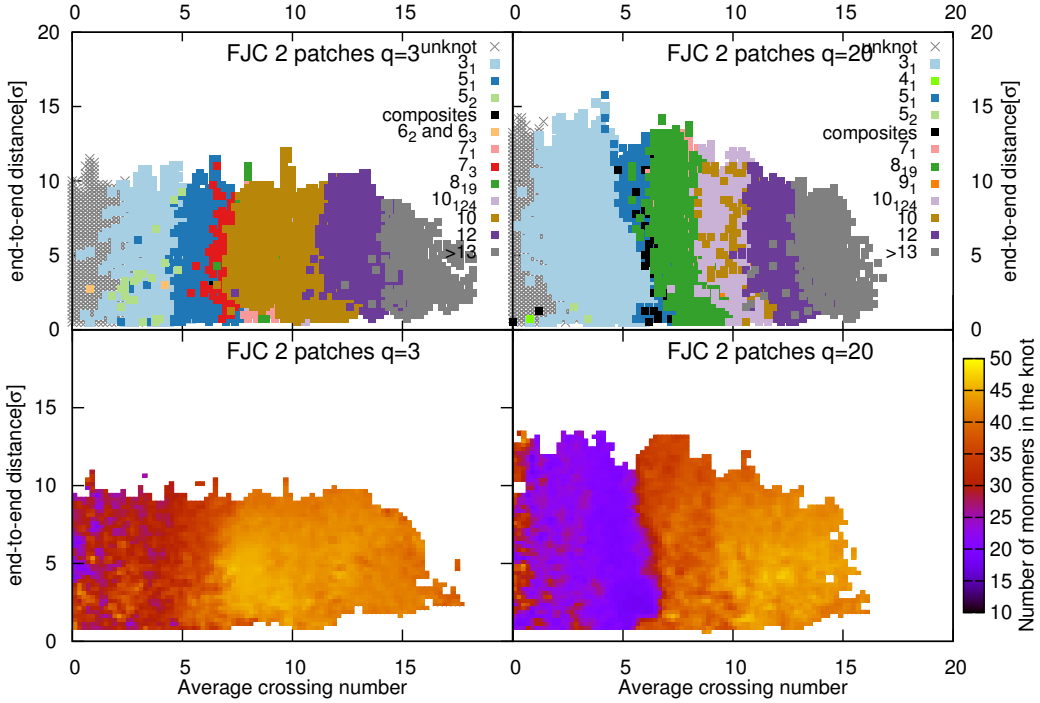

Figure C.2: FJC model with 2 patches and with $q = 3$ and $q = 20$ Top: Topological 'phase diagram' of the most probable knot type for each pair of ACN, R_{ee}. Bottom: spectra of the average number of monomers involved in the knot for each pair ACN, R_{ee}.

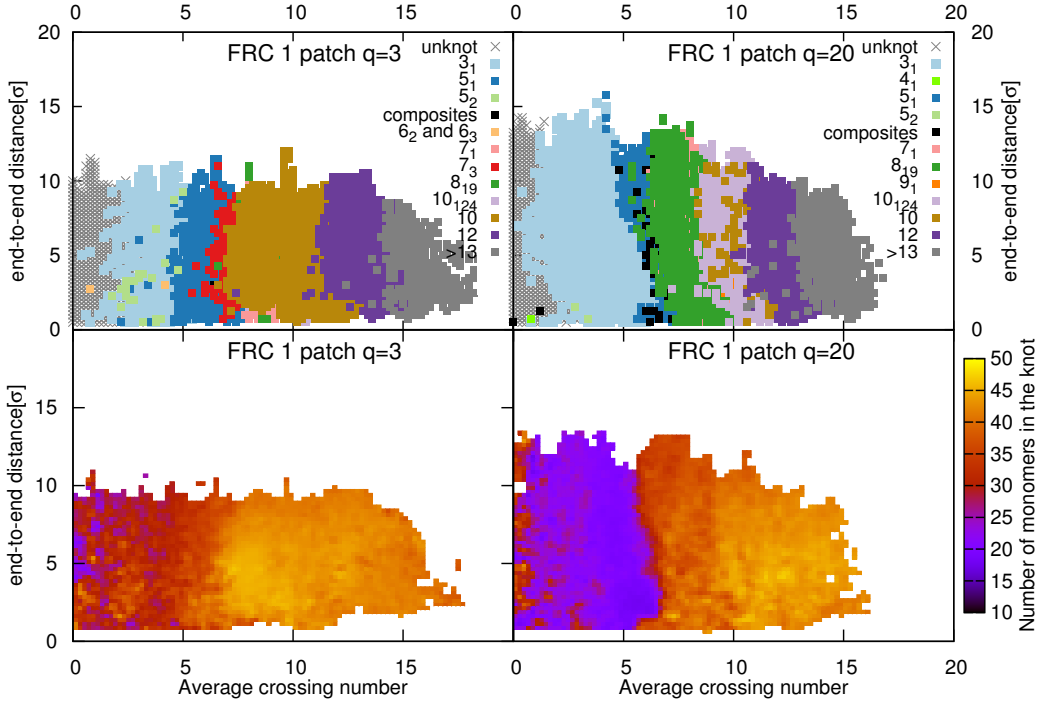

Figure C.3: FRC model with 1 patch and with $q = 3$ and $q = 20$ Top: Topological 'phase diagram' of the most probable knot type for each pair of ACN, R_{ee}. Bottom: spectra of the average number of monomers involved in the knot for each pair ACN, R_{ee}.

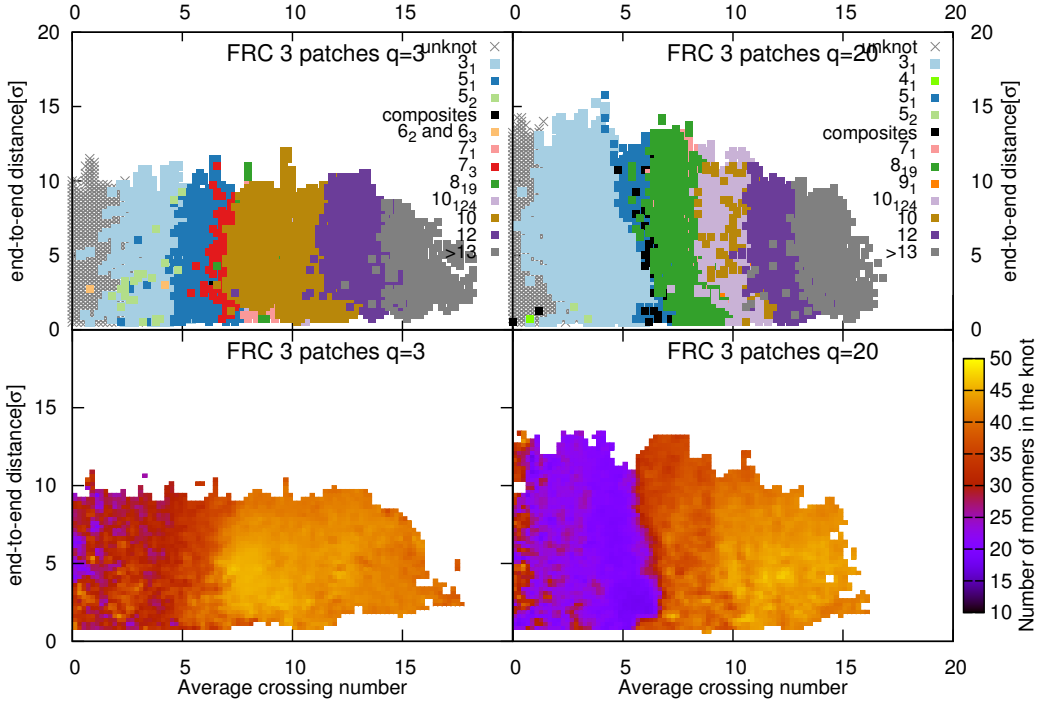

Figure C.4: FRC model with 3 patches and with $q = 3$ and $q = 20$ Top: Topological 'phase diagram' of the most probable knot type for each pair of ACN, R_{ee}. Bottom: spectra of the average number of monomers involved in the knot for each pair ACN, R_{ee}.

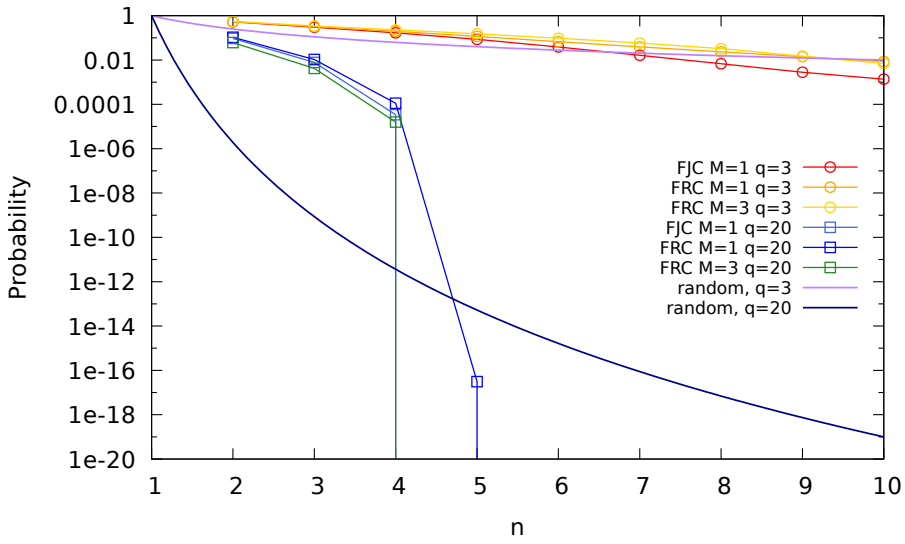

Figure C.5: Probability of having words of the same letters as a function of the word length n for the considered models. For each sequence produced by SEEK, we evaluate the frequency of homogeneous words by counting the number of repetitions of length n and dividing it by the number of possible words with the same length, on a sequence of length $N = 50$. The total probabilities over the whole $ACN \times R_{ee}$ spectrum have then been computed as described in the main text for a variable a. The theoretical curves refer to random distributions of q letters and are given by $\frac{q}{q^n}$.

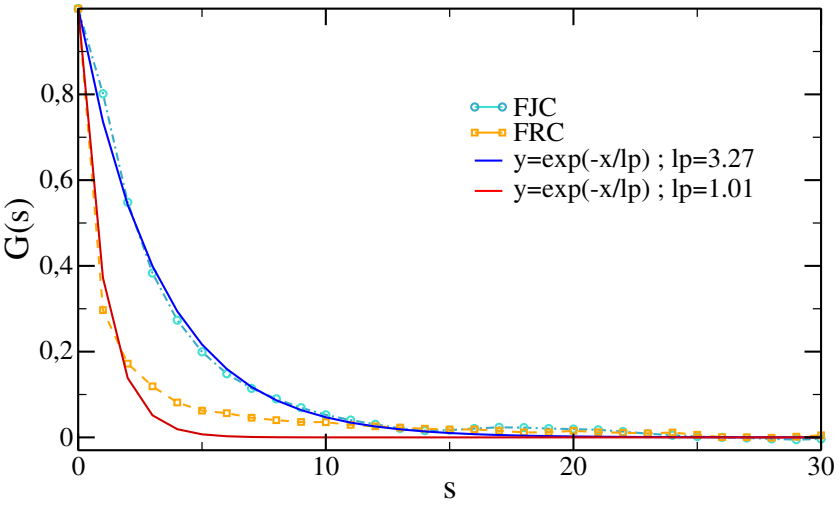

Figure C.6: Bond-bond correlation function $G(s)$ for pure self-avoiding versions of the FJC and FRC models, in which all interactions aside for backbone bonds and steric repulsion were turned off. s correspond to the distance in bond units. Continuous lines correspond to exponential fits $\exp(-s/l_p)$, from which we obtain the persistence length l_p.

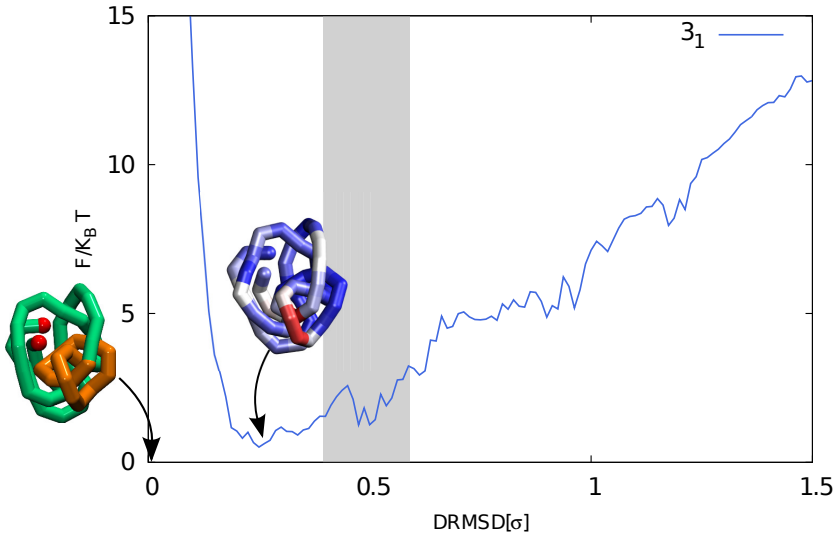

Figure C.7: Example of FOLDING free energy landscape projected onto the root mean square displacement of the inter-particle distance ($DRMSD$, see Methods for definition) between the target structure and each sampled structure, for FJC 1 patch q=20. $DRMSD = 0$ corresponds uniquely to the target structure, we define as molten globule (misfolded) region (grey area) the DRMSD range spanning 1 k_BT from the position of the global minimum of the FOLDING free energy of the corresponding bare polymer without patches, as shown in Ref. [6]. All global minima with DRMSD values below the lower boundary of the grey area are considered correctly folded. The colors of the folded structure represent the distance of each bead from the target structure, in a scale from 0 (blue) to 1 σ (red).

C.1 Isotropic heterogeneous interactions matrices

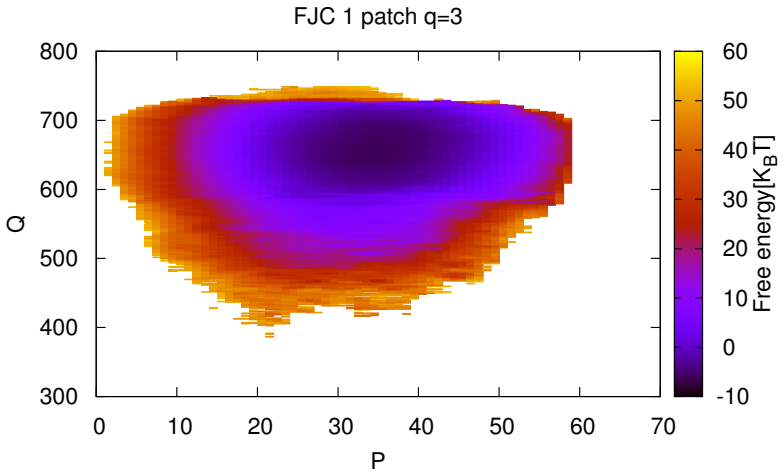

Figure C.8: SEEK free energy landscape projected on the number P of patch-patch contacts and the number Q of isotropic contacts, for FJC with 1 patch $q = 3$.

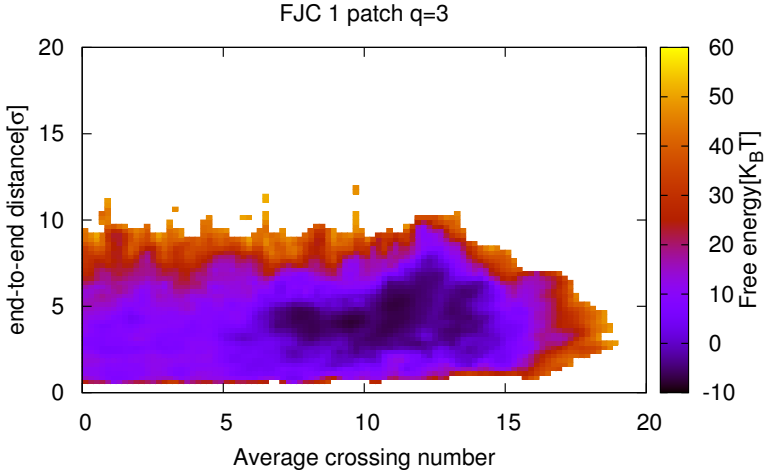

Figure C.9: SEEK free energy landscape projected on the average crossing number ACN and the end-to-end distance, for FJC with 1 patch q=3.

Interaction matrix for alphabet size 3

	1	2	3
1	0.405427	0.399805	-0.408428
2		0.083519	-0.185762
3			-0.294699

Interaction matrix for alphabet size 20

	1	2	3	4	5	6
1	0.081553	0.426631	0.568848	-0.502789	-0.12822	-0.435738
2		0.353419	-0.097927	0.389741	0.243083	-0.336347
3			-0.056157	-0.179468	0.111268	-0.135234
4				0.701963	-0.041826	0.388943
5					-0.073145	-0.186174
6						0.553226

Interaction matrix for alphabet size 20

7	8	9	10	11	12	13
-0.348879	0.128948	0.045576	-0.28945	0.048781	-0.354339	0.273338
0.166491	-0.072994	0.017063	-0.324963	0.169753	0.027633	-0.870442
-0.029389	-0.079183	-0.235617	0.197951	-0.698542	0.131127	0.061888
0.534033	0.275594	-0.097178	0.125153	0.246242	-0.430059	-0.530827
0.016316	0.411457	-0.402362	0.49164	-0.031645	0.034439	-0.129541
0.485911	0.142607	0.089138	-0.26214	-0.30248	0.237108	-0.528338
-0.078432	0.291858	0.03107	-0.038642	0.480908	-0.198523	0.357243
	0.392551	-0.059625	0.200538	-0.468885	-0.05802	-0.284558
		-0.117411	0.081203	-0.173082	-0.186838	-0.626013
			0.084789	0.045141	-0.120325	-0.037095
				-0.161365	-0.352376	0.066567
					-0.000727	0.179239
						0.355301

Interaction matrix for alphabet size 20

14	15	16	17	18	19	20
0.900379	-0.305614	0.11782	0.141584	0.194476	-0.021256	-0.111908
0.125427	0.342717	0.099973	-0.244877	-0.357507	0.319781	-0.236908
-0.619739	0.36774	-0.161619	-0.18599	-0.678679	-0.142067	-0.003966
-0.093322	-0.433012	0.29192	0.010121	-0.627608	0.323705	0.174866
0.026183	-0.129373	-0.034842	-0.494044	0.381081	0.323671	0.046249
0.500438	-0.027833	0.357806	-0.29547	-0.524512	-0.240258	-0.504283
0.1966	-0.153644	0.765287	0.442432	-0.286014	0.227902	0.466241
-0.096826	0.311937	0.000396	0.147358	0.395159	0.372012	0.362592
0.220451	0.073642	0.060668	0.383668	-0.593914	-0.303292	0.074573
0.01467	0.217459	-0.630983	-0.432735	-0.09005	-0.205353	0.056339
0.287437	-0.296898	-0.029977	-0.803003	0.255198	-0.546963	-0.122039
-0.169458	0.109036	0.237348	0.381594	0.008841	-0.102676	-0.091471
-0.186524	0.253168	-0.127866	0.090694	0.031939	-0.376384	-0.007257
-0.152785	-0.266961	-0.094049	-0.276912	-0.452899	0.025443	-0.4211
	-0.275975	0.667552	0.354217	-0.136487	0.086701	0.174718
		0.003189	0.086487	0.135358	-0.01824	0.346642
			-0.223676	-0.102268	0.362454	0.215395
				-0.067085	0.173723	0.679878
					0.33719	-0.109486
						0.084862

Bibliography

[1] Andersen, E. S. *et al.* Self-assembly of a nanoscale dna box with a controllable lid. *Nature* **459**, 73 (2009).

[2] Khalatur, P. G., Khokhlov, A. R. & Krotova, M. K. Evolutionary Approach in Copolymer Sequence Design. *Macromolecular Symposia* **252**, 36–46 (2007).

[3] Moreno, A. J. *et al.* Advantages of orthogonal folding of single polymer chains to soft nanoparticles. *Macromolecules* **46**, 9748–9759 (2013).

[4] Lodge, T. P. Block copolymers: Past successes and future challenges. *Macromolecular Chemistry and Physics* **204**, 265–273 (2003).

[5] Pande, V. S., Grosberg, A. Y. & Tanaka, T. Heteropolymer freezing and design: Towards physical models of protein folding. *Reviews of Modern Physics* **72**, 259–314 (2000).

[6] Cardelli, C. *et al.* The role of directional interactions in the designability of generalized heteropolymers. *Scientific Reports* **7**, 4986 (2017).

[7] Whitesides, G. M. & Grzybowski, B. Self-assembly at all scales. *Science* **295**, 2418–2421 (2002).

[8] Nie, Z., Petukhova, A. & Kumacheva, E. Properties and emerging applications of self-assembled structures made from inorganic nanoparticles. *Nature nanotechnology* **5**, 15 (2010).

[9] Glotzer, S. C. Nanoscale science: complex rules for soft systems. *Nature materials* **2**, 713–4 (2003).

[10] Di Michele, L. & Eiser, E. Developments in understanding and controlling self assembly of dna-functionalized colloids. *Physical Chemistry Chemical Physics* **15**, 3115–3129 (2013).

[11] Sacanna, S., Pine, D. J. & Yi, G.-r. Engineering shape: the novel geometries of colloidal self-assembly. *Soft Matter* (2013).

[12] Bianchi, E., Largo, J., Tartaglia, P., Zaccarelli, E. & Sciortino, F. Phase Diagram of Patchy Colloids: Towards Empty Liquids. *Physical Review Letters* **97**, 4 (2006).

[13] Zaccarelli, E. Colloidal gels: equilibrium and non-equilibrium routes. *Journal of Physics: Condensed Matter* **19**, 323101 (2007).

[14] Bianchi, E., Blaak, R. & Likos, C. N. Patchy colloids: state of the art and perspectives. *Physical Chemistry Chemical Physics : PCCP* **13**, 6397–410 (2011).

[15] Yi, G.-R., Pine, D. J. & Sacanna, S. Recent progress on patchy colloids and their self-assembly. *J. Phys.: Condens. Matter* **25**, 193101 (2013).

[16] Ruzicka, B. *et al.* Observation of empty liquids and equilibrium gels in a colloidal clay. *Nature materials* **10**, 56 (2011).

[17] Romano, F., Sanz, E. & Sciortino, F. Phase diagram of a tetrahedral patchy particle model for different interaction ranges. *The Journal of Chemical Physics* **132**, 184501 (2010).

[18] Giacometti, A., Lado, F., Largo, J., Pastore, G. & Sciortino, F. Effects of patch size and number within a simple model of patchy colloids. *The Journal of Chemical Physics* **132**, 174110 (2010).

[19] Foffi, G. & Sciortino, F. On the possibility of extending the Noro-Frenkel generalized law of correspondent states to nonisotropic patchy interactions. *The Journal of Physical Chemistry. B* **111**, 9702–5 (2007).

[20] Romano, F., Sanz, E. & Sciortino, F. Role of the range in the fluid-crystal coexistence for a patchy particle model. *The Journal of Physical Chemistry. B* **113**, 15133–6 (2009).

[21] Bianchi, E., Tartaglia, P., Zaccarelli, E. & Sciortino, F. Theoretical and numerical study of the phase diagram of patchy colloids: ordered and disordered patch arrangements. *The Journal of chemical physics* **128**, 144504 (2008).

[22] Wang, Y. *et al.* Colloids with valence and specific directional bonding. *Nature* **491**, 51–55 (2012).

[23] Seeman, N. C. & Sleiman, H. F. Dna nanotechnology. *Nature Reviews Materials* **3**, 17068 (2018).

[24] Rothemund, P. W. K. Folding DNA to create nanoscale shapes and patterns. *Nature* **440**, 297–302 (2006).

[25] Lund, K. *et al.* Molecular robots guided by prescriptive landscapes. *Nature* **465**, 206 (2010).

[26] Kay, E. R., Leigh, D. A. & Zerbetto, F. Synthetic molecular motors and mechanical machines. *Angewandte Chemie International Edition* **46**, 72–191 (2007).

[27] Berman, H. M. *et al.* The Protein Data Bank. *Nucleic acids research* **28**, 235–42 (2000).

[28] Darling, S. Directing the self-assembly of block copolymers. *Progress in Polymer Science* **32**, 1152–1204 (2007).

[29] Zhao, D. *et al.* Triblock copolymer syntheses of mesoporous silica with periodic 50 to 300 angstrom pores. *science* **279**, 548–552 (1998).

[30] Wang, G., Luo, X., Liu, C. & Huang, J. Synthesis of abcd 4-miktoarm star-shaped quarterpolymers by combination of the "click" chemistry with multiple polymerization mechanism. *Journal of Polymer Science Part A: Polymer Chemistry* **46**, 2154–2166 (2008).

[31] Iatrou, H. & Hadjichristidis, N. Synthesis and characterization of model 4-miktoarm star co-and quaterpolymers. *Macromolecules* **26**, 2479–2484 (1993).

[32] Hirao, A., Higashihara, T. & Inoue, K. Successive Synthesis of Well-Defined Asymmetric Star-Branched Polymers up to Seven-Arm, Seven-Component ABCDEFG Type by an Iterative Methodology Based on Living Anionic Polymerization. *Macromolecules* **41**, 3579–3587 (2008).

[33] Kochendoerfer, G. G. *et al.* Design and chemical synthesis of a homogeneous polymer-modified erythropoiesis protein. *Science* **299**, 884–887 (2003).

[34] Hawker, C. J. & Wooley, K. L. The convergence of synthetic organic and polymer chemistries. *Science* **309**, 1200–1205 (2005).

[35] Anderson, J. A., Sknepnek, R. & Travesset, A. Design of polymer nanocomposites in solution by polymer functionalization. *Phys. Rev. E* **82**, 021803 (2010).

[36] Hoshino, Y. *et al.* The rational design of a synthetic polymer nanoparticle that neutralizes a toxic peptide in vivo. *Proceedings of the National Academy of Sciences* **109**, 33–38 (2012).

[37] Lebouille, J. G. J. L., Leermakers, F. A. M., Cohen Stuart, M. A. & Tuinier, R. Design of block-copolymer-based micelles for active and passive targeting. *Phys. Rev. E* **94**, 042503 (2016).

[38] Clementi, C., Maritan, A. & Banavar, J. Folding, Design, and Determination of Interaction Potentials Using Off-Lattice Dynamics of Model Heteropolymers. *Physical Review Letters* **81**, 3287–3290 (1998).

[39] Maritan, A., Micheletti, C., Trovato, A. & Banavar, J. R. Optimal shapes of compact strings. *Nature* **406**, 287–90 (2000).

[40] Poletto, C., Giacometti, A., Trovato, A., Banavar, J. R. & Maritan, A. Emergence of secondary motifs in tubelike polymers in a solvent. *Physical Review E* **77**, 061804 (2008).

[41] Hoang, T. X., Trovato, A., Seno, F., Banavar, J. R. & Maritan, A. Geometry and symmetry presculpt the free-energy landscape of proteins. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 7960–4 (2004).

[42] Martin Karplus, Michael Levitt, Arieh Warshel. The Nobel Prize in Chemistry 2013- For the development of multiscale models for complex chemical systems. Nobelprize.org (2013).

[43] Coluzza, I. Computational protein design: a review. *Journal of Physics: Condensed Matter* **29**, 143001 (2017).

[44] Dahiyat, B. I. & Mayo, S. L. Probing the role of packing specificity in protein design. *Proceedings of the National Academy of Sciences of the United States of America* **94**, 10172–7 (1997).

[45] Desjarlais, J. R. & Handel, T. M. De novo design of the hydrophobic cores of proteins. *Protein science : a publication of the Protein Society* **4**, 2006–18 (1995).

[46] Desjarlais, J. R. & Handel, T. M. Side-chain and backbone flexibility in protein core design. *Journal of molecular biology* **290**, 305–18 (1999).

[47] Hellinga, H. W. & Richards, F. M. Construction of new ligand binding sites in proteins of known structure. *Journal of Molecular Biology* **222**, 763–785 (1991).

[48] Distasio, R. A., von Lilienfeld, O. A. & Tkatchenko, A. Collective many-body van der Waals interactions in molecular systems. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 14791–5 (2012).

[49] Rothlisberger, D. *et al.* Kemp elimination catalysts by computational enzyme design. *Nature* **453**, 190–U4 (2008).

[50] Dahiyat, B. I. De Novo Protein Design: Fully Automated Sequence Selection. *Science* **278**, 82–87 (1997).

[51] Kuhlman, B. *et al.* Design of a novel globular protein fold with atomic-level accuracy. *Science (New York, N.Y.)* **302**, 1364–8 (2003).

[52] Kellogg, E. H., Lange, O. F. & Baker, D. Evaluation and optimization of discrete state models of protein folding. *The journal of physical chemistry. B* **116**, 11405–13 (2012).

[53] Bryngelson, J. D. & Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proceedings of the National Academy of Sciences of the United States of America* **84**, 7524–8 (1987).

[54] Shakhnovich, E. I. & Gutin, A. M. Engineering of stable and fast-folding sequences of model proteins. *Proceedings of the National Academy of Sciences of the United States of America* **90**, 7195–9 (1993).

[55] Yue, K. & Dill, K. A. Inverse protein folding problem: designing polymer sequences. *Proceedings of the National Academy of Sciences* **89**, 4163–4167 (1992).

[56] Coluzza, I., Muller, H. G. & Frenkel, D. Designing refoldable model molecules. *Physical Review E* **68**, 46703 (2003).

[57] Coluzza, I. & Frenkel, D. Designing specificity of protein-substrate interactions. *Physical Review. E* **70**, 51917 (2004).

[58] Coluzza, I. & Frenkel, D. Monte Carlo study of substrate-induced folding and refolding of lattice proteins. *Biophysical Journal* **92**, 1150–6 (2007).

[59] Coluzza, I. Transferable coarse-grained potential for de novo protein folding and design. *PloS one* **9**, e112852 (2014).

[60] Coluzza, I. A coarse-grained approach to protein design: learning from design to understand folding. *PloS one* **6**, e20853 (2011).

[61] Coluzza, I., van Oostrum, P., Capone, B., Reimhult, E. & Dellago, C. Sequence Controlled Self-Knotting Colloidal Patchy Polymers. *Physical Review Letters* **110**, 075501 (2013).

[62] Coluzza, I., van Oostrum, P. D., Capone, B., Reimhult, E. & Dellago, C. Design and folding of colloidal patchy polymers. *Soft Matter* **9**, 938–944 (2013).

[63] Gonen, S., DiMaio, F., Gonen, T. & Baker, D. Design of ordered two-dimensional arrays mediated by noncovalent protein-protein interfaces. *Science* **348**, 1365–1368 (2015).

[64] Craig, A. & Terentjev, E. Folding of polymer chains with short-range binormal interactions. *Journal of Physics A: Mathematical and General* **39**, 4811 (2006).

[65] Banavar, J. R., Cieplak, M., Hoang, T. X. & Maritan, A. First-principles design of nanomachines. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 6900–3 (2009).

[66] Škrbić, T., Badasyan, A., Hoang, T. X., Podgornik, R. & Giacometti, A. From polymers to proteins: the effect of side chains and broken symmetry on the formation of secondary structures within a Wang–Landau approach. *Soft Matter* **12**, 4783–4793 (2016). arXiv:1601.00424v1.

[67] Škrbić, T., Hoang, T. X. & Giacometti, A. Effective stiffness and formation of secondary structures in a protein-like model. *The Journal of Chemical Physics* **145**, 084904 (2016).

[68] Gutin, A. M. & Shakhnovich, E. Ground-state of random copolymers and the discrete Random Energy-model. *The Journal of Chemical Physics* **98**, 8174–8177 (1993).

[69] Pande, V. S., Grosberg, A. Y. & Tanaka, T. Statistical mechanics of simple models of protein folding and design. *Biophysical journal* **73**, 3192–3210 (1997).

[70] Go, N. & Taketomi, H. Respective roles of short-range and long-range interactions in protein folding. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **75**, 559–563 (1978).

[71] Frank-Kamenetskii, M. D., Lukashin, A. V. & Vologodskii, A. V. Statistical mechanics and topology of polymer chains. *Nature* **258**, 398 (1975).

[72] Sumners, D. & Whittington, S. Knots in self-avoiding walks. *Journal of Physics A: Mathematical and General* **21**, 1689 (1988).

[73] Pippenger, N. Knots in random walks. *Discrete Applied Mathematics* **25**, 273–278 (1989).

[74] Micheletti, C., Marenduzzo, D. & Orlandini, E. Polymers with spatial or topological constraints: Theoretical and computational results. *Physics Reports* **504**, 1–73 (2011).

[75] Koniaris, K. & Muthukumar, M. Knottedness in ring polymers. *Physical review letters* **66**, 2211 (1991).

[76] Matthews, R., Louis, A. A. & Likos, C. N. Effect of bending rigidity on the knotting of a polymer under tension. *ACS macro letters* **1**, 1352–1356 (2012).

[77] Poier, P., Likos, C. N. & Matthews, R. Influence of rigidity and knot complexity on the knotting of confined polymers. *Macromolecules* **47**, 3394–3400 (2014).

[78] Orlandini, E., Baiesi, M. & Zonta, F. How local flexibility affects knot positioning in ring polymers. *Macromolecules* **49**, 4656–4662 (2016).

[79] Miller, M. A. & Wales, D. J. Novel structural motifs in clusters of dipolar spheres: knots, links, and coils. *The journal of physical chemistry. B* **109**, 23109–12 (2005).

[80] Liu, L. F., Depew, R. E. & Wang, J. C. Knotted single-stranded DNA rings: a novel topological isomer of circular single-stranded DNA formed by treatment with Escherichia coli ω protein. *Journal of molecular biology* **106**, 439–452 (1976).

[81] Dean, F. B., Stasiak, A., Koller, T. & Cozzarelli, N. R. Duplex dna knots produced by escherichia coli topoisomerase i. structure and requirements for formation. *Journal of Biological Chemistry* **260**, 4975–4983 (1985).

[82] Du, S. M., Stollar, B. D. & Seeman, N. C. A synthetic dna molecule in three knotted topologies. *Journal of the American Chemical Society* **117**, 1194–1200 (1995).

[83] Arsuaga, J. *et al.* Dna knots reveal a chiral organization of dna in phage capsids. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 9165–9169 (2005).

[84] Virnau, P., Mirny, L. A. & Kardar, M. Intricate knots in proteins: Function and evolution. *PLoS computational biology* **2**, e122 (2006).

[85] Wagner, J. R., Brunzelle, J. S., Forest, K. T. & Vierstra, R. D. A light-sensing knot revealed by the structure of the chromophore-binding domain of phytochrome. *Nature* **438**, 325 (2005).

[86] Boutz, D. R., Cascio, D., Whitelegge, J., Perry, L. J. & Yeates, T. O. Discovery of a thermophilic protein complex stabilized by topologically interlinked chains. *Journal of molecular biology* **368**, 1332–1344 (2007).

[87] Taylor, W. A deeply knotted protein structure and how it might fold. *Nature* **406**, 8–11 (2000).

[88] Bölinger, D. *et al.* A stevedore's protein knot. *PLoS computational biology* **6**, e1000731 (2010).

[89] Jamroz, M. *et al.* Knotprot: a database of proteins with knots and slipknots. *Nucleic acids research* **43**, D306–D314 (2014).

[90] Millett, K. C., Rawdon, E. J., Stasiak, A. & Sułkowska, J. I. Identifying knots in proteins (2013).

[91] Tubiana, L., Orlandini, E. & Micheletti, C. Probing the Entanglement and Locating Knots in Ring Polymers: A Comparative Study of Different Arc Closure Schemes. *Progress of Theoretical Physics Supplement* **191**, 192–204 (2011).

[92] King, N. P., Jacobitz, A. W., Sawaya, M. R., Goldschmidt, L. & Yeates, T. O. Structure and folding of a designed knotted protein. *Proceedings of the National Academy of Sciences* **107**, 20732–20737 (2010).

[93] Mallam, A. L. & Jackson, S. E. Knot formation in newly translated proteins is spontaneous and accelerated by chaperonins. *Nature chemical biology* **8**, 147 (2012).

[94] Najafi, S. & Potestio, R. Folding of small knotted proteins: Insights from a mean field coarse-grained model. *The Journal of chemical physics* **143**, 12B606_1 (2015).

[95] Wigley, D. B. Structure and mechanism of dna topoisomerases. *Annual review of biophysics and biomolecular structure* **24**, 185–208 (1995).

[96] Arsuaga, J., Vázquez, M., Trigueros, S., Sumners, D. W. & Roca, J. Knotting probability of DNA molecules confined in restricted volumes: DNA knotting in phage capsids. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 5373–7 (2002).

[97] Lim, N. C. & Jackson, S. E. Molecular knots in biology and chemistry. *Journal of Physics: Condensed Matter* **27**, 354101 (2015).

[98] Horner, K. E., Miller, M. A., Steed, J. W. & Sutcliffe, P. M. Knot theory in modern chemistry. *Chemical Society Reviews* **45**, 6432–6448 (2016).

[99] Fielden, S. D., Leigh, D. A. & Woltering, S. L. Molecular knots. *Angewandte Chemie (International ed. in English)* **56**, 11166 (2017).

[100] Sułkowska, J. I., Sułkowski, P., Szymczak, P. & Cieplak, M. Stabilizing effect of knots on proteins. *Proceedings of the National Academy of Sciences* **105**, 19714–19719 (2008).

[101] Clark, R. J. *et al.* The engineering of an orally active conotoxin for the treatment of neuropathic pain. *Angewandte Chemie (International ed. in English)* **49**, 6545–8 (2010).

[102] Sayre, T. C., Lee, T. M., King, N. P. & Yeates, T. O. Protein stabilization in a highly knotted protein polymer. *Protein Engineering, Design & Selection* **24**, 627–630 (2011).

[103] Soler, M. A., Nunes, A. & Faísca, P. F. Effects of knot type in the folding of topologically complex lattice proteins. *The Journal of chemical physics* **141**, 07B607_1 (2014).

[104] Newland, B. *et al.* Single cyclized molecule versus single branched molecule: a simple and efficient 3d "knot" polymer structure for nonviral gene delivery. *Journal of the American Chemical Society* **134**, 4782–4789 (2012).

[105] Qiu, L. Y. & Bae, Y. H. Polymer architecture and drug delivery. *Pharmaceutical research* **23**, 1–30 (2006).

[106] Fox, M. E., Szoka, F. C. & Fréchet, J. M. Soluble polymer carriers for the treatment of cancer: the importance of molecular architecture. *Accounts of chemical research* **42**, 1141–1151 (2009).

[107] Colgrave, M. L. & Craik, D. J. Thermal, chemical, and enzymatic stability of the cyclotide kalata b1: the importance of the cyclic cystine knot. *Biochemistry* **43**, 5965–5975 (2004).

[108] Taylor, W. R. Protein knots and fold complexity: some new twists. *Computational biology and chemistry* **31**, 151–162 (2007).

[109] Marcos, V. *et al.* Allosteric initiation and regulation of catalysis with a molecular knot. *Science* **352**, 1555–1559 (2016).

[110] Dietrich-Buchecker, C. O., Sauvage, J. P. & Kern, J. M. Templated synthesis of interlocked macrocyclic ligands: The catenands. *Journal of the American Chemical Society* **106**, 3043–3045 (1984).

[111] Ayme, J.-F. *et al.* A synthetic molecular pentafoil knot. *Nature Chemistry* **4**, 15–20 (2012).

[112] Leigh, D. A., Pritchard, R. G. & Stephens, A. J. A star of david catenane. *Nature chemistry* **6**, 978 (2014).

[113] Ponnuswamy, N., Cougnon, F. B., Pantos, G. D. & Sanders, J. K. Homochiral and meso figure eight knots and a solomon link. *Journal of the American Chemical Society* **136**, 8243–8251 (2014).

[114] Danon, J. J. *et al.* Braiding a molecular knot with eight crossings. *Science* **355**, 159–162 (2017).

[115] Wüst, T., Reith, D. & Virnau, P. Sequence determines degree of knottedness in a coarse-grained protein model. *Physical review letters* **114**, 028102 (2015).

[116] Knot atlas. http://katlas.org/.

[117] Cesario, M., Dietrich-Buchecker, C., Guilhem, J., Pascard, C. & Sauvage, J. Molecular structure of a catenand and its copper (i) catenate: complete rearrangement of the interlocked macrocyclic ligands by complexation. *Journal of the Chemical Society, Chemical Communications* 244–247 (1985).

[118] Orlandini, E., Polles, G., Marenduzzo, D. & Micheletti, C. Self-assembly of knots and links. *Journal of Statistical Mechanics: Theory and Experiment* **2017**, 034003 (2017).

[119] Umena, Y., Kawakami, K., Shen, J.-R. & Kamiya, N. Crystal structure of oxygen-evolving photosystem II at a resolution of 1.9 Å. *Nature* **473**, 55–60 (2011).

[120] Furumi, S. Active lasing from organic colloidal photonic crystals. *Journal of Materials Chemistry C* **1**, 6003–6012 (2013).

[121] Yang, Y. *et al.* Combinatorial polymer scaffold libraries for screening cell-biomaterial interactions in 3d. *Advanced Materials* **20**, 2037–2043 (2008).

[122] Capone, B., Hansen, J.-P. & Coluzza, I. Competing micellar and cylindrical phases in semi-dilute diblock copolymer solutions. *Soft Matter* **6**, 6075 (2010).

[123] Yang, L., Zhou, H., Shi, G., Wang, Y. & Pan, C.-Y. Synthesis of ABCD 4-miktoarm star polymers by combination of RAFT, ROP, and "Click Chemistry". *Journal of Polymer Science Part A: Polymer Chemistry* **46**, 6641–6653 (2008).

[124] Conde, J., Oliva, N., Atilano, M., Song, H. S. & Artzi, N. Self-assembled rna-triple-helix hydrogel scaffold for microrna modulation in the tumour microenvironment. *Nature materials* (2015).

[125] Chino, M. *et al.* Artificial Diiron Enzymes with a De Novo Designed Four-Helix Bundle Structure. *European Journal of Inorganic Chemistry* **2015**, 3371–3390 (2015).

[126] Huang, P.-S. *et al.* High thermodynamic stability of parametrically designed helical bundles. *Science* **346**, 481–485 (2014).

[127] Hamley, I. W. *Block copolymers in solution: fundamentals and applications* (John Wiley & Sons, 2005).

[128] Leibler, L. Theory of microphase separation in block copolymers. *Macromolecules* **13**, 1602–1617 (1980).

[129] Matsen, M. W. & Schick, M. Stable and unstable phases of a diblock copolymer melt. *Physical Review Letters* **72**, 2660 (1994).

[130] Fredrickson, G. *The equilibrium theory of inhomogeneous polymers* (Oxford University Press, USA, 2006).

[131] Martínez-Veracoechea, F. J. & Escobedo, F. A. Bicontinuous phases in diblock copolymer/homopolymer blends: simulation and self-consistent field theory. *Macromolecules* **42**, 1775–1784 (2009).

[132] Eurich, F., Karatchentsev, A., Baschnagel, J., Dieterich, W. & Maass, P. Soft particle model for block copolymers. *The Journal of chemical physics* **127**, 134905 (2007).

[133] Lodge, T. P., Bang, J., Li, Z., Hillmyer, M. A. & Talmon, Y. Introductory Lecture Strategies for controlling intra-and intermicellar packing in block copolymer solutions: Illustrating the flexibility of the self-assembly toolbox. *Faraday discussions* **128**, 1–12 (2005).

[134] Shakhnovich, E. I. Proteins with selected sequences fold into unique native conformation. *Phys. Rev. Lett.* **72**, 3907–3910 (1994).

[135] Seno, F., Vendruscolo, M., Maritan, A. & Banavar, J. R. Optimal protein design procedure. *Phys. Rev. Lett.* **77**, 1901–1904 (1996).

[136] Micheletti, C., Seno, F., Maritan, A. & Banavar, J. R. Design of proteins with hydrophobic and polar amino acids. *Proteins: Structure, Function and Genetics* **32**, 80–87 (1998). 9712124.

[137] Magee, J., Vasquez, V. & Lue, L. Helical Structures from an Isotropic Homopolymer Model. *Physical Review Letters* **96**, 2078028 (2006).

[138] Glotzer, S. C. & Solomon, M. J. Anisotropy of building blocks and their assembly into complex structures. *Nature materials* **6**, 557–562 (2007).

[139] Leonhard, K., Prausnitz, J. & Radke, C. Solvent–amino acid interaction energies in 3-d-lattice mc simulations of model proteins. aggregation thermodynamics and kinetics. *Physical Chemistry Chemical Physics* **5**, 5291–5299 (2003).

[140] Irbäck, A., Sjunnesson, F. & Wallin, S. Three-helix-bundle protein in a Ramachandran model. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 13614–8 (2000).

[141] Finkelstein, A. V. & Ptitsyn, O. *Protein physics: a course of lectures* (Academic Press, 2002).

[142] Seeman, N. C. DNA in a Material World. *Nature* **421**, 427–431 (2003).

[143] Lee, J. B. *et al.* A Mechanical Metamaterial Made From a DNA Hydrogel. *Nature Nanotechnology* **7**, 816–820 (2012).

[144] Vutukuri, H. R. *et al.* Colloidal analogues of charged and uncharged polymer chains with tunable stiffness. *Angewandte Chemie (International ed. in English)* **51**, 11249–53 (2012).

[145] Durek, T. & Becker, C. F. Protein semi-synthesis: new proteins for functional and structural studies. *Biomolecular engineering* **22**, 153–172 (2005).

[146] Gannavaram, S., Sirin, S., Sherman, W. & Gadda, G. Mechanistic and computational studies of the reductive half-reaction of tyrosine to phenylalanine active site variants of d-arginine dehydrogenase. *Biochemistry* **53**, 6574–6583 (2014).

[147] Sirin, S. *et al.* A computational approach to enzyme design: Predicting ω-aminotransferase catalytic activity using docking and mm-gbsa scoring. *Journal of chemical information and modeling* **54**, 2334–2346 (2014).

[148] Sacanna, S., Irvine, W. T. M., Chaikin, P. M. & Pine, D. J. Lock and key colloids. *Nature* **464**, 575–578 (2010).

[149] van der Meulen, S. A. & Leunissen, M. E. Solid colloids with surface-mobile dna linkers. *Journal of the American Chemical Society* **135**, 15129–15134 (2013).

[150] Feng, L., Pontani, L.-L., Dreyfus, R., Chaikin, P. & Brujic, J. Specificity, flexibility and valence of dna bonds guide emulsion architecture. *Soft Matter* **9**, 9816–9823 (2013).

[151] Miyazawa, S. & Jernigan, R. L. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *Journal of Molecular Biology* **256**, 623–644 (1996).

[152] Betancourt, M. R. & Thirumalai, D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Science* **8**, 361–9 (1999).

[153] Helling, R. *et al.* The designability of protein structures. *Journal Of Molecular Graphics & Modelling* **19**, 157–167 (2001).

[154] Coluzza, I. & Frenkel, D. Virtual-move parallel tempering. *Chemphyschem : a European journal of chemical physics and physical chemistry* **6**, 1779–83 (2005).

[155] Nocera, D. G. The artificial leaf. *Accounts of chemical research* **45**, 767–776 (2012).

[156] Kamenetzky, E. A., Magliocco, L. G. & Panzer, H. P. Structure of solidified colloidal array laser filters studied by cryogenic transmission electron microscopy. *Science* **263**, 207–210 (1994).

[157] Shih, W. M., Quispe, J. D. & Joyce, G. F. A 1.7-kilobase single-stranded DNA that folds into a nanoscale octahedron. *Nature* **427**, 618–621 (2004).

[158] Zanchetta, G., Bellini, T., Nakata, M. & Clark, N. A. Physical Polymerization and Liquid Crystallization of RNA Oligomers. *J. Am. Chem. Soc.* **130**, 12864–12865 (2008).

[159] DeGrado, W. F., Summa, C. M., Pavone, V., Nastri, F. & Lombardi, A. DE NOVO DESIGN AND STRUCTURAL CHARACTERIZATION OF PROTEINS AND METALLOPROTEINS. *Annual Review of Biochemistry* **68**, 779–819 (1999).

[160] Betz, S. F., Raleigh, D. P. & DeGrado, W. F. De novo protein design: from molten globules to native-like states. *Current Opinion in Structural Biology* **3**, 601–610 (1993).

[161] Whitehead, T. A. *et al.* Optimization of affinity, specificity and function of designed influenza inhibitors using deep sequencing. *Nature biotechnology* **30**, 543–8 (2012).

[162] King, N. P. *et al.* Computational design of self-assembling protein nanomaterials with atomic level accuracy. *Science (New York, N.Y.)* **336**, 1171–4 (2012).

[163] Lawrence, D. S., Jiang, T. & Levett, M. Self-Assembling Supramolecular Complexes. *Chemical Reviews* **95**, 2229–2260 (1995).

[164] Damasceno, P. F., Engel, M. & Glotzer, S. C. Predictive self-assembly of polyhedra into complex structures. *Science* **337**, 453–457 (2012).

[165] Rogers, W. B., Shih, W. M. & Manoharan, V. N. Using DNA to program the self-assembly of colloidal nanoparticles and microparticles. *Nature Rev. Mater.* **1**, 16008 (2016).

[166] Bianchi, E., Capone, B., Coluzza, I., Rovigatti, L. & van Oostrum, P. D. J. D. J. Limiting the valence: advancements and new perspectives on patchy colloids, soft functionalized nanoparticles and biomolecules. *Phys. Chem. Chem. Phys.* **19**, 19847–19868 (2017).

[167] Bianco, V., Franzese, G., Dellago, C. & Coluzza, I. Role of Water in the Selection of Stable Proteins at Ambient and Extreme Thermodynamic Conditions. *Physical Review X* **7**, 21047 (2017).

[168] Derrida, B. Random-energy model: An exactly solvable model of disordered systems. *Physical Review B* **24**, 2613–2626 (1981).

[169] Vissers, T., Preisler, Z., Smallenburg, F., Dijkstra, M. & Sciortino, F. Predicting crystals of Janus colloids. *Journal of Chemical Physics* **138** (2013).

[170] Ronti, M. *et al.* Free energy calculations for rings and chains formed by dipolar hard spheres. *Soft Matter* (2017).

[171] Coluzza, I., Jackson, S. E., Micheletti, C. & Miller, M. A. Knots in soft condensed matter. *Journal of Physics Condensed Matter* **27**, 301 (2015).

[172] Tkalec, U., Ravnik, M., Copar, S., Zumer, S. & Musevic, I. Reconfigurable Knots and Links in Chiral Nematic Colloids. *Science* **333**, 62–65 (2011). URL http://www.sciencemag.org/cgi/doi/10.1126/science.1205705.

[173] Seč, D., Čopar, S. & Žumer, S. Topological zoo of free-standing knots in confined chiral nematic fluids. *Nature Communications* **5**, 3057 (2014). URL http://www.nature.com/doifinder/10.1038/ncomms4057.

[174] Machon, T. & Alexander, G. P. Knotted Defects in Nematic Liquid Crystals. *Physical Review Letters* **113**, 027801 (2014). URL https://link.aps.org/doi/10.1103/PhysRevLett.113.027801.

[175] Erbas-Cakmak, S. *et al.* Rotary and linear molecular motors driven by pulses of a chemical fuel. *Science* **358**, 340 LP – 343 (2017).

[176] Polles, G., Marenduzzo, D., Orlandini, E. & Micheletti, C. Self-assembling knots of controlled topology by designing the geometry of patchy templates. *Nature communications* **6** (2015).

[177] Polles, G., Orlandini, E. & Micheletti, C. Optimal self-assembly of linked constructs and catenanes via spatial confinement. *ACS Macro Letters* **5**, 931–935 (2016).

[178] Miyazawa, S. & Jernigan, R. L. Estimation of effective interresidue contact energies from protein crystal-structures - quasi-chemical approximation. *Macromolecules* **18**, 534–552 (1985).

[179] Betancourt, M. R. & Onuchic, J. N. Kinetics of proteinlike models: The energy landscape factors that determine folding. *The Journal of Chemical Physics* **103**, 773–787 (1995). URL http://aip.scitation.org/doi/10.1063/1.470109.

[180] Dill, K. A. *et al.* Principles of Protein-Folding - a Perspective From Simple Exact M Odels. *Protein Sci.* **4**, 561–602 (1995).

[181] Vendruscolo, M. Modified configurational bias monte carlo method for simulation of polymer systems. *The Journal of Chemical Physics* **106**, 2970–2976 (1997).

[182] Seno, F., Trovato, A., Banavar, J. & Maritan, A. Maximum Entropy Approach for Deducing Amino Acid Interactions in Proteins. *Physical Review Letters* **100**, 1–4 (2008).

[183] Coluzza, I., Pisignano, D., Gentili, D., Pontrelli, G. & Succi, S. Ultrathin Fibers from Electrospinning Experiments under Driven Fast-Oscillating Perturbations. *Physical Review Applied* **2**, 054011 (2014).

[184] Diao, Y., Dobay, A., Kusner, R. B., Millett, K. & Stasiak, A. The average crossing number of equilateral random polygons. *Journal of Physics A: Mathematical and General* **36**, 11561–11574 (2003).

[185] Hoste, J. & Thistlethwaite, M. knotscape. URL http://www.math.utk.edu/~morwen/knotscape.html.

[186] Alexander, J. W. & Briggs, G. B. On Types of Knotted Curves. *The Annals of Mathematics* **28**, 562 (1926). URL https://www.jstor.org/stable/1968399?origin=crossref.

[187] Marenz, M. & Janke, W. Knots as a topological order parameter for semiflexible polymers. *Physical review letters* **116**, 128301 (2016).

[188] Lua, R. C. & Grosberg, A. Y. Statistics of knots, geometry of conformations, and evolution of proteins. *PLoS computational biology* **2**, e45 (2006).

[189] Baiesi, M., Orlandini, E. & Stella, A. L. The entropic cost to tie a knot. *Journal of Statistical Mechanics: Theory and Experiment* **2010**, P06012 (2010).

[190] Wüst, T., Reith, D. & Virnau, P. Sequence determines degree of knottedness in a coarse-grained protein model. *Physical review letters* **114**, 28102 (2015).

[191] Micheletti, C., Di Stefano, M. & Orland, H. Absence of knots in known RNA structures. *Proceedings of the National Academy of Sciences* **112**, 2052–2057 (2015).

[192] Tinoco Jr, I. & Bustamante, C. How RNA folds. *Journal of molecular biology* **293**, 271–281 (1999).

[193] Greenleaf, W. J., Frieda, K. L., Foster, D. A. N., Woodside, M. T. & Block, S. M. Direct observation of hierarchical folding in single riboswitch aptamers. *Science* **319**, 630–633 (2008).

[194] Repsilber, D. *et al.* Formation of metastable RNA structures by sequential folding during transcription: time-resolved structural analysis of potato spindle tuber viroid (-)-stranded RNA by temperature-gradient gel electrophoresis. *Rna* **5**, 574–584 (1999).

[195] Liu, L. & Hyeon, C. Contact Statistics Highlight Distinct Organizing Principles of Proteins and RNA. *Biophysical Journal* **110**, 2320–2327 (2016). URL http://dx.doi.org/10.1016/j.bpj.2016.04.020.

[196] Coronel, L., Orlandini, E. & Micheletti, C. Non-monotonic knotting probability and knot length of semiflexible rings: the competing roles of entropy and bending energy. *Soft matter* **13**, 4260–4267 (2017).

[197] Tubiana, L., Orlandini, E. & Micheletti, C. luca-tubiana/kymoknot: Kymoknot initial release (2018). URL https://doi.org/10.5281/zenodo.1239859.

[198] Mallam, A. L. & Jackson, S. E. Folding studies on a knotted protein. *Journal of molecular biology* **346**, 1409–1421 (2005).

[199] Lim, N. C. H. & Jackson, S. E. Mechanistic Insights into the Folding of Knotted Proteins In Vitro and In Vivo. *Journal of molecular biology* **427**, 248–258 (2015).

[200] Najafi, S. & Potestio, R. Folding of small knotted proteins: Insights from a mean field coarse-grained model. *The Journal of chemical physics* **143**, 243121 (2015).