TOPICAL REVIEW • OPEN ACCESS

Computational methods in the study of self-entangled proteins: a critical appraisal

To cite this article: Claudio Perego and Raffaello Potestio 2019 *J. Phys.: Condens. Matter* **31** 443001

OPEN ACCESS
IOP Publishing
Journal of Physics: Condensed Matter
J. Phys.: Condens. Matter **31** (2019) 443001 (30pp)
https://doi.org/10.1088/1361-648X/ab2f19

Topical Review

Computational methods in the study of self-entangled proteins: a critical appraisal

Claudio Perego[1,4] **and Raffaello Potestio**[2,3]

[1] Max Panck Institute for Polymer Research, Ackermannweg 10, Mainz 55128, Germany
[2] Physics Department, University of Trento, via Sommarive, 14 38123, Trento, Italy
[3] INFN-TIFPA, Trento Institute for Fundamental Physics and Applications, 38123 Trento, Italy

E-mail: perego@mpip-mainz.mpg.de and raffaello.potestio@unitn.it

Received 21 October 2016, revised 3 June 2019
Accepted for publication 3 July 2019
Published 12 August 2019

Abstract

The existence of self-entangled proteins, the native structure of which features a complex topology, unveils puzzling, and thus fascinating, aspects of protein biology and evolution. The discovery that a polypeptide chain can encode the capability to self-entangle in an efficient and reproducible way during folding, has raised many questions, regarding the possible function of these knots, their conservation along evolution, and their role in the folding paradigm. Understanding the function and origin of these entanglements would lead to deep implications in protein science, and this has stimulated the scientific community to investigate self-entangled proteins for decades by now. In this endeavour, advanced experimental techniques are more and more supported by computational approaches, that can provide theoretical guidelines for the interpretation of experimental results, and for the effective design of new experiments. In this review we provide an introduction to the computational study of self-entangled proteins, focusing in particular on the methodological developments related to this research field. A comprehensive collection of techniques is gathered, ranging from knot theory algorithms, that allow detection and classification of protein topology, to Monte Carlo or molecular dynamics strategies, that constitute crucial instruments for investigating thermodynamics and kinetics of this class of proteins.

Keywords: protein knots, computational methods, self-entangled polymers

(Some figures may appear in colour only in the online journal)

1. Introduction

The notion of *knot* is part of everyone's cultural background as it is rather ubiquitous in everyday experience. From tying shoelaces to unraveling earphone cords, we often have to deal with tightening strings by means of specific, 'artificial' knots, or removing spontaneous unwanted ones. Entanglements, self-entanglements, and braids are to be found in all areas of human activity: knots keep bags closed, hold fence poles together, secure sails to boats and tie-up boats to piers; knots make up the weaves of fabric as well as of geometric patterns that decorate sacred temples and public buildings. Whatever the type and the scope, tangles are strongly tied to culture, science, technology, and art.

A large part of the fascination that knots exert on us comes from the observation of the order that underlies them. Particularly striking is the fact that the *type* of a knot is largely independent of its precise form and shape—a consequence of the *topological* nature of knots. This distinct feature, namely the possibility to deform a closed and tangled piece of rope

[4] Author to whom any correspondence should be addressed.

without ever changing its knotted state, is qualitatively different from other geometrical or physical ones, such as structure and interactions, and is intrinsically resilient and permanent, provided that the rope is not cut.

This property can be quantified, for example, in terms of the smallest number of crossings that a knotted loop has once flattened onto a surface. This observation makes it possible to classify and organise knots in a hierarchy. Starting with the most trivial knot of all—a circular loop without any crossing—one can construct lists and tables of more and more complex entanglements, each defined by a set of properties and relations with respect to the others (e.g. the number of passages of a cord strand through another it takes to transform a knot of a given type into a different one). Along decades, knot theory has become a rich and ubiquitous field of research and, still nowadays, the analysis of knots and their role in nature remains a challenge for the scientific community [1–4].

The qualitative and quantitative properties of knots have inspired many in the search for their natural occurrences. One of the most imaginative attempts of rationalising natural phenomena in terms of knots has been pursued by Thomson and Tait, who tried to explain the discrete light absorption and emission spectra of atoms describing the latter as closed strings, and their excitation levels as different knotted states. It is only in relatively recent times, however, that it has been demonstrated how frequent and widespread knots are in living matter. Life on Earth, in fact, relies on molecules such as RNA, DNA, and proteins, which are natural heteropolymers, i.e. linear chains composed by fundamental units (nitrogen-containing bases or amino acids) whose sequence determines their biochemical function as well as their three-dimensional arrangement in space. Given that the most fundamental building blocks of life at the molecular level are, in essence, strings, it should come with little surprise that knots play a crucial role there as well.

Exact mathematical results show that a polymer at thermal equilibrium is to be found knotted with a probability that approaches unity exponentially fast as its length increases. The incredibly long DNA filaments that constitute the genetic payload of viruses or the chromosomes of cells are thus prone to self-entangle, with important consequences on their biological function. A knot on the DNA of a virus, for example, could clog the fiber during its injection into a target cell, thus compromising the infection process. Correspondingly, tight tangles within chromatin could be responsible for malfunctions in the condensation into chromosomes and, eventually, cell replication. Several different strategies have developed that allow cells to overcome the potential problems created by entanglements in DNA fibers. For example, the particular spool-like arrangement of DNA in viruses represents a passive method (i.e. one not determined by ad hoc molecular machinery) to avoid the problems that a tight entanglement might determine in the ejection of the genome [5]; knots, in fact, are indeed present, yet they are delocalised and distributed throughout long stretches of the filament. Active systems which resolve entanglements are present, too: a prominent example is given by topoisomerases [6], specific enzymes which resolve torsional strains and entangled strands particularly during DNA replication. The reader interested in studying this fascinating topic is referred to the articles and reviews available in the literature.

In the realm of biological polymers, another class of molecules where knots can be found are proteins. At present, protein knots represent a small yet non negligible fraction of the structures stored in the PDB, as much as $\sim$1% [7–11]. These self-entangled molecules represents a twofold puzzle. It is reasonable to assume that the folding process that leads a polypeptide chain into a knotted conformation entails specific features necessary to cope with the extra-degree of complexity represented by the tangle; nonetheless, these features cannot be expected to lie in the knotted proteins' *physics*, as these molecules are composed by the same amino acids that make up the remaining 99%. The first big question is thus: *how* do these proteins fold?

The existence of these proteins clearly demonstrates that the *sequence* $\rightarrow$ *structure* paradigm is powerful enough as to dictate also complex topologies, however the small relative abundance of topologically nontrivial proteins suggests that natural selections tends to disfavour them. The second big question hence reads: *why* are there knotted proteins? The folding process of a knotted protein is necessarily more complex and error-prone than that of a topologically trivial molecule with comparable length. If a small yet substantial amount of proteins has overcome all sieves and hard walls imposed by natural selection, what looks at the first glance as a clear-cut handicap must either entail an advantage, or at least be much less limiting than what one would intuitively assume.

During the past 25 years, both problems have been tackled by many authors from the experimental as well as theoretical (i.e. numerical) point of view. In this review, we concentrate on the computational tools—standard as well as tailored techniques—that have been developed and employed to unravel the tangled problem of knots in proteins. The aim of this review is to present the reader with a comprehensive, albeit likely not complete, account of the available computational methods developed to study knotted proteins; to provide the basic knowledge necessary to make one's way through their implementation and usage; and to supply a sufficiently broad and accessible collection of resources where this and further knowledge can be retrieved.

The article is structured as follows. In section 2 we provide an overview of the current knowledge of topologically complex proteins, with a particular focus on the crucial questions raised by the existence of these peculiar native structures. In section 3 we dive into the computational methodologies, introducing the mathematical tools used to identify and classify protein entanglements, the algorithms employed to implement and apply these tools, and the existing databases gathering different classes of entangled structures. Next, in section 4, we discuss the computational approaches aimed at simulating protein dynamics, focusing on the system-specific models developed for the study of self-entangled proteins. Finally, in section 5, standing on the high mountain of this knowledge, we look in the direction of the developments to come.

2. Self-entanglements in proteins

Proteins are heteropolymers, that is, linear, unbranched chains whose basic units belong to a repertoire of 20 different building blocks, the amino acids [12]. The most remarkable property of these chains is that the majority of them (with the important exception of intrinsically disordered proteins) fold in a well-defined three-dimensional structure. Unfortunately, this structure is available to us for a fairly limited fraction of all existing proteins, due to the complexity and limitations of the experimental procedures employed to determine the three-dimensional arrangement of the molecule's atoms. Hence, it is often the case that the sequence, and possibly the biological function, of a protein are much better known than its shape.

Once protein chains are arranged in their biologically active native conformation, one can easily marvel before the immense variety of shapes they assume, functions they perform, dimensions they attain; the common nickname depicting proteins as the 'workhorses' of life, albeit not wrong, cannot account for the broad spectrum of roles they play on the biological stage. And yet, before curtains lift and the play starts, a possibly even more fascinating process must take place, that is folding. Newly synthesised proteins leave the ribosome as floating cords fluctuating in the cytoplasm, then collapse onto themselves to attain the crystal-like arrangement that enables their biological function. Skating on a thin cliff between stochasticity and determinism, these polymeric molecules turn into solid particles with features of elasticity, flexibility and plasticity which vary depending on the part they play. Their collapse can occur autonomously, by the sole means of the interactions within the protein and between the protein and the surrounding solvent, or they can be aided in doing so by other proteins, the chaperones, which confine the polypeptide and shield it from external disturbances; they attain their native conformation through a process resembling self-assembly, abruptly crashing onto themselves in a two-state transition or starting the folding simultaneously on several distinct points of the chain, growing the final structure as a crystal forms from several merging domains. Whatever the beaten path, whatever the complexity of the process, we observe in awe a filamentous fleck of matter dance its way from being a piece of rope whipped by the turbulent waters in which it is immersed to become the robust construction of an efficient chemical machine.

And yet, we do not face substantial difficulties in accepting that this process can take place. In spite of its complexity, protein folding is comprehensible phenomenon whose fine, case-specific details might be hard to decipher but whose general features are understandable and largely understood. It is common thought that the major barrier one has to overcome to reproduce *in silico* the folding process of a protein is a computational, not conceptual one. If sufficient computer power is provided, protein folding can be fairly easily simulated.

A much tougher pill to swallow is the idea that a protein can *knot*. This largely depends on the anthropocentric perspective we have on knots—at least those knots we tie ourselves. Many threads in nature knot by themselves—hair, umbilical cords, DNA, earphones. These entanglements, however, are stochastic and irreproducible: their occurrence depends on chance and, if the 'experiment' is performed multiple times, they do not manifest repeatedly in the same manner—position, knot complexity, shape, arrangement. It is possible to measure or compute a probability distribution for a given type of knot to occur on a polymer, but a specific realisation of that knot will not appear systematically [13].

Proteins behave differently. Their function can be carried out only if a well-defined three-dimensional arrangement is attained (with the remarkable exception of intrinsically disordered proteins [14–18]). The necessity of collapsing into such a precise structure is incompatible with the possibility of randomly knotting: if different copies of the same protein tied different knots before folding, they would hardly reach the same native conformation. Only one precise kind of knot, always tied in the same manner, would be compatible with a functional protein structure.

One of the closest examples of such a knot is provided by shoelaces: same position, same topology, same arrangement—and clearly always the same function, that is, keep shoes tight. This analogy pairs the one between a knotted DNA filament and the tangled wire of your earphones. A major difference exists between the two cases though, that is *intentionality*. It is often the case that one would like to *avoid* the latter tangle, which happens nonetheless and always in different ways; on the contrary, the art and craft of tying your shoes is something that requires effort, practice, and a sharp determination to be carried out: very seldom do shoelaces *happen* to be tied in the precise way you want them to (even though they quite often tend to *untie* whether you want it or not; the cause of this has been identified in a combination of the impact of the shoe on the floor, which loosens the knot, and the whipping portion of the free ends [19]). It appears thus understandable why the idea of a self-knotting protein, capable of doing so in a predicable, deterministic, reproducible manner, has been looked at as biologically impossible until very recently.

Indeed, evidence has for a long time pointed in the direction of a complete absence of knots from the realm of protein structures. Until 1994, the totality of resolved structures featured no complex topology, at least as far as the protein backbone *alone* was considered. Different, slightly more complex forms of entanglement had been looked at with a keener eye, namely the knotted loops that the backbone forms if disulphide bridges are present. Indeed, Crippen [20] speculated as early as in 1974 about the properties of knots in such loops, the probability of their occurrence, and the consequences that these have on the molecule's folding and denaturation. However, the current picture of protein folding, depicting the chain progressively collapsing as a polymer in a bad solvent, ruled out the possibility of a complex pathway as the one required to tie a knot, and so did structure prediction software which deemed knotted structures to be 'impossible', thus discarding these results [21–23]. A knotted backbone was taken as a signature of a folding process gone bad, or of a mistake in the structure determination via crystallography.

In 1994, Mansfield [24] raised the question of the existence of protein knots with some data at hand. In fact, he identified a protein, carbonic anhydrase (CAB), which *might* have

contained a knot if the C-terminal domain could have been considered to penetrate, albeit by a few residues, a distal loop. This particular case did not represent strong evidence in favour of the existence of knotted proteins, as the definition of such knot was too loose, strictly dependent on possible inaccuracies in the structural determination of the terminal residues, and on the perspective (the particular projection employed to look at the backbone path). Furthermore, the observation of a knot in CAB had been done already in 1977 by Richardson [25], apparently without echoes. However, the work by Mansfield sparked renovated interest on the topic, and in 1997 he further elaborated on that [26] discussing CAB and S-adenosylmethionine synthetase (MAT).

The main difficulty in the identification of knotted backbones in protein structures is technical: visual inspection is difficult, time-consuming, and error-prone. Algorithmic procedures thus have to be put in place in order to process the vast number of available structures (see section 3 for further details). The first, effective approach in this sense was developed by Taylor in [27]: the method relies on the systematic simplification of the protein chain -initially constituted by the C_α trace- through the iterative removal of a point and the direct connection of the neighbouring ones, all done while keeping the terminal points fixed. If a straight line is obtained between the termini without operating a chain crossing, the protein is unknotted; alternatively, a knot is found whose structural complexity is sufficiently low to allow for a visual inspection and identification. In this work, Taylor speculated about the folding process of one of the identified knotted proteins, PDB code 1yveI, whose entangled core was suggestive of an internal duplication of a sequence stretch into two identical alpha-helices loops which compenetrate.

From this pioneering investigation several powerful algorithms have been developed to automatically identify and characterise knotted structures, which have opened the Pandora box of self-entangled proteins. Indeed, a proper knot can be defined only on a closed loop, while open chains such as proteins are by construction unknotted; however, appropriate strategies combining chain closure methods and knot identification schemes have enabled researchers to spot a wide spectrum of self-entanglement in proteins, ranging from self-evident conformations with buried knots and exposed termini to evanescent links between portions of otherwise unknotted chains. Since 2000, a steadily growing number of instances have been found [7, 10, 28–31] which contain knots of complexity ranging from the simplest trefoil knot to the most complex known to date, a Stevedore knot with six crossings [32]. About 1% of the available PDB entries feature a knotted topology [10].

The relative abundance of different knots is inversely proportional to the complexity, and the knotted portions of these proteins is much larger than the minimal length the corresponding entanglement can have: this property can be reasonably attributed to the difficulty of squeezing several amino acids one against the other, not to mention the large forces that would be required to attain such a densely packed conformation during the folding process. Notably, all known protein knots are of the same *twist* type [10], that is, knots obtained twisting a U-turn shaped strand and closing the chain passing a terminus through the loop; in contrast, only the simplest protein knots belong to the other main class of knots, the *torus* knots, which are defined embedded in a two-dimensional toroidal surface. Indeed, the vast majority of protein knots are of the trefoil type, that is, the sole knot which falls in both twist and torus classes. None of the protein knots with more than three crossings can be classified as a torus knot. This absence might be due to the relative simplicity of knotting a twist knot, which is obtained threading a terminus through a loop only once, in contrast to torus knots which require two or more passages.

In addition to protein knots, one should also account for other kinds of entanglements. Examples of these alternative topological states are *slipknots*, conformations where the backbone makes a U-turn re-entering an already pierced loop, thereby 'undoing' the knot, or *complex lassos*, in which a cysteine bond closes a backbone loop which is pierced by the rest of the chain. A detailed overview of the non-trivial topologies found in protein structures is provided in section 3.4; here, we just underline that, taking into account all these structures, the amount of self-entangled proteins sums up to 6% of the PDB [11].

The observations hitherto reported raise a number of questions, which, up to date, have been only partially answered. These questions pertain the where, the why, and the by what means, that is: are knots 'watermarked' in the protein's sequence in some special, otherwise uncommon manner? What is the functional advantage of a self-entangled topology? And how is this complex conformation attained during the folding process?

The first source of perplexity is the difference, or the lack thereof, between the sequences of knotted and unknotted proteins. How is the topology encoded in the sequence? Are there special and/or specific 'words' or 'phrases' that determine the qualitative difference between a protein with simple, topologically trivial fold and one with a knot or a lasso? In the latter case, the presence of cysteine residues clearly correlates with the peculiar self-entangled state of the molecule; however, one might identify similar properties also in absence of such an evident marker: an example is provided by those proteins having sub-chain loops which entangle one with another, as highlighted by Baiesi *et al* [33, 34]. Here, loops are not physically closed by disulphide bonds, however a geometrical criterion of proximity is sufficient to spot a quasi-continuum of entanglement degrees.

Evidence collected so far does not point towards uncommon, specific, or otherwise extraordinary elements that take part in making the native conformation of a protein knotted. Indeed, sequence analyses have been carried out, which have not been capable of identifying unusual traits in the primary sequence of self-entangled proteins. In some cases, short stretches of the sequence of a protein have been pinpointed as particularly relevant for the formation of the entanglement [30, 35]; however, the sequences of these *knot-promoting loops* did not show any statistically relevant deviation from the average features.

It is thus reasonable to expect proteins to be perfectly capable of encoding a complex topology in their sequence. If

knotted proteins are 'nothing special', though, why are they so rare? And—are they rare for real? A landmark work by Lua and Grosberg [36] indeed showed that knotted proteins are substantially less frequent than one would expect in a collection of collapsed polymers with comparable lengths and structural statistics. This discrepancy is suggestive of evolutionary mechanisms acting either at the level of sequence mutation, thereby preventing or largely reducing the occurrence of knotted mutants, or at the level of selection, sieving out self-entangled conformations as adverse. This last aspect further bifurcates the possibilities, in that the detrimental impact of a knot in a protein's backbone might manifest itself as a barrier to efficient and proper folding, or as a distorted, functionally ineffective conformation—or both.

Given this state of things, one is led to wonder what makes existing protein knots so special. In fact, not only *there are* knotted proteins, whose folding process and biological activity does not seem (too) hampered by the complex topology, rather these instances seem to be scrupulously conserved throughout evolution. Potestio *et al* [35], for example, employed phylogenetic analysis methods to show that in the family tree of N-succinyl-L-ornithine transcarbamylase (SOTCase) all instances belonging to a given branch are knotted; this was the case, for some of them, in spite of a smaller sequence identity with other knotted proteins than with unknotted members of the family on other, unknotted ramifications. In general, conservation of topological motifs in proteins has been observed across a wide range of organisms, with a typically low sequence similarity [11, 30].

If a knotted protein survives the sieve of natural selection, then, it seems to conserve its distinctive topological trait throughout its offspring. This might be due to the small phase space offered by structural and functional constraints to untie the knot by means of a relatively small mutation. However, the survival of these few instances and their obstinate perseverance in the kingdom of Life hints at a correlation between the self-entanglement and a substantial functional advantage [30]. What is, then, the positive impact of a knotted backbone on the biological activity a protein has to carry out?

To attempt an answer to this question one can first look at the protein families where topologically nontrivial structures are more frequent. The majority of these proteins are enzymes which catalyse chemical reactions [37]; for example, trefoil knots are found in carbonic anhydrases, S-adenosylmethionine synthetases, methyltransferases, and N-succinylornithine transcarbamylases, as well as in metal-binding protein essential Rds3p and among sodium/calcium exchanger membrane proteins. The figure-of-eight protein knot is present in ketol-acid reductoisomerases and in the chromophore-binding domain of a red/far-red photoreceptor phytochrome from bacterium D. radiodurans. Ubiquitin carboxyl-terminal hydrolases (UCHs) features a five-crossing Gordian knot, while the Stevedore knot, the most complex topology discovered in proteins so far, is tied on DehI, a α-haloacid dehalogenase.

Why these proteins require a knotted backbone for their catalytic activity is still object of active research. Among the most popular hypotheses one counts the enhanced structural/mechanical stability and resistance to denaturation that the entanglement endows the molecule with. In some specific cases [37], the particular arrangement of the polypeptide chain determined by the knot has been deemed responsible for the precise structure of the molecule's active site as well as its chemical targets and activity. Nonetheless, smoking-gun-clear evidence of the necessity of a tangled backbone to achieve these conformations, instead of a different but unknotted one, is still largely missing [10, 37].

Finally, we question ourselves about the mechanisms put in place by nature to fold these peculiar proteins. As earlier pointed out, the most remarkable difference between the knots found in viral DNA strands and the ones in proteins is reprodu-cibility: always the same knot, always in the same place. By analogy with human-sized cords, this is the difference between a messy tangle of earphones wire and the precise, elegant knot on a tie or a pair of handmade shoes—the pivotal role missing in the first case being played by a skilled, well-intentioned human.

As iconography and clichés depict ties to be knotted on clumsy gentlemen's necks by the hands of their patient and supportive partners, knotted proteins can have their own version of a helpful companion, too, namely chaperones. In many cases, in fact, self-entangled proteins fold with the assistance of molecular machineries such as GroEL/GroES [9], which protect the freshly synthesised chain from the environment and facilitates the process of folding. If, on the one hand, the presence of chaperones represents a considerable aid to the folding process, on the other hand it should be noted that, in many cases, knotted proteins do not require them to spontaneously attain the self-entangled conformation. Albeit inefficiently, though, knotted proteins are sufficiently emancipated.

Unfolding and refolding experiments have largely supported this observation [8, 38–47], thereby highlighting the fact that not only can a knotted topology be fully embedded in an otherwise unspecific amino acid sequence, rather it can also be attained by the sole means that this sequence commonly relies on—intra-protein and protein-solvent interactions. It is reasonable to expect that the deeper/more complex the knot, the harder it will be for the chain to fold it by itself; in fact, the gain in efficiency provided by the presence of chaperones can be 20-fold [9]. Nonetheless, relatively small knotted proteins can snap into a tied native structure without helping hands. Recent work [48–51] has highlighted the potential impact and importance of the protein synthesis itself on knotting. In fact, simulations [48–50] have shown that the chain folding and the knot formation can contextually occur with the polymerisation the chain in the ribosome, a process dubbed *cotranslational folding* [52]. The chain, in fact, transolcates through a pore, so that the newly synthesised stretch of the sequence has a lower conformational entropy than it would have if the remainder of the protein were present. This, as well as other factors, can favour the self-entanglement of the protein.

Differences with respect to the folding pathways followed by unknotted proteins, however, are present. In general, folding can proceed through various routes, characterised by several milestones and intermediate states among which the molecule can interconvert before landing onto the native, biologically functional conformation. The existence of a

'topological bottleneck', on the contrary, forces self-entangled proteins to avoid all those intermediate steps which are incompatible with the topology of the native state and might prevent the achievement of the latter [44]. Backtracking is required in these cases, that is, a partial unfolding necessary to solve the undesired tangle and attempt anew to obtain the desired one. The formation of the native topology is thus identified in the most relevant rate-limiting step in the folding process of self-entangled proteins. This fact has been most clearly highlighted by the experiments carried out in the Jackson group by means of *de novo* folding [8].

When available, chaperones help knotted proteins to fold correctly, reaching the aforementioned $20\times$ factor in speedup. Other self-entangled polypeptides, however, do not take advantage of this kind of aid, and have by default to manage knotting by themselves. This is especially the case of small knotted proteins, whose folding pathways have been thoroughly studied. Backtracking is avoided in these mainly through a fairly polarised free energy landscape, which resembles a funnel-shaped highway rather than a mountain pass [44, 53–55]. Many relevant cases have shown folding mechanisms involving the formation of loops through which a terminus penetrates, thereby establishing the native topology and the native structure almost in a single step. The piercing terminus can happen to be 'straight' or bent in a hairpin-like conformation; in this second case, the knotting event takes place first through the formation of a *slipknot* [7, 30, 37, 56–61], which subsequently opens up into a regular knot. Larger proteins, as anticipated, can feature more complex folding pathways, involving intermediate steps [42, 46, 47, 62, 63] and inspiring novel schemes for the description of the knotting process [64].

The experimental characterisation of knotted proteins and other types of self-entangled polypeptides has achieved remarkable successes. These have required a broad spectrum of techniques, such as mechanical stretching of proteins by means of optical tweezers and AFM, *in vitro* translation-transcription, recombinant and cell-free protein expression, SAXS, fluorescence etc. Several different tools had to be put in place to synthesise wild type and mutant knotted proteins, create new ones, determine and characterise their structure, investigate their response to mechanical stresses, and above everything pinpoint their topological state. Important pieces of knowledge have been obtained by these means.

However, it appears evident that the experimental tools alone are not sufficient. This is more and more true in every facet of science, and the investigation of self-entangled proteins makes no exception. The insight contributed by *in silico* studies—be that through accurate and realistic all-atom models or simplified, effective coarse-grained representations—is of paramount importance to comprehend the fundamental properties and mechanisms that underlie the formation of protein knots. The flexibility given by this instrument in constructing ad hoc models endowed with specific features and studying their properties is an invaluable help in understanding how does a protein tie a knot, what biological role such an entanglement might play, and which inescapable features its sequence must have in order to entail the capability of doing all this.

The objective of the following chapters is to present the reader a list of the most popular, effective, and efficient techniques that have been employed in this endeavour so far. The theoretical basis underlying all computational methods and models is illustrated with the aim of being clear and informative rather than detailed and comprehensive, in order to provide an agile resource to refer to when in search of the appropriate tool to tackle a given problem involving self-entangled proteins. The vastness of this yet rather young field of research, combined with the rapidity with which it evolves, makes it difficult to imagine that this resource will stand the proof of time, the latest edge-cutting development surely being only a few months away from the time of this writing. However, the construction lies and relies on a solid bedrock, and it is the intent and hope of the writers to present the readers with a sufficiently good guide of the old town so as to allow them to confidently explore the modern neighbourhoods of the city—and why not, motivate them to build a new block.

3. Classification of protein entanglements

Knots in proteins represent an example of *physical knots*, topological entanglements of linear objects, characterized by physical properties such as thickness, friction, or flexibility. These properties distinguish such objects from the immaterial curves considered by the formal knot theory. Nonetheless, the study of physical knots mutuates several concepts from knot theory, crucial to define and classify the topological states of proteins [65–67]. In the present section we report these theoretical concepts, costituting the necessary background for the study of entanglements in proteins.

According to theory, knots can be rigorously defined only as a property of closed curves [1–4]. However, as mentioned before, objects such as proteins, whose geometry can be represented by an open curve, are found in stable, deeply entangled states. The entangled configurations of open curves can be traced back to a well-defined knot if a *closure* operation is performed, namely if its two ends are artificially connected by extending the curve. The definition of the closure represents therefore a crucial step in the detection and classification of knotted polymers and proteins.

Recently, Turaev has proposed the mathematical definition of *knotoids* [68], which generalizes the concept of knots including open curves entanglements. As such, these topological objects are well-suited to characterize the topology of proteins, without requiring the definition of a closure [69–71]. A further approach to the classification of topology in open curves is adopted in [72], where protein structures are analized as *virtual* knots. Both these classification methods build on a statistical treatment of *all possible* planar projections of three-dimensional curves. While knotoids do not require closure by definition, in virtual knots a so-called *virtual* closure is performed on each planar projections, keeping trace of the possible topological ambiguities introduced while closing the curve. Despite the existence of these novel, more general concepts, ordinary knot theory is used in most of the literature

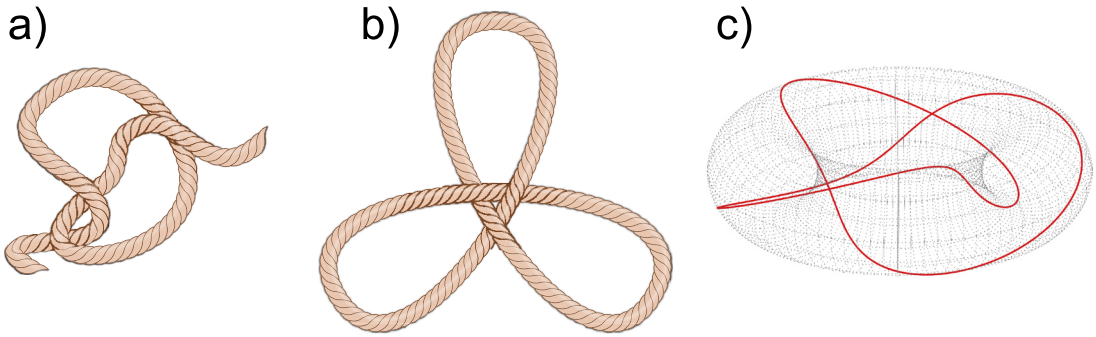

Figure 1. Pictorial representation of a knotted open string (a), of a knotted closed string (b), and of a knotted closed curve (in red) in the three dimensional space. To render the three-dimensionality of the curve it is embedded on the surface of a torus (gray dotted mesh).

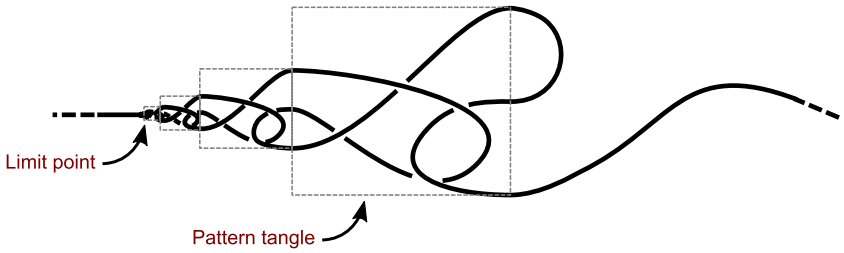

Figure 2. A wild knot, featuring a pattern tangle (indicated by the dashed boxes) that is repeated and rescaled infinite times, towards the limit where the tangle reduces to a point. The repeated tangle is trivial, but the whole curve is not (see e.g. [2, 76]). To represent the three dimensional character of the curve we have used the knot diagram notation, explained in section 3.1.2.

about entangled proteins, therefore we shall rather focus on the knot-classification than on knotoids or virtual knots.

We first consider immaterial closed curves, introducing the knot theory insights required to classify their topological state, in section 3.1. Then, in section 3.2, the main numerical methods used to identify knots are reviewed. After that, in section 3.3, closure techniques are summarized. The definition and detection of entangled states other than knots is addressed in section 3.4, while the existing databases that gather all known self-entangled proteins are discussed in section 3.5.

3.1. Knots in closed curves

3.1.1. Knot definition. As mentioned before, knot theory defines the topological state of closed curves. By common experience we know that, while a knot on an open string can be undone by proper manipulation, this is impossible if the ends of the string are attached to form a loop (see figures 1(a) and (b)). Any possible deformation in space of this closed loop preserves its topological state. This suggests that the knotted state of a closed curve can be operationally defined by means of spatial deformations.

Let us mathematically represent the knotted closed string of figure 1(b) as a closed curve embedded in the three-dimensional euclidean space $X \subset \mathbb{R}^3$, as shown in figure 1(c). In mathematics, all the possible continuous deformations of X in space are called *homotopies* (see e.g. [2] for further details). To define the topological state of X we need to restrict the class of transformations to those homotopies that prevent the curve from passing through itself, named *isotopies*. This is still not sufficient, since X has no thickness, any entanglement hosted by the curve can be continuously reduced to a single point, transforming the curve into an un-knotted loop, also called *trivial knot*. This leads to the definition of an *ambient isotopy* (AI), which deforms X through the continuous transformation of its embedding space, in this case $\mathbb{R}^3$. The action of an AI on X does not change its topological state, thus the AIs are the mathematical analogous of the string manipulation mentioned before.

The topological state of a curve is defined through AIs as an equivalence class, named *knot type*. Two curves that can be transformed into each other by AIs belong to the same knot type, namely they are topologically equivalent. This definition is related to the concept of knot *complement*, namely $K = S - N(X)$, where S is a compact region embedding the curve and $N(X)$ is a tubular region that indicates the neighboring space of X. Indeed, K defines the knot type, as equivalent knots have homeomorphic complements [73]. For example, all the curves that can be transformed into a circle belong to the trivial knot type, and are said to be unknotted. There exist infinitely many possible knot types, the known ones being classified in catalogs [74, 75].

From this general definition of knots, we shall restrict to those knot types that can describe a physical objects, which are named *tame* knots. Tame knots, as demonstrated in [2], can always be represented by closed, non-intersecting and finite polygonal curves, called *polygonal knots*. An example of a knot type excluded from this definition is the so-called *wild* knot, shown in figure 2, which features an infinitely recursive character and cannot represent a physical knot.

Moreover, since most of the techniques reviewed in the following have been conceived for the topological analysis of polymers, it will be sometimes useful to directly refer to

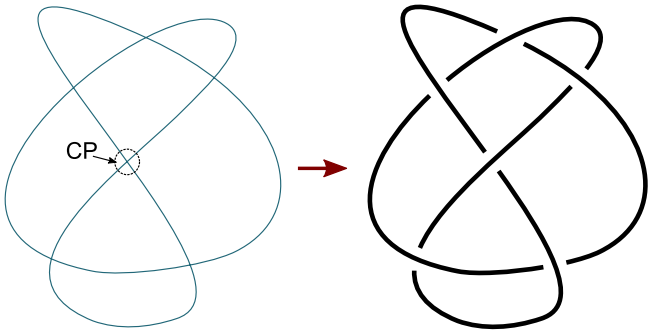

Figure 3. Representation of a two-dimensional, regular projection of a 6_2 knot (left side) via knot diagram (right side).

a bead-chain structure rather than to an abstract polygonal curve.

3.1.2. Knot diagrams. In the process of topological classification, a schematic visualization of knots, capable to underline differences between knot types, provides a significant help. A natural, compact way to visualize a curve embedded in $\mathbb{R}^3$ is to project it on a plane. The mapping of a three dimensional curve on a plane naturally generates singular points, in which more than one point of the curve is mapped. The relevant projection for knot identification is the so-called *regular* projection, in which the singularities are *ordinary double points*, also called *crossing points* (CP). In CPs two strands of the curve overlap with different tangents, as shown in figure 3. It can be demonstrated that physical knots can always be represented through a regular projection [2]. To preserve the topological information in the two-dimensional projection it is also necessary to assign the information of relative depth of the crossing branches in each double point. This is normally done as shown in figure 3, by interrupting the branch that underpasses the CP. The resulting representation is named *knot diagram*.

Knot diagrams can be extremely complex, in particular if they represent a collapsed globular structure, such as that of proteins. However, a complex knot diagram can be transformed to reach a simpler representation, without changing its topological state. Intuitively, the transformations that allow one to reduce the complexity of a knot diagram are planar projections of AIs. These consist in all planar deformations of the diagram that do not affect the CPs, plus the three Reidemeister moves (RM) [77]. The latter ones act on the number of CPs of a diagram without affecting its topology (a pictorial representation of RMs is reported in figure 4). These transformations generate the class $\mathcal{D}(X)$ of topologically equivalent diagrams of X. One can thus use Reidemeister moves to reduce the number n of double points in a diagram down to the minimum number $\tilde{n} = \min_{\mathcal{D}(X)} n$, named *crossing number*, obtaining the so-called *minimal diagram*. We underline that the minimal diagram is generally the simplest representation of a knot, but sometimes it hides specific knot properties (for further insights on this we refer to specific knot theory books, such as [2, 65]).

3.1.3. Knot classification. The crossing number $\tilde{n}$ of a knot diagram is a *topological invariant*, namely a property that depends only on the topological state of the curve and not on a specific three-dimensional realization or planar projection. More precisely, $\tilde{n}$ is a weak invariant, as different knots can have the same crossing number. The known knots are classified using the notation $\tilde{n}_i$, in which i indexes different knots with the same crossing number. The trivial knot is classified as 0_1, while the simplest known knot, namely the trefoil knot, is referred to as 3_1. According to these conventions known knots are classified in tables such as the one reported in figure 5(a), including all knot types found in protein structures. A single curve can also form *composite* knots, that contain more than one knot connected as in figure 5(b). The components of a composite knot are called *factor* knots, while *prime* knots cannot be decomposed in two or more non-trivial factors (the trivial knot is always a factor of another knot). Knot tables, as that reported in figure 5(a) usually include only prime knots.

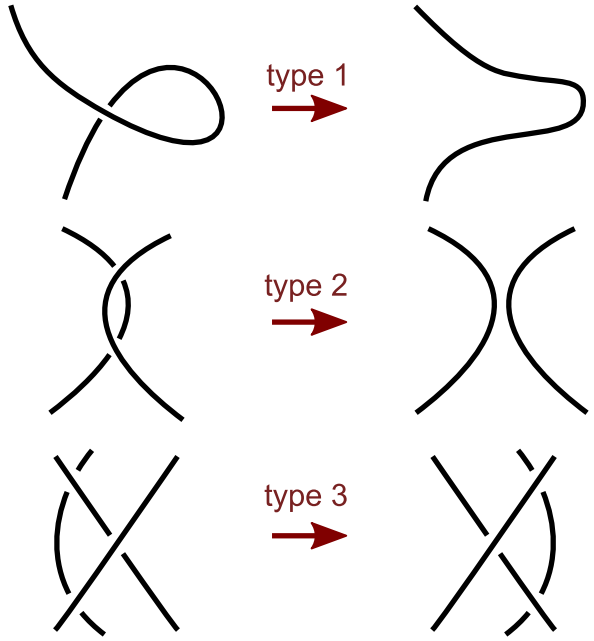

Figure 4. The three Reidemeister moves.

A further topological specification is the *chirality*, or handedness, of a knot. A knot is chiral if there is no AI that can map it to its mirror image. The two mirror images of a chiral knot are named enantiomers. The simplest example of chiral knot is the 3_1, while the 4_1 is an example of a-chiral knot.

3.2. Identification of knots

The identification of a knot in a closed chain consists in determining the equivalence of its diagram, or of its factors diagrams, to a minimal tabulated diagram. For simple knots, and projections, this can be achieved by algorithms implementing geometrical deformations and RMs. However, as the complexity of the knot diagrams under consideration increases, this strategy becomes impractical and more efficient techniques are necessary.

3.2.1. Topological invariants. For decades mathematicians have searched for topological invariants which can distinguish between different knots, and at the same time be efficiently

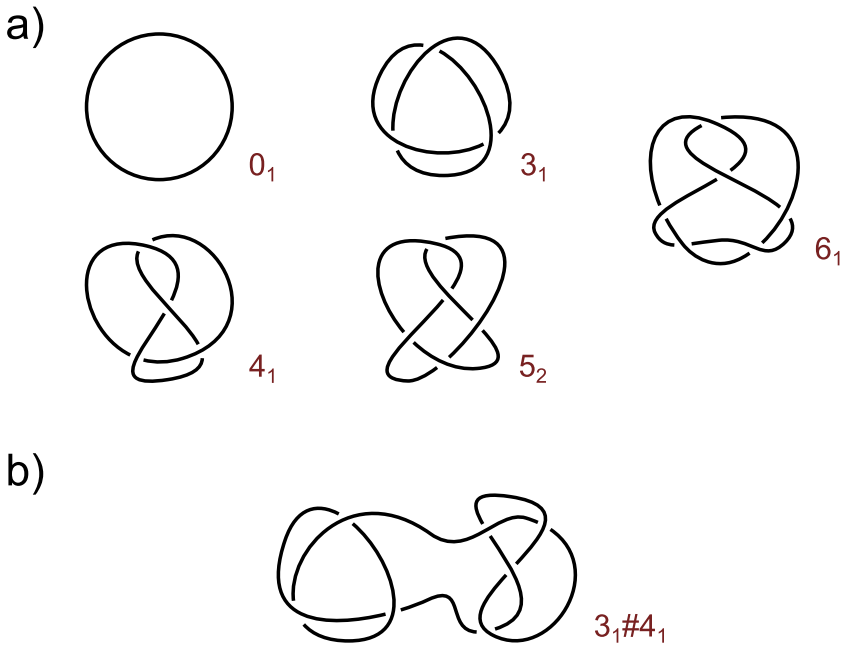

Figure 5. (a) Minimal diagrams of the non-trivial knot types found in protein structures. (b) Example diagram of a composite knot, formed by connecting a 3_1 with a 4_1 diagram.

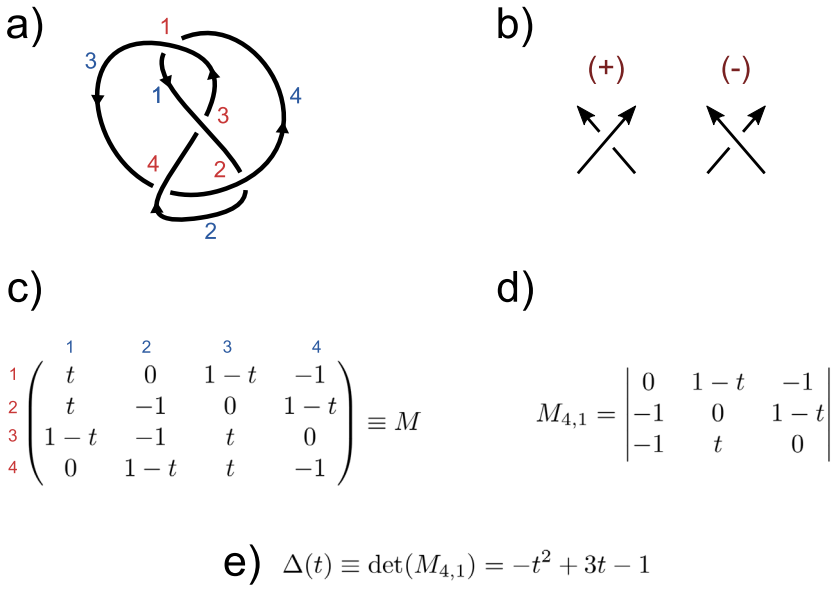

Figure 6. Alexander polynomial calculation for a 4_1 knot: first (a) an orientation is assigned to the diagram, and the CP and arcs are numbered (with red and blue labels, respectively). By using the convention for the sign of the CP (b), the matrix M is constructed (c). An arbitrary minor $M_{i,j}$ is chosen (d), the determinant of which gives the Alexander polynomial (e).

computed also for very complex embeddings of curves. As mentioned before, the crossing number, on which the knot classification is built, is the simplest topological invariant. However its calculation requires by definition the determination of the minimal knot diagram, and it is therefore a complex operation. Nowadays many invariants have been discovered, ranging from Gauss winding number [78], to the knot group [79, 80].

The most widely used invariants for identifying knots in polymers are the polynomial invariants, that are constructed defining combinatorial rules that apply to any knot diagram, not requiring the construction of the minimal diagram. These invariants include the Alexander polynomial [78], Jones polynomial [81], and HOMFLY polynomial [82] (from the names of its co-discoverers Hoste, Ocneanu, Millett, Freyd, Lickorish and Yetter). A detailed review of these invariants is beyond the purposes of present work, the interested reader can refer to [2, 80]. As an example on how these polynomials are constructed, we report the procedure for the calculation of the Alexander polynomial $\Delta(t)$ [66, 83]:

(i) First, an arbitrary orientation is assigned to the diagram. This orientation defines the sign of the crossings according to the convention represented in figure 6(b).

(ii) Following the orientation, a progressive number $x = 1 \ldots n$ is assigned to the crossing points of the diagram, starting from an arbitrary point.
(iii) Also the n arcs, the diagram branches separated by underpasses, are numbered progressively following the orientation.
(iv) An $n \times n$ matrix M is defined. Each row is associated to a crossing point, while each column is associated to an arc. The nonzero elements of the xth row correspond to the three arcs that meet in the xth crossing point. Let us define that in x the ith arc overpasses the two consecutive j and k arcs. The xth row elements are then defined according to the following rules:

- if the crossing is positive: $M(x,i) = 1 - t$, $M(x,j) = -1$ and $M(x,k) = t$,
- if the crossing is negative: $M(x,i) = 1 - t$, $M(x,j) = t$ and $M(x,k) = -1$,
- if $i = k$, or $i = j$, $M(x,j) = 1$ and $M(x,k) = -1$.

(v) An arbitrary minor of M is selected to obtain a $(n-1) \times (n-1)$ matrix, the so-called Alexander matrix.
(vi) The determinant of the Alexander matrix is computed to obtain the Alexander polynomial $\tilde{\Delta}(t)$.

The $\tilde{\Delta}(t)$ obtained in this way is not a topological invariant, but all the $\tilde{\Delta}(t)$'s corresponding to the same knot (computed from different diagrams) differ by a factor $\pm t^m$, with $m \in \mathbb{Z}$. Therefore one can obtain the so-called irreducible Alexander polynomial $\Delta(t) = \pm t^m \tilde{\Delta}(t)$, where the sign and $m \in \mathbb{Z}$ are chosen so that the lowest order term of $\Delta(t)$ is a positive constant. The irreducible Alexander polynomial is an actual topological invariant. In the following, as it is customary in the literature, we will drop the 'irreducible' and simply indicate it as Alexander polynomial. In figure 6 an example of the algorithm application is displayed.

It is clear from this example that the calculation of polynomial invariants depends on the number n of crossing points associated to the particular projection considered, thus making it desirable to simplify the knot diagram before the calculation, in order to operate on a smaller—if not the smallest possible—number of crossings. It can be shown that Jones polynomial calculation time scales as 2^n, while Alexander polynomial calculation time scales as $(n-1)^3$ [84]. For its scaling properties, and for its relatively simple implementation, the Alexander polynomial is largely used in the literature of protein topology (see e.g. [24, 85–87]). However, since Alexander polynomial cannot distinguish between two enantiomers, a more general invariant, such as the HOMFLY polynomial, has to be computed when the chirality information is required [87]. We also stress that polynomial invariants cannot always distinguish among different knots, for example the Kinoshita–Terasaka knot, with crossing number $\tilde{n} = 11$, has $\Delta(t) = 1$, the same as the trivial knot [2] (while HOMFLY polynomial can differentiate among the two). However, of all knots with $\tilde{n} < 11$ only six cannot be differentiated by the Alexander polynomial. Thus $\Delta(t)$ represents a valuable tool to discriminate between relatively simple knots, which are more likely to occur in nature.

3.2.2. Curve smoothing. In polymer and bio-polymer physics, knot theory is usually applied to analyze globular configurations, as for example in the study of proteins. In such cases the polymer is represented by a dense polygonal curve, whose projections may have a large number of CPs n, even if the topology of the curve is trivial. This makes the calculation of the Alexander polynomial, or of other polynomial invariants, extremely demanding. To overcome this limitation the analyzed curve should be modified to reduce n, without changing its topology. In other words one has to implement an algorithm that mimics AIs.

An effective algorithm that performs this curve smoothing, or rectification, was proposed Koniaris and Muthukumar in [84]. In this method the polygonal curve is modified by progressively analyzing the triplets of consecutive vertexes. If no curve segment crosses the triangle formed by the considered triplet, then its central vertex is removed, otherwise the algorithm moves to the next triplet. The computation time of this procedure scales as $M \log(M)$, where M is the initial number of vertexes of the polygonal. This calculation allows to minimize the crossings n in any possible projection of the curve, significantly accelerating the computation of polynomial invariants. Algorithms based on the same principle of curve smoothing have lately been proposed in [27, 88]. In many cases curve rectification allows also the visual detection of knots, so it can be considered itself as a knot finding technique [27]. Currently, curve smoothing algorithms are included in all the common procedures for finding knots in proteins [85, 87, 89, 90].

3.2.3. Knot localization. A crucial property of physical knots is their size, namely the length of the shortest curve portion hosting the entanglement. In knotted polymers the size of the knots can deeply modify the equilibrium and dynamics properties of the polymeric chain (see e.g. [91, 92]). The qualification of knot size and localization on the backbone (or depth) requires the topological analysis of all possible portions of the considered chain, or *arcs*, that is all the combinations of consecutive vertices of the polygonal curve [27, 30, 56, 93]. Once an arc is selected, a closure operation is performed, and then its topology is assessed. The results of such an operation can be collected by means of *knot matrices*, or fingerprints [56, 94, 95], which allow to effectively visualize the size and location of knots [96]. As shown in figure 7, each entry of the knot matrix indicates (via color or shading) the topology of an arc, whose end vertices are indicated by the row and column indexes. The smallest portion of the chain that embeds a specific topology is the so-called *knot core*, while the two chain segments excluded by the knot core are called *knot tails*. Recently, an alternative approach was proposed to display the topology of a knotted closed chain and all its subchains [97], named disk matrix. Based on a longitude-latitude map, the disk matrices are meant to reproduce the periodicity of circularized chain, but can provide also useful insights on the topology of open chains such as proteins.

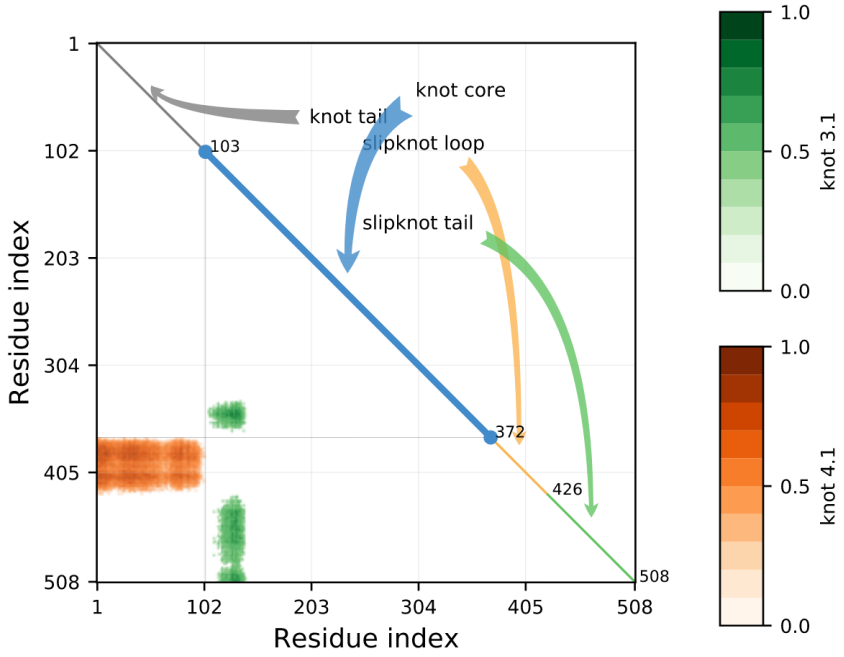

Figure 7. Knot matrix representation of carnitine transporter chain A (PDB id: 2wswA), as displayed by KnotProt server (https://knotprot.cent.uw.edu.pl/, see section 3.5 for more details). 2wswA features a slipknotted topology, that contain a 3_1 (the area shaded with green shades) and a 4_1 (orange shades) knot. The intensity of the color indicates the probability of obtaining a knot with stochastic closure of the relative sub-chain (see the legends on the right). Along the diagonal of the matrix the components of the 4_1 slipknotted topology are indicated, the knot core and tail, and two other components characterizing slipknots, the *slipknot loop*, that winds back undoing the knotted configuration, and the *slipknot tail*, formed by those residues that follow the loop, completing the unknotted topology.

3.3. Closure methods

As mentioned multiple times, only closed curves have a well-defined topological state. A *closure* operation is thus necessary to apply knot theory to polypeptide chains and polymers in general. With the word closure we mean an artificial extension of the open polygonal curve representing the polymer chain under study, so that its ends are connected to form a closed curve. Since there are infinite ways of performing this operation, closure can be a source of ambiguities in defining the topological state of a curve. For example, the closing extension can interfere with the part of the curve representing the actual polymer chain. In general one wants to avoid this interference, aiming at a closure arc that joins the two polymer ends without intersecting any of the existing features. In the following we review the main closure techniques that have been proposed and applied in the literature of polymer knots.

Direct bridging: The polygonal is closed by connecting its ends with a straight line (as shown in figure 8(a)). It has the easiest implementation but, in particular when globular polymers or proteins are considered, the closure segment is likely to interfere with the actual chain topology [36, 98].

External closure: Each end of the curve is prolonged outwards and connected at large distance, by a polygonal or a circular arc (as shown in figure 8(b)). The idea of a closure far outside from the volume occupied by the open polygonal is introduced in [26], while the virtual closing loop has been formalised in different ways. Each end can be prolonged along the direction of its vector distance from the centre of mass of the polygonal $\mathbf{r}_{\rm com}$, to reach the surface of a sphere of radius R, much larger than the size of the curve. The two resulting ends are then joined by a polygonal or circular arc of radius R [28, 36, 91]. Another possibility is to attach a large planar loop, forming an almost complete circle of radius R, to the two ends [99]. In [56] the outward extension of the ends is performed by small incremental steps, choosing at each step the direction that maximizes the distance between the new virtual end point and the vertexes of the polygonal. In all these approaches the virtual extension of the curve can interfere with its topological state, in particular when the ends are enclosed in the volume occupied by the polygonal. However, external closure is particularly convenient in the framework of protein knots, as the termini are typically located on the surface of the polymer globule. This trend is taken into account in [35], where a selective external closure is performed on those proteins that have both termini exposed on the surface of their native structure. When this condition is satisfied for both protein ends, then external closure can be performed without ambiguities.

Stochastic closure: This approach is based on a 'statistical definition of knottedness', by which an ensemble of closures is defined, and then the most statistically relevant state is adopted as the curve topology. Two variants of this scheme have been defined. In [100] a set of N points $\{\mathbf{r}_i\}$ is generated, uniformly distributed on a sphere that encloses the whole curve. Typically the radius of this sphere is few times larger than the smallest sphere enclosing the polygonal. N different closures of the curve are then defined by connecting each $\mathbf{r}_i$ to both ends. This results in a spectrum

of topologies, in which the most populated state is chosen as the assigned topology. In those cases in which no clear dominant state emerges, no knot type is assigned. The randomized direction of the chain extensions implies that interference with the polymer globule can occur in a relevant number of cases. This is a drawback with respect to those methods that choose this direction in order to avoid interference (as e.g. the minimally interefering closure, presented in the following). However, if enough statistics is collected, the spectrum of topology should solve most of the ambiguities.

In the second variant, proposed by Mansfield [24], the procedure is the same, with the difference that N pairs of $\mathbf{r}_i$'s are generated on the enclosing sphere. Each point of the pair is connected by a straight segment to one of the two curve ends, and then they are joined with an arc on the enclosing sphere [36]. Both these strategies can provide detailed topological information, but they also demand more computational time with respect to the other, 'deterministic' schemes. The two variants of external closure are represented in figure 8(c).

Minimally interfering closure: In this scheme the closure is performed trying to minimize the distance travelled by the closing segment across the volume occupied by the polygonal [93], as this constitutes the main source of arbitrariness on the topological state. To this purpose, the convex hull enclosing the polygonal is first computed. The closure is then chosen depending on the values of d_{out}, the sum of the distances between the termini and the closest face of the convex hull, and d_{in}, the distance between the termini. If $d_{\text{in}} \leqslant d_{\text{out}}$ the closure is performed via direct bridging, if instead $d_{\text{in}} > d_{\text{out}}$ each end is prolonged outwards, perpendicularly to the closest face of the convex hull, and then the resulting extensions are connected at large distance by a circular arc (as displayed in figure 8(d)). The resulting technique avoids the drawbacks of the first two schemes, avoiding the large computational requirement of stochastic closure.

These are the most common closure strategies proposed to determine the topological state of an open polygonal, mainly used for defining the topology of knotted proteins [24, 36] and for determining the size of knots in polymers [91, 93, 99, 101]. Reviews and comparisons of different closure methods are available in the literature (see e.g. [36, 102]). In summary one can conclude that in most of the cases external closure is enough to define the protein topology, but stochastic closure methods are more precise and reliable, and they should be used to solve the most ambiguous cases. We also underline that the reliability of a closure process can depend on its combination with curve smoothening algorithms.

3.4. Further topological states

As mentioned in section 2, native conformations of proteins can feature other kinds of entangled states that, while not satisfying the criteria defined in the previous pages, can be definitely included in the family of topologically complex structures.

We have already introduced *slipknots*, topological states in which a sub-section of the curve is found to be knotted, while the full-length curve is not. These motifs are commonly associated to protein knots and have been considered in different surveys and reviews on protein topology [29, 30, 56, 104]. A slipknot state can be detected by means of the knot matrix, which provides information about the topology of all possible protein sub-chains (see e.g. figure 7). Slipknots play also a relevant role in the study of knotted protein folding, as a slipknotted structure could represent a populated intermediate in the formation of the native topology [57, 105].

Up to now we have considered the topological state featured by a polygonal curve representing the protein backbone. In many native structures, however, residues can also form non-sequential covalent (or ion-based) bonds, typically disulphide bridges. Since these bonds provide a physical closure to a subsection of the peptidic chain, they can determine unambiguously defined knots, sometimes referred to as *covalent* knots [106, 107] or *deterministic* knots, if also non-covalent bonds are accounted for [90]. Already hypothesized in the '70s [20], these knots appear somehow less problematic than backbone knots, since their formation does not imply an intricate folding pathway.

The formation of disulphide bridges, connecting cysteine residues along the backbone, is central also to the formation of other topological motifs. A known example is that of *Cysteine knots* [104, 108], in which a covalent loop formed by two disulphide bridges is pierced by a third bridge (see figure 9). It must be stressed that cysteine knots, despite the closure provided by the disulphide bonds, do not represent actual knots, in a topological sense, as they can be untangled by continuous deformation of the chain [90]. Nonetheless, they are biologically interesting motifs as they provide exceptional stability to the protein structure [59, 108].

Another class of entangled proteins defined by the formation of cysteine bridges, is that of *complex lassos* [109], already introduced in section 2. As mentioned, in these structures a disulphide bridge seals a covalent loop, the surface of which is pierced one or more times by the protein backbone. This motif was first observed in mini-proteins named lassopeptides [110], and in Leptin [111], before it was found to be relatively common among PDB structures, appearing in about 18% of the proteins with a cysteine bridge [112]. These topological motifs are suspected to play a role in the signalling activity [113, 114] and, through control of the bridge formation, they can represent a useful testing ground for investigating topologically complex folding [115]. As complex lassos differ from standard knots, one cannot rely on knot theory to detect them and other strategies are required. In [109] a technique named minimal surface analysis is proposed to unambiguously identify lasso structures. It consists in determining the surface of minimal area spanned by the covalent loop, and then locating the intersections of the backbone with this surface. By means of this technique, different lasso types could

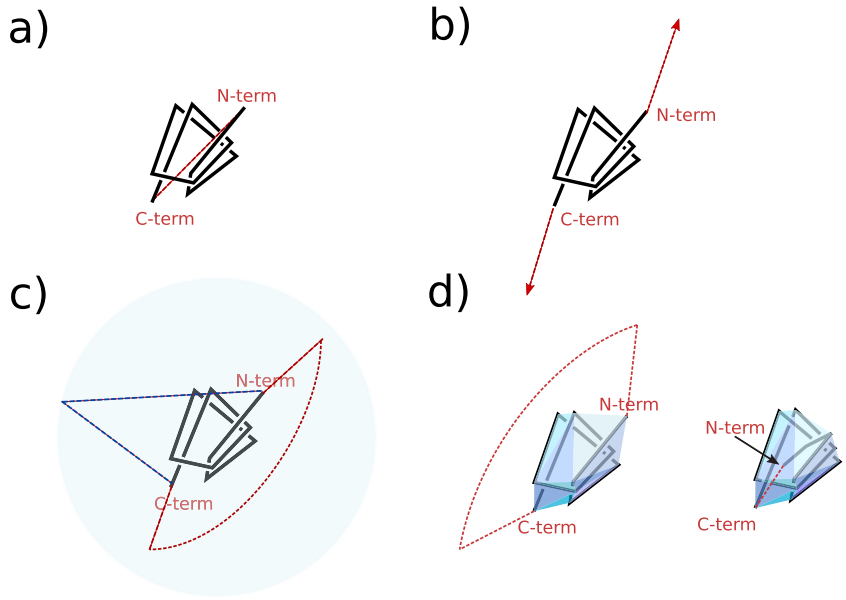

Figure 8. Representation of the closure schemes presented in the text: (a) direct bridging, (b) external closure, (c) stochastic closure, with a single or two external points, (d) minimal interfering closure, where the case $d_{\text{in}} > d_{\text{out}}$ is shown on the left, and the case $d_{\text{in}} < d_{\text{out}}$ is shown on the right.

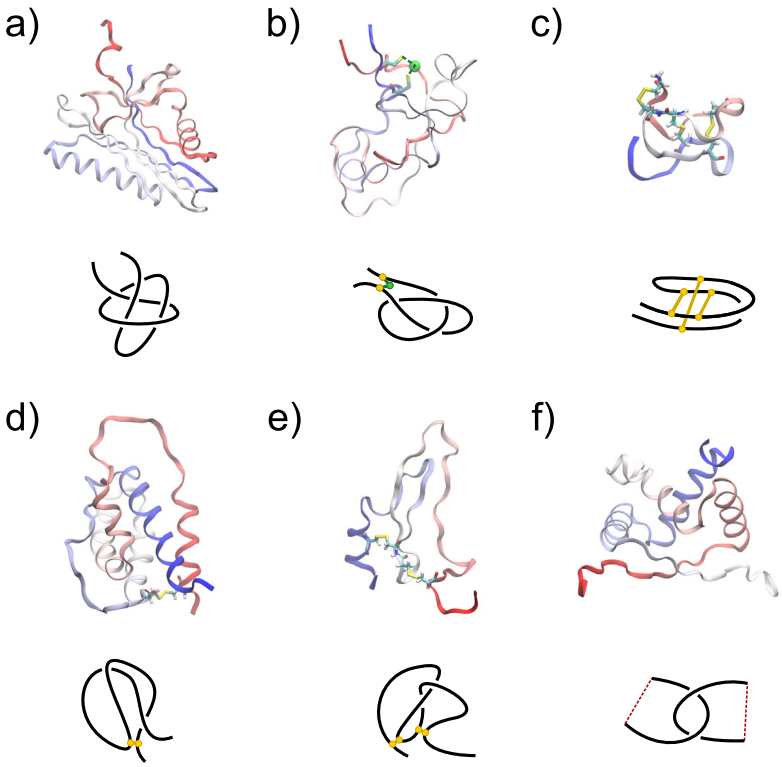

Figure 9. Self-entangled protein states other than knotted proteins. Each state is indicated by a representative structure (top) and diagram (bottom). (a) Slipknot (PDB: 2QQDc), (b) deterministic knot, (PDB: 5ZYAd) (c) cysteine knot (PDB: 2ml7), (d) complex-lasso (PDB: 1jli), (e) protein link (PDB: 2lfk), (f) linked protein dimer (PDB: 1arr). Structures are depicted with VMD [103], using a ribbon-like representation for the backbone and atomistic detail for crucial bonds such as disulphides or ionic bridges. These bonds are also highlighted in the diagram representation.

be classified, depending on the number and direction of the piercings.

The minimal surface analysis has been employed to identify yet another class of protein entanglements: *protein links* [116]. Links are topological objects that generalize the concept of knots, considering the possible entanglements between more than one closed curve (see for example [2]). Links can indeed be embedded along a single protein chain, presenting two or more interlinked loops closed via disulphide or ionic bonds. Moreover, it is possible to observe links at the *quaternary* structural level, that is involving protein compounds. In this case the relevant loops can be again sealed by disulphides, but also obtained circularizing the protein chains by joining the N- and C-terminals via the closure techniques described in 3.3. Finally, links can be present at the 'macromolecular' level, that is in multi-component structures that form intertwined complexes, such as viral capsidis [117] or protein catenanes [118]. As mentioned earlier, the identification of protein links can exploit minimal surface analysis, but also topological invariants, such as the HOMFLY polynomial, can be used [119]. Another descriptor that can be employed to assess the linking between polypeptides is the Gauss linking number [120]. This quantity is defined as a double line integral along two closed curves, that results in an integer number, indicating the linking state of the two curves. The Gauss linking number has been employed in [33] to identify linked proteins in domain-swapped dimers and, very recently, to assess the presence of evolutionary patterns in self-entangled proteins [51] and to correlate folding rates to the global topology of proteins [121]. An alternative, practical method to identify linking among protein pairs is that of using an MD model of the polypeptide chains and simulate the mechanical stretching of the proteins, as shown in [122].

3.5. Databases of self-entangled proteins

Since the discovery of the first knotted protein in 1994, the accuracy of structural biology has substantially improved and, together with it, the potentialities of bio-informatics tools for the detection of protein entanglements. As a consequence, surveys over the multitude of structures deposited in the RCSB Protein Data Bank were performed (see e.g. [27, 30, 32, 36, 56], just to name a few), to attain a comprehensive classification of protein topologies. Nowadays, large variety of self-entanglements have been found, spacing from simple knots to links, passing by complex lassos, and the remarkable amount of information about protein topology has been collected in a set of databases, each dedicated to a specific class of entanglements. These databases, together with the software for the analysis of user-provided structures, are typically hosted on online servers, which make the topological data available to the public.

The first public database on knotted proteins, the KNOTS server (http://knots.mit.edu), was released in [85]. It consisted in a tool for the topological analysis of protein structures, coupling external closure, KMT reduction and Alexander polynomial calculation, to provide the type and size of eventual knots. The server maintained also an up-to-date collection of known protein knots. Slightly later Lai *et al* released the pKNOT server, with similar purposes of KNOTS, employing the curve smoothing algorithm of Taylor [27] as a central method for both simplifying structures and detecting entanglements. pKNOT was later upgraded to a second version [86] including also the possibility of providing input sequences for analysis, by means of a structure prediction algorithm [123].

KNOTS and pKNOT are currently not updated any more, and the reference database for the protein knot community is represented by KnotProt [87]. KnotProt is a comprehensive collection of knotted and slipknotted proteins, resulting from the analysis of all the structures gathered in the PDB. On a weekly basis, the newly deposited structures are automatically analyzed, and eventually included in the database. For what concerns the methodology, KnotProt performs a stochastic closure of the protein chain, and of all its subchains, analyzing the resulting closed curves by means of Alexander and HOMFLY polynomials (the latter to identify the chirality information). The KMT algorithm is employed to simplify the polygonals and reduce the cost of computing the invariants. The topological information is stored in the form of knot matrices (see figure 7), that indicate the size and depth of knots/slipknots along the chain, but also the probabilistic character of topological states, obtained via the stochastic closure method. The database provides also an extensive set of biological information about the protein, including sequence and structure similarities with other KnotProt or PDB entries.

KnotProt has been recently upgraded to the 2.0 version [90], available online at https://knotprot.cent.uw.edu.pl/. This new version includes further analysis to provide a more complete assessment of the chain topology. Besides knots and slipknots, the database includes two other entanglement types: cysteine knots, and deterministic knots. The latter include those knots uniquely defined by a chemical closure, that is the formation of a closed loop via non-sequential bonds, either covalent (disulfides, post-translational amide bonds, or aromatic residue concatenation) or ionic bridges. The name 'deterministic' is chosen in opposition to the 'probabilistic' character of standard knots, defined via stochastic closure of the chain terminals. Moreover, KnotProt 2.0 implements the notion of 'knotoid', mentioned at the beginning of this section. This concept associates a topological state to an open curve, by means of its projection on a surface. As such, different knotoid types can be associated to a protein, depending on the direction of this projection. KnotProt employs the tools implemented in [71] to formulate a 'probabilistic' classification of knotoid types in a protein chain, accumulating a set of random projections, similarly to the procedure adopted for the stochastic closure. As in the case of the older databases, KnotProt is also designed for the analysis of user-provided structures.

Apart from knots, databases of other protein entanglements are also available to the community. For example, the data on Inibitor Cysteine Knots, a numerous family of miniproteins featuring the cysteine knot motif, is gathered in the KNOTTIN database [124] (www.dsimb.inserm.fr/KNOTTIN/), even though these structures are also included in KnotProt2.0.

The *LassoProt* database (http://lassoprot.cent.uw.edu.pl/) [112] collects all the complex-lasso structure detected to date. Analogously to KnotProt, LassoProt is built on the PDB, and it is automatically updated with the newly deposited structures. The minimal surface analysis technique proposed in [109] is the reference method for the detection of complex lassos. LassoProt includes the possibility of selecting closing bridges other than disulphides, using e.g. Amide, Ester and Thioester bonds. The database provides a great deal of information for each structure, including the geometry of the minimal surface spanned by the loops, and biological/structural information collected from the PDB.

Finally, structures featuring a linked topology are collected by LinkProt (http://linkprot.cent.uw.edu.pl), a database that gathers the linking state of structures formed by up to four chains. As mentioned earlier, links are topological objects formed by more than one closed curve, which can be constituted either by peptidic loops circularized through a non-sequential bond (as e.g. cysteine or amide bridges), or by separate peptidic chains with terminals joint via closure methods. In analogy with the terminology introduced for KnotProt, the first are named 'deterministic' links and the second 'probabilistic' links. After the closed curves are identified, the minimal surface analysis is performed, combined with the calculation of the HOMFLY polynomial. This provides all the topological information about the linking state, such as chirality and orientation. Besides deterministic and probabilistic links, Linkprot includes also a separate section with known macromolecular linked structures, such as capsidis and catenanes. As for KnotProt and LassoProt, also this database is built on the PDB, being automatically updated to feature new deposited structures. Moreover, LinkProt provides useful biological information on the analyzed entry, such as structural or sequential similarity families.

As a last remark on databases, we underline that KnotProt, LassoProt and LinkProt are all maintained by the University of Warsaw and are compatible among each other, allowing the easy combination and comparison of topological motifs in protein structures.

4. Simulation of entangled protein dynamics

In the present section we review the computational approaches devised and employed to investigate self-entangled protein dynamics. As for many other realms in biophysics, molecular simulations represent a tool of crucial importance to reach a better understanding of the biological phenomena that involve proteins. The sub-molecular resolution and the fine parameter control available with computational models offer a chance to observe aspects of the protein dynamics that are inaccessible to the available experimental techniques, such as the precise folding pathway, or the dynamics of stretching of a polypeptide. In this light, it is easy to comprehend the crucial role that molecular simulations play in the study of topologically complex proteins, being able e.g. to unveil how a simple sequence of amino-acid can encode the ability to reproducibly and efficiently form complex self-entanglements, or to provide a detailed framework for the interpretation of spectroscopy experiments.

Numerical simulations associate a theoretical model to the real, biological system under consideration, including a representation of the protein chain and its environment, and a description of the relevant physical laws governing this system. Such models can range from detailed atomistic descriptions to minimalistic simple models, depending on the accuracy required and the available computational resources.

The accurate numerical study of a process such as knotted protein folding, implies a thorough sampling of the conformational space accessible to the protein model, a task that can easily fall beyond the possibilities of the computational resources available nowadays. For this reason a large number of simplified models have been proposed in the literature, trying to isolate the crucial factors of the complex system under study, trading off the accuracy of the description with feasible calculations and a more straightforward interpretation of the results.

In the next pages we shall provide a comprehensive summary of the computational models and techniques employed for the simulation of self-entangled proteins, trying to underline the impact of the key assumptions and referring to the main results obtained by means of different approaches. We will start by discussing Coarse-grained lattice representations of the protein, and then move to continuous space models, that still rely on a coarse resolution of the polypeptide. After this part, which is by far the most rich in content, we will discuss those approaches that employ a fully atomistic resolution in describing the system. Finally, we will present techniques that aim at describing some crucial factors interacting with the folding and functioning of self-entangled proteins, such as interfaces, chaperonins, or other proteins.

4.1. Coarse-grained models on lattice

The discretisation of space is a crucial step in the definition of countless theoretical models, in all branches of natural sciences. Polymer physics does not represent an exception in this sense. In lattice models a flexible polymer chain is described as a succession of nearest neigbouring vertexes of a lattice, as displayed in figure 10. Each of these vertexes represents a monomer or, since we are interested in proteins, an amino acid residue. This coarse-grained (CG), monomer-to-site representation is referred to as *simple lattice model*. This approximated description substantially reduces the conformational space accessible to the polymer, being naturally prone to numerical simulations.

Lattice polymer simulations consist in a Monte Carlo (MC) sampling in the space of possible conformation of the polymer. The generated set of configurations is supposed to be representative of a statistical (e.g. equilibrium) ensemble, from which expectation values are extracted. Despite the substantial simplifications, lattice models have been demonstrated to reproduce essential universal properties of polymeric systems. For these reasons, since the first, seminal simulations

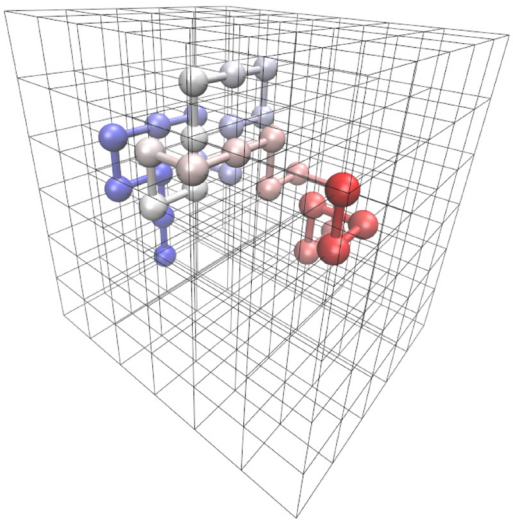

Figure 10. Cubic lattice representation of a polymer (graphics produced with VMD [103]).

(see e.g. [125]), lattice models have been extensively used in all branches of polymer physics.

The algorithms used for sampling the chain configurations depend on the physical properties associated to the polymer description, like e.g. the persistence length, or the solubility of the monomers. The simplest possibility is that of generating random walks on the lattice sites, made by $N-1$ unitary segments, where N is the number of monomers composing the chain. No restriction is imposed on the occupation of the lattice sites, so that any generated path is accepted, and the method efficiently samples independent conformations. This way, the description only retains the chain connectivity of the polymer, resulting in non-realistic observables [126]. A first step towards a more realistic model is that of considering the excluded volume of each monomer, generating the so-called self-avoiding walks (SAWs) on a lattice [127], in which only a single monomer is allowed on a lattice site. This reduces the acceptance rate of the generated paths, and the sampling efficiency strongly decreases with the chain length. This behaviour, named attrition, represents a major obstacle in lattice polymer simulations, in particular if one is interested in compact configurations, as in the protein folding problem [128].

Different lattice configurations can be generated independently from each other, but also as a succession, operating a set of minimal changes (MC moves) to an existing conformation. This sequential generation is particularly efficient and it allows also to estimate 'dynamic' properties of the polymer, such as relaxation times and correlation functions. Different sets of MC moves have been proposed in the literature [126, 129–131], with extensive discussions about their efficiency and ergodicity, that is their ability to visit all the allowed configuration space, or about their capability to preserve the topology of a lattice polymer. In the specific case of knotted protein folding one has to rely on a set of appropriate MC moves that can mimic the process without violating the topological barriers, as e.g. in the algorithm presented in [129].

If only the steric effect of occupied lattice points is accounted for, the newly generated configuration is accepted with a binary probability, 0 or 1, depending on whether it violates or not the excluded volume condition. However, the polymer model can be enriched, for instance by introducing an interaction energy among the monomers. In this case the acceptance probability depends on the configuration energy, as e.g. in the widespread Metropolis algorithm [132].

The modelling of monomer interactions is a crucial step to develop a suitable description for the study of protein folding. A first, simple possibility is that of adding an attractive interaction among the monomers that are not sequentially connected [133, 134], but occupy neigbouring sites (*contact potential*), thus favoring the collapse of the chain into a compact configuration. While this simple model can already provide insights on the statistical behaviour of homopolymers [131], it is not suitable for heteropolymers such as proteins, the behaviour of which is driven by the specific interactions occurring among the different amino acids, and between the latter ones and the solvent molecules. The purpose of capturing these specific correlations has lead to the formulation of several lattice models of heteropolymer interactions [135–138], which stimulated important advancements in the field of protein folding (see e.g. [139] and reference therein).

In the following we mainly focus on two lattice descriptions, which have been employed for the study of entangled proteins, namely the HP model [137] and the Gō model [135]. The HP model aims at reproducing the hydrophobic behaviour of specific protein residues, which is considered to play a prominent role in the process of protein folding [139]. Indeed, when a polypeptide is immersed in water it tends to minimize the exposure of its hydrophobic residues to the solvent, reaching a globular compact conformation, similarly to homo-polymers in poor solvent. This tendency represents an important driving force in the realization of the native state.

In order to mimic this behaviour the HP model represents the lattice polymer as a sequence of two types of amino acids, hydrophobic (H) and polar (P). The non-connected neigbouring H residues interact with an attractive *contact potential*, while the P residues are inert. The energy of an N residues HP polymer is given by:

$$E\left(\{r_i\}\right) = \sum_{i,j}^{N} \epsilon_{ij} d(r_i, r_j), \tag{1}$$

where $\{r_i\}$ represents a chain configuration, ϵ_{ij} is determined by the H/P type of i and j residues, that is $\epsilon_{\mathrm{HH}} = -1$ and $\epsilon_{\mathrm{PP}} = \epsilon_{\mathrm{HP}} = 0$. $d(r_i, r_j) = 1$ if i and j are non-sequential residues in contact, that is r_i and r_j are first-neigbouring sites. Otherwise, $d(r_i, r_j) = 0$. In this description the solvent is represented by the unoccupied sites in the lattice. The attractive interaction between H residues minimizes their exposure to the solvent sites, favoring the formation of a collapsed hydrophobic core.

The HP model has been extensively employed to investigate the thermodynamic features of protein folding (see e.g. [139–142]). In the framework of knotted proteins the HP model has been employed by Wüst *et al* to asses the statistical rarity of knotted native states [143]. In their work the authors compared the probability of finding knots in low-energy,

compact conformations of HP lattice polymers with random or designed sequences, as opposed to the behaviour of H-type homopolymers. Exploiting the simplicity of the HP model, the authors explored the space of possible residue patterns, unveiling important correlations between the sequence information and the folding and knotting processes. The results demonstrated that the sequence of hydrophobic-polar monomers is crucial in the determination of the knotting probability of a polymer chain. We underline that the purpose of analyzing native-like conformations requires the efficient exploration of the rugged free-energy landscape associated to the system, that can hardly be attained with classical metropolis sampling. In [143] the sampling was *enhanced* via the Wang–Landau method combined with two specific MC moves [131, 144].

In HP models the ground state of a protein representation is not known *a priori*, unless the sequence is specifically designed to favor some compact configuration. In Gō models, instead, the minimum of the potential energy corresponds to a specific pre-defined conformation, representing the native structure of the protein model [135, 145]. For this reason Gō models are also referred to as structure based models (SBM). The chosen interaction is again a contact potential, which now favors the vicinity of residues that are in contact in the native conformation $\{r_{0i}\}$. Using the notation of equation (1), the Gō model energy is given by:

$$E(\{r_i\}) = \sum_{i,j}^{N} C_{ij} d(r_i, r_j), \tag{2}$$

where the matrix C:

$$C_{ij} = \begin{cases} \epsilon & \text{if } |r_{0i} - r_{0j}| = 1 \\ 0 & \text{otherwise} \end{cases}, \tag{3}$$

indicates if two residues are in contact ($|r_{0i} - r_{0j}| = 1$, in units of lattice spacing) in the native configuration. ϵ is a negative constant that determines the attraction between non-sequential native contacts. This description allows to study the thermodynamic and kinetic stability of pre-designed native structures, and also to enlighten the folding pathways preferred by the protein model [146–150]. The theory underlining the formulation of Gō potential is the so-called energy landscape theory, according to which the existing protein sequences have been selected by evolution so that their folding free-energy landscape is 'funneled'. This means that a strong natural bias exists towards the sampling of native-like conformations, which drives the collapse of the polymer chain towards the native fold, determining an efficient and reproducible folding process. More details about energy landscape theory can be found in [151]. One can also interpret the potential of equation (2) as a model for the hydrogen bonds that stabilize the native structure of a protein.

In the framework of knotted proteins Gō models have been extensively used, both with lattice and off-lattice representation. We focus here on the former, while off-lattice models are treated in the following. In general, Gō models on a lattice do not aim at representing existing structures, so the contact potentials are defined to promote a designed native conformation. This, in the framework of knotted proteins, represents an advantage, allowing to design and test specific topologies. Such a strategy has been applied by Faisca *et al* in [152], where the folding thermodynamics and process of a model protein, with a designed native structure containing a shallow trefoil knot, were investigated with MC lattice simulations. The possibility of defining *a priori* the native target conformation allows also to simulate and compare structural homologues, whose native structures mostly overlap. Exploiting this option, in [152] the authors compared the behaviour of a knotted protein model to an unknotted homologue, sharing 90% of the native conformation with the former. This comparison made it possible to assess the effects of the knotted topology on the folding and unfolding processes.

The simplicity of this model allows to perform a thorough sampling in the configuration space, similarly to the case of HP model. It is indeed possible to gather a large number of folding trajectories, so that folding and knotting probabilities can be computed, and possible intermediate states retrieved. This strategy allowed to enlighten key features of the folding and unfolding kinetics. For example lattice Gō models have shown that the average folding time of a knotted protein model is considerably longer than that of its unknotted homologue [152, 153], or that the knotting process typically occurs in the late phases of the folding, when most of the contacts are formed [152, 153]. The possibility of designing the target native structure enables also the study of more complex topologies, such as that of a 5_2 knot [154], or to gradually increase the knot depth [153].

Another interesting methodological possibility is that of analysing the effect of mutations, that is modifying the interaction of some specific contacts. In [154] the authors demonstrated that, by turning off some particular contact interaction, located on the knotting loop and on the threading terminus, the folding success rate could be remarkably enhanced. Indeed, as mentioned in section 2, the early creation of native contacts can form topological bottlenecks, or kinetic traps, that considerably slow down the folding. The protein then has to backtrack, breaking the untimely formed contacts in order to retrace a correct folding path. The introduction of mutations prevents the early creation of contacts, reducing possible topological bottlenecks and speeding up the folding. This can be seen as the simplest way to model non-native interactions which, as underlined in the following, have been shown to play a relevant role in the folding of knotted proteins [155–157].

In [152] the MC moves were accepted/rejected by means of the classical Metropolis algorithm, properly choosing the simulation temperature to simulate either the folding or the unfolding of the protein. However, as in the case of [143], the large free-energy barriers involved in the conformational space, make the use of an enhanced sampling approach desirable. In [158] this issue was addressed by applying replica exchange (RE) of MC simulations of different temperatures [153, 154, 158].

4.2. Coarse-grained models in continuous space

In this section the detail of the protein description is increased, dropping the discretized space, and redefining the polymer

representation 'off-lattice', namely in a continuous three-dimensional space (among the seminal works about off-lattice protein and polymer modelling the reader can refer to [159–163]). The sampling in a continuous space can be performed via MC, e.g. redefining the MC moves as discussed in [156], but also with *molecular dynamics* (MD), that is by numerically solving the Hamilton equations of the system. MD can in principle represent the actual dynamics of the protein and it is therefore a crucial tool to study the kinetics of folding. However, if CG is operated, special attention should be paid when retrieving kinetics[163–165].

We start by considering those off-lattice protein models in which each amino-acid is represented by an interacting bead situated at the position of the C_α atom. The internal degrees of freedom of each residues are neglected, as it is neglected the presence of solvent molecules. To mimic the effect of the solvent environment, and to sample the canonical ensemble at a desired temperature T, Langevin dynamics is solved. Namely, the equation of motion for the chain is given by:

$$m\ddot{\mathbf{R}} = -\nabla_{\mathbf{R}} U - \gamma\dot{\mathbf{R}} + \Gamma, \tag{4}$$

where $\mathbf{R} = (\mathbf{r}_1, \ldots, \mathbf{r}_N)$ is the generalized vector containing the coordinates of the N beads, $\nabla_{\mathbf{R}}$ is the gradient in the coordinate space, U is the interaction potential of the model, γ is the viscosity coefficient and Γ is the random force term that introduces thermal fluctuations. γ is tipically chosen large enough that inertial effects are damped, but still lower than the equivalent viscosity of amino acids in water solution [160]. It has been shown that, in this regime, the timing of folding events, such as β-sheet formation, scales linearly with γ [163–165], allowing the choice of a lower viscosity to speed-up the calculations. The stochastic term Γ is drawn from a Gaussian distribution with variance related to the temperature by the relation:

$$\langle \Gamma(t)\Gamma(0)\rangle = 2\gamma k_{\mathrm{B}} T\delta(t), \tag{5}$$

where k_{B} is Boltzmann's constant and δ is Dirac delta distribution. Equation (4) does not fully account for the action of the solvent, as no description of hydrophobicity is provided. However, in the considered CG models, the hydrophobic response of amino-acids is implicitly incorporated within the interaction among the different beads, as in the HP lattice model. An example of off-lattice CG model driven by hydrophobic potentials can be found in [160, 166].

In the framework of self-entangled folding structure-based CG models have been widely used. In continuous space it is indeed possible to construct Gō-models that have the experimentally known knotted protein conformations as energy minimum. The general form of the potential energy for an off-lattice Gō-model can be written as follows:

$$U = U_{\mathrm{bond}} + U_{\mathrm{bb}} + U_{\mathrm{nat}} + U_{\mathrm{non-nat}}, \tag{6}$$

in which U_{bond} is the bond energy between consecutive beads of the chain, U_{bb} restrains the relative directions of the chain vectors to mimic backbone stiffness, U_{nat} is the interaction among residues in contact in the native state (analogous to equation (2) in MC descriptions), and $U_{\mathrm{non-nat}}$ generates the excluded volume and other interactions among the remaining residues. To define U_{nat} one constructs the 'contact map' from the experimentally detected native structure of a protein. The contact map is the off-lattice counterpart of C matrix in equation (3), that is an $N \times N$ matrix indicating which non-consecutive residues are in contact and, in some cases, the specific nature of the interaction (e.g. hydrogen or aromatic bonding, ionic bridges). The simplest way to construct a contact-map is by defining a cut-off distance between the C_α atoms belonging to different amino acids. If the native distance is lower than this cut-off the two residues are considered in contact. Another approach, that introduces further structural detail, was proposed by Tsai [167], usually referred to as Tsai contact map. In this method the position of all non-hydrogen atoms of the native structure are represented by spheres, of radii equal to the relative atomic van der Waals radius, rescaled by a factor 1.24. Whenever spheres belonging to different residues overlap, the two residues are considered in contact. By default, pairs of consecutive or next-to-consecutive residues, are excluded from contact map definitions, as their interactions are already accounted by U_{bond} and U_{bb}. Several definitions of contact maps have been proposed and tested, details and references can be found e.g. in [168–170] and, in the framework of knotted folding, in [171]. Once the contact map is defined, U_{nat} is constructed as a sum of attractive potentials among all pairs of residues natively in contact.

The potential terms in equation (6) can be specified in many different ways, as shown e.g. in the study of [168]. In the following, we review those descriptions employed for the computational study of self-entangled proteins.

4.2.1. Clementi et al model. This first model was introduced in the framework of small globular protein folding by Clementi *et al* [172], in Onuchic group. The polypeptide is represented by a chain of N spherical beads centered at the C_α atom positions $\mathbf{R} = \mathbf{r}_1, \ldots, \mathbf{r}_n$. The potential energy of the model is given by equation (6), in which the different potential contributions are defined as follows. Let $\mathbf{d}_i = |\mathbf{r}_{i+1} - \mathbf{r}_i|$ be the Euclidean distance vector between two consecutive beads, and d_i its magnitude. The bond energy between consecutive beads is:

$$U_{\mathrm{bond}} = \sum_{i=1}^{N-1} k_{\mathrm{b}}(d_i - d_{0i})^2, \tag{7}$$

where d_{0i} is the native distance between the ith and the $i+1$th residues. The backbone stiffness is represented by restraining the bending and dihedral angles of the protein chain to their native values, namely:

$$U_{\mathrm{bb}} = U_{\mathrm{bending}} + U_{\mathrm{dihedral}}, \tag{8}$$

where U_{bending} is a harmonic potential involving triplets of consecutive beads:

$$U_{\mathrm{bending}} = \sum_{i=1}^{N-2} k_\theta(\theta_i - \theta_{0i})^2, \tag{9}$$

in which θ_i is the ith angle, formed by beads i, $i+1$ and $i+2$, and θ_{0i} is its respective native value. The dihedral term is given by:

$$U_{\rm dihedral} = \sum_{i=1}^{N-2} [k_{1\phi}(1+\cos(\phi_i - \phi_{0i})) + k_{3\phi}(1+\cos(3\phi_i - 3\phi_{0i}))], \tag{10}$$

where ϕ_i is the ith dihedral angle of the chain, i.e. the angle between two intersecting planes, one containing i, $i+1$ and $i+2$ beads and the other containing $i+1$, $i+2$ and $i+3$ beads. ϕ_{0i} is the respective native value. Equation (8) has been used also in other protein models, employing standard, non-specific equilibrium angles (see e.g. [160]).

The $U_{\rm nat}$ term involves 12–10 Lennard-Jones pair potentials:

$$U_{\rm nat} = \sum_{i<j}^{N-1} C_{ij}\epsilon \left[5\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 6\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10}\right], \tag{11}$$

where the contact map C_{ij} is equal to 1 for residues in contact in the native state, and 0 otherwise. Since $U_{\rm bb}$ regulates the interaction among consecutive quadruplets of beads, pairs of sequential distance $|i-j| \leqslant 3$ are excluded from the native contact map (i.e. $C_{ij} = 0$). The length scale σ_{ij} is equal to the distance r_{0ij} between the ith and jth residues in the native state, so that the minimum of the potential is at $r_{ij} = r_{0ij}$. The non-native term $U_{\rm non-nat}$ contains only a repulsive contribution:

$$U_{\rm non-nat} = \sum_{i<j}^{N-1} \tilde{C}_{ij}\epsilon \left(\frac{\sigma}{r_{ij}}\right)^{12}, \tag{12}$$

where $\tilde{C}_{ij} = 1 - C_{ij}$ is the complement of the contact map, and includes all possible residue pairs except the native contacts. In equation (12) the length scale is constant, typically $\sigma = 4$ Å. A common choice for the parameters in Clementi's model is: $k_{\rm b} = 100\epsilon$, $k_\theta = 40\epsilon$, $k_{1\phi} = 1.0\epsilon$, $k_{3\phi} = 0.5\epsilon$, where ϵ is the energy unit corresponding to the depth of the native contact well in equation (11). Since its publication, this model was further developed, for example by implementing the shadow contact map [169], or by introducing a Gaussian-shaped well potential for the native contacts [173]. All these methodologies are available to the public by means of the SMOG [174] and SMOG2 [175] interfaces.

Clementi's description has been widely used in protein folding, and it is widespread also in self-entangled protein studies. To provide few examples, it has been employed to investigate the folding of MJ0366 [105, 171], the smallest knotted protein, or even more complex systems, such as the laboratory engineered 2ouf-knot [176] and the 5_2-knotted Ubiquitin C-terminal Hydrolases [177], and other kind of backbone entanglements, including complex lassos [111, 113, 114] and links [116, 178].

4.2.2. Cieplak et al model. A widely used Gō-like description in the framework of self-entangled protein folding has been initially proposed by Cieplak and Hoang to study universality of protein folding times [179]. The main feature of this model is that the $U_{\rm bb}$ potential is implemented as a four body term that favors the native chirality of the backbone:

$$U_{\rm bb} = \sum_{i=2}^{N-2} \frac{1}{2}\kappa\epsilon\left(\chi_i - \chi_{0i}\right)^2. \tag{13}$$

χ_i indicates the chirality associated to the residue i, that is defined, in terms of the distance vectors $\mathbf{d}_i$:

$$\chi_i = \frac{(\mathbf{d}_{i-1} \times \mathbf{d}_i)\cdot \mathbf{d}_{i+1}}{d_{0i}^3}. \tag{14}$$

χ_{0i} in equation (13) indicates the value of the native conformation. This form of $U_{\rm bb}$ was first proposed in [180], and it is a simpler and more numerically efficient version of the original potential proposed in [179]. The coupling constant κ is usually chosen equal to 1. The chirality potential of equation (14) plays an analogous role as the angular terms in Clementi's model, but it has in general a weaker action. It can be demonstrated [168] that equation (14) is approximately equivalent to the dihedral term in equation (8). The bending stiffness is here unrestrained, so that the chain is less stiff than in the previous case. To compensate for this, native contacts between i and $i+3$ residues are included in the contact map. A further difference with previous description consists in the use of a 12–6 Lennard-Jones potential for the native contacts:

$$U_{\rm nat} = \sum_{i<j}^{N-1} C_{ij} 4\epsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], \tag{15}$$

where the σ_{ij} are chosen so that the distance of the energy minimum ($2^{1/6}\sigma_{ij}$) corresponds to the native distance r_{0ij}.

In the framework of self-entangled proteins, Cieplak model has been used to investigate the folding of the smallest knotted protein [49], of deeply knotted proteins YibK and YbeA [48, 57], and of protein dimers [122, 177], and to explore thermal unfolding [181, 182]. Moreover, as discussed in the following (section 4.4), this description has been employed in combination with methods to simulate protein dynamics under specific conditions, such as the presence of interface environment [182], the translocation through a proteasome channel [183], or co-translational folding [48].

Cieplak's model has been also extensively employed to simulate protein stretching experiments, in which single molecules are manipulated by means of atomic force microscopy [185] or optical tweezers [186], and stretched pulling two selected residues apart (see figure 11(A). The resulting force-extension ($F-d$) diagrams show complex peak patterns (an example is shown in figure 11(B), that can reveal useful information about the structure and stability of the analyzed proteins [187, 188]. MD simulations represent a useful tool for the interpretation of these experimental results [189], generating theoretical $F-d$ diagrams that can be reconducted to the experimental ones to clarify the nature of the different force peaks. The stretching force is generated by constraining the position of a selected residue, and by imposing a moving harmonic potential to a second residue. Both constant velocity and constant force pulling procedures have been implemented. An interesting aspect of this simulation protocol is that the simulated velocity of stretching can be set orders of magnitude larger than in experiments, still being able to extrapolate to the experimental timescales. As a result, the computational time required for stretching simulations is relatively shorter and allows also the use of atomistic resolution MD (see section 4.3). Nonetheless, CG models such as

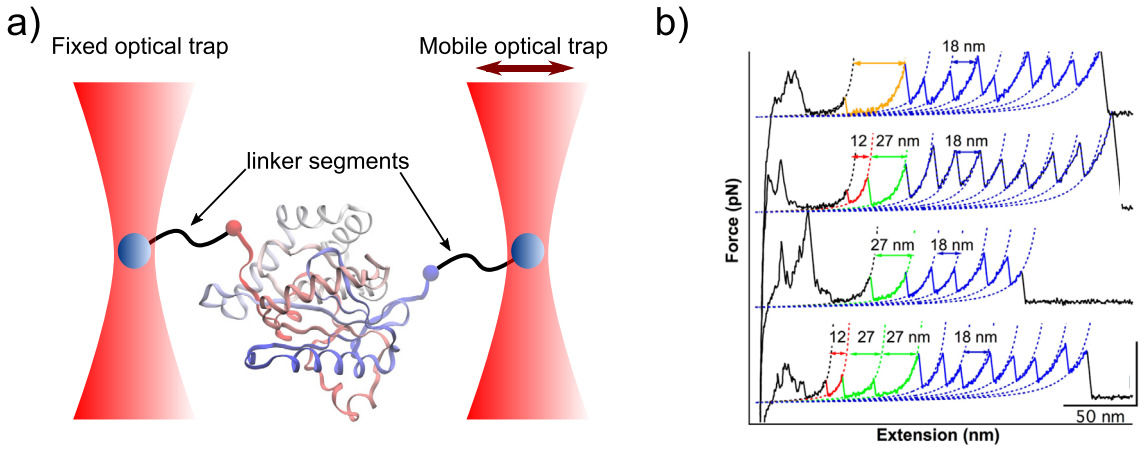

Figure 11. (A) Scheme of typical protein stretching experiment, with fixed and movable optical traps, linked to the terminals of the protein. (B) Typical $F-d$ diagrams, obtained from mechanical unfolding of a slipknot protein. Reprinted with permission from [184]. Copyright © 2012, American Chemical Society.

Cieplak's have shown to be useful for large surveys of stretching simulations, that could help generalize the mechanical properties of proteins [190, 191]. In the field of self-entangled protein studies mechanical stretching represents a crucial tool of analysis, being capable of revealing possible intermediate conformations, or providing information on the stability of backbone entanglements [63]. Cieplak's model has been widely employed to simulate the stretching of self-entangled proteins [58, 59, 181, 192–194] and protein complexes [122, 177].

Comparisons between the two presented variants of CG Gō-models can be found in the literature (see e.g. [168, 195] and, in the realm of entangled folding, [49]), showing e.g. the lower stability of chirality potential with respect to the angular potential of equation (8), or indicating the different ranges of folding temperatures for the two models.

4.2.3. Prieto et al model. An example of off-lattice Gō-model that has been used in combination with MC sampling is that proposed by Prieto *et al* in [196, 197]. In this description, the protein is represented as a chain of hard spheres centered on the C_α atoms. The new conformations are generated by MC moves, which are constrained by the imposition of excluded volume of the spheres, and by fixed bond distances $d = 3.8$ Å. The energy of a configuration is then determined by native and backbone interactions, which are both described by a harmonic well potential, namely:

$$U_{\mathrm{bb+nat}} = \sum_{i<j} u_w(r_{ij}), \tag{16}$$

where the pair interaction is given by:

$$u_w(r_{ij}) \begin{cases} w_{ij}[(r_{ij}-r_{0ij})^2 - a^2], & \text{if } r_{0ij}-a < r_{ij} < r_{0ij}+a \\ 0, & \text{otherwise} \end{cases}, \tag{17}$$

in which $a = 0.6$ Å, r_{0ij} is the native distance of the residues i and j and w_{ij} defines the contacts. w_{ij} corresponds to the native contact map but, since there are no angular interactions, also contacts separated by two or three bonds are here considered. In order to preserve the native chirality, when contacts between i and $i+3$, a sign is assigned to r_{ij}. This description has been employed to assess the effects of local flexibility and steric confinement on the knotted folding of MJ0366, in comparison with the results of a lattice model [198].

4.2.4. Heterogeneous interactions. The models presented up to here are characterized by sequentially homogeneous interaction, meaning that parameters such as the LJ energy scale ϵ, or the angular stiffnesses k_θ, $k_{1\phi}$ and $k_{3\phi}$, are independent of the residue index. In these descriptions the sequence information is all represented by the native structure information. This can be an useful simplification to reproduce protein folding only on an approximate, qualitative level. However, in some cases, the lack of interaction specificity can prevent the folding, in particular when the native conformation exhibits a complex topology [48]. In order to improve the effectiveness of protein modelling, many choices of heterogeneous native-contact potentials have been explored, some examples can be found in the systematic study of [193]. One of the simplest choices is that of rescaling the native contacts corresponding to cysteine-cysteine bridges, increasing the depth of the potential well so that the rupture of the bond becomes unlikely [193]. This method has been used in the study of self-entangled proteins for which cysteine bridges are topologically relevant, such as complex lasso structures[111, 113–115] and protein dimers [177]. In some other cases cysteine bridges are treated equally to peptidic bonds (e.g. equation (7)) [59, 111, 113, 114].

An example of how interaction heterogeneity can be crucial for knotted folding can be found in [199], where the atomic interaction-based coarse grained (AICG) model, initially proposed to study allosteric proteins [200], is applied to self-entangled proteins. The AICG approach is built on the potential energy of Clementi's model, in which the coupling constants are made residue dependent. This means that AICG has heterogeneous bond energy stiffnesses ($k_\mathrm{b} \to k_{\mathrm{b}i}$), angular stiffnesses ($k_\theta \to k_{\theta i}$, $k_{1\phi} \to k_{1\phi i}$ and $k_{3\phi} \to k_{3\phi i}$), and native contacts ($\epsilon \to \epsilon_{ij}$). The first step to construct the AICG model is the definition of the relative strength of native contacts.

The coupling constant is decomposed as $\epsilon_{ij} = \epsilon_{\rm av} w_{ij}$, where $\epsilon_{\rm av}$ describes the average strength of residue-residue contacts, and w_{ij} is the relative weight. The w_{ij}'s are computed by means of all-atom (AA) implicit solvent MD in the native state, defining an energy decomposition protocol to retrieve CG contact energies from atomistic interactions [200, 201]. One can either perform an all-atom native state simulation for each protein to be studied, or use a linear regression model proposed in [200], trained over an ensemble of different proteins. After the w_{ij}'s are set, the remaining free parameters of the potential are tuned. The multiplicity of bending and angular parameters is first reduced, assigning each residue to a specific secondary structure such as β-strands or α-helixes, to which a unique set $\{k_{\rm b}, k_\theta, k_{1\phi}, k_{3\phi}\}$ is associated. The reduced set of parameters is then optimized iteratively by matching the average equilibrium fluctuations of CG MD with those of AA MD. Again, one can perform this tuning for any specific system, or use a set of trained parameters obtained from an ensemble of optimizations.

In [199] the AICG approach has been applied by using a backbone potential differing from equation (8), named flexible local potential, constructed from a library of angular distributions via Boltzmann inversion [202], with the aim of reproducing more realistic backbone flexibility. This variant of the model, named AICG2 has been employed to study the folding of the engineered knotted protein 2ouf-knot, of the smallest knotted protein MJ0366, and of the deeply knotted YibK. In the first two cases the AICG2 model obtained a much larger propensity of folding with respect to the homogeneous Gō-model, demonstrating that interaction specificity can play a crucial role in entangled folding. In the case of YibK the folding was instead inaccessible, suggesting that crucial interactions in this system are still neglected at the level of this description.

4.2.5. Non-native interactions. The models presented until now are purely *native-centric*, meaning that non-native interactions (the $U_{\rm non-nat}$ term of equation (6)) are described only by excluded-volume contributions, implicitly assuming that these interactions play a negligible role in determining the folding dynamics. This represents a radical approximation, which is nonetheless supported by accurate MD simulation studies [203]. On the other hand, multiple computational studies have shown that non-native interactions might play a relevant role [204], especially when entangled proteins are considered. As seen earlier, the effects of structural mutations on the lattice model of [154] demonstrated that moving away from the purely native-centric picture can improve the folding propensity of knotted lattice proteins.

Non-native interactions were proven to be crucial in entangled folding by means of a CG off-lattice model in the pioneering work of Wallin *et al* [155]. In this work the folding of the deeply knotted protein YibK was investigated by means of the previously discussed Clementi Gō-model, equipped with an extra non-native attractive potential term. Let us decompose the non-native contact matrix (see equation (12)) as $\tilde{C}_{ij} = R_{ij} + A_{ij}$, where R_{ij} and A_{ij} represent repulsive and attractive contacts, respectively. The non-native potential is then given by:

$$U_{\rm non-nat} = U_{\rm steric} + 0.8\epsilon \sum_{i<j}^{N-1} A_{ij} \mathrm{e}^{(r_{ij}-\sigma_{\rm nn})^2/2}, \qquad (18)$$

where $U_{\rm steric}$ is given by equation (12) and $\sigma_{\rm nn} = 4.0$ Å is the attractive contact length. In [155] these contacts are chosen ad-hoc within residues that are involved in the entanglement formation. The inclusion of this attractive term has a striking effect on the folding propensity of the model: while the bare native-centric model was unable to reach the correct topology in all the presented simulations, with non-native interaction the model could always reach the native conformation. These results supported the idea that non-native contacts can regulate the kinetic accessibility of the entangled state, and increase the folding probability of the protein.

Further studies on the impact of non-native interactions in the folding of entangled proteins were obtained by means of a more physically detailed model in [53, 156, 157]. This model, presented in detail in [156] is based on a Gō-like CG description, that can be described by equation (6). Apart from the common harmonic bond potential (equation (7)), this description features heterogeneous angular and native contact interactions [205, 206]. The angular potentials do not include native angles as in equations (9) and (10), but depend on the secondary structure associated to the respective residues. The bending term is given by a double-well potential:

$$U_{\rm bending} = \sum_{i=1}^{N-2} -\frac{1}{\gamma} \ln\left[\mathrm{e}^{-\gamma k_\alpha (\theta_i-\theta_\alpha)^2 - \gamma\epsilon_\alpha} + \mathrm{e}^{-\gamma k_\beta (\theta_i-\theta_\beta)^2}\right], \qquad (19)$$

where $\theta_\alpha = 92°$ and $\theta_\beta = 130°$ are equilibrium values for helical and extended chain respectively, and k_α,ϵ_α and k_β are specific parameters, typical values are reported in [207]. The dihedral term is a generalization of equation (10):

$$U_{\rm dihedral} = \sum_{i=1}^{N-2} \sum_{n=1}^{4} k_{n\phi} \left[1 + \cos(n\phi_i - \delta_n)\right], \qquad (20)$$

where $k_{n\phi}$ and δ_n depend on the secondary structure associated to the two central residues of the ith dihedral. The employed values are indicated in [205]. The native contacts are described by the following heterogeneous term:

$$U_{\rm nat} = \sum_{i<j}^{N-1} C_{ij}\epsilon_{ij} \left[13\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{12} - 18\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{10} + 4\left(\frac{\sigma_{ij}}{r_{ij}}\right)^{6}\right], \qquad (21)$$

where ϵ_{ij} is set as the specific hydrogen bond energy, if a native hydrogen bond is present. Otherwise, ϵ_{ij} is chosen proportional to the quasi-chemical contact potentials of Miyazawa and Jernigan [208], properly renormalized to the hydrogen bond energy scale. As shown in figure 12, the pair term in equation (21) qualitatively differs from the other LJ-like potentials in that a small energy barrier must be overcome to establish the contact, mimicking desolvation energy cost.

On top of this native-centric potential, a non-native term and a long-range electrostatic term are introduced. As in the previous case, the non-native contact matrix is divided in

attractive and repulsive contributions, defining the following non-native potential term:

$$U_{\text{non-nat}} = \sum_{i<j}^{N-1} [A_{ij} U_A(r_{ij}) + R_{ij} U_R(r_{ij}) + U_{\text{el}}(r_{ij})], \quad (22)$$

where U_A and U_R are the attractive and repulsive contact potentials, and U_{el} describes the long-range electrostatics. Attractive contacts interact via a heterogeneous 12-6 LJ potential:

$$U_A(r_{ij}) = 4|\epsilon_{ij}| \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right], \quad (23)$$

and repulsive contacts interact with the following functional form (displayed in figure 12):

$$U_R(r_{ij}) = \begin{cases} 4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right] + 2\epsilon_{ij}, & \text{if } r_{ij} < 2^{1/6}\sigma_{ij} \\ -4\epsilon_{ij} \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right], & \text{if } r_{ij} \geqslant 2^{1/6}\sigma_{ij} \end{cases}. \quad (24)$$

The length-scale of the contacts is $\sigma_{ij} = (\sigma_i + \sigma_j)/2$ where σ_i and σ_j are the van der Waals radii of the interacting residues. The interaction strength is set as $\epsilon_{ij} = \lambda(e_{ij} - e_0)$, where e_{ij} are negative values taken from the Miyazawa and Jernigan residue interaction matrix [208] and λ and e_0 are free parameters, obtained in [207] by fitting calculated binding affinities with experimental values. The average non-native interaction strength resulting from equations (23) and (24) is approximately 1/10 of native one. Finally, the electro-static term is given by Debye–Huckel potential:

$$U_{\text{el}}(r_{ij}) = \frac{q_i q_j}{4\pi\epsilon_0 D r_{ij}} \exp\left(-\frac{r_{ij}}{\xi} \right), \quad (25)$$

in which q_i and q_j are the residue charges, ϵ_0 is the dielectric constant in vacuum, D is the relative dielectric constant of water and $\xi \sim 10$ Å is the Debye screening length. While the resulting model is definitely richer of physical ingredients than the usual Gō-models, non-native interactions introduce frustration, slowing down the diffusion of the protein along the folding funnel. For this reason, in [53, 156] the sampling is performed via MC, applying a set of local moves that allow to retrieve dynamic properties with a significantly lower cost than with MD [129, 209].

By comparing the results obtained by this model with and without equation (22), it had been possible to enlighten the key role of non-native interaction in favoring early knot formation in carbamoyltransferases [156], and in determining the knotted folding of MJ0366 [53].

4.2.6. Elastic folder model. As mentioned before, in the study of entangled protein folding one of the issues of Gō-like potentials is the untimely formation of native contacts, that can entrap the chain in a misfolded configuration. These bonds need then to be ruptured in order to reach the native topology. This backtracking mechanism amplifies the simulation time

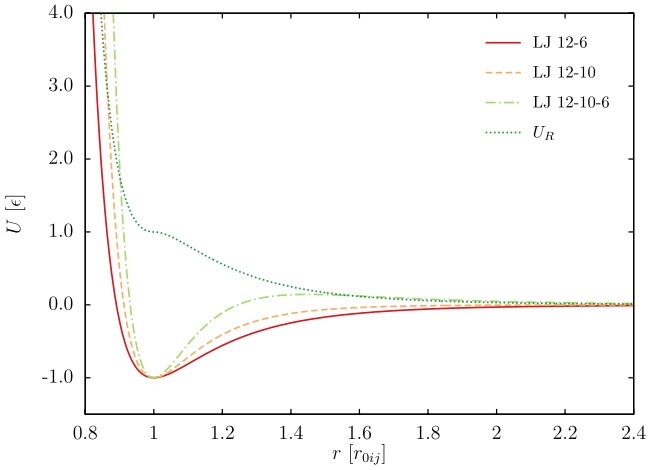

Figure 12. Comparison of LJ-like functional forms employed in different pair potentials presented in the text: the LJ 12-6 interaction of equation (15) (red solid line), the LJ 12-10 form of equation (11) (yellow dashed line), the LJ 12-10-6 potential of equation (21), and the non-native repulsive potential U_R defined in equation (24). The potential is in units of ϵ, and the length scale is r_{0ij} so that the stationary points of the different functions coincide. In the native potentials, r_{0ij} is the native distance between residues i and j. This corresponds to $2^{1/6}\sigma$ in LJ 12-6, and to σ in LJ 12-10 and LJ 12-10-6. For the non-native interaction U_R, $r_{0ij} = 2^{1/6}\sigma$, where the length scale σ is the mean of the residues van der Waals radii.

required to reach the folded state, resulting in very low folding probabilities. Nonetheless, it is believed that topologically complex proteins have evolved to fold in a reproducible and efficient way, so their sequence should encode sufficient information to avoid kinetic traps and misfolded configurations. Building on this idea, Najafi and Potestio have proposed a CG off-lattice description, dubbed elastic folder model (EFM), which is constructed to fold along optimal pathways, that is in the most reproducible and rapid way.

To achieve this the EFM employs a typical CG C_α description of the protein, in which native contact potentials are absent and the folding is governed by the back-bone angular interactions alone. The energy can be written as:

$$U_{\text{tot}} = U_{\text{bond}} + U_{\text{steric}} + U_{\text{bb}}, \quad (26)$$

where the Kremer–Grest model [210] is used to represent peptidic bonds and steric interactions. The former are described with the finitely extensible non-linear elastic potential:

$$U_{\text{bond}} = -\sum_{i=1}^{N-1} \frac{k_{\text{FENE}}}{2} \left(\frac{R_0}{\sigma} \right)^2 \ln \left[1 - \left(\frac{r_{i,i+1}}{R_0} \right)^2 \right], \quad (27)$$

in which $k_{\text{FENE}} = 30.0\epsilon$ is the interaction strength parameter, in units of an energy constant ϵ, $R_0 = 1.5\sigma$ and the length-scale $\sigma = 3.8$ Å is the typical extension of a peptidic bond. The Weeks–Chandler–Anderson interaction [211] is used for the steric part:

$$U_{\text{steric}} = \sum_{i<j}^{N} U_{\text{WCA}}(r_{ij}), \quad (28)$$

with:

$$U_{\mathrm{WCA}} = \begin{cases} 4\epsilon\left[\left(\frac{\sigma}{r_{ij}}\right)^{12} - \left(\frac{\sigma}{r_{ij}}\right)^{6}\right] + \epsilon & \text{if } r < 2^{1/6}\sigma \\ 0 & \text{otherwise.} \end{cases} \tag{29}$$

The structural information of the protein is included in the back-bone angular potential, which has the same functional form used by Clementi model (equation (8)). The driving force of the folding is thus provided by the bending and torsion terms. The idea of folding pathway optimality is implemented by using heterogeneous angular stiffnesses $k_{\theta i}, k_{1\phi i}, k_{3\phi i}$, which are tuned through a stochastic optimization approach, aimed at maximizing the folding propensity of the model. This optimization procedure is based on iteratively tuning the parameters $k_{\theta i}, k_{1\phi i}, k_{3\phi i}$ building on the results of several folding simulations. In [55] this is done by a MC sampling of the parameter space, in which a mutation of the stiffness is accepted or rejected according to its effect on the folding probability. This technique has been used to investigate the folding routes of two small knotted proteins [55], observing a qualitative agreement with the results of more detailed models. The optimization procedure has been recently improved via an evolutionary, parallelized strategy that allows a more efficient exploration of the parameter space [115].

4.2.7. More detailed CG. During decades of computational protein studies distinct CG descriptions of the polypeptide chain have been proposed, naturally increasing the detail starting from the simple C_α representation. However, mainly for computational requirements, these models have not been extensively employed in the field of entangled proteins.

An example of a more detailed description of the protein chain is found in [212], where the thermal unfolding of YibK is simulated with different techniques. Among these, wo CG models were employed, both increasing the detail of C_α representation by describing further elements of the amino-acid chain. The first one, described in [213], is a MC approach which includes side-chain and C_β interaction sites. The Hamiltonian contains heterogeneous backbone and native contact interactions, that combine general potential forms for secondary structures, native structure information, and statistical potentials built over a database globular proteins [214]. The second model as well includes a side-chain representation, and employs interaction potentials parametrized over known protein structures and *ab initio* calculations [215–217]. This second description is employed with MD sampling.

A second example of detailed CG in knotted proteins can be found in [218], where the folding of tRNA methyltransferase, presenting a deep native trefoil knot, is addressed. Actually only the entangled fragment of the protein chain is considered in the presented calculations. The used representation is inspired from the associative memory Hamiltonian used in structure-prediction studies [219, 220], that describes the structure retaining the position of C_α, C_β and O atoms. The energy function is the sum of a back-bone term, including self-avoiding and stiffness interactions, and a heterogeneous gaussian Gō-like term for $C_\alpha - C_\alpha$, $C_\alpha - C_\beta$ and $C_\beta - C_\beta$ pairs. The detailed form of the Hamiltonian can be found in [218, 220]. Another peculiar aspect of this work is the use of structure prediction methods to investigate the kinetics of the model. In these techniques the sampling of the folding space is obtained by generating a complex network of the possible local energy minima, using a combination of MC moves, minimum energy path, and energy optimization algorithms [221–223]. The transition rates across this network of conformations are then computed using discrete path sampling method [224], and the global kinetics of the model is inferred.

4.3. All-atom simulations

In this section we further increase the resolution of the modelling, focusing on those methods that employ an all-atom (AA) protein representation. This level of detail implies a significant increase of computational time with respect to CG simulations, considerably limiting the accessible timescales. Even using state-of-the-art super-computers, or dedicated computational architectures [225], a gap still exists between the timescales accessible with AA modelling of biological systems (roughly within the range μs–ms), and the typical timescales of biological processes (that can easily exceed minutes [8]). These limitations are critical in the study of self-entangled protein folding, motivating the popularity of simple CG models. Nonetheless, exploiting specific strategies to reduce the computational times, AA methods have been also applied to the study of topologically complex proteins.

Before discussing the usage of AA modelling in self-entangled proteins, we briefly outline few crucial features of this methodology. The AA representation provides a more accurate description of the molecular geometry, and of the realistic packing that can play a crucial role in processes such as protein folding. Beside this, a full atomistic description makes it possible to use physics-based potentials, that include a realistic treatment of interactions. These potentials, usually referred to as 'force-fields' (FFs), aim at reproducing the potential energy surfaces derived from quantum mechanics, in a transferable way throughout large sets of molecules. In general, atomistic FFs are based on additive terms, describing bonded and non-bonded interactions [226]. The former include bond, bending and dihedral energies, already encountered in the CG models, and the latter include electrostatic and van der Waals forces. Each contribution is typically considered as an additive term, however also cross-terms can be introduced, describing the interdependence of different degrees of freedom [227]. Once the functional form of each term is fixed, a set of free parameters regulating the strength of each additive term has to be chosen, typically fitting the results of *ab initio* calculations on specific chemical groups and accounting for crystallography and spectroscopy data. Nowadays, several FFs are available, such as Charmm, Amber or OPLS, each adopting different functional forms and parametrization techniques, that result in specific domains of applicability (liquids, nucleic acids, proteins and so on). A review on the topic can be found in [226].

The mentioned FFs describe the interaction of those atoms belonging to the biological complex under study, while for

the solvent environment, typically water, a specific description is employed. The AA modelling of water, also referred to as 'explicit solvent' treatment, has a long history, and many approaches are available, each using a particular interaction site geometry and parametrization protocol [228]. A popular alternative to AA description of the solvent is the so-called 'implicit solvation', in which the solvent is accounted for as a continuum medium, approximating the average effect of water on the solute system. An example of this technique is to apply the generalized born approximation to solve the electrostatics, while the remaining effects are accounted for with an estimate of the solvent accessible surface area [229–231]. The use of implicit solvation forces a radical approximation of the AA treatment, but it can still provide accurate results, at the same time significantly reducing the computational load.

Let us now review how the mentioned methodologies could be applied in the study of entangled proteins. As mentioned earlier, a proper framework for AA MD of proteins is that of mechanical stretching simulations, that involve shorter timescales than in the folding. The stretching protocol is equivalent to that described in section 4.2.2, namely fixing or restraining the position of a selected atom, typically the C_α of a terminal residue, and then pulling another atom apart, typically at the other end of the chain, by means of a moving harmonic potential [189, 232]. The velocity of the moving potential is several (∼6–8) orders of magnitude larger than the pulling speed in experiments, but the resulting $F - d$ patterns are compatible to those resulting from AFM measurments, allowing the interpretation of the latter. The chosen velocity of pulling allows to reach significant stretching of the protein within nanoseconds, timescales that are nowadays accessible by AA simulations. The first AA stretching simulations on an entangled protein were performed already in 2004 in [233], where a theoretical interpretation was provided to previous stretching experiments on the trefoil knotted carbonic anhydrase [234]. Here, both implicit and explicit solvent representations were utilized, reproducing in both cases the experimental force peaks. These results were successively complemented by further atomistic simulations [235]. Other works on stretching of topologically complex proteins can be found, together with experimental results, in [236], where the figure-of-eight knot in phytocrome is tightened while unfolding the polypeptide, or in [61, 184], where the slipknotted protein AFV3-109 is manipulated.

As said, when biological timescales come into play, the use of AA simulations can lead to unreasonable computational times. For this reason only few groups have tried to study the folding of self-entangled proteins by using fully atomistic representations. A viable strategy to obtain indications on the folding transition, avoiding the time-scale issue, is that of simulating the high-temperature unfolding process starting from the equilibrated native conformation. The higher temperature accelerates the conformational changes, and time requirements can be strongly reduced. In [212] the untying of deep trefoil knot of YibK was simulated with AA MD in explicit water, setting the temperature to $T = 900$ K. These calculations gave insights on the conformational change linking the native entangled state to a fully denaturated state, and on the stability of the knotted structure.

When the actual folding process is considered, one cannot escape the time-scale requirements. For this reason, by now, only the folding of the smallest knotted protein, MJ0366, has been studied with full atomistic detail. Moreover, to achieve this result, different strategies aimed at reducing the computational requirements have been proposed. One possible approach is that of combining the AA representation with ELT, employing Gō-like interaction potentials, the calculation of which is much less expensive than that of realistic force-fields [237]. This is the method chosen in [105], where an AA model is employed to sample the folding free-energy landscape of MJ0366. The employed model, presented in [174, 238], includes only non-hydrogen atoms, featuring the standard harmonic bond and angular terms, and the dihedral potential indicated by equation (10), which regulates also the dihedral angles of the side-chains. An extra harmonic term is added to restrain the improper dihedrals of the structure, typically to preserve planar arrangements. Non-native interactions are purely repulsive, while an atomistic contact map is built. Gaussian potentials were used to generate the pair interactions [173]. This description allowed to observe the effect of the realistic atomistic packing and protein geometry on the folding of MJ0366, in comparison with the results of ordinary CG simulations [105].

In a crucial work by Beccara *et al* [53], the folding of MJ0366 has been simulated by means of a fully atomistic description of the protein, together with a realistic interaction potential (Amber99SB [239]). Here the problem of time limitation has been tackled by employing an implicit solvent representation [231], together with an enhanced sampling technique. More in detail, the authors applied the dominant reaction pathway approach [240], combining a ratchet-and-pawl [241] biasing protocol, to favor the evolution towards the folded state, together with an a-posteriori scoring method, to estimate the relative probability of the biased trajectories [242, 243]. By means of this technique the full folding transition of MJ0366, starting from an initially denatured conformation up to the knotted, native state, was observed, and useful insights on the preferential folding pathways were collected.

Finally, the folding of MJ0366 was studied with an even higher detail of modelling in [244], where the Amber99SB force-field was employed together with an explicit solvent model (TIP3P water [245]). The authors performed unbiased MD calculations of this system, composed by about 1.9×10^4 atoms, using the special-purpose facility ANTON [225]. The results showed the dynamics of the native knot formation, starting from slipknotted conformations previously generated with the AA Gō-model of [105]. These are the only unbiased, AA, explicit solvent calculations of a protein knotting events available up-to-date. However, with this level of detail, the simulation of the full entangled folding process still remains out-of-reach.

4.4. External factors

In this section we review simulation techniques that aim at modelling those external factors influencing the dynamics of self-entangled proteins. The picture of an isolated protein,

spontaneously folding in water at physiological conditions is far from being general, as interactions with other cellular components or proteins are the standard. Moreover, the environment plays a crucial role also after the folding has occurred, influencing the biological functions of proteins. For this reason, several simulation techniques have been proposed to account for specific external factors, some of key interest in the realm of topologically complex structures.

4.4.1. Protein complexes. Many of the studied proteins constitute functional homo-dimers. In such cases it can be relevant to simulate the cooperative folding of two protein specimens, and observe their interactions and dimerisation process [173]. As previously mentioned, in [105] the folding of MJ0366 was simulated via an AA Gō-model. In this paper, the authors also investigated the possible dimerisation along or after the folding process, modelling the native contacts between two MJ0366 monomers via Gō-potential, with the same interaction strength and functional form (Gaussian) employed for the intramonomer native contacts. The monomers were held together by an harmonic potential acting on their centers of mass. Varying the curvature of this potential, the effect of different crowding on folding and dimerisation was probed. The interaction between protein dimers was simulated also in [122, 178, 246], where the study of both folding and stretching was addressed, this time using the CG Gō-model of Cieplak. Once again, the contacts between monomers were treated like the intra-chain native contacts. Using the same approach, in [122], the folding of known homo-dimeric knotted proteins, such as MJ0366 and YibK, was simulated, assessing the impact of the dimerisation potential on the process.

Cieplak Gō-model has been employed to simulate also other dimeric structures, forming protein links. In [122], the authors propose a four-terminal pulling simulation protocol, that detects the intertlinking among homo and hetero-dimer structures. This technique has been applied to more than 10^4 n-meric structures, detecting about 9% of entangled complexes. Four of these linked dimers have been studied in [178], assessing their folding and thermodynamic properties. In general the full dimerisation is found to occur as a late step, subsequent to the proper folding of the monomers.

4.4.2. Co-translational folding. Another aspect that plays a crucial role in protein folding *in vivo*, is the translation [52]. Indeed, the folding dynamics can take place already while the polypeptide chain is emerging from the ribosome, and this possibility can completely alter the picture of spontaneous folding adopted in standard simulations. This 'co-translational' folding can be particularly relevant for self-entangled proteins, as the sequential character of topology formation can be regulated by the time-scale of translation. This possibility is addressed in [48], where a simple model of on-ribosome protein folding is proposed. In this model the protein is represented by means of the Cieplak CG description, introduced before, while the ribosome is represented as a plane that generates a uniform potential:

$$U_{\rm r} = \frac{3\sqrt{3}}{2}\epsilon\left(\frac{\sigma_0}{z}\right)^9, \tag{30}$$

where z is the distance from the plane and $\sigma_0 = 4\times(2)^{-1/6}$. The protein chain emerges from a fixed coordinate on the plane, giving birth to a new C_α residue every t_w, starting from the N-terminal. This simple representation of the process aims at modelling the sequential generation of the polypeptide, and the excluded volume determined by the ribosome. This technique was used to simulate the folding of knotted proteins [48, 49], observing an improvement in the folding propensity. A simpler, but relatable study was performed in [158], where one terminal of the lattice protein chain was constrained to a chemically inert plane (determining only excluded volume). This method can provide a rough approximation of the entropic limitations present in nascent folding, but also in single molecule experiments. An alternative, more detailed, approach to simulate co-translational folding was recently proposed in [50], where the folding of the deeply knotted protein Tp0624 is simulated by modeling the ribosome channel with a cylindrical 10 Å tunnel with a funneled exit located on a planar wall. Initially, the stretched protein is contained in the tunnel, and a constant force is applied to its residues in order to push them through the channel exit. The protein description adopted is the Clementi model, with Gaussian native contacts. The interaction of the wall with the residues can be repulsive (WCA potential, equation (29)) or attractive (LJ 12-6, equation (23)). By assigning attractive interactions to a pre-selected set of residues this scheme could promote the formation of a loop on the ribosome wall at the channel exit. Starting from this configuration the authors demonstrated that the tying of the deep knot could be strongly favored by the ribosome action.

4.4.3. Air-water interface. An external factor that can play a key role in protein dynamics is the interaction with an air-water interface [247]. The vicinity of such interfaces had been also found to influence the topology of proteins [182], e.g. favoring non-native entanglements, or untying shallow knots. The method used here is again based on Cieplak's CG Gō model, where the air-water interface is effectively described by a field of forces coupled to the hydropathy of amino acids. The force acting on the ith residue is defined as follows:

$$F_i^{\rm wa} = q_i A\frac{\exp\left(-z_i^2/2W^2\right)}{\sqrt{2\pi}W}, \tag{31}$$

where q_i is the hydropathy index associated to the amino-acid [248], z_i its coordinate on the axis perpendicular to the interface, and $A = 10\epsilon$ and $W = 5$ Å are energy and length-scale of the interaction.

4.4.4. Chaperonin cage. It is well known that the folding of proteins can be assisted, and accelerated, by the bacterial chaperonin GroEL-GroES, a complex of two large heptameric units that form a compartment capable of accomodating a folding polypeptide [249]. The encapsulated protein exhibits a

faster folding rate, which can be the result of an active action of the chaperonin, via e.g. destabilisation of misfolded configurations, and a passive action, via the steric confinement determined by the compartment. The GroEL-GroES complex is of interest also in the field of self-entangled folding, as it has been shown to assist the folding of knotted proteins such as YibK and YbeA [8, 43]. To better understand the action of the chaperonin, computational techiques that simulate folding under spatial confinement have been proposed [250]. Of interest for topologically complex folding is [198], where the confinement effect was studied by means of both lattice (see section 4.1) and off-lattice CG Gō models (see section 4.2.3). The methodology simply consists in limiting the space accessible to the protein chain to a cubic or spherical cavity, the size of which is chosen accordingly to the typical chaperonin size (few nm of radius). Moreover, in [177, 251], the chaperonin cage was modeled by means of a cylindric repulsive cavity, following [250]. The spatially confined folding of different entangled proteins was here studied using Clementi's CG Gō model. Overall, the steric effect of the cage is found to enhance the capability of folding of the protein models, reducing the folding time and promoting different routes with respect to bulk simulations.

4.4.5. Pore translocation. Crucial biological processes, such as mithocondrial import, or degradation, require proteins to be translocated through nanopores of 12–14 Å of minimum diameter. This size is too narrow to accomodate a native folded structure, and the protein needs to be unfolded by the action of unfoldases, that employ energy from ATP hydrolysis to perform this operation. It has been proposed that topologically complex proteins could prevent such a process, as it was shown that a mechanically tightened protein knot has a size comparable or larger than the pore diameter [28, 252]. It is therefore of interest to understand how the cellular machineries that operate protein translocation can cope with the presence of knots, untying the chain before it jams the pore channel.

In simulations this is attempted by means of an external repulsive potential that models the pore channel, and by applying a pulling force on one extremity of the protein model, to drive its translocation through the pore potential. Once again, because of computational time requirements, the preferential protein descriptions for these simulations, are simple C_α representations of self-entangled proteins. As a first example, in [253], specifically designed knotted polypeptides were represented by the model of [160, 166], and pulled by a constant force directed along the axis of the pore. The latter was represented by a cylindrical repulsive potential of length $L \approx 50$ Å and radius $\rho \approx 6.5$ Å. Different topologies were tested, showing that the knotted polypeptide can still translocate through the pore, but the rate of the process is significantly reduced.

A similar strategy was employed in [254], where models of real knotted proteins such as YibK, YbeA and MJ0366 were tested. Cieplak Gō-model was used for the representation of the protein, and a cylindrical repulsive potential modeled the pore, while a constant pulling force was used to drive the translocation dynamics. To mimic the transport into mitochondria, which is mediated by a short polypeptide sequence attached to the protein end, a 10-bead unstructured chain was added to the end of the protein to be pulled through the pore potential. The effective radius of the modeled pore was similar to the previous case, $\rho \approx 7$ Å, but the results demonstrated how deep protein knots can get stuck at the pore entrance. In [255] the same techniques and test systems were employed, but a periodic pulling force was applied to the protein terminals. This cyclic force, which is more realistic in reproducing an ATP-hydrolisis activated behaviour, could resolve the translocation jam by letting the knot slide off the chain.

A different pore-model has been introduced in [256], to represent the translocation into the proteasome. In this model, the entrance ring of the proteasome is described as a repulsive torus with major and minor radius of 13 and 6 Å respectively, placed at the end of a cylindric channel of radius $\rho \approx 7.5$ Å, which mimics the proteasome chamber. These two elements share the longitudinal axis, determining a cylindrical channel with a smooth, funneled entrance of 7 Å radius (the torus hole). Also here, proteins were described by means of Cieplak structure-based model, while the pulling of the protein inside the channel was performed either with a constant force or a constant velocity protocol. In [183] this methodology was used to simulate translocation of knotted globular proteins such as YibK and MJ0366 and transiently knotted polyglutamine tracts, assessing the dependence of the translocation process on protein topology, pulling force and pore model. Also a more rugged pore entrance was tested, by substituting the torus with 12 overlapping spheres of 6 Å of radius, arranged in circle. This model was also dynamic, implementing the temporary shrinking of the entrance spheres, three at a time. This behaviour is meant to reproduce the allosteric transformations of the proteins that compose the proteasome entrance ring, providing a more realistic description of the biological process.

5. Summary and conclusions

In the previous pages we have reviewed numerous computational techniques, proposed during decades of scientific research to shed light on a complex and fascinating topic such as that of self-entanglements in proteins. Here, we have focussed in particular on those exquisitely computational methodologies that have been developed and employed to comprehend the nature of these entanglements—their role, their occurrence, their formation, and their impact on a biomolecule's life cycle.

Albeit limited on the one hand by the vastness of the argument, and on the other hand by the necessity and intention to provide the Reader with a sufficiently agile document for a first experience in the field, the coverage and depth of this review is certainly far from completeness and high-resolution accuracy. It is nonetheless our hope that the present review will spark interest in researchers, particularly young ones, and motivate them to pursue a research activity aimed at the discovery of the many different facets encountered when tackling the study of self-entangled proteins.

The wide phenomenology of these systems, and the correspondingly broad range of computational methods required to make sense of their properties, are strong indications of two crucial aspects. Firstly, we notice that the community active in this field is in fast development, and that the number of available instruments keeps steadily increasing, thereby opening a spectrum of possibilities that were unthinkable only a few years ago; this is largely due not only thanks to the quick advancements in computer science, rather also, and maybe most prominently, to the coordinated efforts of biologists, chemists, and physicists. Secondly, it is easy to see how the different models employed for the study of self-entangled proteins constitute a very heterogeneous substrate, from which a variety of results could be obtained, sometimes reaching also jarring contradictions. This is in part due to computational limitations, which impose radical approximations to meet the temporal requirements of simulating complex biophysical processes such as knotted protein folding. Consequently, the questions raised by the existence of native protein entanglements, despite the undoubtedly remarkable progresses attained insofar, are far from being answered, and a long road still needs to be travelled.

In fact, the computational limitations will not be solved, at least for few years, by the progress of computer hardware alone, since the gap between the biological timescales of interest and the ones accessible with accurate MD models still ranges for few orders of magnitude. The greatest advancements will thus come from improved algorithms, that is, smarter and more efficient ways of modelling these systems, sampling their conformations, simulating their behaviour, and extracting relevant information from simplified *in silico* models.

To this end, one of the most promising strategies—in this particular subfield as well as in many others of computational biophysics—is to employ a *multi-scale* description of the system, combining models and methods that use different level of detail in a hierarchical and/or concurrent manner. This approach enables one to attain an improved qualitative and often quantitative picture of the process as a whole, taking advantage of the efficiency of coarse-grained models as well as extracting accurate physico-chemical information thanks to the atomistic or, in general, higher-resolution detail.

In this respect, this review has aimed at summarising a picture of the whole set of instruments currently available to undertake this endeavour, providing the appropriate references to combine different techniques and tackle a specific problem. An encompassing overview of computational models and methods also entails the chance to facilitate the community in spotting possible weaknesses in the current stage of the simulation and modelling technologies, thereby better directing the efforts to resolve them and complement the arsenal of available techniques with novel, sharper ones.

Acknowledgments

The authors would like to acknowledge networking support by the COST Action CA17139. CP acknowledges funding from the European Unions Horizon 2020 research and innovation programme under the GOKNOT Marie Skłodowska-Curie Grant Agreement No. 796969. The authors thank L Tubiana and Y Zhao for a critical reading of the manuscript.

ORCID iDs

Claudio Perego https://orcid.org/0000-0001-8885-3080
Raffaello Potestio https://orcid.org/0000-0001-6408-9380

References

[1] Adams C C 2004 *The Knot Book: an Elementary Introduction to the Mathematical Theory of Knots* (Providence, RI: American Mathematical Society)
[2] Cromwell P R 2004 *Knots and Links* (Cambridge: Cambridge University Press)
[3] Lickorish W R 2012 *An Introduction to Knot Theory* vol 175 (New York: Springer)
[4] Livingston C 1993 *Math. Assoc. Am.* **106** 1–255
[5] Marenduzzo D, Orlandini E, Stasiak A, Sumners D W, Tubiana L and Micheletti C 2009 *Proc. Natl Acad. Sci.* **106** 22269–74
[6] Pommier Y, Sun Y, Huang S Y N and Nitiss J L 2016 *Nat. Rev. Mol. Cell Biol.* **17** 703
[7] Virnau P, Mallam A and Jackson S 2011 *J. Phys.: Condens. Matter* **23** 033101
[8] Lim N C and Jackson S E 2015 *J. Mol. Biol.* **427** 248–58
[9] Schmid F X 2015 *J. Mol. Biol.* **427** 225–7
[10] Jackson S E, Suma A and Micheletti C 2017 *Curr. Opin. Struct. Biol.* **42** 6–14
[11] Dabrowski-Tumanski P and Sułkowska J I 2017 *Polymers* **9** 454
[12] Finkelstein A V and Ptitsyn O B (ed) 2016 *Protein Physics: A Course of Lectures (Second Edition)* 2nd edn (Amsterdam: Academic) p iv
[13] Dario Meluzzi D E S and Arya G 2010 *Annu. Rev. Biophys.* **39** 349–66
[14] Tompa P 2012 *Trends Biochem. Sci.* **37** 509–16
[15] Oldfield C J and Dunker A K 2014 *Annu. Rev. Biochem.* **83** 553–84
[16] Uversky V N 2014 *Chem. Rev.* **114** 6557–60
[17] van der Lee R *et al* 2014 *Chem. Rev.* **114** 6589–631
[18] Wright P E and Dyson H J 2015 *Nat. Rev. Mol. Cell Biol.* **16** 18–29
[19] Daily-Diamond C A, Gregg C E and O'Reilly O M 2017 *Proc. R. Soc.* A **473** 20160770
[20] Crippen G M 1974 *J. Theor. Biol.* **45** 327–38
[21] Tramontano A, Leplae R and Morea V 2001 *Proteins: Struct. Funct. Bioinform.* **45** 22–38
[22] Rohl C A, Khatib F and Weirauch M T 2006 *Bioinformatics* **22** e252–9
[23] Mallam A L 2009 *FEBS J.* **276** 365–75
[24] Mansfield M L 1994 *Nat. Struct. Mol. Biol.* **1** 213–4
[25] Richardson J S 1977 *Nature* **268** 495–500
[26] Mansfield M L 1997 *Nat. Struct. Mol. Biol.* **4** 166–7
[27] Taylor W R 2000 *Nature* **406** 916–9
[28] Virnau P, Mirny L A and Kardar M 2006 *PLOS Comput. Biol.* **2** 1–6
[29] Taylor W R 2007 *Comput. Biol. Chem.* **31** 151–62
[30] Sułkowska J I, Rawdon E J, Millett K C, Onuchic J N and Stasiak A 2012 *Proc. Natl Acad. Sci. USA* **109** E1715–23
[31] Faísca P F 2015 *Comput. Struct. Biotechnol. J.* **13** 459–68
[32] Bölinger D, Sułkowska J I, Hsu H P, Mirny L A, Kardar M, Onuchic J N and Virnau P 2010 *PLoS Comput. Biol.* **6** e1000731

[33] Baiesi M, Orlandini E, Trovato A and Seno F 2016 *Sci. Rep.* **6** 33872
[34] Baiesi M, Orlandini E, Seno F and Trovato A 2017 *J. Phys. A: Math. Theor.* **50** 504001
[35] Potestio R, Micheletti C and Orland H 2010 *PLoS Comput. Biol.* **6** 1–10
[36] Lua R C and Grosberg A Y 2006 *PLoS Comput. Biol.* **2** 1–8
[37] Lim N C H and Jackson S E 2015 *J. Phys.: Condens. Matter* **27** 354101
[38] Mallam A L and Jackson S E 2005 *J. Mol. Biol.* **346** 1409–21
[39] Mallam A L and Jackson S E 2006 *J. Mol. Biol.* **359** 1420–36
[40] Mallam A L and Jackson S E 2007 *J. Mol. Biol.* **366** 650–65
[41] Mallam A L, Onuoha S C, Grossmann J G and Jackson S E 2008 *Mol. Cell* **30** 642–8
[42] King N P, Jacobitz A W, Sawaya M R, Goldschmidt L and Yeates T O 2010 *Proc. Natl Acad. Sci.* **107** 20732–7
[43] Mallam A L and Jackson S E 2012 *Nat. Chem. Biol.* **8** 147–53
[44] Wang I, Chen S Y and Hsu S T D 2015 *J. Phys. Chem.* B **119** 4359–70
[45] Wang L W, Liu Y N, Lyu P C, Jackson S E and Hsu S T D 2015 *J. Phys.: Condens. Matter* **27** 354106
[46] Lou S C, Wetzel S, Zhang H, Crone E W, Lee Y T, Jackson S E and Hsu S T D 2016 *J. Mol. Biol.* **428** 2507–20
[47] Wang I, Chen S Y and Hsu S T D 2016 *Sci. Rep.* **6** 31514
[48] Chwastyk M and Cieplak M 2015 *J. Phys.: Condens. Matter* **27** 354105
[49] Chwastyk M and Cieplak M 2015 *J. Chem. Phys.* **143** 045101
[50] Dabrowski-Tumanski P, Piejko M, Niewieczerzal S, Stasiak A and Sulkowska J I 2018 *J. Phys. Chem.* B **122** 11616–25
[51] Baiesi M, Orlandini E, Seno F and Trovato A 2018 (arXiv:1809.02173)
[52] Kaiser C M, Goldman D H, Chodera J D, Tinoco I and Bustamante C 2011 *Science* **334** 1723–7
[53] Beccara S A, Škrbić T, Covino R, Micheletti C and Faccioli P 2013 *PLoS Comput. Biol.* **9** 1–9
[54] Sułkowska J I, Noel J K, Ramírez-Sarmiento C A, Rawdon E J, Millett K C and Onuchic J N 2013 *Biochem. Soc. Trans.* **41** 523–7
[55] Najafi S and Potestio R 2015 *J. Chem. Phys.* **143** 243121
[56] King N P, Yeates E O and Yeates T O 2007 *J. Mol. Biol.* **373** 153–66
[57] Sułkowska J I, Sułkowski P and Onuchic J 2009 *Proc. Natl Acad. Sci.* **106** 3119–24
[58] Sułkowska J I, Sułkowski P and Onuchic J N 2009 *Phys. Rev. Lett.* **103** 268103
[59] Sikora M, Sułkowska J I and Cieplak M 2009 *PLoS Comput. Biol.* **5** 1–15
[60] Millet K C 2010 *J. Knot Theory Ramifications* **19** 601–15
[61] He C, Lamour G, Xiao A, Gsponer J and Li H 2014 *J. Am. Chem. Soc.* **136** 11946–55
[62] Capraro D T and Jennings P A 2016 *Biophys. J.* **110** 1044–51
[63] Ziegler F, Lim N C H, Mandal S S, Pelz B, Ng W P, Schlierf M, Jackson S E and Rief M 2016 *Proc. Natl Acad. Sci.* **113** 7533–8
[64] Flapan E, He A and Wong H 2019 *Proc. Natl Acad. Sci.* **116** 9360–9
[65] Kauffman L 2001 *Knots and Physics* (*Series on Knots and Everything*) (London: World Scientific)
[66] Orlandini E and Whittington S G 2007 *Rev. Mod. Phys.* **79** 611–42
[67] Micheletti C, Marenduzzo D and Orlandini E 2011 *Phys. Rep.* **504** 1–73
[68] Turaev V 2012 *Osaka J. Math.* **49** 195–223
[69] Goundaroulis D, Gügümcü N, Lambropoulou S, Dorier J, Stasiak A and Kauffman L 2017 *Polymers* **9** 444
[70] Goundaroulis D, Dorier J, Benedetti F and Stasiak A 2017 *Sci. Rep.* **7** 6309
[71] Dorier J, Goundaroulis D, Benedetti F and Stasiak A 2018 *Bioinformatics* **34** 3402–4
[72] Alexander K, Taylor A J and Dennis M R 2017 *Sci. Rep.* **7** 42300
[73] Luecke J 1989 *J. Am. Math. Soc.* **2** 371–415
[74] Rolfsen D 1976 *Knots and Links* vol 346 (Providence, RI: American Mathematical Society)
[75] Hoste J, Thistlethwaite M and Weeks J 1999 Knotscape (www.math.utk.edu/∼morwen/knotscape.html)
[76] Fox R H and Artin E 1948 *Ann. Math.* 979–90
[77] Reidemeister H 1932 *Knotentheorie* (Berlin: Springer)
[78] Alexander J W 1928 *Trans. Am. Math. Soc.* **30** 275–306
[79] Crowell R H and Fox R H 1963 *Introduction to Knot Theory* (*Sociedad Colombiana de Matematicas*) (New York: Springer)
Crowell R H and Fox R H 2012 *Springer Science & Business Media*
[80] Wu F Y 1992 *Rev. Mod. Phys.* **64** 1099–131
[81] Jones V F 1985 *Bull. Am. Math. Soc.* **12** 103–11
[82] Lickorish W R and Millett K C 1987 *Topology* **26** 107–41
[83] Vologodskii A, Lukashin A, Kamenetskii M and Anshelevich V 1974 *Sov. J. Exp. Theor. Phys.* **39** 1059
[84] Koniaris K and Muthukumar M 1991 *J. Chem. Phys.* **95** 2873–81
[85] Kolesov G, Virnau P, Kardar M and Mirny L A 2007 *Nucleic Acids Res.* **35** W425–8
[86] Lai Y L, Chen C C and Hwang J K 2012 *Nucleic Acids Res.* **40** W228–31
[87] Jamroz M, Niemyska W, Rawdon E J, Stasiak A, Millett K C, Sułkowski P and Sułkowska J I 2015 *Nucleic Acids Res.* **43** D306–14
[88] Khatib F, Weirauch M T and Rohl C A 2006 *Bioinformatics* **22** e252
[89] Lua R C 2012 *Bioinformatics* **28** 2069
[90] Dabrowski-Tumanski P, Sułkowska J I, Rubach P, Stasiak A, Goundaroulis D, Dorier J, Sułkowski P, Millett K C and Rawdon E J 2018 *Nucleic Acids Res.* **47** D367–75
[91] Marcone B, Orlandini E, Stella A L and Zonta F 2007 *Phys. Rev.* E **75** 041105
[92] Orlandini E, Stella A L and Vanderzande C 2009 *Phys. Biol.* **6** 025012
[93] Tubiana L, Orlandini E and Micheletti C 2011 *Prog. Theor. Phys. Suppl.* **191** 192
[94] Taylor W R 2005 *Physical and Numerical Models in Knot Theory* (Singapore: World Scientific) pp 171–202
[95] Norcross T S and Yeates T O 2006 *J. Mol. Biol.* **362** 605–21
[96] Rawdon E J, Millett K C, Sułkowska J I and Stasiak A 2013 *Biochem. Soc. Trans.* **41** 538–41
[97] Rawdon E J, Millett K C and Stasiak A 2015 *Sci. Rep.* **5** 8928
[98] Virnau P, Kantor Y and Kardar M 2005 *J. Am. Chem. Soc.* **127** 15102–6
[99] Katritch V, Olson W K, Vologodskii A, Dubochet J and Stasiak A 2000 *Phys. Rev.* E **61** 5545–9
[100] Millett K, Dobay A and Stasiak A 2005 *Macromolecules* **38** 601–6
[101] Marcone B, Orlandini E, Stella A L and Zonta F 2005 *J. Phys. A: Math. Gen.* **38** L15
[102] Millett K C, Rawdon E J, Stasiak A and Sułkowska J I 2013 *Biochem. Soc. Trans.* **41** 533–7
[103] Humphrey W, Dalke A and Schulten K 1996 *J. Mol. Graph.* **14** 33–8
[104] Yeates T O, Norcross T S and King N P 2007 *Curr. Opin. Chem. Biol.* **11** 595–603
[105] Noel J K, Sułkowska J I and Onuchic J N 2010 *Proc. Natl Acad. Sci.* **107** 15403–8
[106] Liang C and Mislow K 1994 *J. Am. Chem. Soc.* **116** 11189–90
[107] Taylor W R and Lin K 2003 *Nature* **421** 25

[108] Daly N L and Craik D J 2011 *Curr. Opin. Chem. Biol.* **15** 362–8
[109] Niemyska W, Dabrowski-Tumanski P, Kadlof M, Haglund E, Sułkowski P and Sułkowska J I 2016 *Sci. Rep.* **6** 36895
[110] Rebuffat S, Blond A, Destoumieux-Garzón D, Goulard C and Peduzzi J 2004 *Curr. Protein Peptide Sci.* **5** 383–91
[111] Haglund E, Sułkowska J I, He Z, Feng G S, Jennings P A and Onuchic J N 2012 *PLoS One* **7** e45654
[112] Dabrowski-Tumanski P, Niemyska W, Pasznik P and Sułkowska J I 2016 *Nucleic Acids Res.* **44** W383
[113] Haglund E, Sułkowska J I, Noel J K, Lammert H, Onuchic J N and Jennings P A 2014 *PLoS Comput. Biol.* **10** 1–11
[114] Haglund E, Pilko A, Wollman R, Jennings P A and Onuchic J N 2017 *J. Phys. Chem.* B **121** 706–18
[115] Perego C and Potestio R 2019 *Biophys. J.* accepted (https://doi.org/10.1016/j.bpj.2019.05.025)
[116] Dabrowski-Tumanski P and Sułkowska J I 2017 *Proc. Natl Acad. Sci.* **114** 3415–20
[117] Duda R L 1998 *Cell* **94** 55–60
[118] Cao Z, Roszak A W, Gourlay L J, Lindsay J G and Isaacs N W 2005 *Structure* **13** 1661–4
[119] Dabrowski-Tumanski P, Sułkowska J I, Jarmolinska A I, Niemyska W, Rawdon E J and Millett K C 2016 *Nucleic Acids Res.* **45** D243–9
[120] Panagiotou E, Kröger M and Millett K C 2013 *Phys. Rev.* E **88** 062604
[121] Panagiotou E and Plaxco K W 2018 (arXiv:1812.08721)
[122] Zhao Y, Chwastyk M and Cieplak M 2017 *J. Chem. Phys.* **146** 225102
[123] Chen C C, Hwang J K and Yang J M 2009 *BMC Bioinform.* **10** 366
[124] Postic G, Gracy J, Périn C, Chiche L and Gelly J C 2017 *Nucleic Acids Res.* **46** D454–8
[125] Montroll E W 1950 *J. Chem. Phys.* **18** 734–43
[126] Kremer K and Binder K 1988 *Comput. Phys. Rep.* **7** 259–310
[127] Wall F T, Hiller L A Jr and Wheeler D J 1954 *J. Chem. Phys.* **22** 1036–41
[128] Shakhnovich E and Gutin A 1990 *J. Chem. Phys.* **93** 5967–71
[129] Quake S R 1995 *Phys. Rev.* E **52** 1176–80
[130] Binder K and Paul W 2008 *Macromolecules* **41** 4537–50
[131] Wüst T and Landau D P 2009 *Phys. Rev. Lett.* **102** 178101
[132] Metropolis N, Rosenbluth A W, Rosenbluth M N, Teller A H and Teller E 1953 *J. Chem. Phys.* **21** 1087–92
[133] Meirovitch H and Lim H A 1989 *J. Chem. Phys.* **91** 2544–54
[134] Chang I and Meirovitch H 1993 *Phys. Rev.* E **48** 3656–60
[135] Gō N and Taketomi H 1978 *Proc. Natl Acad. Sci.* **75** 559–63
[136] Miyazawa S and Jernigan R L 1985 *Macromolecules* **18** 534–52
[137] Lau K F and Dill K A 1989 *Macromolecules* **22** 3986–97
[138] Shakhnovich E I and Gutin A M 1993 *Proc. Natl Acad. Sci.* **90** 7195–9
[139] Dill K A, Bromberg S, Yue K, Chan H S, Ftebig K M, Yee D P and Thomas P D 1995 *Protein Sci.* **4** 561–602
[140] Chan H S and Dill K A 1991 *J. Chem. Phys.* **95** 3775–87
[141] Chan H S and Dill K A 1998 *Proteins: Struct. Funct. Bioinform.* **30** 2–33
[142] Dill K A 1999 *Protein Sci.* **8** 1166–80
[143] Wüst T, Reith D and Virnau P 2015 *Phys. Rev. Lett.* **114** 028102
[144] Wüst T and Landau D P 2012 *J. Chem. Phys.* **137** 064903
[145] Taketomi H, Ueda Y and Gō N 1975 *Int. J. Peptide Protein Res.* **7** 445–59
[146] Cieplak M, Hoang T X and Li M S 1999 *Phys. Rev. Lett.* **83** 1684–7
[147] Pande V S and Rokhsar D S 1999 *Proc. Natl Acad. Sci.* **96** 1273–8
[148] Faisca P and da Gama M T 2005 *Biophys. Chem.* **115** 169–75
[149] Travasso R D M, da Gama M M T and Fasca P F N 2007 *J. Chem. Phys.* **127** 145106
[150] Travasso R D M, Fasca P F N and Rey A 2010 *J. Chem. Phys.* **133** 125102
[151] Onuchic J N and Wolynes P G 2004 *Curr. Opin. Struct. Biol.* **14** 70–5
[152] Fasca P F N, Travasso R D M, Charters T, Nunes A and Cieplak M 2010 *Phys. Biol.* **7** 016009
[153] Soler M A and Fasca P F N 2013 *PLoS One* **8** 1–10
[154] Soler M A, Nunes A and Fasca P F N 2014 *J. Chem. Phys.* **141** 025101
[155] Wallin S, Zeldovich K B and Shakhnovich E I 2007 *J. Mol. Biol.* **368** 884–93
[156] Škrbić T, Micheletti C and Faccioli P 2012 *PLoS Comput. Biol.* **8** e1002504
[157] Covino R, Škrbi T, Beccara S A, Faccioli P and Micheletti C 2014 *Biomolecules* **4** 1–19
[158] Soler M A and Fasca P F N 2012 *PLoS One* **7** 1–13
[159] Honeycutt J D and Thirumalai D 1992 *Biopolymers* **32** 695–709
[160] Veitshans T, Klimov D and Thirumalai D 1997 *Folding Des.* **2** 1–22
[161] Irbäck A, Peterson C and Potthast F 1997 *Phys. Rev.* E **55** 860–7
[162] Irbäck A, Peterson C, Potthast F and Sommelius O 1997 *J. Chem. Phys.* **107** 273–82
[163] Klimov D K and Thirumalai D 1997 *Phys. Rev. Lett.* **79** 317–20
[164] Hoang T X and Cieplak M 2000 *J. Chem. Phys.* **112** 6851–62
[165] Hoang T X and Cieplak M 2000 *J. Chem. Phys.* **113** 8319–28
[166] Sorenson J M and Head-Gordon T 1999 *Proteins: Struct. Funct. Bioinform.* **37** 582–91
[167] Tsai J, Taylor R, Chothia C and Gerstein M 1999 *J. Mol. Biol.* **290** 253–66
[168] Sułkowska J I and Cieplak M 2008 *Biophys. J.* **95** 3174–91
[169] Noel J K, Whitford P C and Onuchic J N 2012 *J. Phys. Chem.* B **116** 8692–702
[170] Wołek K, Gòmez-Sicilia À and Cieplak M 2015 *J. Chem. Phys.* **143** 243105
[171] Dabrowski-Tumanski P, Jarmolinska A I and Sułkowska J I 2015 *J. Phys.: Condens. Matter* **27** 354109
[172] Clementi C, Nymeyer H and Onuchic J N 2000 *J. Mol. Biol.* **298** 937–53
[173] Lammert H, Schug A and Onuchic J N 2009 *Proteins: Struct. Funct. Bioinform.* **77** 881–91
[174] Noel J K, Whitford P C, Sanbonmatsu K Y and Onuchic J N 2010 *Nucleic Acids Res.* **38** W657–61
[175] Noel J K, Levi M, Raghunathan M, Lammert H, Hayes R L, Onuchic J N and Whitford P C 2016 *PLoS Comput. Biol.* **12** 1–14
[176] Sułkowska J I, Noel J K and Onuchic J N 2012 *Proc. Natl Acad. Sci.* **109** 17783–8
[177] Zhao Y, Dabrowski-Tumanski P, Niewieczerzal S and Sułkowska J I 2018 *PLoS Comput. Biol.* **14** 1–20
[178] Zhao Y and Cieplak M 2018 *Proteins: Struct. Funct. Bioinform.* **86** 945–55
[179] Cieplak M and Hoang T X 2003 *Biophys. J.* **84** 475–88
[180] Kwiecińska J I and Cieplak M 2005 *J. Phys.: Condens. Matter* **17** S1565
[181] Sułkowska J I, Sułkowski P, Szymczak P and Cieplak M 2010 *J. Am. Chem. Soc.* **132** 13954–6
[182] Zhao Y, Chwastyk M and Cieplak M 2017 *Sci. Rep.* **7** 39851
[183] Wojciechowski M, Gómez-Sicilia À, Carrión-Vázquez M and Cieplak M 2016 *Mol. Biosyst.* **12** 2700–12
[184] He C, Genchev G Z, Lu H and Li H 2012 *J. Am. Chem. Soc.* **134** 10428–35
[185] Florin E, Moy V and Gaub H 1994 *Science* **264** 415–7

[186] Kellermayer M S Z, Smith S B, Granzier H L and Bustamante C 1997 *Science* **276** 1112–6
[187] Erickson H P 1997 *Science* **276** 1090–2
[188] Mitsui K, Hara M and Ikai A 1996 *FEBS Lett.* **385** 29–33
[189] Grubmüller H, Heymann B and Tavan P 1996 *Science* **271** 997–9
[190] Sułkowska J I and Cieplak M 2007 *J. Phys.: Condens. Matter* **19** 283201
[191] Sułkowska J I and Cieplak M 2008 *Biophys. J.* **94** 6–13
[192] Sułkowska J I, Sułkowski P, Szymczak P and Cieplak M 2008 *Proc. Natl Acad. Sci.* **105** 19714–9
[193] Sułkowska J I, Sułkowski P, Szymczak P and Cieplak M 2008 *Phys. Rev. Lett.* **100** 058106
[194] Sikora M and Cieplak M 2011 *Proteins: Struct. Funct. Bioinform.* **79** 1786–99
[195] Wołek K and Cieplak M 2016 *J. Chem. Phys.* **144** 185102
[196] Prieto L, de Sancho D and Rey A 2005 *J. Chem. Phys.* **123** 154903
[197] Prieto L and Rey A 2007 *J. Chem. Phys.* **127** 175101
[198] Soler M A, Rey A and Faísca P F 2016 *Phys. Chem. Chem. Phys.* **18** 26391–403
[199] Li W, Terakawa T, Wang W and Takada S 2012 *Proc. Natl Acad. Sci.* **109** 17789–94
[200] Li W, Wolynes P G and Takada S 2011 *Proc. Natl Acad. Sci.* **108** 3504–9
[201] Li W and Takada S 2010 *Biophys. J.* **99** 3029–37
[202] Terakawa T and Takada S 2011 *Biophys. J.* **101** 1450–8
[203] Best R B, Hummer G and Eaton W A 2013 *Proc. Natl Acad. Sci.* **110** 17874–9
[204] Fasca P F N, Nunes A, Travasso R D and Shakhnovich E I 2010 *Protein Sci.* **19** 2196–209
[205] Karanicolas J and Brooks C L 2002 *Protein Sci.* **11** 2351–61
[206] Best R B, Chen Y G and Hummer G 2005 *Structure* **13** 1755–63
[207] Kim Y C and Hummer G 2008 *J. Mol. Biol.* **375** 1416–33
[208] Miyazawa S and Jernigan R L 1996 *J. Mol. Biol.* **256** 623–44
[209] Jorgensen W L and Tirado-Rives J 1996 *J. Phys. Chem.* **100** 14508–13
[210] Grest G S and Kremer K 1986 *Phys. Rev.* A **33** 3628–31
[211] Weeks J D, Chandler D and Andersen H C 1971 *J. Chem. Phys.* **54** 5237–47
[212] Tuszynska I and Bujnicki J M 2010 *J. Biomol. Struct. Dyn.* **27** 511–20
[213] Boniecki M, Rotkiewicz P, Skolnick J and Kolinski A 2003 *J. Comput.: Aided Mol. Des.* **17** 725–38
[214] Kolinski A, Jaroszewski L, Rotkiewicz P and Skolnick J 1998 *J. Phys. Chem.* B **102** 4628–37
[215] Liwo A, Arłukowicz P, Czaplewski C, Ołdziej S, Pillardy J and Scheraga H A 2002 *Proc. Nat.l Acad. Sci.* **99** 1937–42
[216] Liwo A, Khalili M and Scheraga H A 2005 *Proc. Natl Acad. Sci.* **102** 2362–7
[217] Liwo A, Khalili M, Czaplewski C, Kalinowski S, Ołdziej S, Wachucik K and Scheraga H A 2007 *J. Phys. Chem.* B **111** 260–85
[218] Prentiss M C, Wales D J and Wolynes P G 2010 *PLoS Comput. Biol.* **6** 1–12
[219] Friedrichs M S and Wolynes P G 1989 *Science* **246** 371–3
[220] Prentiss M C, Hardin C, Eastwood M P, Zong C and Wolynes P G 2006 *J. Chem. Theory Comput.* **2** 705–16
[221] Li Z and Scheraga H A 1987 *Proc. Natl Acad. Sci.* **84** 6611–5
[222] Henkelman G and Jnsson H 2000 *J. Chem. Phys.* **113** 9978–85
[223] Prentiss M C, Wales D J and Wolynes P G 2008 *J. Chem. Phys.* **128** 225106
[224] Wales D J 2002 *Mol. Phys.* **100** 3285–305
[225] Shaw D E *et al* 2008 *Commun. ACM* **51** 91–7
[226] Ponder J W and Case D A 2003 *Protein Simulations* (*Advances in Protein Chemistry* vol 66) (New York: Academic) pp 27–85
[227] Mackerell A D Jr, Feig M and Brooks C L III 2004 *J. Comput. Chem.* **25** 1400–15
[228] Jorgensen W L and Tirado-Rives J 2005 *Proc. Natl Acad. Sci.* **102** 6665–70
[229] Kollman P A *et al* 2000 *Acc. Chem. Res.* **33** 889–97
[230] Tsui V and Case D A 2000 *Biopolymers* **56** 275–91
[231] Onufriev A, Bashford D and Case D A 2004 *Proteins: Struct. Funct. Bioinform.* **55** 383–94
[232] Lu H, Isralewitz B, Krammer A, Vogel V and Schulten K 1998 *Biophys. J.* **75** 66271
[233] Ohta S, Alam M T, Arakawa H and Ikai A 2004 *Biophys. J.* **87** 4007–20
[234] Alam M T, Yamada T, Carlsson U and Ikai A 2002 *FEBS Lett.* **519** 35–40
[235] Dzubiella J 2013 *J. Phys. Chem. Lett.* **4** 1829–33
[236] Bornschlögl T, Anstrom D M, Mey E, Dzubiella J, Rief M and Forest K T 2009 *Biophys. J.* **96** 1508–14
[237] Zhou Y and Linhananta A 2002 *Proteins: Struct. Funct. Bioinform.* **47** 154–62
[238] Whitford P C, Noel J K, Gosavi S, Schug A, Sanbonmatsu K Y and Onuchic J N 2009 *Proteins: Struct. Funct. Bioinform.* **75** 430–41
[239] Hornak V, Abel R, Okur A, Strockbine B, Roitberg A and Simmerling C 2006 *Proteins: Struct. Funct. Bioinform.* **65** 712–25
[240] Beccara S A, Škrbić T, Covino R and Faccioli P 2012 *Proc. Natl Acad. Sci.* **109** 2330–5
[241] Paci E and Karplus M 1999 *J. Mol. Biol.* **288** 441–59
[242] Faccioli P, Sega M, Pederiva F and Orland H 2006 *Phys. Rev. Lett.* **97** 108101
[243] Sega M, Faccioli P, Pederiva F, Garberoglio G and Orland H 2007 *Phys. Rev. Lett.* **99** 118102
[244] Noel J K, Onuchic J N and Sułkowska J I 2013 *J. Phys. Chem. Lett.* **4** 3570–3
[245] Jorgensen W L, Chandrasekhar J, Madura J D, Impey R W and Klein M L 1983 *J. Chem. Phys.* **79** 926–35
[246] Rżycki B, Mioduszewski Ł and Cieplak M 2014 *Proteins: Struct. Funct. Bioinform.* **82** 3144–53
[247] Cieplak M, Allan D B, Leheny R L and Reich D H 2014 *Langmuir* **30** 12888–96
[248] Kyte J and Doolittle R F 1982 *J. Mol. Biol.* **157** 105–32
[249] Mayhew M, da Silva A C, Martin J, Erdjument-Bromage H, Tempst P and Hartl F U 1996 *Nature* **379** 420
[250] Takagi F, Koga N and Takada S 2003 *Proc. Natl Acad. Sci.* **100** 11367–72
[251] Niewieczerzal S and Sułkowska J I 2017 *PLoS One* **12** 1–23
[252] Szymczak P 2013 *Biochem. Soc. Trans.* **41** 620–4
[253] Huang L and Makarov D E 2008 *J. Chem. Phys.* **129** 121107
[254] Szymczak P 2014 *Eur. Phys. J. Spec. Top.* **223** 1805–12
[255] Szymczak P 2016 *Sci. Rep.* **6** 21702
[256] Wojciechowski M, Szymczak P, Carrión-Vázquez M and Cieplak M 2014 *Biophys. J.* **107** 1661–8