POLISH ACADEMY OF SCIENCES
INSTITUTE OF PHYSICS

Established in 1920 by
the Polish Physical Society

ACTA PHYSICA POLONICA

- General Physics
- Atomic and Molecular Physics
- Condensed Matter
- Optics and Quantum Optics
- Quantum Information
- Biophysics
- Applied Physics

Special Issue — in Memory of Professor Marek Cieplak (1950–2021)

RECOGNIZED BY THE EUROPEAN PHYSICAL SOCIETY

Volume 145 — Number 3, Warsaw, March 2024

PL ISSN 0587-4246 ATPLB6 145 (3) S1–S70 (2024)

In Memory

Professor Marek Cieplak

(1950–2021)

Marek Cieplak, born December 8, 1950, passed away on December 31, 2021. He was a Professor of Physics, a lecturer of summer courses in physics at Rutgers University and Johns Hopkins University in the USA, a member of the Scientific Council of the Institute of Physics of the Polish Academy of Sciences (PAS), initiator of biophysical topics at the Institute of Physics PAS, founder and head of the Laboratory of Biological Physics, spiritus movens of the series of international scientific conferences *"Biomolecules and Nanostructures"*, **member of scientific societies and editorial boards of scientific journals such as** *Journal of Physics: Condensed Matter* **and** *Acta Physica Polonica A*.

Figure 1: The commemorated Marek. Photo taken from the family resources with permission.

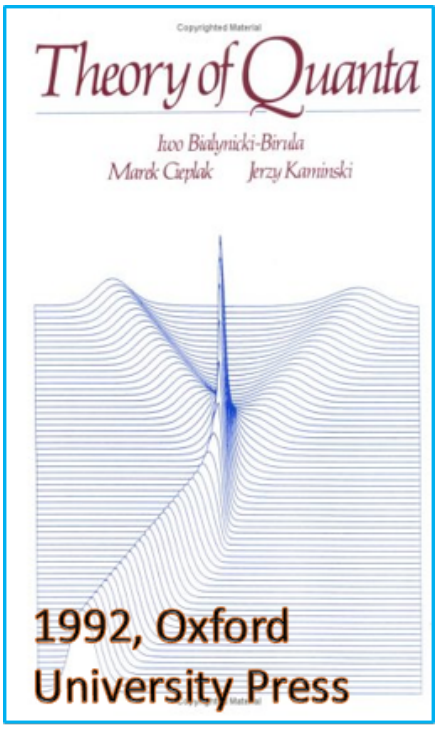


Figure 2: (Top panel) Marta Gieburowska (Cieplak) and Marek Cieplak, students of the Faculty of Physics, University of Warsaw in the Large Experimental Hall (SDD). (Bottom panel) Marek — as a student — got involved in helping Professor Białynicki-Birula to prepare notes for his lecture on quantum mechanics. These notes eventually evolved into a book, which Marek co-authored with Professor Białynicki-Birula and Professor Jerzy Kamiński, many years later. Photos taken from the family resources with permission.

This issue collects several articles that aim to symbolically commemorate Marek Cieplak, not only because of his profession. All articles went through the peer review process, and the authors and reviewers deserve thanks.

The unusual and unique feature of this issue is not only its scientific content, but also the inclusion of several personal recollections. Below are short texts about Marek, as he is remembered by colleagues, former PhD students (now full-fledged scientists) and his friends. The editors appreciate their presence and sharing! It is worth reading these personal recollections because they outline the broader profile of Marek, his human qualities for which he was appreciated.

The editors

◇ ◇ ◇

Marek Cieplak was a man full of passion — both in physics and beyond. You could feel the strength of his personality already as you approached his office, which stood out against the gray and somewhat colorless corridors of the Institute of Physics of the Polish Academy of Sciences — the door was always open, and inside it was filled with flowers, paintings, posters, and figurines. Once you stepped in, Marek would tear himself away from his inseparable computer and bombard you with a vast number of scientific ideas, intertwined with the latest cultural curiosities, books he had read, movies he had seen, or travels, which he was passionate about. Marek's research was just as eclectic — from work on spin glasses, to the analysis of river networks, optimal paths in disordered systems, fluid invasion into the porous media, and the microscopic origins of friction. In the last 25 years, his main focus has been biological physics, in particular numerical models of protein folding and unfolding, interpretation of genetic microarray data or knots and entanglements in biomolecules. We collaborated on many of these subjects, publishing 15 papers together.

I will miss his energy, passion, and drive to keep moving forward. I am grateful for the time we shared and for the inspiration he brought to our work.

Piotr Szymczak

◇ ◇ ◇

Professor Marek Cieplak, as we all know, was an extraordinary character, a person who always had his own opinion, always had substantive arguments and was able to defend them to the end. This was also Marek's whole life... and this state of mind was contagious to others — to me certainly. I had the pleasure of being a doctoral student of Professor Cieplak. Marek's knowledge and behavior had a huge impact on my scientific career and on my decision to follow this path. Marek was 110% committed to every doctorate, and he demanded the same! He was at work every day, he came every day to talk about both science and life. His room at the Institute of Physics was full of flowers, paintings and souvenirs brought from his travels around the world, as well as two, sports items: a tennis racket and a bicycle. Marek cycled to work whenever he could. His wife and daughters painted the paintings. This room looked like a real museum — it always encouraged me, even when Marek asked rhetorically: "Why don't computers count when it's cold outside, or why don't I have the conscience to freeze them like that?" He did it jokingly, but you could still feel a shiver of fear. Yes, he definitely made sure that the learning progressed and that every moment was well-planned and used. I remember well that before I left for trekking in the Himalayas, he

made sure that I had a copy of my data and that I had planned the calculations appropriately so that the computers would not "freeze" during my 3-week trip. On the other hand, after returning I saw how happy he was, not only that the data had been calculated, but that I had returned in one piece. Marek was caring; he was hot-tempered, but he always meant well and strived for objective truth.

From a scientific point of view, I think it should be strongly emphasized that the topic of non-trivial topology in proteins, which I continue to develop (as do a lot of scientists around the world), was born by accident, but this accident was created by Marek. One day in 2006, just before the holidays, Marek came up with the idea to determine the free energy landscape of proteins from the point of view of mechanical resistance. This idea probably resulted from the fact that at that time single-molecule optical tweezers, which trapped micron-sized silica beads (of diameter range of 0.2–5 m) to exert forces on the system of interest, achieved very high precision in measurements, showing that titin (the protein that makes up our muscles) has a mechanical resistance of about 210 pN. On the other hand, the mechanical resistance of the calcium-binding C2A protein has been found to be much weaker, i.e., the peak force is only of order 60 pN. Marek's goal was to learn the limits of the mechanical strength of proteins and to understand whether this feature correlates with the biological function or perhaps the spatial structure of proteins.

The idea was brilliant (more on that below), but what was worse, Marek wanted to do it in his own style, i.e., as best as possible. That meant writing a program to stretch (possibly in many direction) all known protein structures deposited in the Protein Structure Database in 2006. At that time, there were over 50000 structures. This idea was in the style of current big data, although such an approach was not yet used at that time. Marek always did things ahead of his time. So the man had to bite the bullet and rise to the challenge. The idea was also very successful from the point of view of the so-called Go-like model, which Marek created with his PhD student (Professor T.X. Hoang). In the Go-type model that Marek used, two factors played a major role: native contacts (defined based on the geometry of the protein in its native state) and the spatial structure of the protein (alpha-helices, beta hairpins). Therefore, the model performed is ideal for studying mechanical properties, starting from the native structure of the protein. After 6 months of research, it was possible to find proteins whose resistance force was over 1000 pN. It turned out that among these proteins there were proteins with non-trivial topology knotted proteins. Our first reaction was that this must be a mistake — none of us knew or expected knots in the proteins to exist.

There are proteins that show such high mechanical resistance. However, proteins with non-trivial topology show lower resistance, and in such a case the observed force comes from the knot tying, but thanks to that, we were able to find them. Today, the study of non-trivial topologies in proteins has become a separate, dynamic research field at the intersection of biophysics, biochemistry, and mathematics. Marek's works are among the best cited on this topic, and a collection of his works forms the basis of a review entitled "Topology in soft and biological matter" published this year in *Physics Reports* **1075**, 1 (2024).

Joanna Sułkowska

◇ ◇ ◇

The outstanding role that Professor Marek Cieplak played in theoretical physics, condensed matter physics, and biological physics is beyond dispute. However, I personally believe, and I would like to emphasize, that all these achievements were possible thanks to his exceptional personality.

Marek was a visionary, an inspirer and a titan of hard work, constantly seeking new inspirations with tireless energy, determination and scientific passion, along with many humanistic interests. He was always on the lookout for scientific developments, driven by the energy and determination to bring ideas to fruition. The word that best describes Marek is "passion" and he embodied this spirit, as lively and tireless as a great fire.

Marek had one mantra that he often repeated and consistently implemented: "Don't give up without a fight!" This was his Golden Thought. He always took the bull by the horns. I remember him as a person for whom nothing was too difficult to at least try. That is why he had a remarkable ability to motivate colleagues to push beyond their limits.

He was the initiator of introducing biological physics at the Institute of Physics of the Polish Academy of Sciences. Fueled by Marek's enthusiasm and determination, in 2004 we began our activity as a small Biological Physics Group SL-1.5. Our first challenge was to secure funding and set up experimental laboratories. We acquired space in an old transformer station and storage area at our Institute, where I started building laboratories equipped with the essentials to commence experimental work, with great support from Marek. Over time, many people joined the group, new topics emerged, including nanotechnology, and our group was upgraded to the Laboratory of Biological Physics SL-4. However, we still needed significant changes and investments, as our initial setup was quite basic.

I must emphasize that Marek's tireless enthusiasm for taking on risky initiatives was one of the strongest motivational impulses I have ever experienced. I am certain that the energy derived from daily interactions with Marek was vital in undertaking such ambitious projects. It ultimately led me to establish a consortium to create seventeen new laboratories in Poland, including our own microspectroscopy laboratory and a second computer cluster for Marek's group within the POIG ERDF NanoFun Project.

One of the most significant contributions by Marek to sharing high-level science with society is the "Biomolecules and Nanostructures" (BioNano) conference series. Marek envisioned this event as an opportunity for scientific friends to meet and discuss topics in an informal atmosphere. It resembled a group retreat, yet extended to around one hundred participants in a secluded area, far from large cities. The great success of these conferences stemmed from the fact that Marek's friends and collaborators from all over the world came together to meet him in person. He deeply valued his scientific friendships.

Over the years, we have transformed the meeting from an extended group retreat called the "Workshop on Structure and Function of Biomolecules" (2004, 2006) into an international scientific conference series "Biomolecules and Nanostructures" (2011, 2013, 2015, 2017, 2019), while maintaining the informal character of the meetings. The broader framework of the BioNano conferences also aligned with Marek's further involvement in nanobiotechnology studies. The venues changed, always close to nature and often in spartan accommodation, but the priority was still on the quality of the lectures and scientific discussions, fostering connections and collaboration among attendees rather than focusing on comfort. The conference grew, and Marek remained close to and engaged with the participants.

The great added value of the BioNano conferences lay in crossing barriers between the exact and natural sciences and the humanities — this was interdisciplinarity and multidisciplinarity in the best sense of the word. The opening lectures led the audience from linguistic methodology to molecular biology and from fundamental physics to physiology and evolution. This was made possible because, in addition to his great commitment to scientific matters, Marek moved us with his humanistic sensitivity. He often noticed little things that helped us feel more integrated with the world.

Marek was also always accompanied by his family. It was truly remarkable that despite his immense commitment to his professional career, he always maintained close ties with his loved ones.

Professor Marek Cieplak exemplified a rare combination of great ambition and versatile competencies, constantly seeking new inspirations. I remember him sitting in his armchair, going through the latest issues of *Nature* or *Science*, selecting topics that were new to him and to which he could apply the methods he had developed in order to join a new field.

Marek was always "on the ball", catching everything and hitting the mark. He will remain in my deepest memories.

Anna Niedźwiecka

In memory of Professor Marek Cieplak

REVIEW

The GōMartini Approach: Revisiting the Concept of Contact Maps and the Modelling of Protein Complexes

L.F. Cofas-Vargas[a], R.A. Moreira[b], S. Poblete[c,d], M. Chwastyk[e] and A.B. Poma[a,*]

[a] *Biosystems and Soft Matter Division, Institute of Fundamental Technological Research, Polish Academy of Sciences, Pawińskiego 5B, 02-106 Warsaw, Poland*
[b] *BCAM, Basque Center for Applied Mathematics, Mazarredo 14, 48009 Bilbao, Bizkaia, Spain*
[c] *Centro Científico y Tecnológico de Excelencia Ciencia & Vida, Fundación Ciencia & Vida, Avenida del Valle Norte 725, 8580702 Santiago, Chile*
[d] *Facultad de Ingeniería, Arquitectura y Diseño, Universidad San Sebastián, Bellavista 7, 8420524 Santiago, Chile*
[e] *Institute of Physics, Polish Academy of Sciences, al. Lotników 32/46, PL-02668 Warsaw, Poland*

Doi: 10.12693/APhysPolA.145.S9 *e-mail: apoma@ippt.pan.pl

We present a review of a series of contact maps for the determination of native interactions in proteins and nucleic acids based on a distance threshold. Such contact maps are mostly based on physical and chemical construction, and yet they are sensitive to some parameters (e.g., distances or atomic radii) and can neglect some key interactions. Furthermore, we also comment on a new class of contact maps that only requires geometric arguments. The contact map is a necessary ingredient to build a robust GōMartini model for proteins and their complexes in the Martini 3 force field. We present the extension of a popular structure-based Gō-like approach to the study of protein–sugar complexes, and the limitations of this approach are also discussed. The GōMartini approach was first introduced by Poma et al. (*J. Chem. Theory Comput.* **13**, 1366 (2017)) in Martini 2 force field, and recently, it has gained the status of gold standard for protein simulation undergoing conformational changes in Martini 3 force field. We discuss several studies that have provided support for this approach in the context of the biophysical community.

topics: Martini 3, structure-based coarse-graining, single-molecule force microscopy, biomolecules

1. Introduction

Structural biology has made significant strides in recent years, fueled by advancements in experimental techniques like nuclear magnetic resonance (NMR), X-ray crystallography, and cryo-electron microscopy (cryo-EM). These techniques provide detailed insights into the three-dimensional structures of biomolecules, shedding light on their functional mechanisms. However, static structural data alone fails to capture the dynamic aspects of molecular biology. To bridge the gap between static structural data and dynamic experimental data, robust and versatile computational models capable of accurately describing the dynamics of biomolecular complexes are essential. The GōMartini approach [1] for proteins offers versatility by combining the latest Martini 3 force field [2] for proteins and other biomolecules (e.g., lipids, carbohydrates, nucleic acids, etc.) and its cost-effective edge renders this approach ideal for large-scale applications in cellular environments [3]. Structure-based (SB) model offers a promising approach, utilising coarse-grained (CG) representations to capture the essence of a biomolecule structure and dynamics.

The typical time scales of biological processes involving, e.g., unfolding of proteins and protein recognition, among other events, are in the range of 10^{-6}–10^{-3} s, and, thus, they are orders of magnitude slower than typical molecular motion (i.e., 10^{-15}–10^{-12} s) simulated in all-atom (AA) molecular dynamics (MD). The length scales of conformational rearrangements are also much smaller in AA-MD simulation than they would be for studying processes involving large structural changes in biological systems. In this regard, the SB model and CG approaches of biomolecular systems are ideal tools to overcome such limitations. The replacement of the position of each amino acid by its C^{α} atom is a common choice. In this approach, several degrees of freedom of the system are removed, which enables reaching the experimental time and length

scales while maintaining a molecular-level model of the systems under consideration. In particular, CG approaches are used to infer the Young modulus and confront it with atomic force microscopy (AFM) experiments. Importantly, the mechanism of deformation that gives rise to the linear force-displacement response can be characterised in the CG simulation. Several CG models are not sensitive to pH or ionic strength, and they also do not consider the electrostatic interactions, post-translational covalent modifications of amino acids, etc. Those factors have been demonstrated to be important in, for example, the recognition of cell receptors by pathogens and control of the assembly of protein complexes. In addition, standard AA-MD simulation can target system sizes on the scale of $\sim$ 500 million atomistic particles in the latest SARS-CoV-2 full virion in aerosol droplet [4], which is only possible in a few high-performance computing clusters around the world. However, analogous systems formulated using CG force fields, such as Martini 3 [2], SIRAH [5], and UNRES [6], are an order of magnitude smaller. In CG-MD simulation, these systems can be studied in a moderate-sized computing cluster. Moreover, due to the large time-step used in CG-MD (e.g., MD simulation with Martini 3 employs $\Delta t_{\mathrm{CG}} = 20$ fs in comparison to AA-MD with a $\Delta t_{\mathrm{MD}} = 2$ fs), CG simulations are expected to reach longer time scales than AA-MD. At the core of the SB approach in proteins lies the concept of native interactions, also known as "native contacts" (NC), which provides a simple form to understand the important interactions in equilibrium; it represents the close spatial proximity between residues or atoms in the native state. Defining native interactions poses a challenge, as simple cut-off distance-based definitions can lead to two incompatible outcomes: (i) the exclusion of relevant contacts beyond 6 Å, and (ii) the introduction of non-physical next-nearest neighbour contacts. To address these limitations, various methods have been developed to define native contacts, including atomic overlap map, shadow map, CSU contact map, and Voronoi maps (to be discussed in the next section). Each method offers unique advantages and limitations, and the optimal choice depends on the specific application. In the past, we combined both the semi-atomistic approach (e.g., Martini 3 force field) and the SB approach, and as such, we developed an alternative strategy to study conformational changes of proteins, and through this review work, we plan to show the extension to protein complexes.

Hence, we first discuss here several contact maps that employ distance cut-off and chemical and physical information, secondly, we briefly introduce the extension of the popular SB model developed for protein–sugar complex and, lastly, the more robust model, the so-called GōMartini approach that is based on the SB model of proteins developed by Professor Marek Cieplak (e.g., a C^{α}-based Gō-like model) and others. This simple model turned out to be efficient in capturing the long-time behaviour of certain biomolecular systems under mechanical forces and under high temperatures [7–11]. Most importantly, the Martini force field with an almost atomic resolution can use a backmapping protocol to recover an AA representation from the CG representation with an almost atomic resolution.

2. Contact maps for determination of interaction and topological aspect in proteins and nucleic acids

2.1. Contact maps based on distance threshold and geometric principles

A simple protein contact map (CM) based on a distance cut-off that allows for the calculation of protein interactions depends essentially on the atomic positions (see Fig. 1). This method considers the interaction of any pair of atoms in different residues that are within a certain distance of each other. For example, in protein studies, contacts have often been defined based on atomic geometry by selecting the heavy atoms in a given amino acid residue — an atomic contact is found, if two heavy atoms associated with distance residues are within a specific cut-off distance (i.e., 4.0–6.5 Å) [12]. Despite its simplicity, the cut-off CM suffers from several issues that render it less accurate and reliable in determining native contacts [13], especially in the context of SB models that require accurate determination of the native contacts to examine the emerging protein dynamics from the underlying geometry. One of the problems with the cut-off CM is a high sensitivity to the cut-off distance. This means that even slight adjustments of this parameter can result in substantial changes in the number of identified contacts. This sensitivity can impede the comparison of results across different studies. Additionally, the cut-off CM often identifies contacts between atoms that are not physically in contact with each other, leading to erroneous conclusions about molecular structure and function. Furthermore, it fails to account for occluded contacts or structural elements, overlooking the accessibility of atoms and their embedding within the larger-scale structure, potentially overestimating the number of identified contacts.

In contrast, the shadow CM [14] fixes some of the previous limitations of cut-off CM and offers a more advanced approach to determining atomic contacts within a protein. It considers the concept of "shadows" cast by other atoms. In this method, two atoms are only considered to be in contact if there are no other atoms blocking the line of sight between them. The process of obtaining contacts involves the following steps:

(i) Calculating the distances between all pairs of atoms in the protein and creating a list of pairs within a specified cut-off distance.

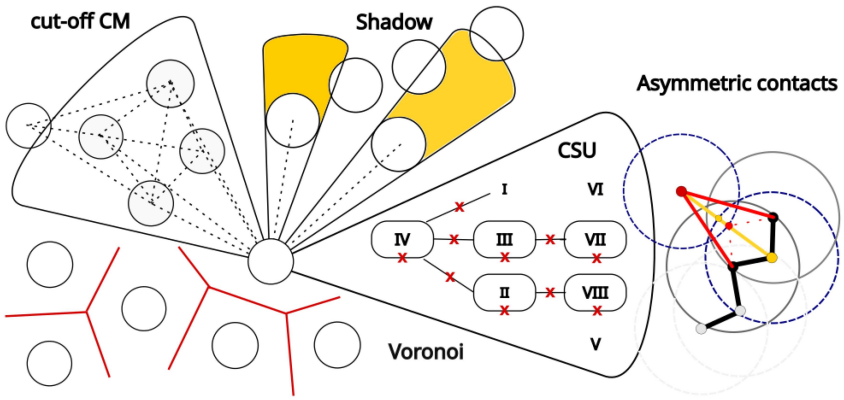


Fig. 1. Representations of contact maps (CM) in the literature. The cut-off CM used only the distance between centres, while the shadow methodology improved the former by not discarding centres in the yellow areas. The CM can be improved by including structural categories, the contacts of structural units (CSU) methodology, represented by the Roman numbers, and defining illegitimate (marked with red crosses) interaction. However, the CSU CM uses an extended sphere that accounts for solvent effects (dashed circles at right) as a first guess for contacts, which can be asymmetric (as shown by the red and yellow lines at right) due to shadowing effects. Cut-off-free pure geometric strategy can use a Voronoi diagram, as shown in the panel below, to create the CM based on the Delaunay triangularization.

(ii) Utilising a spherical screening radius, typically ≤ 0.5 Å, centred at each atom, and for each pair of atoms, excluding contacts if one atom is obscured by the shadow of another atom and intuitively captures only the "visible" atoms from the perspective of a reference atom.

This procedure is visually depicted in Fig. 1. The shadow CM has demonstrated superior accuracy compared to the cut-off CM for several reasons. First, it exhibits less sensitivity to the choice of cut-off distance, as this parameter is primarily used to establish the initial set of potential contacts and the occluded contacts are subsequently removed. Second, the shadow CM aligns better with experimental data by capturing contacts between atoms separated by intervening atoms, such as water molecules. Despite its advantages, the shadow CM does come with certain limitations. Notably, it introduces increased computational complexity and retains some sensitivity to the cut-off distance, albeit to a lesser extent than the cut-off CM. This sensitivity to noise may result in the identification of false positive contacts, particularly in molecules with flexible structures (i.e., loops and coils) or when dealing with noisy experimental data. Furthermore, the shadow CM cannot capture solvent-mediated contacts or indirect contacts in general.

An entirely different strategy for defining protein contacts involves a cut-off-free methodology that relies solely on geometric principles [12, 15, 16]. The Voronoi tessellation [17] is a technique for partitioning the physical space into convex polyhedrons, called Voronoi cells, with each cell associated with a specific site. Typically, the C^{α} atom of each amino acid is chosen in the context of protein structure analysis. The Delaunay triangulation [17] complements Voronoi tessellation by connecting a set of points with a network of triangles, ensuring that no point lies inside the circumcircle of any triangle. In the case of protein structure analysis, Delaunay triangulation is often used in conjunction with Voronoi tessellation to define protein contacts. To define these contacts using Voronoi tessellation and Delaunay triangulation, the following steps are typically followed:

(i) Constructing the Voronoi tessellation of the protein structure.
(ii) Determining for each pair of adjacent Voronoi cells whether a corresponding Delaunay edge exists.
(iii) If the Delaunay edge exists, then the two sites are considered to be in contact.

There are several advantages of using Voronoi tessellation and Delaunay triangulation in protein structure for contact definition and topological properties studies. This method is computationally efficient, robust to data noise, and capable of capturing both direct and indirect contacts.

2.2. Contact maps based on chemical and physical information

The contacts of structural units (CSU) method [18] is a structure-based approach that leverages geometric and chemical information to identify contacts between amino acid residues within a protein. It involves three main steps:

(i) Identifying pairs of heavy atoms that are in close proximity, typically within a defined distance threshold.

(ii) Assigning each atom to a specific class based on its chemical properties, such as its element (O, N, C, S) and its connectivity to other atoms.

(iii) Establishing contacts between residues based on the presence of specific interactions between their individual atoms, including hydrogen bonds, aromatic interactions, and hydrophobic interactions. Any interactions that do not fit into these specific categories — labelled as "non-specific contacts" — are excluded from consideration in the CSU method. This exclusion is based on the premise that these non-specific interactions may not convey the structural or functional relevance that the method aims to capture.

In essence, the CSU method focuses on recognizing and emphasizing interactions with well-defined chemical characteristics, enhancing the specificity and relevance of the identified contacts within the protein structure. However, the CSU method has certain limitations, e.g., it only accounts for attractive interactions and neglects repulsive interactions, potentially leading to the inclusion of contacts destabilised by repulsive forces. Additionally, this method may identify contacts between residues that are not physically in contact, as it relies on the presence of at least one specific contact, and can also return asymmetric contacts, as depicted in Fig. 1, due to shadowing effects of neighbouring units. Moreover, it might miss important contacts, particularly in helical structures, due to its focus on specific interactions. Overall, the CSU method provides a valuable approach for identifying contacts in proteins, but it does have limitations stemming from its exclusion of repulsive interactions and potential lack of selectivity.

The repulsive CSU (rCSU) methodology [7] addresses these shortcomings by incorporating repulsive interactions and refining contact identification, offering a more accurate and reliable approach to a new form of contact map generation. The rCSU methodology extends the CSU approach by considering repulsive interactions between charged atoms. It aims to provide a more precise representation of inter-residue contacts by accounting for both attractive and repulsive forces. The rCSU algorithm proceeds in a manner similar to CSU:

(i) Initially, it identifies pairs of heavy atoms in close proximity.

(ii) Subsequently, it classifies atoms based on their chemical properties, akin to CSU.

(iii) The determination of whether there is a contact between two residues is dependent on the overall balance or net outcome of interactions at the atomic level, calculated as the difference between the number of attractive contacts (e.g., hydrogen bonds, aromatic interactions, ionic bridges, and hydrophobic interactions) and the number of repulsive contacts (Coulombic repulsions between charged atoms).

(iv) If the net contact is positive, a contact between the residues is established.

The rCSU methodology offers several advantages over CSU. It provides more accurate contact predictions by considering repulsive interactions, reducing the likelihood of contacts destabilised by repulsive forces. Furthermore, it enhances contact selectivity by evaluating the net contact between residues, lowering the probability of false positives. This methodology also captures a wider range of interactions, including ionic bridges, resulting in a more comprehensive representation of inter-residue contacts. In summary, the rCSU methodology presents a more accurate and reliable approach to CM determination compared to CSU, as it incorporates more chemical information and improves contact selectivity.

The OV+rCSU method [7] combines the strengths of the overlap (OV) method and the rCSU method to identify contact maps in proteins. The OV method identifies contacts based on the overlap of enlarged van der Waals spheres around the heavy atoms, while the rCSU method incorporates repulsive interactions between atoms with charges to refine contact identification. In contrast, the shadow CM method relies on a fixed distance cut-off, independent of atomic size, and removes contacts with intervening atoms. OV+rCSU is superior to the shadow CM method because it considers atomic sizes derived from experimental studies [19], and repulsive interactions, enabling the capture of a broader range of interactions while maintaining selectivity and a decrease in false positives.

The CSU method is simpler, only considering attractive interactions and disregarding repulsive interactions. This method can simply lead to false positives, as some contacts may be destabilised by repulsive forces. OV+rCSU addresses this limitation by incorporating repulsive interactions to refine contact identification. While rCSU is an improvement over CSU, it may miss some true contacts due to its focus on net contact between residues. The OV+rCSU method complements rCSU by considering overlaps of van der Waals spheres, potentially capturing additional contacts. OV may identify false positives due to its reliance on overlaps without considering repulsive interactions. OV+rCSU and rCSU address this limitation by incorporating repulsive interactions to refine contact identification.

The Voronoi/Delaunay [15] methodology provides a cut-off-free approach [12] to CM determination, relying on geometric constructs to define contacts based on the proximity and connectivity of residues. It involves partitioning space into polyhedra, known as Voronoi cells, with each cell

associated with a residue. The faces of these polyhedra define the closest contacts between residues, offering a geometric foundation for contact definition. This methodology does not require a fixed cut-off distance, eliminating the need for arbitrary cut-off/parameter selection. It delivers a geometric basis for contact definition, ensuring consistency and robustness, capturing both local and global contact patterns, and providing a comprehensive view of the protein structural connectivity. This method can be used to define both residue–residue and atom–atom contacts, offering flexibility in granularity.

The choice between the OV+rCSU and Voronoi/Delaunay methodologies depends on the specific application's requirements for accuracy, efficiency, and robustness. For applications demanding high accuracy and comprehensiveness, such as protein folding simulations or detailed structural analysis, the OV+rCSU methodology may be the preferred/recommended choice in SB models. The explicit consideration of atomic sizes and repulsive interactions provides a more detailed and realistic representation of native contacts in proteins. For applications requiring a fast and efficient method for capturing local and global contact patterns, such as network analysis or large-scale structural comparisons, the Voronoi/Delaunay methodology may be the better choice. Its cut-off-free nature and geometric foundation make it computationally efficient and less sensitive to arbitrary cut-off selections. In general, the OV+rCSU methodology is well-suited for applications where a high level of accuracy and detail is crucial, while the Voronoi/Delaunay methodology is well-suited for applications where efficiency and robustness are primary considerations.

Overall, the OV+rCSU methodology offers a more accurate and comprehensive approach to contact map determination compared to rCSU, OV, and shadow CMs individually. It combines the strengths of the OV and rCSU methods to identify a broader range of interactions while maintaining selectivity and reducing the number of false positives.

A recent method that considers the equilibrium dynamics of a protein, such as the differential/dynamic contact map (dCM) [20], offers an alternative solution. It can identify the most structurally relevant contacts in a protein using AA-MD simulations. This method relies on contact frequency and the definition of stability. Frequency measures the number of times a contact was observed between two residues. High contact frequencies indicate more stable contacts. Stability, on the other hand, is determined by considering the chemical characteristics of residues involved in a contact. For instance, hydrophobic interactions are generally more stable than polar–polar and electrostatic interactions. To obtain a more detailed view of the set of protein contacts, the OV+rCSU approach is used with the dCM analysis. The OV+rCSU considers the chemical character of each residue and the respective contacts between a pair of residues, classifying them into categories to count the number of stabilising and destabilising contacts per residue, defining a contact when both residues have a net stabilising character. The dCM and OV+rCSU methodologies together form a robust contact map technique known as differential contact map that has been validated in the study of the dynamics of large protein complexes. For example, the dCM analysis identifies the high-frequency (> 0.9) contacts between amino acids in the SARS-CoV-2 trimeric spike protein [20]. It reveals that flexible loops are the source of contact fluctuations, comprising approximately 1772 amino acids based on secondary structural analysis, while helices and strands are roughly represented by 712 and 819 residues, respectively. The entire spike protein has 3363 residues. This indicates that the methodology is feasible even for large protein complexes.

2.3. Contact maps for intrinsically disordered proteins

Creating a contact map for intrinsically disordered proteins (IDPs) presents challenges due to their lack of well-defined tertiary structure, which evolves over time. Furthermore, the energetic landscape of these proteins significantly differs from those with stable structures that possess a singular energetic minimum [21], as opposed to the shallow energetic wells between which the protein's conformation fluctuates [22]. This necessitates defining the contact map temporally and updating it at every simulation step. Given the absence of a fixed protein structure, a specialized algorithm is essential for determining this contact map.

One feasible approach is an algorithm based on three criteria: distance between amino acids, orientation of specific residues' side groups, and the potential number of contacts a given residue can establish. The algorithm categorizes contacts into three types: sidechain–sidechain (ss), backbone–backbone (bb), or backbone–sidechain (bs), each utilising slightly different criteria.

The distance criterion serves as the foundational parameter governing the onset of a contact. Contacts break when the distance between the centres of C^{α} atoms of particular residues exceeds a defined limit, $f\sigma_{i,j}$, where $\sigma_{i,j} = r_{\min}(0.5)^{1/6}$. Here, $r_{\min}$ indicates the position of the energetic potential minimum. The values of $r_{\min}$, determined from an analysis of distances between residues in 21,090 non-redundant proteins from the CATH database, varied based on the interaction type. The $r_{\min}$ values for bb and ss contacts were determined as the mean values from the collected data, resulting in 5.0 Å and 6.8 Å for bb and bs contacts, respectively.

However, for ss contacts, $r_{\min}$ was individually calculated for each pair of residues. This value ranged from 6.42 Å for the Ala–Ala interaction up to 10.85 Å for the Trp–Trp interaction, and comprehensive details about other pairs of residues capable of forming contacts are presented in [23, 24]. As mentioned earlier, contacts fluctuate during the simulation, and contact is broken when the distance between residues exceeds $f\sigma_{i,j}$, where $f = 1.5$. Different values of this factor were also considered and are well described in [23, 24].

Another crucial criterion is the orientation of residues. Implementing this criterion is not straightforward, as each residue is treated as a spherical bead. Therefore, neighbouring residues must be considered for direction implementation, as detailed in [23, 24]. Determining the orientation of a backbone hydrogen bond or a sidechain C^{β} atom relies on the positions of three consecutive C^{α} atoms. This criterion is essential because we assume that a bb contact can occur if the N-atom on the backbone part of the i-th residue can establish a hydrogen bond with the O-atom on the backbone part of residue j, or vice versa. However, this interaction is permissible only when both atoms are oriented toward each other. The same requirement for residue orientation applies to ss and bs contacts. Detailed descriptions of the mathematical formulas that enable the implementation of these requirements are provided in [23, 24].

The final criterion involves the residue types that are essential for defining contacts. They are categorized into six classes: (1) Gly, (2) Pro, (3) hydrophobic, (4) polar, (5) negatively charged, and (6) positively charged. The solvent being implicit in the program restricts the simulation of interactions between polar residues and water molecules. Employing the one-bead-per-residue model leads to a less dense protein representation. To compensate for this, we restrict each amino acid's capacity to form a limited number of contacts. The formula $z_s = \mathrm{n_b} + \min(s, \mathrm{n_H} + \mathrm{n_P})$ determines this number, where $\mathrm{n_b}$ signifies the allowable count of backbone contacts, s represents the maximum quantity of sidechain contacts, $\mathrm{n_H}$ denotes the upper limit for contacts with hydrophobic residues, and $\mathrm{n_P}$ signifies the limit for contacts with polar side chains. Detailed values of these parameters can be found in [23].

It is crucial to note that the contact map, irrespective of its method of creation, can be utilised with any potential energy function, whether it is a spherical potential like Lennard–Jones (LJ) or one that integrates directional criteria. The initial step always involves validating the distance criterion, followed by evaluating the ability to create specific contacts. The above-mentioned methods for defining contacts are necessary not only for the description of the dynamics of single protein chains but also for research on the aggregation of IDPs or even the creation of protein droplets.

2.4. Contact maps for nucleic acids: RNA structures

SB models based on a Gō-like approach have also been used to study RNA molecules in CG descriptions. The CG model considers a nucleotide by a single bead, and then a contact map is built on the basis of the distances between the interaction sites in the native structure. Such is the approach used in the self-organized polymer (SOP) model proposed by Hyeon and Thirumalai [25], designed to analyse the dynamics of RNA unfolding under constant force.

As in the CSU method, additional physicochemical details can be introduced by considering the interaction type between nucleotides. The main forces that stabilize RNA structures are due to stacking interactions and base pairing. The former is present when two nucleobases are close enough and lie on parallel planes exhibiting an overlap between their faces. On the other hand, base pairs originate from hydrogen bonds formed between the edges of the nitrogenous bases, yielding a relatively large number of possible geometrical arrangements between the four nucleobases that characterize RNA molecule: adenine (A), uracil (U), guanine (G), and cytosine (C). In particular, A–U or C–G base pairs and stacking interactions give rise to the well-known A-form double helix, a motif of extreme importance in RNA structure. Electrostatics can also be considered explicitly, regardless of the proximity of the nucleotides in the native structure. For this purpose, a point charge is generally placed on the phosphorus atom on the backbone, which interacts with other charged particles through an implicit solvent approach.

Some Gō models have employed specific terms or functional forms for stacking and base pairs contacts using this information. The three-interaction site model of Hyeon and Thirumalai [25, 26] defines a nucleotide by three point particles representing the nucleobase (A, U, C, G), sugar ring (i.e., ribose: $C_5H_{10}O_5$), and phosphate group (PO_4^{3-}), which allows the introduction of directional interactions. The model has been parameterized with melting temperatures of small RNA fragments to study RNA folding thermodynamics under several ion concentrations and temperatures, and its interactions are also specific for contacts belonging to a double helix. Later versions of the model, however, are capable of introducing complementary base pairs between non-native contacts and stacking interactions between non-consecutive nucleotides [27]. This combination makes it possible to deal with a more complex free-energy landscape while introducing contacts that stabilize the native structure and taking care of describing properly the thermodynamics of the double helices, which have an important contribution to the overall stability. In addition, the model of Hori and Takada [28], designed for the study of structural deformations

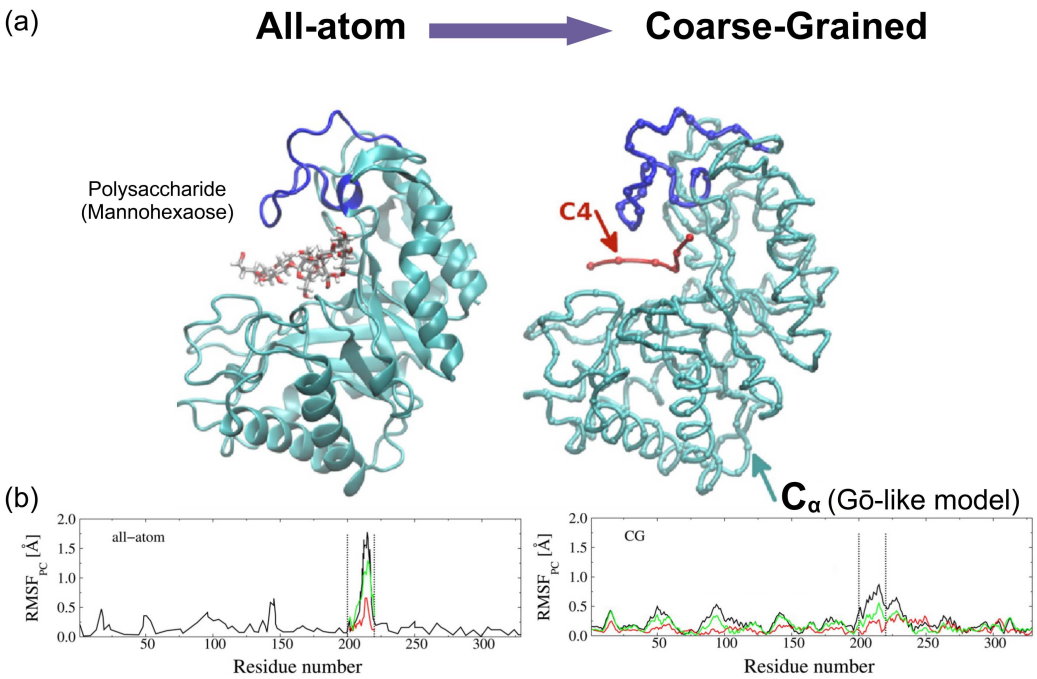


Fig. 2. Panel (a) shows the all-atom MD and SB CG modelling of the sugar–Man5B complex (PDB:3W0K). The blue protein segment comprises the residues (200–220), which is considered the active loop responsible for the cleavage of the O-glycosidic bond in polysaccharides. The CG model employs the C_α positions for protein and C4 atoms for the sugar hexamer. Panel (b) shows the fluctuations of the complex under AA and CG simulations. The RMSF shows fluctuations of the protein segment undocked (black line) and docked with mannohexaose and cellohexaose in green and red solid lines, respectively. This figure was adapted with the permission from [35].

of RNA and protein–RNA complexes, also uses a three-point representation of a nucleotide and a parametrization from MD simulations and distinguishes stacking from base-pairs in their Gō-like approach.

The large number of non-complementary base pairs and the possibility of forming hydrogen bonds between nucleobases and phosphate groups or sugar rings increases the complexity of the interaction network of RNA molecules. Despite this, several tools such as ClaRNA [28, 29], FR3D [30], or x3DNA-DSSR [30, 31] can be used to annotate structures and identify the most relevant interactions in the system of interest, which can help to build Gō models able to capture the essentials of the phenomena to study under simulations.

3. The structure-based model: A Gō-like approach for protein–sugar complexes

In nature, proteins and polysaccharides can exist separately and also form complexes, for instance, the degradation of cellulose fibrils by fungi or bacteria involves the processing of the biomaterial by enzymes (e.g., endo- and exo-glucanases) and thus, a relevant biotechnological process that has been improved for the biofuel production [32, 33]. At the molecular level, enzymes recognize the cellulose chain ends or broken chains, and after attaching to them, the cleavage of the O-glycosidic bond is carried out, releasing several small oligomers that can be the source of energy for several microorganisms. Also, glycosylation of proteins by sugar moieties (i.e., N-glycan or O-glycan) can induce conformational changes via allosteric communication. Such an effect was reported in the conformational transition from closed to open state in the SARS-CoV-2 spike (S) glycoprotein [34]. The relevance of describing such events by molecular simulation can lead us to the development of novel therapeutics against pathogens such as viruses and bacteria. In this regard, the study of protein–sugar complexes remains an active field of research in the biomolecular community.

The extension of the Gō-like approach for the study of protein–sugar complexes was carried out in [35]. In this work, the Cα-based Gō-like model for proteins was coupled with a structure-based coarse-grained (SB CG) model for polysaccharides. Each sugar oligomer was formed by D-glucose units connected by the β $(1 \rightarrow 4)$ glycosidic bonds in cellohexaose and the α $(1 \rightarrow 4)$ glycosidic bonds in amylohexaose case. Then, each sugar monomer in the CG description was represented by one CG bead centred on the position of the carbon atom. We considered the position of the C4 atom for comparison

with respect to the C1 position. Alternatively, we also derive parameters for C1 and the centre-of-mass of the monomer. The new set of CG values for bonded and non-bonded parameters was determined by AA-MD simulations using two statistical methods. One of these methods was the Boltzmann inversion (BI) method [36], while the other was denoted by the energy-based (EB) approach. The non-bonded parameters for the protein–sugar complex were mapped by the Lennard–Jones (12–6) potential according to the Gō-like approach. These two methods were employed to calculate the stiffness of sugar oligomers and protein secondary structures. Protein Man5B comprises 330 residues, with PDB ID 3W0K. This study shows the stiffness of α-helices, which, on average, are stiffer than β-strands. Also note that in proteins, the secondary structures are generally stiffer (based on elastic contacts) by a factor of 5 than in sugars. This CG model was validated for the case of the hexaose-Man5B catalytic complex (see Fig. 2). The large fluctuations calculated by principal component analysis (PCA) of the active loop in Man5B were retained in the CG description. The main trend between AA-MD [37] and CG-MD simulations regarding the binding activity was also captured in the CG model.

These results highlighted the energetic differences between protein–sugar interactions and native interaction in proteins ($\epsilon_{\mathrm{PP}} \sim 1.5$ kcal/mol). It was reported that the strength of sugar–protein energy value (ϵ_{SP}) was in the range of 3 to 6 kcal/mol. This SB model for protein–sugar complex is constructed under implicit solvent conditions, and no detailed chemistry of residues is included, thus, ligand recognition associated with long-range interaction or the effect of single point mutations that induce conformational changes cannot be captured by this simple model. Furthermore, atomistic backmapping is not doable under this representation because of the level of CG description based on C^{α} atoms. In the next section, we present an alternative approach that is based on the Gō-like model for proteins and circumvents several limitations of this SB model.

4. Overview of the GōMartini approach for protein complexes

The GōMartini approach was first introduced by Poma et al. [1], and it coupled the Martini 2 with the SB model for protein (i.e., Gō-like type) developed by Cieplak's lab. The protocol for the GōMartini approach is depicted in Fig. 3 (see also [38]); the first step begins with an experimental 3D structure of a globular protein. From this structure, the OV+rCSU contact map is created from the server http://pomalab.ippt.pan.pl/GoContactMap [39, 40]. The next step involves the transformation of AA to CG representation using the ./martinize2 script [41]. In the case of protein complexes, each of their chains must be isolated

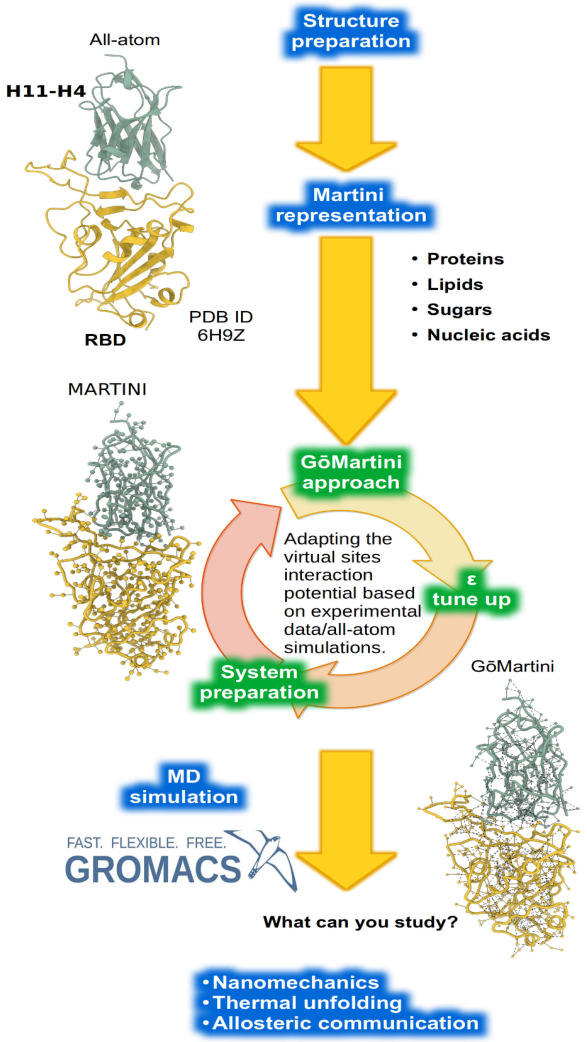


Fig. 3. The GōMartini approach workflow for protein complexes. The study of oligomeric complexes involves building a Martini CG representation, creating a contact map, and introducing Gō bonds mapped as LJ potential through the virtual site implementation (denoted as dummy beads in GROMACS versions before 3.3). The resulting complex is solvated and neutralized at ambient conditions. MD simulations are performed using the GROMACS MD package [38]. By iteratively modifying $\varepsilon_{\mathrm{G\bar{o}}}$ value in the range of 9.4 and 12.0 kJ/mol, CG-Martini simulations are capable of reproducing experimental results.

in individual PDB files. The GōMartini approach is applied to obtain a CG structure with the "create_gomartini.py" script. The strength of the Gō potential mapped by a Lennard–Jones (12–6) potential can be adjusted to match results from experiments or atomistic simulations. The GōMartini approach allows the sampling of large-scale conformational changes, a limitation in AA-MD simulation [1]. This holds particular significance in the investigation of proteins that undergo significant structural transitions, such as unfolding events

under a mechanical force, inter-domain motions, and catalytic rearrangements, as well as in a better description of protein complex stability [39, 42–46].

The Martini force field allows for the simulation of biomacromolecules at a faster rate than the conventional AA representation. For proteins, an old SM model based on an elastic network (EN) model was used to preserve protein native structure by adding harmonic bonds between C^{α} atoms. This has the drawback of not being able to explore the full conformational landscape of the protein, as well as overestimating the number of contacts between nearby residues [1]. The GōMartini approach involves replacing the contacts obtained by the EN model based on a simple cut-off distance by the native contacts built based on the OV and chemistry-based rCSU CMs. Then, virtual sites are placed near the BB particles. Such implementation in GROMACS 2020 (or above) enhances the production stage. The only energy scale in this approach ($\varepsilon_{G\bar{o}}$) can be modified iteratively. The default value is 9.414 kJ/mol, which corresponds to the energy of the hydrogen bond in proteins [35]; however, alternative values have been employed to replicate experimental results, including 12 kJ/mol, and in certain instances, values of 100 kJ/mol over certain pairs of contacts were necessary [40]. Below is a brief review of successful GōMartini studies with Martini 2 [47] and Martini 3 [2] force fields.

4.1. Martini 2

The first applications of the GōMartini approach were carried out using the Martini 2 force field. It examined the folding of small α- and β-peptides, the conformational flexibility of a set of proteins, and the nanomechanics of a titin domain using different definitions of contact maps and different $\varepsilon_{G\bar{o}}$ values. Folding simulations were done for an α-helix segment of the histidine-containing phosphocarrier protein and a β-strand of the G protein. Native contacts were calculated from the PDB structures, and the coordinates for the unfolded conformers were obtained from a CG simulation at 500 K with implicit solvent. The results showed that both peptides refolded in almost all simulations. The equilibrium dynamics of the type I cohesin domain, the domain of I27 from titin, and ubiquitin (PDB IDs: 1AOH, 1TIT, and 1UBQ, respectively) were examined at both atomistic and CG resolutions. RMSD analyses revealed that the proteins were stable along GōMartini simulations with deviations smaller than 0.2 nm and that characteristic residue fluctuations were captured during the MD simulation, in agreement with the previous EN model (i.e., ELNEDIN approach [48]) and AA-MD simulations. A principal component analysis indicated that the GōMartini approach was able to capture the opening and closing motion of the Man5B glycoside hydrolase. The amplitudes were comparable to those observed in AA-MD simulations [37]. Finally, nanomechanical studies employing the GōMartini approach on the domain of I27 from titin showed the nanomechanics can be captured at slower pulling speeds than AA-MD simulations, and unfolding forces were similar to experimental values when extrapolated to low loading rates.

In another study [39], the GōMartini approach was used to describe the membrane remodelling dynamics of the F-Bin/Amphiphysin/Rvs (F-BAR) protein PACSIN-1. The conformation of PACSIN-1 was not maintained using the original definition of native contacts, namely OV+rCSU, due to the overstabilization of contacts between neighbouring residues. The native contacts have been redefined. Consequently, their definition included all i-th and $(i+3)$-th residue pairs. Also, if the minimum distance between all heavy elements (i.e., N, C, and O atoms) is shorter than a distance threshold, a pair of residues, i-th and $j > (i+3)$-rd, is considered to have a native contact. Throughout the simulations, lateral PACSIN-1:PACSIN-1 interactions were observed and correlated with the solved 3D structure. This optimization reproduced the structural and local fluctuations observed in AA-MD simulations.

The GōMartini approach has been used to investigate the association of lipids with various proteins [49–52]. In one of these studies, the conformational dynamics and the effect of oligomerization of npq2 Light-harvesting complex II (LHCII) on the association with lipids [52]. Another study examined the stability of LHCII in its monomeric and trimeric forms, the cofactor flexibility, and the impact of membrane composition [51]. Both studies demonstrate the usefulness of the GōMartini approach for describing the conformational flexibility of proteins, with results comparable to those obtained by experimental techniques or AA-MD simulations.

The stability and enzyme flexibility of proteomimetics in the presence of zinc metalloproteinase thermolysin was studied using GōMartini. The simulation results were consistent with experimental observations [50]. In another study [53], the structural stability of PET-degrading enzyme (i.e., PETase) in a complex with copolymers at high temperatures was examined. The results obtained from the GōMartini simulation were in agreement with the temperature-dependent conformation observed in AA-MD simulations [53].

One of the most notable applications of the GōMartini approach is in the nanomechanics of proteins that requires the use of steered molecular dynamics (SMD) simulations. The level of CG in this approach has the advantage of reaching experimental time and length scales while maintaining a detailed description of the system at the molecular level[26]. In this particular aspect, several studies have dealt with the nanomechanics of $A\beta_{40}$, $A\beta_{42}$, α-synuclein, and other self-assembly

peptides [43, 45, 46]. These investigations have shed light on the stability of biological fibrils [54] and their significance in the progression of neurodegenerative diseases, as well as on the mechanical properties that can be used to develop new materials with industrial uses [43, 45, 46]. In another study, the unbinding pathways of the complex anticalin:CTLA-4 and its nanomechanics under various pulling geometries, which led to diverse force–distance profiles, were investigated using AFM single-molecule force spectroscopy (SMFS) experiments and GōMartini simulations. As a result, this approach explained the observed experimental patterns of mechanical stability that were attributed to pulling geometries and to the loss of native contacts between secondary motifs [40, 44].

4.2. Martini 3

The recent version of Martini 3 for proteins [2] and other polysaccharides [55–57] has improved the ability of GōMartini to study large conformational transitions in protein under several environmental conditions. The protein copper, zinc superoxide dismutase was the first use of GōMartini and Martini 3 for the study of conformational events [42]. The authors captured the allosteric effect of the G93A mutation on the electrostatic loop (EL). Note that the larger flexibility of EL causes the opening of this loop, which further destabilises the zinc-binding site of this enzyme via an increase in the hydration levels. In accordance with hybrid quantum-mechanical/molecular-mechanical MD simulations, the opening of the EL was reproduced using simulations, as well as its conformational flexibility. This study paved the way for the utilisation of the GōMartini methodology in the comprehensive examination of mutations and their allosteric effects on the structure and function of proteins.

The implementation of CG-MD simulations employing the GōMartini strategy resulted in the identification of a second phosphatidylinositol 4,5-bisphosphate binding site on the C-terminal domain of the tubby protein. The validation of this new binding site was carried out by mutating charged residues to alanine, both *in silico* and in living cells. It was shown that the affinity for phosphoinositide was reduced in both experiments [58].

In a study of the accessory factors UbiJ and UbiK, the GōMartini approach was used to improve the sampling process. Also, the absorption of a trimeric protein in the membrane was studied. A contact profile along an AA-MD simulation between the protein and the membrane was necessary to tune the CG model, which improved the accuracy of the interactions [59]. Small bifunctional molecules capable of modulating protein-membrane interactions were studied by GōMartini [60]. The CorA transport system asymmetric gating mechanism was investigated using the same method. For this purpose, both AA-MD and CG-MD simulations with different conformations of the protein chain were performed. The highly dynamic conformational changes observed in the set of simulations were consistent with recent structural studies. Based on previously reported information and results from the CG simulations, the authors proposed a patent on the novel asymmetric gating model for this protein system [61].

The giant mechanical stability of the adhesion bone sialoprotein-binding protein (Bbp) of *Staphylococcus aureus* and its role in biofilm formation have been recently investigated using AA-MD and GōMartini SMD simulations. Single-molecule force spectroscopy [44] has given evidence of such a high degree of mechanostability in Bbp. Additional experiments on the Bbp-fibrinogen-α complex revealed that this is one of the most mechanostable protein complexes studied so far. These results agreed with experimental SMFS data [44].

The GōMartini approach has proven to be useful for the study of diverse protein systems, revealing details about their nanomechanics, allosteric effects, and a deeper appreciation of their conformational flexibility. It is possible to extend this approach to the study of protein complexes with diverse oligomeric states by using the workflow depicted in Fig. 3. GōMartini can compensate for protein–protein interactions that cannot be fully captured by the Martini 3 force field, enhancing our tools for the analysis of these complexes.

5. Conclusions

An interesting alternative for the study of biomacromolecular events at the nanometric scale and with a temporal resolution closer to experimental studies is presented by the GōMartini approach. The scope of its application will be widened with its future expansion to encompass other types of molecules, such as carbohydrates, lipids, and nucleic acids. The GōMartini has become the gold standard in Martini 3, as it offers flexibility by combining physical and chemical information in the construction of the contact map in protein. This is a particular advantage that renders the information stored in the Gō interaction crucial for the understanding of the mechanism of protein–protein dissociation, as well as during the nanomechanical deformation, as one can track directly the rupture of Gō bonds as it will be in a continuum system. Martini 2 is used to overestimate the protein–protein interaction, and in Martini 3, the protein interface requires the contribution of additional Gō bonds that can be obtained by the OV+rCSU CM. We anticipate that a combination of statistical potentials and machine learning approaches can assist the contact map determination in protein complexes.

Acknowledgments

A.B.P. acknowledges Marek Cieplak for inspiring the research on structure-based models in protein–sugar complexes. R.A.M acknowledges Marek Cieplak in sharing the source code of the OV+rCSU contact map in Fortran. A.B.P. acknowledges financial support from the National Science Center, Poland, under grant No. 2022/45/B/NZ1/02519. R.A.M acknowledges Basque Modelling Task Force (BMTF) from the Basque Government and BCAM Severo Ochoa accreditation CEX2021-001142-S/MICIN/AEI/10.13039/501100011033. M.Ch. acknowledges Marek Cieplak for encouraging the investigation on structures containing cavities, knots, and various unstructured systems, explored using coarse-grained models. M.Ch. acknowledges financial support from the National Science Centre (NCN), Poland, under grant No. 2018/31/B/NZ1/00047 and the European H2020 FETOPEN-RIA-2019-01 grant PathoGelTrap No. 899616. A.B.P. and M.Ch. acknowledge the support of the computer resources by the PL-GRID infrastructure.

References

[1] A.B. Poma, M. Cieplak, P.E. Theodorakis, *J. Chem. Theory Comput.* **13**, 1366 (2017).

[2] P.C.T. Souza, R. Alessandri, J. Barnoud et al., *Nat. Methods* **18**, 382 (2021).

[3] J.A. Stevens, F. Grünewald, P.A.M. van Tilburg et al., *Front. Chem.* **11**, 1106495 (2023).

[4] L. Casalino, A.C. Dommer, Z. Gaieb et al., *Int. J. High Perform Comput. Appl.* **35**, 432 (2021).

[5] M.R. Machado, E.E. Barrera, F. Klein, M. Sóñora, S. Silva, S. Pantano, *J. Chem Theory Comput.* **15**, 2719 (2019).

[6] K.M. Ocetkiewicz, C. Czaplewski, H. Krawczyk, A.G. Lipska, A. Liwo, J. Proficz, A.K. Sieradzan, P. Czarnul, *Bioinformatics* **39**, btad391 (2023).

[7] K. Wołek, Ŕ. Gómez-Sicilia, M. Cieplak, *J. Chem Phys.* **143**, 243105 (2015).

[8] J.I. Sułkowska, M. Cieplak, *Biophys J.* **95**, 3174 (2008).

[9] Y. Zhao, M. Chwastyk, M. Cieplak, *J. Chem. Phys.* **146**, 225102 (2017).

[10] Y. Zhao, M. Chwastyk, M. Cieplak, *Sci. Rep.* **7**, 39851 (2017).

[11] A.B. Poma, M.S. Li, P.E. Theodorakis, *Phys. Chem. Chem. Phys.* **20**, 17020 (2018).

[12] C.H. da Silveira, D.E.V. Pires, R.C. Minardi et al., *Proteins* **74**, 727 (2009).

[13] P.G. Wolynes, J.N. Onuchic, D. Thirumalai, *Science* **267**, 1619 (1995).

[14] J.K. Noel, P.C. Whitford, J.N. Onuchic, *J. Phys. Chem. B* **116**, 8692 (2012).

[15] A. Poupon, *Curr. Opin. Struct. Biol.* **14**, 233 (2004).

[16] F. Dupuis, J.-F. Sadoc, R. Jullien, B. Angelov, J.-P. Mornon, *Bioinformatics* **21**, 1715 (2005).

[17] F. Aurenhammer, R. Klein, D.-T. Lee, *Voronoi Diagrams and Delaunay Triangulations* World Scientific Publishing Company, 2013.

[18] V. Sobolev, A. Sorokine, J. Prilusky, E.E. Abola, M. Edelman, *Bioinformatics* **15**, 327 (1999).

[19] J. Tsai, R. Taylor, C. Chothia, M. Gerstein, *J. Mol. Biol.* **290**, 253 (1999).

[20] R.A. Moreira, H.V. Guzman, S. Boopathi, J.L. Baker, A.B. Poma, *Materials* **13**, 5362 (2020).

[21] M. Chwastyk, M. Cieplak, *J. Phys. Chem. B* **124**, 11 (2020).

[22] B.R.H. de Aquino, M. Chwastyk, Ł. Mioduszewski, M. Cieplak, *Phys Rev Res.* **2**, (2020).

[23] Ł. Mioduszewski, M. Cieplak, *Phys. Chem. Chem. Phys.* **20**, 19057 (2018).

[24] Ł. Mioduszewski, J. Bednarz, M. Chwastyk, M. Cieplak, *Comput. Phys. Commun.* **284**, 108611 (2023).

[25] C. Hyeon, D. Thirumalai, *Biophys J.* **92**, 731 (2007).

[26] N.A. Denesyuk, D. Thirumalai, *J. Phys. Chem B.* **117**, 4901 (2013).

[27] N.A. Denesyuk, D. Thirumalai, *Nat. Chem.* **7**, 793 (2015).

[28] N. Hori, S. Takada, *J. Chem. Theory Comput.* **8**, 3384 (2012).

[29] T. Waleń, G. Chojnowski, P. Gierski, J.M. Bujnicki, *Nucleic Acids Res.* **42**, e151 (2014).

[30] M. Sarver, C.L. Zirbel, J. Stombaugh, A. Mokdad, N.B. Leontis, *J. Math. Biol.* **56**, 215 (2008).

[31] X.-J. Lu, H.J. Bussemaker, W.K. Olson *Nucleic Acids Res.* **43**, e142 (2015).

[32] L. Artzi, E.A. Bayer, S. Morad'S, *Nat. Rev. Microbiol.* **15**, 83 (2017).

[33] Y. Shida, T. Furukawa, W. Ogasawara, *Biosci. Biotechnol. Biochem.* **80**, 1712 (2016).

[34] T. Sztain, S.-H. Ahn, A.T. Bogetti et al., *Nat. Chem.* **13**, 963 (2021).

[35] A.B. Poma, M. Chwastyk, M. Cieplak, *J. Phys. Chem. B.* **119**, 12028 (2015).

[36] D. Reith, M. Pütz, F. Müller-Plathe, *J. Comput Chem.* **24**, 1624 (2003).

[37] R.C. Bernardi, I. Cann, K. Schulten, *Biotechnol. Biofuels.* **7**, 83 (2014).

[38] M.J. Abraham, T. Murtola, R. Schulz et al., *SoftwareX* **1–2**, 19 (2015).

[39] M.I. Mahmood, A.B. Poma, K.-I. Okazaki, *Front Mol. Biosci.* **8**, 619381 (2021).

[40] Z. Liu, R.A. Moreira, A. Dujmović et al., *Nano Lett.* **22**, 179 (2022).

[41] P.C. Kroon, F. Grunewald, J. Barnoud, M. van Tilburg, P.C.T. Souza, T.A. Wassenaar, S.J. Marrink, *eLife* **12**, RP90627 (2023).

[42] P.C.T. Souza, S. Thallmair, S.J. Marrink, R. Mera-Adasme, *J. Phys. Chem. Lett.* **10**, 7740 (2019).

[43] F. Fontana, F. Gelain, *Nanoscale Adv.* **2**, 190 (2020).

[44] P.S.F.C. Gomes, M. Forrester, M. Pace, D.E.B. Gomes, R.C. Bernardi, *Front. Chem.* **11**, 1107427 (2023).

[45] A.B. Poma, H.V. Guzman, M.S. Li, P.E. Theodorakis, *Beilstein J. Nanotechnol.* **10**, 500 (2019).

[46] A.B. Poma, T.T.M. Thu, L.T.M. Tri, H.L. Nguyen, M.S. Li, *J. Phys. Chem. B* **125**, 7628 (2021).

[47] L. Monticelli, S.K. Kandasamy, X. Periole, R.G. Larson, D.P. Tieleman, S.-J. Marrink, *J. Chem. Theory Comput.* **4**, 819 (2008).

[48] X. Periole, M. Cavalli, S.-J. Marrink, M.A. Ceruso, *J. Chem. Theory Comput.* **5**, 2531 (2009).

[49] K.A. Wilson, L. Wang, Y.C. Lin, M.L. O'Mara, *BBA Adv.* **1**, 100010 (2021).

[50] H. Sun, B. Qiao, W. Choi et al., *ACS Cent. Sci.* **7**, 2063 (2021).

[51] S. Thallmair, P.A. Vainikka, S.J. Marrink, *Biophys. J.* **116**, 1446 (2019).

[52] F. Azadi-Chegeni, S. Thallmair, M.E. Ward et al., *Biophys J.* **121**, 396 (2022).

[53] C. Waltmann, C.E. Mills, J. Wang et al., *Proc. Natl. Acad. Sci. USA* **119**, e2119509119 (2022).

[54] R.A. Moreira, J.L. Baker, H.V. Guzman, A.B. Poma, *Methods Mol. Biol.* **2340**, 357 (2022).

[55] R.A. Moreira, S.A.L. Weber, A.B. Poma, *Molecules* **27**, 976 (2022).

[56] V. Lutsyk, P. Wolski, W. Plazinski, *J. Chem. Theory Comput.* **18**, 5089 (2022).

[57] F. Grünewald, M.H. Punt, E.E. Jefferys et al., *J. Chem. Theory Comput.* **18**, 7555 (2022).

[58] V. Thallmair, L. Schultz, W. Zhao, S.J. Marrink, D. Oliver, S. Thallmair, *Sci Adv.* **8**, eabp9471 (2022).

[59] R. Launay, E. Teppa, C. Martins et al., *Int. J. Mol. Sci.* **23**, 10323 (2022).

[60] J. Morstein, R. Shrestha, Q.N. Van et al., *ACS Chem. Biol.* **18** 2082 (2023).

[61] M. Nemchinova, J. Melcr, T.A. Wassenaar, S.J. Marrink, A. Guskov, *J. Chem. Inf. Model.* **61**, 2407 (2021).

In memory of Professor Marek Cieplak

REVIEW

Key Factors Controlling Fibril Formation of Proteins

T.T.M. Thu[a,b], H.N.T. Phung[c], N.T. Co[d], A. Kloczkowski[e,f] and M.S. Li[g,*]

[a] *Faculty of Materials Science and Technology, University of Science — VNU HCM, 227 Nguyen Van Cu Street, District 5, Ho Chi Minh City, 700000 Vietnam*
[b] *Vietnam National University, Area 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, 700000 Vietnam*
[c] *Faculty of Natural Sciences and Technology, Tay Nguyen University, 567 Le Duan street, Buon Me Thuot, Vietnam*
[d] *Faculty of Chemistry, University of Gdańsk, Fahrenheit Union of Universities, Wita Stwosza 63, 80-308 Gdańsk, Poland*
[e] *The Steve and Cindy Rasmussen Institute for Genomic Medicine, Nationwide Children's Hospital, 575 Children's Crossroad, Columbus, Ohio 43215, USA*
[f] *Department of Pediatrics, The Ohio State University College of Medicine, Columbus, Ohio 43215, USA*
[g] *Institute of Physics, Polish Academy of Sciences, al. Lotników 32/46, PL-02668, Warsaw, Poland*

Doi: 10.12693/APhysPolA.145.S21 *e-mail: masli@ifpan.edu.pl

Fibril formation resulting from protein aggregation is a hallmark of a large group of neurodegenerative human diseases, including Alzheimer's disease, type 2 diabetes, amyotrophic lateral sclerosis, and Parkinson's disease, among many others. Key factors governing protein fibril formation have been identified over the past decades to elucidate various facets of misfolding and aggregation. However, surprisingly little is known about how and why fibril structure is achieved, and it remains a fundamental problem in molecular biology. In this review, we discuss the relationship between fibril formation kinetics and various characteristics, including sequence, mutations, monomer secondary structure, mechanical stability of the fibril state, aromaticity, hydrophobicity, charge, and population of fibril-prone conformations in the monomeric state.

topics: protein fibril formation, aggregation rate, neurodegenerative diseases, amyloid beta peptides

1. Introduction

The protein folding takes place in an environment crowded with other biological macromolecules. As a result, proteins are exposed to intermolecular interactions that may lead to aggregation [1]. There are about 50 human diseases characterized by aggregation of proteins [2, 3]. A large number of diseases that can be attributed to amyloidosis are due to the fact that aggregation of pathogenic proteins occurs both in the extracellular space and in the cytoplasm and nucleus. The list of diseases associated with protein aggregation continues to grow. Recently, preeclampsia, a pregnancy-specific disorder, was shown to share pathophysiological features with recognized protein aggregation disorders [4, 5]. Although proteins may vary in sequence, their disease-associated aggregates share a common fibrillar structure known as amyloid fibrils, which have a typical diameter of 7–10 nm and an X-ray diffraction pattern of about 5 Å on the meridian. Those diseases have common pathogenic pathways, in which protein self-assembly results in irreversible loss of normal structure and function along with the gain of aberrant and debilitating functions.

A large body of evidence suggests that amyloid fibrils and associated oligomeric intermediates are related to several neurodegenerative diseases, including Alzheimer's, Parkinson's, Huntington's, prion diseases, type II diabetes, and amyotrophic lateral sclerosis, among others [2, 6]. The most extensively studied case is Alzheimer's disease, which is thought to be associated with abnormal aggregation of so-called beta-amyloid (Aβ) peptides. Aβ peptides, cleaved from the amyloid precursor protein [7], mainly adopt two forms: Aβ_{40} and Aβ_{42} peptides, containing 40 and 42 amino acids, respectively. For illustration, in this review we will focus on the aggregation of Aβ peptides. However, the key factors governing the kinetics of fibril formation should be applied to any protein because they are based on general principles. For example,

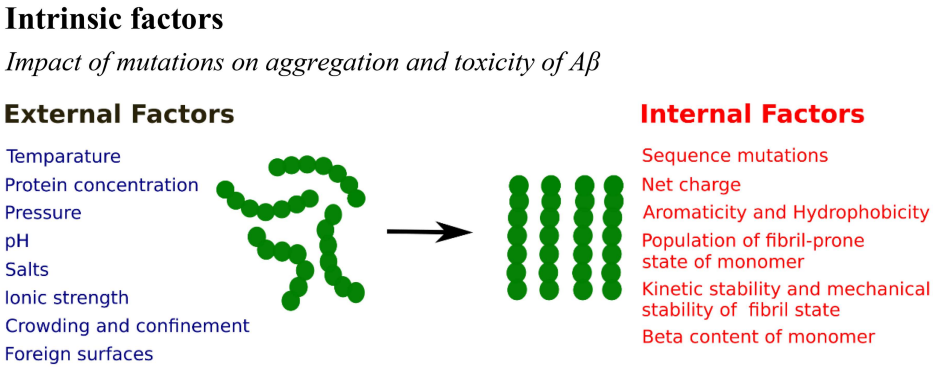


Fig. 1. Internal (group 1) and external (group 2) factors controlling protein aggregation process.

hydrophobicity, which is one such factor, is clearly universal since the stronger the hydrophobic interactions, the faster protein folding and self-assembly occurs.

Although many theoretical and experimental studies have been carried out in recent decades, our understanding of the protein aggregation process remains incomplete. It is not clear why all amyloid fibrils have the common structural feature that is a cross-β structure stabilized by backbone hydrogen bonds oriented parallel to the fibril axis. An important question then arises: What factors play a decisive role in the formation of amyloid fibrils? Many review articles have been devoted to this issue [6, 8–13], but none of them fully reflects the overall picture. The purpose of this review is primarily to present the results achieved over the years by our group, as well as the latest achievements of other groups. There are many factors that influence the aggregation process, and they can be divided into two main groups: (i) internal factors related to the intrinsic characteristics of proteins and (ii) external/environmental factors (Fig. 1). The first group refers to the properties of a polypeptide chain, including sequence, ability to resist mechanical forces (mechanical stability), aromaticity, charge, hydrophobicity, and population of the so-called fibril-prone conformation ($\boldsymbol{N}^*$) in a monomeric state [14]. The second group involves external conditions that cause aggregation, such as temperature, pH, salt concentration, and crowding. Here, we focus on the first group of factors, namely the role of mutations, mechanical stability, secondary structures, population of fibril-prone conformations, etc.

2. Intrinsic factors

2.1. Impact of mutations on aggregation and toxicity of Aβ

Alzheimer's disease (AD) is a multifactorial disease with 70% genetic and 30% environmental causes. Among genetic factors are genes associated with a family history of the disease: familial AD (FAD) and sporadic AD (SAD). Amyloid precursor protein (APP), presenilin 1 (PSEN1), and presenilin 2 (PSEN2) genes are responsible for the occurrence of FAD, while the apolipoprotein E (APOE) gene is responsible for SAD.

We focus on mutations of Aβ peptides, which are related to FAD because these peptides are cleaved from APP by β- and γ-secretases. Mutations can change the morphology of aggregates and toxicity (Table I [15–45]), and their study is, therefore, important for understanding the molecular mechanism of AD.

Experimental [30, 46–49] and theoretical [50–57] studies revealed that mutations in the turn region such as Flemish (A21G), Dutch (E22Q), Italian (E22K), Arctic (E22G), Iowa (D23N), and Osaka (E22Δ, remove glutamic acid) mutants can change aggregation properties. While most of them enhance the toxicity and self-assembly of Aβ, the Flemish mutant reduces not only the aggregation rate but also toxicity of Aβ_{40} and Aβ_{42} [30, 31, 58]. Aβ_{40}(A21G) behaves like the wild-type (WT), but with a slower expansion phase [58]. In line with the conclusion of Betts et al. [58], Murakami et al. [31] observed that the aggregative potency of the Flemish mutant was lowest among the mutants at the 21–23 region, and the thioflavin (ThT) dye fluorescence of this peptide was weaker than WT. In contrast to A21G, Aβ_{40}(E22G) is more neurotoxic and aggregates faster than the wild-type during both the lag phase and saturation phase. The Arctic mutation also changes the formation of Aβ40 wild-type (Aβ_{40}-WT) from network-like to annular protofibrils [31, 59]. In Aβ_{42}, the E22G mutation aggregates slightly slower than WT but increases protofibril formation [31, 32, 59]. In addition, Lo et al. [33] have shown that Arctic (E22G) mutation increases the aggregation rate of Aβ in micelle solution by decreasing helical structure in the 15–25 segment. Liang et al. [34] studied the three-point mutation of Aβ_{40} (L17A/F19A/E22G) and found that Aβ_{40}(E22G) can reduce the toxicity when combined with L17A and F19A by reducing the β-content and by enhancing the α-helix structure.

Mutations of Aβ peptides and their effect on aggregation rate, toxicity, and aggregate morphology. TABLE I

Mutation	Reference	Aggregation rate	Toxicity	Morphology
Aβ_{40}(D1Y)	Maji et al. [15]	reduce	reduce	oligomer long, unbranched fibrils with smooth margin
Aβ_{42}(D1Y)	Maji et al. [15]	reduce	reduce	oligomer long, unbranched fibrils with smooth margin
Aβ_{40}(D1E-A2V)	Qahwash et al. [16]	reduce	increase (slightly)	short protofibrils
Aβ_{42}(A2F)	Luheshi et al. [17]	increase	increase	
Aβ(pE3-42)	Jawhar et al. [18]	increase	increase	
Aβ_{40}(A2V)	Di Fede et al. [19]	increase	increase	straight, unbranched,s 8-nm-diameter fibril
Aβ_{42}(A2V)	Di Fede et al. [19], Messa et al [20]	increase	increase	annular oligomer
Aβ_{40}(A2T)	Jonsson et al. [21], De Strooper et al. [22]	reduce	reduce	
Aβ_{42}(A2T)	Jonsson et al. [21], De Strooper et al. [22]	reduce	reduce	
Aβ_{40}(H6R)	Janssen et al. [23], Hori et al. [24]	increase	increase	
Aβ_{42}(H6R)	Janssen et al. [23], Hori et al. [24]	increase	increase	
Aβ_{40}(D7H)	Hori et al. [24]	increase	increase	mature fibril
Aβ_{42}(D7H)	Hori et al. [24]	increase	increase	oligomer
Aβ_{40}(D7N)	Hori et al. [24], Wakutani et al. [25], Ono et al. [26]	increase	increase	
Aβ_{42}(D7N)	Hori et al. [24], Wakutani et al. [25], Ono et al. [26]	increase	increase	
Aβ_{42}(E11K)	Zhou et al. [27]	increase	increase	
Aβ_{42}(K16N)	Kaden et al. [28]	increase	reduce	random globular structures
Aβ_{40}(K16N)	Kaden et al. [28]	increase	increase	random globular structures
Aβ_{40}(K16A)	Kaden et al. [28], Sinha et al. [29]	reduce	reduce	long unbranched fibrils
Aβ_{42}(K16A)	Kaden et al. [28], Sinha et al. [29]	reduce	reduce	
Aβ_{40}(K28A)	Sinha et al. [29]	reduce	reduce	Long unbranched fibrils
Aβ_{42}(K28A)	Sinha et al. [29]	reduce	reduce	
Aβ_{40}(A21G)	Hendricks et al. [30], Murakami et al. [31]	reduce	reduce	
Aβ_{42}(A21G)	Hendricks et al. [30], Murakami et al. [31]	reduce	reduce	
Aβ_{40}(E22Q)	Murakami et al. [31]	increase	increase	ribbon-like structure
Aβ_{42}(E22Q)	Murakami et al. [31]	increase	increase	
Aβ_{40}(E22G)	Murakami et al. [31], Nilsberth et al. [32], Lo et al. [33]	increase	increase	annular protofibril
Aβ_{42}(E22G)	Murakami et al. [31], Nilsberth et al. [32], Lo et al. [33]	increase	increase	
Aβ_{40}(E22G-L17A-F19A)	Liang et al. [34]	reduce	reduce	
Aβ_{40}(E22Δ)	Ovchinnikova et al. [35], Berhanu et al. [36]	increase	increase	
Aβ_{42}(E22Δ)	Ovchinnikova et al. [35], Berhanu et al. [36]	increase	increase	
Aβ_{40}(E22K)	Murakami et al. [31]	increase	increase	
Aβ_{42}(E22K)	Murakami et al. [31]	increase	increase	
Aβ_{40}(D23N)	Murakami et al. [31], Qiang et al. [37]	NA	increase	

TABLE I cont.

Mutation	Reference	Aggregation rate	Toxicity	Morphology
Aβ_{42}(D23N)	Murakami et al. [31]	reduce	increase	
Aβ_{42}(G25L)	Fonte et al. [38], Hung et al. [39]	NA	reduce	
Aβ_{42}(G29L)	Fonte et al. [38], Hung et al. [39]	NA	reduce	
Aβ_{42}(G33L)	Fonte et al. [38], Hung et al. [39], Decock et al. [40]	increase	reduce	oligomer
Aβ_{42}(G37L)	Fonte et al. [38], Hung et al. [39]	NA	reduce	
Aβ(G33L-G38L)	Decock et al. [40]	increase oligomer	NA	oligomer
Aβ_{42}(G33A)	Hameier et al. [41]	increase	reduce	
Aβ_{42}(G33I)	Hameier et al. [41]	increase	reduce	
Aβ_{40}(A30W)	Estrada-Rodríguez et al. [42]	NA	reduce	
Aβ_{42}(A30W)	Estrada-Rodríguez et al. [42]	NA	reduce	
Aβ_{40}(M35C)	Estrada-Rodríguez et al. [42]	NA	reduce	
Aβ_{42}(M35C)	Estrada-Rodríguez et al. [42]	NA	reduce	
Aβ_{42}(G37V)	Thu et al. [43]	NA	reduce	ellipse-like
Aβ_{42}(V36P-G37P)	Roychaudhuri et al. [44]	reduce	reduce	
Aβ_{40}(G33V-V36P-G38V)	Roychaudhuri et al. [44]	increase	increase	
Aβ_{42}(G33V-V36P-G38V)	Roychaudhuri et al. [44]	increase	increase	
I41K, A42R, I41D-A42Q, I41D-A42S, I41H-A42D, I41E-A42L, I41H-A42N, I41T-A42N, I41T-A42Q, I41L-A42N, I41Q-A42Y, I41Q-A42L, I41T-A42M, I41T-A42I, I41K-A42L, I41R-A42R	Kim et al. [45]	reduce	NA	

In experimental work on E22Q mutation, Miravalle et al. [60] have shown that after 24 hours of incubation in ThS dye, only Aβ_{40}(E22Q) peptide revealed the presence of short filaments with a ribbon-like structure, whereas WT and E22K peptides did not show any presence of the fibrous state. In addition, E22Q has the highest amount of β-structures with the contribution of β-sheets and β-turns. Murakami et al. [31] also revealed that Aβ_{42}(E22Q) has the strongest aggregation in the 21–23 region in comparison to WT, Italian, Arctic, and Iowa mutants. Thus, the Dutch (E22Q) mutation was considered the most toxic in the 21–23 region [31, 60]. Similarly to E22Q, Italian (E22K) mutant aggregates faster than WT and has more toxicity for both Aβ_{40} and Aβ_{42} [31, 60]. These results support the clinical evidence that patients with Dutch and Italian mutations are diagnosed with hereditary cerebral hemorrhage with amyloidosis (HCHWA). Another D23N mutation in the turn region, studied by nuclear magnetic resonance (NMR) spectroscopy by Qiang et al. [37], shows the fibril morphology with the cross β structure. Murakami et al. [31] have shown that Aβ_{42}(D23N) (Iowa) mutant has a 2–3-fold more potent cytotoxicity and slightly slower aggregation rate than wild-type Aβ_{42}. Deletion of glutamic acid at residue 22, i.e., E22Δ mutation, increases the aggregation rate of Aβ_{42} and Aβ_{40} peptides, and this mutation is also more toxic than WT [35]. Berhanu et al. [36] have shown that the fibril structure of E22Δ is more stable than WT. Glutamic acid (E) is an important amino acid, being acidic polar with a negative charge. When E is substituted by Q or G (neutral amino acids), the primary structure of Aβ changes and the secondary structure also changes, resulting in the enhancement of toxicity. Therefore, decreasing the negative charge at residues 22, 23 of Aβ can increase the aggregation rate and toxicity of this peptide. Electrostatic interactions in this region play an important role in the thermodynamic stability and neurotoxicity of Aβ [61, 62].

Previous studies have shown that the N-terminus region (residues 1–8 of Aβ_{40} and 1–16 of Aβ_{42}) is disordered in the fibril state [63–66]. NMR spectroscopy studies revealed that the fibril structure of Aβ_{42} forms the parallel β-sheet like a "hairpin" [63, 65, 67, 68], and the hydrophilic turn region bends to form the U-shape. Some simulations ignored the N-terminus segment, considering only the 17–42 segment of Aβ_{42} or 9–40 of Aβ_{40} [69, 70]. However, experimental studies of the whole

Fig. 2. A2T, A2V, E11K mutations and secretase cleavage sites [27].

structure of Aβ, including the N-terminus region, concluded that residues at the N-terminus region are ordered and play an important role in the whole sequence of Aβ and its toxicity [71–73]. The fibrillar structure of Aβ consists of three β-sheets forming the S-shape [74]. Antibodies bound at the N-terminus interact well with soluble and insoluble Aβ species [75]. Thus, the N-terminal region plays an important role in the Aβ assembly, suggesting that the binding of small molecules in this region may inhibit the Aβ-induced toxicity [76].

The H6R (English) [23, 24], D7H (Taiwanese) [77], and D7N (Tottori) [24–26] mutations were found to stabilize the secondary structure of $A\beta$, which enhances the aggregation rate [24]. Besides, Aβ_{40}(D7H) has a propensity to form mature fibrils, while Aβ_{42}(D7H) prefers to form oligomers [77]. Experimental studies revealed that the D1Y mutation of Aβ_{40}/Aβ_{42} slows the assembly process [15], while the A2F mutation increases its toxicity [17]. The two-point mutation D1E-A2V influences the fibril morphology of Aβ_{40} [16]. While Aβ_{40} WT forms long fibrillar aggregates, Aβ_{40}(D1E-A2V) develops only protofibrillar morphologies. Cellular toxicity assays indicated that Aβ_{40}(D1E-A2V) was slightly more toxic than Aβ_{40} WT to human neuroblastoma SHEP cells and rat primary cortical and hippocampal neurons. Deletion of the two first residues of Aβ_{42} and the substitution of glutamic acid at the 3^{rd} position by pyroglutamic acid, i.e., Aβ(pE3-42), is considered a key factor in the pathology of AD because of the high aggregation propensity of the mutant, its abundance in AD brain, and cellular toxicity [18].

The FAD mutation Aβ_{40}(A2V) causes an early onset of AD [20]. It was revealed that A2V levels up the aggregation kinetics of Aβ_{40}, but the mixture of Aβ_{40} wild-type and Aβ_{40}(A2V) reduces the toxicity of this mutation [19]. For the Aβ_{42} peptide, the A2V mutation has a different fibril morphology and increases the aggregation rate. The fibril of Aβ_{42}(A2V) has the prevalent content of annular structures with higher hydrophobicity and toxicity [20].

In contrast to A2V, the A2T mutation has a strong protective effect, preventing cognitive decline in the elderly without AD [21, 22]. The A2T mutation is located at the second residue in the Aβ peptide, corresponding to residue 673 in APP and nearby β-secretase (residue 672 in APP, see Fig. 2).

Zhou et al. [27] studied E682K mutation (site of β' enzyme in Fig. 2) on APP, corresponding to E11K on Aβ_{42}. They showed that E11K enhances the formation of Aβ from APP by β, β', and γ enzymes [27] (Fig. 2). Therefore, individuals having this mutation can get AD at age 49–53, i.e., earlier than others [27]. Kaden et al. [28] reported that the K16N mutation (site of enzyme α in Fig. 2) enhances the toxicity of Aβ when mixing K16N and WT. The aggregate of this mutation has an oligomeric structure of various sizes. Replacing K with A at residue 16, the K16A mutation reduces the toxicity of Aβ by changing the morphology of aggregates and increasing the content of the α-structure [28, 29]. Sinha et al. [29] studied the K28A mutation, showing that the substitution of lysine by alanine inhibits Aβ toxicity. In addition, the K28A mutation reduces the process of conversion in the secondary structure and enhances the random coil structure. These observations support the hypothesis that Lys28 stabilizes the nucleation phase in the fibrillization process proposed by Lazo et al. [78].

The C-terminus region (residues 31–40/31–42) of Aβ is stable and plays a key role in aggregation and binding with other ligands [79, 80]. Thus, mutations in this area are of great interest. The glycine zipper motif at the C-terminus, including glycine at residues 25, 29, 33, 37, can influence the transformation of a random helix or α-helical structure into a β-sheet and, therefore, fibril formation. Destabilizing this structure by mutations is an effective way to study its role [38–41]. Mutants G25L, G29L, G33L, and G37L, where glycine was replaced by leucine, were shown to be less toxic than Aβ_{42} WT in mouse primary cortical neurons [39]. Research by Fonte et al. [38] supports this idea; in particular, G37L reduces the toxicity of Aβ in all models tested. The G33L mutation enhances the oligomer structure of Aβ [41], but when mixed with G38L, this effect becomes weaker [40]. Thu et al. [43] replaced glycine with valine at residue 37 and found that the G37V mutation did not change the rate of aggregation but reduced the toxicity and changed the fibril morphology from network to ellipse-like shape.

In vitro and *in vivo* experiments showed that Aβ_{42} oligomers with the replacement of glycine 33 by isoleucine and alanine are much less toxic than Aβ_{42} WT [41]. In addition, mutations G33A and G33I promoted aggregation by increasing the

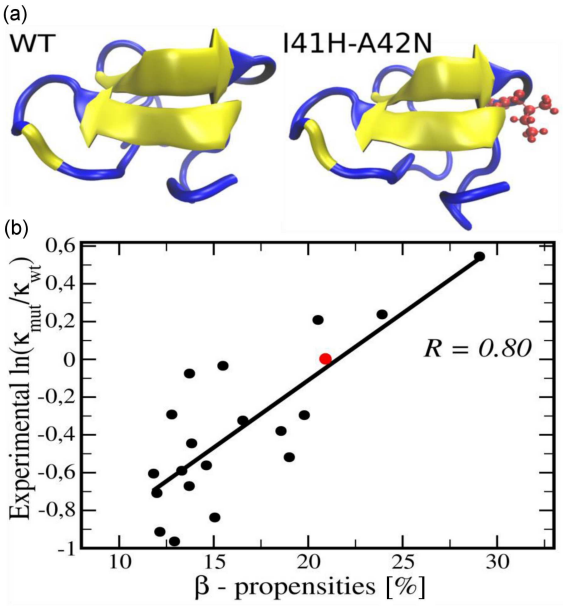


Fig. 3. Initial structure for MD simulation of Aβ_{42}-WT and I41H-A42N mutation; mutated residues are in red (panel a). Dependence of the logarithm of the relative aggregation rate on β-content; the red circle refers to WT. Linear fit is $y = -1.534 + 0.071x$ ($R = 0.80$) (panel b) [83]. References to experimental data shown in this figure can be found in [83].

population of large oligomers (16- to 20-mers) at the expense of small oligomers (2- to 4-mers). Recently, Rodríguez et al. [42] have studied one-point mutations K28A, A30W, and M35C on the 25–35 segment of Aβ. They have found that fibril formation was more dependent on the primary sequence of peptides than on their length. Mutations A30W and M35C reduced toxicity by the reducing production of reactive oxygen species (ROS) but did not affect the aggregation rate. Thus, the primary sequence is most important for the cytotoxicity of Aβ.

In another study, Roychaudhuri et al. [44] found that the β-hairpin motif with a β-turn at residues Val36-Gly37 is highly populated in Aβ_{31-42} but does not exist in Aβ_{31-40}. In addition, the three-point mutation G33V-V36P-G38V (VPV) levels up the β-turn and β-hairpin content at the C-terminus, increases cytotoxicity, and alters the aggregate morphology. The VPV mutation makes Aβ_{40} oligomerization as fast as Aβ_{42}, while Aβ_{42} becomes "super Aβ_{42}". In contrast to VPV, the V36P-G37P two-point mutation of Aβ_{42} produces Aβ_{40}-like oligomers instead of forming hexamers and dodecamers. This study showed that the V36P-G37P mutation leads to the abolishment of β-turn formation at residues 36–37 and reduces the β-content and toxicity of Aβ_{42} [79]. Linh et al. [81] performed all-atom molecular dynamics (MD) simulations of the full-length Aβ_{40} and Aβ_{42} and obtained results different from those of Roychaudhuri et al. [44], indicating that the VPV mutation promotes the β-turn structure at residues 36–37 but is insufficient to make Aβ_{40}(VPV) oligomerization to become like Aβ_{42} WT [81]. Besides, the β-hairpin motif at residues 36–37 present in Aβ_{42} WT does not appear in Aβ_{40}(VPV).

Kim et al. [45] synthesized mutants by replacing I41 and A42 with less hydrophobic amino acids. They showed that substitution of these residues with negatively charged hydrophilic amino acids (I41D-A42Q, I41D-A42S, I41H-A42D, I41E-A42L), neutral hydrophilic amino acids (HN, TN, TQ, LN, QY, QL, TM, TI), or positively charged residues (I41K, KL, RR, A42R), slows aggregation. Thus, the last two residues, namely I41 and A42, play an important role in the aggregation process and toxicity of Aβ_{42}. Table I shows mutations in Aβ peptides and their impact on various properties.

2.2. Beta-content in monomeric state

The influence of secondary structure on the aggregation rate of protein was studied indirectly by Chiti et al. [82] by finding the effect of the free energy change in conversion from the α-helix to β-sheet conformation ($\Delta\Delta G$). The equation representing the correlation between the combination of $\Delta\Delta G$, the change in hydrophobicity (ΔHydr), and the change in charge (ΔCharge) with the aggregation rates for the mutant κ_{mut} and the wild-type κ_{wt} is [82]

$$\ln\left(\frac{\kappa_{mut}}{\kappa_{wt}}\right) = A\Delta\Delta G + B\Delta\text{Hydr} + C\Delta\text{Charge}. \quad (1)$$

By choosing the appropriate fit parameters A, B, and C, a high correlation between these quantities was found. Recently, Thu et al. [83] have calculated the β-content of 19 mutations of Aβ_{42} using replica exchange molecular dynamics simulation in implicit water. They showed that the experimentally measured aggregation rate κ depends on the calculated β-content in monomeric state $\kappa = \kappa_0 \exp(c\beta)$, $c = 0.071$ with the correlation level $R = 0.80$ (Fig. 3b). Thus, the higher the β-propensity, the faster formation of fibrils. It would be interesting to test this conclusion on other systems.

2.3. Population of fibril-prone state in monomeric state

It is known that in the monomeric native state, the protein is compact, and in the fibrillar state, it forms an expanded β-structure, which is called the fibril-prone state **N***. Consequently, **N*** is an excited state in the energy spectrum of the monomer [84]. In lattice models, for a chain with a sufficiently small number of beads, it is possible

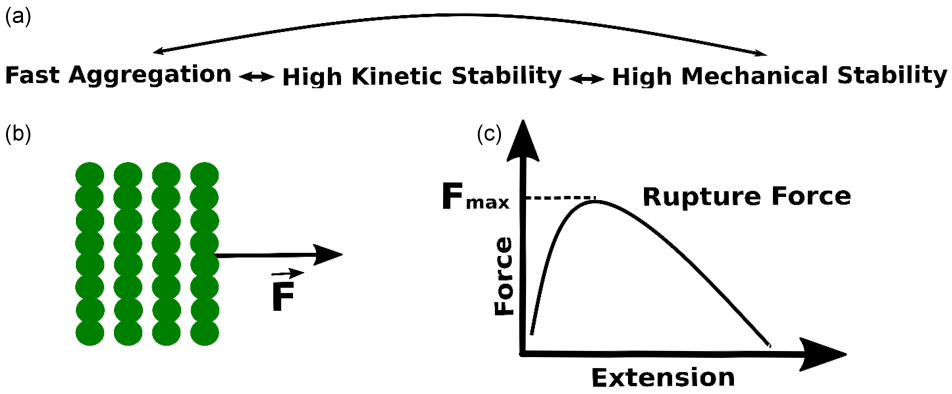


Fig. 4. (a) Relationship between aggregation rate, kinetic stability, and mechanical stability of the fibril state. (b) Pulling a chain from the fibrillar structure to probe its mechanical stability. (c) Force-extension profile with the rupture force defined as $F_{\max}$.

to perform an exact search of all possible conformations and find the full energy spectrum. The fibril-prone state population is then defined as

$$P_{N^*} = \frac{1}{Z} \exp\left(-\frac{E_{N*}}{k_{\mathrm{B}}T}\right), \tag{2}$$

$$Z = \sum_{i=1}^{N} \mathrm{e}^{-E_i/(k_{\mathrm{B}}T)}, \tag{3}$$

where Z is the partition function, E_{N^*} is the energy of the $\boldsymbol{N}^*$ state, E_i is the energy of the i-th state, and N is the total number of states. However, in most cases (continuum models or lattice models with long chains), the exact energy spectrum cannot be found, and (2) cannot be used to calculate P_{N^*}. In this situation the population of fibril-prone state is approximately estimated using simulations. Namely, P_{N^*} is defined as the time of the appearance of the $\boldsymbol{N}^*$ state during the entire simulation divided by the total simulation time.

Li et al. [14] proposed that the higher the population, the faster the rate of fibril formation. The rationale for this hypothesis is that the $\boldsymbol{N}^*$ state can serve as a template for fibril formation, and its high population will promote this process. Using the toy lattice model [84], it was shown that the fibril formation rate decreases exponentially with increasing P_{N^*},

$$\kappa \sim \exp(-cP_{N*}), \tag{4}$$

where the constant c depends on the studied models and systems. The validity of the $\boldsymbol{N}^*$-theory (see (4)) has been confirmed for all-atom [85] and off-lattice coarse-grained models [86]. Recently, using the coarse-grained self-organized polymer-intrinsically disordered protein (SOP-IDP) model [87] and MD simulations, Chakraborty et al. [88] have shown that P_{N^*} of Aβ_{42} is larger than that of Aβ_{40}, which indicates that, in agreement with experiment [89], the former aggregates faster than the latter due to the two terminal hydrophobic residues. It was shown that the population of the $\boldsymbol{N}^*$ state depends on the morphology of fibrils, implying that the shape of the aggregate depends on the time of its formation. In other words, the $\boldsymbol{N}^*$-theory ascertains that fibrillar polymorphism is time-dependent or under kinetic control [90]. Assuming that the fibril formation obeys Ostwald's rule, which states that the least stable polymorph would form first, followed by a subsequent transition to a more stable form, Chakraborty et al. [88] predicted that the S-bend Aβ_{42} fibril is more stable than the U-bend form, as the latter forms faster.

2.4. Mechanical stability of the fibril state

Kouza and co-workers [91] proposed a new definition of the kinetic stability (G_{fib}) of the fibrillar state based on the probability (P_{fib}) of reaching the fibrillar configuration, i.e.,

$$G_{fib} = -k_{\mathrm{B}}T \ln(P_{fib}), \tag{5}$$

$$\tau_{fib} = \exp\left(a\, G_{fib}\right). \tag{6}$$

Their computational study also indicated that the fibril formation time (τ_{fib}) showed no clear correlation with the fibril state energy (E_{fib}) or the free energy of the system. Instead, τ_{fib} displayed an exponential dependence on G_{fib} (see (6)). This relationship between G_{fib} and τ_{fib} can be interpreted as evidence that the kinetic stability of the fibrillar state correlates with the rate of fibril formation. Moreover, this relationship can be qualitatively understood using the framework shown in Fig. 4a. On the one hand, the higher the mechanical stability, the higher the kinetic stability, determined by (5). On the other hand, the higher the kinetic stability, the faster the aggregation occurs. Consequently, the higher the mechanical stability of the fibrillar state, the faster the fibril formation.

The mechanical stability of the fibril can be accessed using steered molecular dynamics (SMD) simulations [91]. Namely, this mechanical stability can be characterized by the rupture force or the maximum force in the force-extension/time profile

obtained by pulling a single chain from the fibril structure (Fig. 4b and c). Using all-atom models to calculate mechanical stability and fibril formation time for short peptides such as KLVFF and FVFLM, the relationship between these two quantities was confirmed [91]. Aβ_{42} has been experimentally shown to form fibril faster than Aβ_{40} [71, 89], which is consistent with SMD simulations that the former is mechanically more stable than the latter [91]. By performing all-atom SMD simulations for 20 Aβ_{42} mutants whose aggregation rates are known from experiments, Thu and Li [92] obtained clear evidence that the aggregation rate correlates with the mechanical stability of the fibrillar structure. Since calculating fibril formation times for relatively large proteins using all-atom models is computationally prohibitive, this relationship is very useful as it allows us to estimate τ_{fib} from the rupture force, which can be easily obtained from SMD simulations.

3. External factors

The process of protein folding, which involves the transformation of proteins into their three-dimensional functional conformations or native states, serves as a core principle in structure biology. However, proteins are also prone to adopting energetically preferential aggregated configurations, a phenomenon known as protein misfolding or aggregation. Numerous variables, including the inherent characteristics of the proteins, the environmental physical conditions, or the overcapacity of the regulatory systems, potentially influence this process. Figure 1 illustrates the standard external factors that contribute to the protein aggregation process.

3.1. Temperature

Thermal variations have a significant impact on the process of associating monomers into higher-ordered structures [9, 93, 94]. A significant enhancement in the aggregation rate of β-lactoglobulin was observed as the temperature shifted from 30 to 50°C [95]. The temperature range of 29–45°C and 4–40°C was reported to elicit an acceleration in the nucleation and elongation phases of self-assembly for Aβ peptides [96, 97]. Elevated temperatures can cause the protein to deviate from its native conformation, resulting in partially or fully denatured states in which hydrophobic regions are exposed to the solvent environment. The presence of such hydrophobic cores, as a result of thermal denaturation, increases the likelihood of intermolecular interactions between proteins, leading to the acceleration of the aggregation process [98, 99]. It is noteworthy that a specific subset of proteins exhibits a phenomenon known as cold-induced denaturation, where the stability of the native structure is lost at

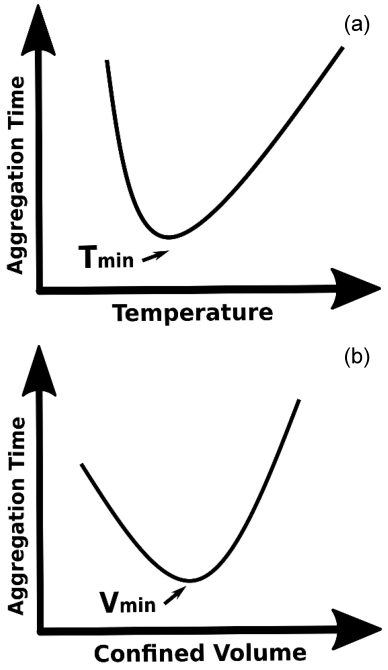


Fig. 5. U-shape dependence of aggregation time on temperature (a) and confined volume (b). The fastest aggregation occurs at $T_{\min}$ and $V_{\min}$.

low temperatures, resulting in the acceleration of the condensation process [100, 101]. Self-assembly of ribosomal protein L9 occurs faster at 4°C than at 25°C [101]. The reduction of temperature from 37 to 5°C catalyzed the monoclonal antibody aggregation [102]. The adherence of temperature-dependent rate of aggregation of certain proteins to the traditional Arrhenius law was observed within a narrow temperature range [103–105]. Nevertheless, the Arrhenius law may not govern the behavior of proteins within a wider temperature range [94, 106–109].

The complex effect of temperature on proteins, both direct and indirect, influences the self-assembly process in multiple ways. However, in general, the temporal dependence of aggregation on temperature can be understood as the trade-off between the entropy and energy of the studied system, which is typically characterized by a U-shaped form (Fig. 5a). The minimum aggregation time is attained at the optimal temperature ($T_{\min}$) that corresponds to the highest rate of aggregation [110].

3.2. Protein concentration

The concentration of proteins plays a pivotal role in modulating protein aggregation propensity because it significantly influences both the intermolecular distances and the interaction among protein molecules. The critical concentration is defined as the concentration above which protein self-assembly occurs. This concentration depends on the specific type of protein and typically fluctuates within

the micro-molar to nano-molar range in biophysical conditions. Polyglutamine (polyQ) [111], β-ovalbumin [112], and α-synuclein [113] initiate their self-assembly at respective concentrations approximating 3 μM, 7 μM, and 0.7 μM. For Aβ_{42} and Aβ_{40} peptides, their assembly thresholds are in the μM range [114, 115] and could potentially reach nM levels [116]. Furthermore, an increase in monomer concentration results in a decrease in both the lag time as well as the overall time required for aggregation owing to the intensification of collision frequencies among the monomers [113]. Nevertheless, high protein concentration can result in the retardation of aggregation [117, 118] due to the trade-off between on-pathway and off-pathway oligomers [119].

In the case of protein self-assembly via the primary nucleation mechanism, the relation between characteristic times τ_F (for instance, lag time or half time) and the concentration c is represented by $\tau_F \approx c^{-(n_c+1)/2}$ [120], where n_c is the size of the critical nucleus. It is worth noting that a distinct dependence on concentration has been discerned in the scenario of secondary nucleation [121].

3.3. Pressure

High hydrostatic pressure exerts influences on the conformation of proteins, the interactions between proteins, and the formation of polymers or aggregates through volume modifications [122]. Some studies suggested that the volumetric fluctuation arising from the exclusion of water from internal cavities [123, 124], the hydration of hydrophobic surfaces [125], the dissociation, and the rupture of associated ion-pair interactions [126] are the underlying causes of pressure-induced protein unfolding and may have a consequential impact on protein aggregation rates under high-pressure conditions [127–129].

3.4. pH

The acidity of a solution (pH) plays a role in the charge density of the protein surface. A highly acidic pH environment causes a concentration of similar charges on the surface of peptides, leading to strong repulsion and hindering the self-assembling of peptide molecules. For instance, the formation of salt bridges between Lys28 and Asp23 is prevented due to the neutralization of residue Lys28 at pH levels greater than 9.5, resulting in the inhibition of the self-assembly of peptides Aβ_{42} [130]. Generally, the tendency for protein to aggregate increases at pH values near the isoelectric point of the protein [131].

3.5. Ionic strength

The kinetics of protein aggregation, as well as the morphological characteristics of aggregated products, are significantly influenced by the ionic strength of the surrounding medium. Multiple deposition forms of α-synuclein were noted by Hoyer et al. [132] within NaCl and $MgCl_2$ solutions. Amyloid fibrillogenesis of β_2-microglobulin was affected by the addition of anions SO_4^{2-}, Cl^-, I^-, ClO_4^- to the solution [133]. Other investigations have explored the impacts of ionic strength on the propensity for aggregation in proteins such as β-lactoglobulin [134], islet amyloid polypeptide (IAPP) [135], Aβ_{40} [136], or Aβ_{42} [137].

3.6. Salts

In solution, the binding of unpaired charged residues or backbones of proteins with the cations and anions generated from salt dissolution can lead to alterations in protein structures and protein dissolution capacity or affect inter-protein interactions, thereby influencing the propensity for protein self-assembly [138–140]. Adding salt ions to the solution of HCA II at temperature 328 K switched HCA II aggregation behavior from a monophasic to a biphasic mechanism [141]. The competitive formation of amyloids versus amorphous aggregates was observed by Adachi et al. [142] as they were studying the effect of varying NaCl concentration on the aggregation rate of β_2-microglobulin — bovine serum albumin kinetics changed from downhill to nucleation-dependent kinetics in the presence of guanidinium hydrochloride (GdmCl) and $CaCl_2$ in the studied solution [143]. NaCl can increase the rhGCSF aggregation rate [144], yet it can also impede the self-assembly capacity of the recombinant factor VIII SQ [145].

3.7. Crowding and confinement

Protein misfolding and aggregation occur in an environment that includes a variety of components called crowders. In biological organisms, crowders, including proteins, sugars, lipid membranes, chaperones, nucleic acids, collagen, and others, can account for up to 40% of living matter [146–148]. *In vitro* settings, crowders can be artificially introduced substances such as nanoparticles [149] or polymers [150]. Crowders can speed up the self-assembly process of proteins, which is primarily explained by their volume exclusion effect, which narrows the spatial region available to proteins and thereby reduces their entropic cost [151–155]. In contrast, in a densely populated environment with sufficiently small particles, the aggregation process may be slowed due to diffusion restrictions imposed on peptides by crowders [149, 156, 157] or the potential deformation of proteins from their aggregation-prone states [158].

Often intertwined in discussions due to their strong correlation, crowding and confinement are distinct yet related concepts in protein aggregation. Crowding refers to the densely populated milieu

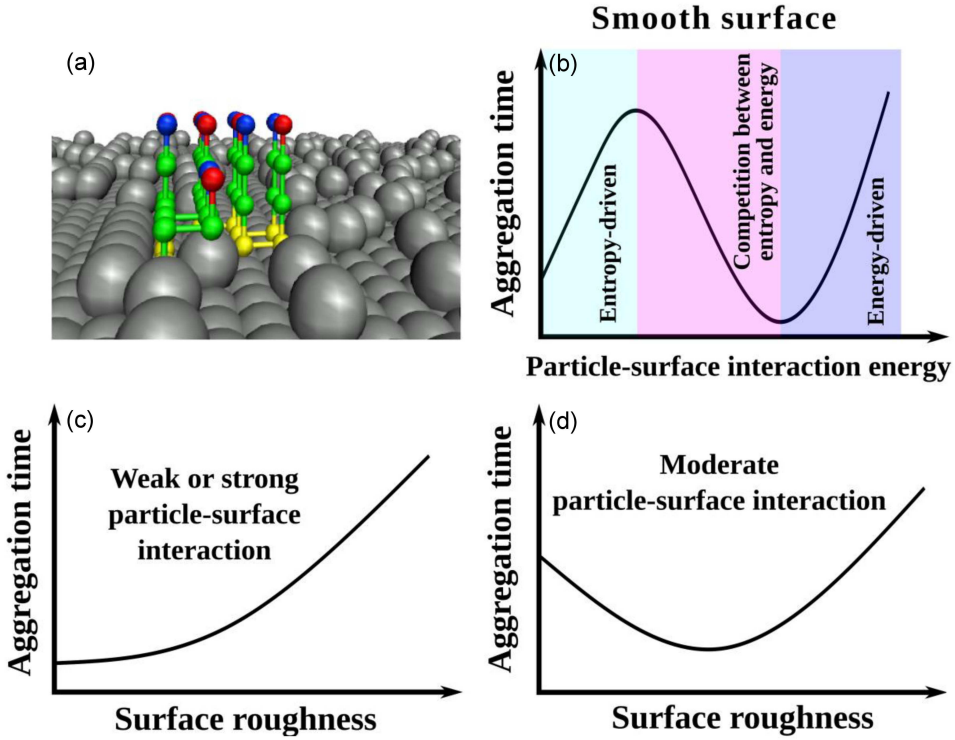


Fig. 6. Deposition of six peptides on the rough surface in the lattice model (a), the effect of varying protein–smooth surface interaction energy on the protein aggregation time (b), the relation between the aggregation time and roughness degree of foreign surfaces under the conditions of small and high (c), and moderate (d) particle–surface interactions. Weak, moderate, and strong particle–surface interactions in (c) and (d) correspond to entropy-driven, entropy—energy competition, and energy-driven regimes in (b).

in which the aggregation occurs, whereas confinement addresses the association of proteins within the fixed or rigid structures, which may include chaperones, ribosome exit tunnels, or cytoskeleton [159]. The interplay between entropy and energy in proteins manifests as a U-shaped curve representing the dependence of protein aggregation rate on confined volume (Fig. 5b). In highly confined spaces, the restriction of the conformational entropy of proteins prevents them from reaching an optimal energy state, hence significantly elongating the aggregation time. Conversely, as the volume of confinement increases, the protein conformational entropy experiences a sharp increase, which in turn leads to a slowdown in the aggregation process [156].

3.8. Foreign surfaces

The phenomena of protein aggregation are not limited to solution environments but are also observed under various surface conditions. Although the presence of foreign surfaces can be perceived as crowding agents, their role extends beyond the traditional crowding concept, which primarily captures the global effects of the environment on protein self-assembly. Indeed, the influence of foreign surfaces on protein aggregation has received considerable attention due to their wide range of applications spanning drug discovery, new materials development, and polymer science [160–162]. Foreign surfaces can potentially expedite the aggregation process; numerous lipid membranes, for instance, have been observed to catalyze fibril growth [163–165]. Other examples include mica and glass surfaces, which have been reported to act as catalysts in the fibrillation of α-synuclein [166] and Aβ_{18-22} [167]. Conversely, certain external surfaces may have a suppressive influence on the assembly of amyloid fibrils; in particular, the deposition of IAPP was notably inhibited in a milieu containing surfaces coated with polymeric nanoparticles [149], as well as the self-assembly of Aβ_{42} into fibril-like structures was slowed down by protein-coated surfaces of graphene oxide [168]. Additionally, the interaction between proteins and surfaces could extend to the modification of fibril morphology [169, 170], sometimes up to the complete alternation of their fibrillar structures [171]. Furthermore, the aggregation propensity of proteins has been demonstrated to be sensitive to changes in surface topography [172] and the degree of surface roughness [173, 174], highlighting the intricate and subtle nature of protein–surface interactions.

Applying a simple lattice model for investigating the aggregation of 8-bead chains on both smooth and rough surfaces (Fig. 6a), Co and Li [174] proposed a general scheme for understanding the

self-assembly in the presence of foreign surfaces. They found that due to the trade-off between entropy and energy, a moderately absorbing smooth surface promoted protein aggregation, while weakly and strongly absorbing surfaces hindered the process (Fig. 6b). For rough surfaces, both weakly and highly absorbent surfaces tended to increase the duration of the aggregation process (Fig. 6c). However, moderately absorbent surfaces showed a dual effect, i.e., at higher roughness levels, these surfaces inhibited protein deposition, whereas at lower roughness levels, they catalyzed the aggregation process (Fig. 6d).

3.9. Other external factors

Here we list other external factors that are not discussed in this review: oxidative stress [175, 176], organic solvent [177], ligands [178], freezing [179], thawing [180], metal ions [181, 182], UV illumination [183], drying [184], pumping [185], surfactants [186], biopolymer [187], interface- and mechanical-force-mediated amyloid formation [188].

4. Conclusions

In this review, we have focused on new developments related to intrinsic and external factors that can influence protein aggregation. For many decades, protein folding research has been dominated by the assumption that thermodynamics determines protein structure and function. However, recent experimental evidence supports the newly emerging paradigm of non-equilibrium control of protein behavior [189]. Specifically, the speed of synthesis of proteins in the ribosome greatly influences their properties, mRNA sequence evolution, and disease. Consequently, studying the effect of translation kinetics on protein misfolding, aggregation [190], and related diseases will be of great interest in the near future.

The relationship between viruses and amyloids is attracting more attention. Aβ aggregation, for example, was found to be promoted by the HSV-1 viral corona both *in vitro* and *in vivo* [191]. The SARS-CoV-2 nucleocapsid protein (N protein) accelerates αS fibrillation through electrostatic interactions, inducing cell death [192]. It has been shown that SARS-CoV-2 proteins can also form aggregates in isolation [193, 194], and similar results have been obtained for other virus species, such as the Hendra and Nipah viruses [195]. Based on the observation that human amyloids can interact with viruses, interfering with their replication, protein aggregation has been proposed as a strategy to discover new antiviral agents [196]. Thus, identification of the underlying factors that control virus–amyloid interactions is an important research direction in life sciences.

Finally, the main topic of this review resonates with the work of Professor Marek Cieplak on protein droplets, amyloid glass phase in systems of disordered homopeptides [197], and liquid–liquid phase separation [198]. The mechanical stability of fibrils discussed here is related to the protein and capsid stability studied by M. Cieplak et al. [199, 200] using simple Go models. His work in this direction had a high impact on the computational community because it showed that many important results could be obtained with simple models that did not require time-consuming simulations. In particular, his contributions influenced the development of some of the ideas presented in this article.

Author's Contribution

All authors contributed to discussion of topics and writing of the review. T.T.M. Thu and H.N.T. Phung contributed equally.

Acknowledgments

TTMT was funded by Vietnam National University, Ho Chi Minh City (VNU-HCM), under grant number C2020-18-19.

References

[1] C.M. Dobson, *Nature* **426**, 884 (2003).

[2] F. Chiti, C.M. Dobson, *Ann. Rev. Biochem.* **75**, 333 (2006).

[3] C.M. Dobson, *Cold Spring Harb. Perspect. Biol.* **9**, a023648 (2017).

[4] I.A. Buhimschi, U.A. Nayeri, G. Zhao, L.L. Shook, A. Pensalfini, E.F. Funai, I.M. Bernstein, C.G. Glabe, C.S. Buhimschi, *Sci. Trans. Med.* **6**, 245ra292 (2014).

[5] M. Kouza, A. Banerji, A. Kolinski, I.A. Buhimschi, A. Kloczkowski, *Phys. Chem. Chem. Phys.***19**, 2990 (2017).

[6] P.H. Nguyen, A. Ramamoorthy, B.R. Sahoo et al., *Chem. Rev.* **121**, 2545 (2021).

[7] D.J. Selkoe, *Nat. Cell Biol.* **6**, 1054 (2004).

[8] M.C. Owen, D. Gnutt, M. Gao, S.K.T.S. Wärmländer, J. Jarvet, A. Gräslund, R. Winter, S. Ebbinghaus, B. Strodel, *Chem. Soc. Rev.* **48**, 3946 (2019).

[9] R. Rajan, S. Ahmed, N. Sharma, N. Kumar, A. Debas, K. Matsumura, *Mater. Adv.* **2**, 1139 (2021).

[10] S. Devi, M. Chaturvedi, S. Fatima, S. Priya, *Toxicology* **465**, 153049 (2022).

[11] T. Sinnige, *Chem. Sci.* **13**, 7080 (2022).

[12] J. Sheng, N.K. Olrichs, B.M. Gadella, D.V. Kaloyanova, J.B. Helms, *Int. J. Mol. Sci.* **21**, 6530 (2020).

[13] P. Alam, K. Siddiqi, S.K. Chturvedi, R.H. Khan, *Int. J. Biol. Macromol.* **103**, 208 (2017).

[14] M.S. Li, N.T. Co, G. Reddy, C.K. Hu, J.E. Straub, D. Thirumalai, *Phys. Rev. Lett.* **105**, 218101 (2010).

[15] S.K. Maji, R.R.O. Loo, M. Inayathullah, S.M. Spring, S.S. Vollers, M.M. Condron, G. Bitan, J.A. Loo, D.B. Teplow, *J. Biol. Chem.* **284**, 23580 (2009).

[16] I. Qahwash, K.L. Weiland, Y. Lu, R.W. Sarver, R.F. Kletzien, R. Yan, *J. Biol. Chem.* **278**, 23187 (2003).

[17] L.M. Luheshi, G.G. Tartaglia, A.-C. Brorsson, A.P. Pawar, I.E. Watson, F. Chiti, M. Vendruscolo, D.A. Lomas, C.M. Dobson, D.C. Crowther, *PLoS Biol.* **5**, e290 (2007).

[18] S. Jawhar, O. Wirths, T.A. Bayer, *J. Biol. Chem.* **286**, 38825 (2011).

[19] G. Di Fede, M. Catania, M. Morbin et al., *Science* **323**, 1473 (2009).

[20] M. Messa, L. Colombo, E. Del Favero, L. Cantů, T. Stoilova, A. Cagnotto, A. Rossi, M. Morbin, G. Di Fede, F. Tagliavini, *J. Biol. Chem.* **289**, 24143 (2014).

[21] T. Jonsson, J.K. Atwal, S. Steinberg, J. Snaedal, P.V. Jonsson, S. Bjornsson, H. Stefansson, P. Sulem, D. Gudbjartsson, J. Maloney, *Nature* **488**, 96 (2012).

[22] B. De Strooper, T. Voet, *Nature* **488**, 38 (2012).

[23] J. Janssen, J. Beck, T. Campbell, A. Dickinson, N. Fox, R. Harvey, H. Houlden, M. Rossor, J. Collinge, *Neurology* **60**, 235 (2003).

[24] Y. Hori, T. Hashimoto, Y. Wakutani, K. Urakami, K. Nakashima, M.M. Condron, S. Tsubuki, T.C. Saido, D.B. Teplow, T. Iwatsubo, *J. Biol. Chem.* **282**, 4916 (2007).

[25] Y. Wakutani, K. Watanabe, Y. Adachi, K. Wada-Isoe, K. Urakami, H. Ninomiya, T. Saido, T. Hashimoto, T. Iwatsubo, K. Nakashima, *J. Neurol. Neurosurg. Psychiatry* **75**, 1039 (2004).

[26] K. Ono, M.M. Condron, D.B. Teplow, *J. Biol. Chem.* **285**, 23186 (2010).

[27] L. Zhou, N. Brouwers, I. Benilova, A. Vandersteen, M. Mercken, K. Van Laere, P. Van Damme, D. Demedts, F. Van Leuven, K. Sleegers, *EMBO Mol. Med.* **3**, 291 (2011).

[28] D. Kaden, A. Harmeier, C. Weise, L.M. Munter, V. Althoff, B.R. Rost, P.W. Hildebrand, D. Schmitz, M. Schaefer, R. Lurz, *EMBO Mol. Med.* **4**, 647 (2012).

[29] S. Sinha, D.H. Lopes, G. Bitan, *ACS Chem. Neurosci.* **3**, 473 (2012).

[30] L. Hendriks, C.M. Van Duijn, P. Cras, M. Cruts, W. Van Hul, F. Van Harskamp, A. Warren, M.G. McInnis, S.E. Antonarakis, J.-J. Martin, *Nat. Genet.* **1**, 218 (1992).

[31] K. Murakami, K. Irie, A. Morimoto, H. Ohigashi, M. Shindo, M. Nagao, T. Shimizu, T. Shirasawa, *J. Biol. Chem.* **278**, 46179 (2003).

[32] C. Nilsberth, A. Westlind-Danielsson, C.B. Eckman et al., *Nat. Neurosci.* **4**, 887 (2001).

[33] C.J. Lo, C.C. Wang, H.B. Huang, C.F. Chang, M.S. Shiao, Y.C. Chen, T.H. Lin, *Amyloid* **22**, 8 (2015).

[34] C.T. Liang, H.B. Huang, C.C. Wang, Y.R. Chen, C.F. Chang, M.S. Shiao, Y.C. Chen, T.H. Lin, *PLoS One* **11**, e0154327 (2016).

[35] O.Y. Ovchinnikova, V.H. Finder, I. Vodopivec, R.M. Nitsch, R. Glockshuber, *J. Mol. Biol.* **408**, 780 (2011).

[36] W.M. Berhanu, E.J. Alred, U.H. Hansmann, *J. Phys. Chem. B* **119**, 13063 (2015).

[37] W. Qiang, W.M. Yau, Y. Luo, M.P. Mattson, R. Tycko, *Proc. Natl. Acad. Sci. USA* **109**, 4443 (2012).

[38] V. Fonte, V. Dostal, C.M. Roberts, P. Gonzales, P. Lacor, J. Magrane, N. Dingwell, E.Y. Fan, M.A. Silverman, G.H. Stein, *Mol. Neurodegener.* **6**, 61 (2011).

[39] L.W. Hung, G.D. Ciccotosto, E. Giannakis, D.J. Tew, K. Perez, C.L. Masters, R. Cappai, J.D. Wade, K.J. Barnham, *Neurosci. J.* **28**, 11950 (2008).

[40] M. Decock, S. Stanga, J.-N. Octave, I. Dewachter, S.O. Smith, S.N. Constantinescu, P. Kienlen-Campard, *Front. Aging Neurosci.* **8**, 107 (2016).

[41] A. Harmeier, C. Wozny, B.R. Rost, L.-M. Munter, H. Hua, O. Georgiev, M. Beyermann, P.W. Hildebrand, C. Weise, W. Schaffner, *Neurosci. J.* **29**, 7582 (2009).

[42] A.E. Estrada-Rodríguez, D. Valdez-Pérez, J. Ruiz-García, A. Treviño-Garza, A.M. Gómez-Martínez, H.G. Martínez-Rodríguez, A.M. Rivas-Estilla, R. Vidal-tamayo, V. Zomosa-Signoret, *Int. J. Pept. Res. Ther.* **25**, 493 (2019).

[43] T.T.M. Thu, S.-H. Huang, L.A. Tu, S.-T. Fang, M.S. Li, Y.-C. Chen, *Neurochem. Int.* **129**, 104512 (2019).

[44] R. Roychaudhuri, M. Yang, A. Deshpande, G.M. Cole, S. Frautschy, A. Lomakin, G.B. Benedek, D.B. Teplow, *J. Mol. Biol.* **425**, 292 (2013).

[45] W. Kim, M.H. Hecht, *J. Biol. Chem.* **280**, 35069 (2005).

[46] E. Levy, M.D. Carman, I.J. Fernandez-Madrid, M.D. Power, I. Lieberburg, S.G. van Duinen, G.T. Bots, W. Luyendijk, B. Frangione, *Science* **248**, 1124 (1990).

[47] K. Kamino, H.T. Orr, H. Payami et al., *Am. J. Hum. Genet.* **51**, 998 (1992).

[48] T.J. Grabowski, H.S. Cho, J.P. Vonsattel, G.W. Rebeck, S.M. Greenberg, *Ann. Neurol.* **49**, 697 (2001).

[49] T. Tomiyama, T. Nagata, H. Shimada et al., *Ann. Neurol.* **63**, 377 (2008).

[50] F. Massi, J.E. Straub, *Biophys. J.* **81**, 697 (2001).

[51] S. Côté, P. Derreumaux, N. Mousseau, *J. Chem. Theory Comput.* **7**, 2584 (2011).

[52] Y.S. Lin, V.S. Pande, *Biophys. J.* **103**, L47 (2012).

[53] A. Huet, P. Derreumaux, *Biophys. J.* **91**, 3829 (2006).

[54] O. Coskuner, O. Wise-Scira, G. Perry, T. Kitahara, *ACS Chem. Neurosci.* **4**, 310 (2013).

[55] S. Mitternacht, I. Staneva, T. Hard, A. Irback, *Proteins* **78**, 2600 (2010).

[56] M.G. Krone, A. Baumketner, S.L. Bernstein, T. Wyttenbach, N.D. Lazo, D.B. Teplow, M.T. Bowers, J.E. Shea, *J. Mol. Biol.* **381**, 221 (2008).

[57] A. Baumketner, M.G. Krone, J.-E. Shea, *Proc. Natl. Acad. Sci. USA* **105**, 6027 (2008).

[58] V. Betts, M.A. Leissring, G. Dolios, R. Wang, D.J. Selkoe, D.M. Walsh, *Neurobiol. Dis.* **31**, 442 (2008).

[59] H.A. Lashuel, D.M. Hartley, B.M. Petre, J.S. Wall, M.N. Simon, T. Walz, P.T. Lansbury Jr.,*J. Mol. Biol.* **332**, 795 (2003).

[60] L. Miravalle, T. Tokuda, R. Chiarle, G. Giaccone, O. Bugiani, F. Tagliavini, B. Frangione, J. Ghiso, *J. Biol. Chem.* **275**, 27110 (2000).

[61] J. Davis, F. Xu, R. Deane, G. Romanov, M.L. Previti, K. Zeigler, B.V. Zlokovic, W.E. Van Nostrand, *J. Biol. Chem.* **279**, 20296 (2004).

[62] M.R. Elkins, T. Wang, M. Nick, H. Jo, T. Lemmin, S.B. Prusiner, W.F. DeGrado, J. Stöhr, M. Hong, *J. Am. Chem. Soc.* **138**, 9840 (2016).

[63] A.T. Petkova, W.-M. Yau, R. Tycko, *Biochem* **45**, 498 (2006).

[64] A.K. Paravastu, R.D. Leapman, W.-M. Yau, R. Tycko, *Proc. Natl. Acad. Sci. USA* **105**, 18349 (2008).

[65] T. Lührs, C. Ritter, M. Adrian, D. Riek-Loher, B. Bohrmann, H. Döbeli, D. Schubert, R. Riek, *Proc. Natl. Acad. Sci. USA* **102**, 17342 (2005).

[66] T. Takeda, D.K. Klimov, *J. Phys. Chem. B* **113**, 6692 (2009).

[67] T. Sato, P. Kienlen-Campard, M. Ahmed, W. Liu, H. Li, J.I. Elliott, S. Aimoto, S.N. Constantinescu, J.-N. Octave, S.O. Smith, *Biochem* **45**, 5503 (2006).

[68] A.T. Petkova, Y. Ishii, J.J. Balbach, O.N. Antzutkin, R.D. Leapman, F. Delaglio, R. Tycko, *Proc. Natl. Acad. Sci. USA* **99**, 16742 (2002).

[69] F. Massi, D. Klimov, D. Thirumalai, J.E. Straub, *Protein Sci.* **11**, 1639 (2002).

[70] S. Verma, A. Singh, A. Mishra, *Biochim. Biophys. Acta* **1834**, 24 (2013).

[71] Z. Lv, R. Roychaudhuri, M.M. Condron, D.B. Teplow, Y.L. Lyubchenko, *Sci. Rep.* **3**, 2880 (2013).

[72] H.A. Scheidt, I. Morgado, S. Rothemund, D. Huster, *J. Biol. Chem.* **287**, 2017 (2012).

[73] B. Sarkar, V.S. Mithu, B. Chandra, A. Mandal, M. Chandrakesan, D. Bhowmik, P.K. Madhu, S. Maiti, *Angew. Chem. Int. Ed. Engl.* **53**, 6888 (2014).

[74] B. Ma, R. Nussinov, *J. Biol. Chem.* **286**, 34244 (2011).

[75] W. Zago, M. Buttini, T.A. Comery, C. Nishioka, S.J. Gardai, P. Seubert, D. Games, F. Bard, D. Schenk, G.G. Kinney, *Neurosci. J.* **32**, 2696 (2012).

[76] H. Li, Z. Du, D.H. Lopes, E.A. Fradinger, C. Wang, G. Bitan, *J. Med. Chem.* **54**, 8451 (2011).

[77] W.-T. Chen, C.-J. Hong, Y.-T. Lin, W.-H. Chang, H.-T. Huang, J.-Y. Liao, Y.-J. Chang, Y.-F. Hsieh, C.-Y. Cheng, H.-C. Liu, *PLos One* **7**, e35807 (2012).

[78] N.D. Lazo, M.A. Grant, M.C. Condron, A.C. Rigby, D.B. Teplow, *Protein Sci.* **14**, 1581 (2005).

[79] J.T. Jarrett, E.P. Berger, P.T. Lansbury Jr., *Biochem* **32**, 4693 (1993).

[80] Y.-J. Chang, N.H. Linh, Y.H. Shih, H.-M. Yu, M.S. Li, Y.-R. Chen, *ACS Chem. Neurosci.* **7**, 1097 (2016).

[81] N.H. Linh, T.T. Minh Thu, L. Tu, C.-K. Hu, M.S. Li, *J. Phys. Chem. B* **121**, 4341 (2017).

[82] F. Chiti, M. Stefani, N. Taddei, G. Ramponi, C.M. Dobson, *Nature* **424**, 805 (2003).

[83] T.T.M. Thu, N.T. Co, L.A. Tu, M.S. Li, *J. Chem. Phys.* **150**, 225101 (2019).

[84] M.S. Li, D.K. Klimov, J.E. Straub, D. Thirumalai, *J. Chem. Phys.* **129**, 175101 (2008).

[85] H.B. Nam, M. Kouza, H. Zung, M.S. Li, *J. Chem. Phys.* **132**, 165104 (2010).

[86] P.I. Zhuravlev, G. Reddy, J.E. Straub, D. Thirumalai, *J. Mol. Biol.* **426**, 2653 (2014).

[87] U. Baul, D. Chakraborty, M.L. Mugnai, J.E. Straub, D. Thirumalai, *J. Phys. Chem. B* **123**, 3462 (2019).

[88] D. Chakraborty, J.E. Straub, D. Thirumalai, *Proc. Natl. Acad. Sci. USA* **117**, 19926 (2020).

[89] S.W. Snyder, U.S. Ladror, W.S. Wade, G.T. Wang, L.W. Barrett, E.D. Matayoshi, H.J. Huffaker, G.A. Krafft, T.F. Holzman, *Biophys. J.* **67**, 1216 (1994).

[90] R. Pellarin, P. Schuetz, E. Guarnera, A. Caflisch, *J. Am. Chem. Soc.* **132**, 14960 (2010).

[91] M. Kouza, N.T. Co, M.S. Li, S. Kmiecik, A. Kolinski, A. Kloczkowski, I.A. Buhimschi, *J. Chem. Phys.* **148**, 215106 (2018).

[92] T.T.M. Thu, M.S. Li, *J. Chem. Phys.* **157**, 055101 (2022).

[93] F. Franks, R.H.M. Hatley, H.L. Friedman, *Biophys. Chem.* **31**, 307 (1988).

[94] W. Wang, C.J. Roberts, *AAPS J.* **15**, 840 (2013).

[95] J.J. Kayser, P. Arnold, A. Steffen-Heins, K. Schwarz, J.K. Keppler, *J. Food Eng.* **270**, 109764 (2020).

[96] R. Sabaté, M. Gallardo, J. Estelrich, *Int. J. Biol. Macromol.* **35**, 9 (2005).

[97] Y. Kusumoto, A. Lomakin, D.B. Teplow, G.B. Benedek, *Proc. Natl. Acad. Sci. USA* **95**, 12277 (1998).

[98] P.L. Privalov, S.J. Gill, in: *Advaces in Protein Chemistry*, Vol. 39, Eds. C.B. Anfinsen, J.T. Edsall, F.M. Richards, D.S. Eisenberg, Academic Press, 1988, p. 191.

[99] R.L. Remmele, S.D. Bhat, D.H. Phan, W.R. Gombotz, *Biochem* **38**, 5241 (1999).

[100] P.L. Privalov, *Crit. Rev. Biochem. Mol. Biol.* **25**, 281 (1990).

[101] B. Luan, B. Shan, C. Baiz, A. Tokmakoff, D.P. Raleigh, *Biochem* **52**, 2402 (2013).

[102] R. Esfandiary, A. Parupudi, J. Casas-Finet, D. Gadre, H. Sathish, *J. Pharm. Sci.* **104**, 577 (2015).

[103] A. Oliva, J.B. Fariña, M. Llabrés, *J. Chromatogr.* **1022**, 206 (2016).

[104] M. Smith, J. Sharp, C. Roberts, *Biophys. J.* **93**, 2143 (2007).

[105] M. Manno, E.F. Craparo, A. Podestŕ, D. Bulone, R. Carrotta, V. Martorana, G. Tiana, P.L. San Biagio, *J. Mol. Biol.* **366**, 258 (2007).

[106] W. Wang, C.J. Roberts, *Int. J. Pharm.* **550**, 251 (2018).

[107] N. Chakroun, D. Hilton, S.S. Ahmad, G.W. Platt, P.A. Dalby, *Mol. Pharm.* **13**, 307 (2016).

[108] Z. Sahin, Y.K. Demir, V. Kayser, *Eur. J. Pharm. Sci.* **86**, 115 (2016).

[109] A. Saluja, V. Sadineni, A. Mungikar, V. Nashine, A. Kroetsch, C. Dahlheim, V.M. Rao, *Pharm. Res.* **31**, 1575 (2014).

[110] N.T. Co, C.K. Hu, M.S. Li, *J. Chem. Phys.* **138**, 185101 (2013).

[111] K. Kar, M. Jayaraman, B. Sahoo, R. Kodali, R. Wetzel, *Nat. Struct. Mol. Biol.* **18**, 328 (2011).

[112] R. Sabaté, J. Estelrich, *Biopolymers* **67**, 113 (2002).

[113] G. Meisl, X. Yang, B. Frohm, T.P.J. Knowles, S. Linse, *Sci. Rep.* **6**, 18728 (2016).

[114] L.O. Tjernberg, A. Pramanik, S. Björling, P. Thyberg, J. Thyberg, C. Nordstedt, K.D. Berndt, L. Terenius, R. Rigler, *Chem. Biol.* **6**, 53 (1999).

[115] R. Sabaté, J. Estelrich, *J. Phys. Chem. B* **109**, 11027 (2005).

[116] M. Novo, S. Freire, W. Al-Soufi, *Sci. Rep.* **8**, 1783 (2018).

[117] T. Deva, N. Lorenzen, B.S. Vad, S.V. Petersen, I. Thřrgersen, J.J. Enghild, T. Kristensen, D.E. Otzen, *Biochim. Biophys. Acta* **1834**, 677 (2013).

[118] E.T. Powers, D.L. Powers, *Biophys. J.* **94**, 379 (2008).

[119] K.L. Zapadka, F.J. Becher, A.L. Gomes dos Santos, S.E. Jackson, *Interface Focus* **7**, 20170030 (2017).

[120] S. Saha, S. Deep, *Curr. Phys. Chem.* **4**, 114 (2014).

[121] S.I.A. Cohen, S. Linse, L.M. Luheshi, E. Hellstrand, D.A. White, L. Rajah, D.E. Otzen, M. Vendruscolo, C.M. Dobson, T.P.J. Knowles, *Proc. Natl. Acad. Sci. USA* **110**, 9758 (2013).

[122] Y.S. Kim, T.W. Randolph, M.B. Seefeldt, J.F. Carpenter, *Method Enzymol.* **413**, 237 (2006).

[123] G.A.P. de Oliveira, M.A. Marques, M.M. Pedrote, J.L. Silva, *High Press Res.* **39**, 193 (2019).

[124] J. Roche, J.A. Caro, D.R. Norberto, P. Barthe, C. Roumestand, J.L. Schlessman, A.E. Garcia, B.E. García-Moreno, C.A. Royer, *Proc. Natl. Acad. Sci. USA* **109**, 6945 (2012).

[125] G.A.P. de Oliveira, J.L. Silva, *Proc. Natl. Acad. Sci. USA* **112**, E2775 (2015).

[126] S.D. Hamann, *Rev. Phys. Chem. Jpn.* **50**, 147 (1980).

[127] R.J. St. John, J.F. Carpenter, T.W. Randolph, *Proc. Natl. Acad. Sci. USA* **96**, 13029 (1999).

[128] D. Foguel, J.L. Silva, *Biochem* **43**, 11361 (2004).

[129] T.W. Randolph, M. Seefeldt, J.F. Carpenter, *Biochim. Biophys. Acta.* **1595**, 224 (2002).

[130] S. Kobayashi, Y. Tanaka, M. Kiyono, M. Chino, T. Chikuma, K. Hoshi, H. Ikeshima, *J. Mol. Struct.* **1094**, 109 (2015).

[131] M. López de la Paz, K. Goldie, J. Zurdo, E. Lacroix, C.M. Dobson, A. Hoenger, L. Serrano, *Proc. Natl. Acad. Sci. USA* **99**, 16052 (2002).

[132] W. Hoyer, T. Antony, D. Cherny, G. Heim, T.M. Jovin, V. Subramaniam, *J. Mol. Biol.* **322**, 383 (2002).

[133] B. Raman, E. Chatani, M. Kihara, T. Ban, M. Sakai, K. Hasegawa, H. Naiki, C.M. Rao, Y. Goto, *Biochem* **44**, 1288 (2005).

[134] C.C. vandenAkker, M.F.M. Engel, K.P. Velikov, M. Bonn, G.H. Koenderink, *J. Am. Chem. Soc.* **133**, 18030 (2011).

[135] P.J. Marek, V. Patsalo, D.F. Green, D.P. Raleigh, *Biochem* **51**, 8478 (2012).

[136] A. Abelein, J. Jarvet, A. Barth, A. Gräslund, J. Danielsson, *J. Am. Chem. Soc* **138**, 6893 (2016).

[137] B. Priyanka, S.K.M. Venkata, *Curr. Chem. Biol.* **14**, 216 (2020).

[138] A.M. Tsai, J.H. van Zanten, M.J. Betenbaugh, *Biotechnol. Bioeng.* **59**, 281 (1998).

[139] T. Arakawa, S.N. Timasheff, *Biochem* **23**, 5912 (1984).

[140] R.A. Curtis, J. Ulrich, A. Montaser, J.M. Prausnitz, H.W. Blanch, *Biotechnol. Bioeng.* **79**, 367 (2002).

[141] P. Gupta, S. Deep, *RSC Adv.* **5**, 95717 (2015).

[142] M. Adachi, M. Noji, M. So, K. Sasahara, J. Kardos, H. Naiki, Y. Goto, *J. Biol. Chem.* **293**, 14775 (2018).

[143] S. Saha, S. Deep, *J. Phys. Chem. B* **118**, 9155 (2014).

[144] E.Y. Chi, S. Krishnan, B.S. Kendrick, B.S. Chang, J.F. Carpenter, T.W. Randolph, *Protein Sci.* **12**, 903 (2003).

[145] A. Fatouros, T. Österberg, M. Mikaelsson, *Int. J. Pharm.* **155**, 121 (1997).

[146] S.B. Zimmerman, S.O. Trach, *J. Mol. Biol.* **222**, 599 (1991).

[147] M. Feig, I. Yu, P.-h. Wang, G. Nawrocki, Y. Sugita, *J. Phys. Chem. B* **121**, 8009 (2017).

[148] R.J. Ellis, *Trends Biochem. Sci.* **26**, 597 (2001).

[149] C. Cabaleiro-Lago, F. Quinlan-Pluck, I. Lynch, K.A. Dawson, S. Linse, *ACS Chem. Neurosci.* **1**, 279 (2010).

[150] S. Mittal, L.R. Singh, *J. Biochem.* **156**, 273 (2014).

[151] R.J. Ellis, A.P. Minton, *Biol. Chem.* **387**, 485 (2006).

[152] D.A. White, A.K. Buell, T.P.J. Knowles, M.E. Welland, C.M. Dobson, *J. Am. Chem. Soc.* **132**, 5170 (2010).

[153] A. Magno, A. Caflisch, R. Pellarin, *J. Phys. Chem. Lett.* **1**, 3027 (2010).

[154] E.P. O'Brien, J.E. Straub, B.R. Brooks, D. Thirumalai, *J. Phys. Chem. Lett.* **2**, 1171 (2011).

[155] Z. Zhou, J.-B. Fan, H.-L. Zhu, F. Shewmaker, X. Yan, X. Chen, J. Chen, G.-F. Xiao, L. Guo, Y. Liang, *J. Biol. Chem.* **284**, 30148 (2009).

[156] N.T. Co, C.-K. Hu, M.S. Li, *J. Chem. Phys.* **138**, 185101 (2013).

[157] G. Gao, M. Zhang, D. Gong, R. Chen, X. Hu, T. Sun, *Nanoscale* **9**, 4107 (2017).

[158] D.C. Latshaw 2nd, C.K. Hall, *Biophys. J.* **109**, 124 (2015).

[159] F. Musiani, A. Giorgetti, in: *International Review of Cell and Molecular Biology, Vol. 329, Early Stage Protein Misfolding and Amyloid Aggregation*, Ed. M. Sandal, Academic Press, 2017 p. 49.

[160] B. Yang, D.J. Adams, M. Marlow, M. Zelzer, *Langmuir* **34**, 15109 (2018).

[161] A. Keller, G. Grundmeier, *Appl. Surf. Sci.* **506**, 144991 (2020).

[162] R. Vacha, S. Linse, M. Lund, *J. Am. Chem. Soc.* **136**, 11776 (2014).

[163] D.J. Lindberg, E. Wesén, J. Björkeroth, S. Rocha, E.K. Esbjörner, *Biochim. Biophys. Acta. Biomembr.* **1859**, 1921 (2017).

[164] A. Morriss-Andrews, F.L.H. Brown, J.-E. Shea, *J. Phys. Chem. B* **118**, 8420 (2014).

[165] A. Rawat, R. Langen, J. Varkey, *Biochim. Biophys. Acta Biomembr.* **1860**, 1863 (2018).

[166] M. Rabe, A. Soragni, N.P. Reynolds, D. Verdes, E. Liverani, R. Riek, S. Seeger, *ACS Chem. Neurosci.* **4**, 408 (2013).

[167] Y.-C. Lin, C. Li, Z. Fakhraai, *Langmuir* **34**, 4665 (2018).

[168] M. Mahmoudi, O. Akhavan, M. Ghavami, F. Rezaee, S.M.A. Ghiasi, *Nanoscale* **4**, 7322 (2012).

[169] S.-g. Kang, T. Huynh, Z. Xia, Y. Zhang, H. Fang, G. Wei, R. Zhou, *J. Am. Chem. Soc.* **135**, 3150 (2013).

[170] F. Zhang, H.-N. Du, Z.-X. Zhang et al., *Angew. Chem. Int. Ed.* **45**, 3611 (2006).

[171] T. Ban, K. Morigaki, H. Yagi, T. Kawasaki, A. Kobayashi, S. Yuba, H. Naiki, Y. Goto, *J. Biol. Chem.* **281**, 33677 (2006).

[172] R. Huang, R. Su, W. Qi, J. Zhao, Z. He, *Nat. Nanotechnol.* **22**, 245609 (2011).

[173] K. Shezad, K. Zhang, M. Hussain, H. Dong, C. He, X. Gong, X. Xie, J. Zhu, L. Shen, *Langmuir* **32**, 8238 (2016).

[174] N.T. Co, M.S. Li, *Biomolecules* **11**, 596 (2021).

[175] M.Y. Aksenov, M.V. Aksenova, D.A. Butterfield, J.W. Geddes, W.R. Markesbery, *Neuroscience* **103**, 373 (2001).

[176] X. Wang, W. Wang, L. Li, G. Perry, H.G. Lee, X. Zhu, *Biochim. Biophys. Acta.* **1842**, 1240 (2014).

[177] M. Vaezzadeh, M. Sabbaghian, P. Yaghmaei, A. Ebrahim-Habibi, *Protein Pept. Lett.* **24**, 955 (2017).

[178] S. Nusrat, R.H. Khan, *Prep. Biochem. Biotechnol.* **48**, 43 (2018).

[179] A.A. Thorat, B. Munjal, T.W. Geders, R. Suryanarayanan, *J. Control Release* **323**, 591 (2020).

[180] K. Jain, N. Salamat-Miller, K. Taylor, *Sci. Rep.* **11**, 11332 (2021).

[181] C. Wallin, M. Friedemann, S.B. Sholts, A. Noormägi, T. Svantesson, J. Jarvet, P.M. Roos, P. Palumaa, A. Gräslund, S. Wärmländer, *Biomolecules* **10**, 44 (2019).

[182] B. Alies, C. Hureau, P. Faller, *Metallomics* **5**, 183 (2013).

[183] Z. Zhao, K. Engholm-Keller, M.M. Poojary, S.G. Boelt, A. Rogowska-Wrzesinska, L.H. Skibsted, M.J. Davies, M.N. Lund, *J. Agric. Food Chem.* **68**, 6701 (2020).

[184] A. Allmendinger, Y. Ni, A. Bernhard, H. Nalenz, *PDA J. Pharm. Sci. Technol.* **76**, 52 (2022).

[185] H. Wu, T.W. Randolph, *J. Pharm. Sci.* **109**, 1473 (2020).

[186] K.B. Vargo, P. Stahl, B. Hwang, E. Hwang, D. Giordano, P. Randolph, C. Celentano, R. Hepler, K. Amin, *Mol. Pharm.* **18**, 148 (2021).

[187] J. Hong, L.M. Gierasch, Z. Liu, *Biophys. J.* **109**, 144 (2015).

[188] F. Tao, Q. Han, P. Yang, *Chem. Commun.* **59**, 14093 (2023).

[189] A.K. Sharma, E.P. O'Brien, *Curr. Opin. Struct. Biol.* **49**, 94 (2018).

[190] P.D. Lan, D.A. Nissley, I. Sitarik, Q.V. Van, Y. Jiang, M.S. Li, E.P. O'Brien, *J. Mol. Biol.* **436**, 168487 (2024).

[191] K. Ezzat, M. Pernemalm, S. PÍlsson et al., *Nat. Commun.* **10**, 2331 (2019).

[192] S.A. Semerdzhiev, M.A.A. Fakhree, I. Segers-Nolten, C. Blum, M.M.A.E. Claessens, *ACS Chem. Neurosci.* **13**, 143 (2022).

[193] T. Bhardwaj, K. Gadhave, S.K. Kapuganti et al., *Nat. Commun.* **14**, 945 (2023).

[194] S. Nyström, P. Hammarström, *J. Am. Chem. Soc.* **144**, 8945 (2022).

[195] J.F. Nilsson, H. Baroudi, F. Gondelaud, G. Pesce, C. Bignon, D. Ptchelkine, J. Chamieh, H. Cottet, A.V. Kajava, S. Longhi, *Int. J. Mol. Sci.* **24**, 399 (2023).

[196] E. Michiels, K. Roose, R. Gallardo et al., *Nat. Commun.* **11**, 2832 (2020).

[197] Ł. Mioduszewski, M. Cieplak, *Phys. Chem. Chem. Phys.* **22**, 15592 (2020).

[198] D.Q.H. Pham, M. Chwastyk, M. Cieplak, *Front. Chem.* **10**, 1106599 (2023).

[199] M. Cieplak, T.X. Hoang, M.O. Robbins, *Proteins* **49**, 114 (2002).

[200] M. Cieplak, M.O. Robbins, *J. Chem. Phys.* **132**, 015101 (2010).

In memory of Professor Marek Cieplak

REVIEW

Recent Advances in Mapping Protein Self-Assembly and Aggregation for Common Proteinopathies

S. Bhattacharya and D. Thompson*

Department of Physics, Bernal Institute, University of Limerick, V94 T9PX, Ireland

Doi: 10.12693/APhysPolA.145.S37 *e-mail: damien.thompson@ul.ie

The accumulation of abnormal conformation by brain peptides and proteins followed by their aberrant self-assembly into insoluble aggregates is the hallmark of "proteinopathies", common across many neurodegenerative disorders. Experiments suggest that soluble low-molecular-weight oligomers formed in the early stages of assembly are neurotoxic, and hence, drug targets. However, the inherent polymorphic nature of these short-lived oligomers restricts their experimental characterisation in pathological protein self-assembly pathways. Here, we shed light on the latest contributions from atomic-level modelling techniques, such as computer-based molecular dynamics simulations in bulk solution and on surfaces, which are guiding experimental efforts to map early stages of protein self-assembly in common proteinopathies, including Alzheimer's and Parkinson's diseases, which could potentially aid in molecular-level understanding of disease pathologies. Predictive computational modelling of amyloid-β and tau protein assemblies in Alzheimer's disease and α-synuclein protein assemblies in Parkinson's disease highlights the potential for identification and characterisation of new therapeutic targets for currently incurable neurodegeneration.

topics: proteinopathies, self-assembly, computational modelling, molecular dynamics simulations

1. Introduction

Deposition of protein fibrillar aggregates is a characteristic shared by > 50 human diseases [1]. Pathological protein self-assembly with the formation of inclusion bodies, such as fibrils, is the hallmark of many neurodegenerative disorders (ND) [2], or broadly, "proteinopathies". NDs are a heterogeneous group of lethal brain disorders that may be characterised by symptomatic gradual decline of the structure and function of central and peripheral nervous systems [3]. They share a significant Global Burden of Disease (GBD) [4], with World Health Organization (WHO) projections that dementia will account for $>$ 1% of total deaths by 2030 [5]. NDs, including Alzheimer's disease (AD), Parkinson's disease (PD), Huntington's disease (HD) [6], prion diseases, amyotrophic lateral sclerosis (ALS), and other systemic amyloidosis diseases [3], exhibit distinct aetiologies but share common pathologies. These disorders could be characterised by amyloidosis or the production of amyloids, where abnormal protein conformations form through spontaneous misfolding and self-assembly starting from their intrinsically disordered native state (intrinsically disordered proteins, IDPs) [7]. AD and PD are the most common proteinopathies [8]. Currently, only five Food and Drug Administration (FDA)-approved drugs are available to treat cognitive symptoms of AD or slow its progression by removal of brain amyloid [9–11], and a handful of FDA-approved treatment options address the motor symptoms associated with PD [12]. Yet, to date, there exists no clinically effective disease-modifying strategy for AD and PD multifactorial diseases, creating a massive burden on the management of symptoms and patient care [13, 14]. To translate disease-modifying strategies into effective clinical targets, urgent re-evaluation of current therapeutic and molecular targets is required.

The development of effective treatment for AD and PD is hampered by an insufficient understanding of the events that trigger the self-assembly of the monomeric IDPs into higher-order assemblies and, eventually, fibrils [15]. Knowledge to date is summarised in Fig. 1 (see also [16]), showing the potential molecular processes from misfolded monomeric proteins to self-assembled aggregates. In addition, other mechanisms of amyloid toxicity are also proposed from a misfolded monomer [6, 17, 18], including investigations of elastic and thermodynamic properties of amyloid-β and α-synuclein fibrils from molecular simulations to understand experimental nanomechanical characterisation techniques [19]. Mechanical properties of fibrillar assemblies can also serve as a diagnostic fingerprint for potential applications or pathology [20–22], supported by co-development

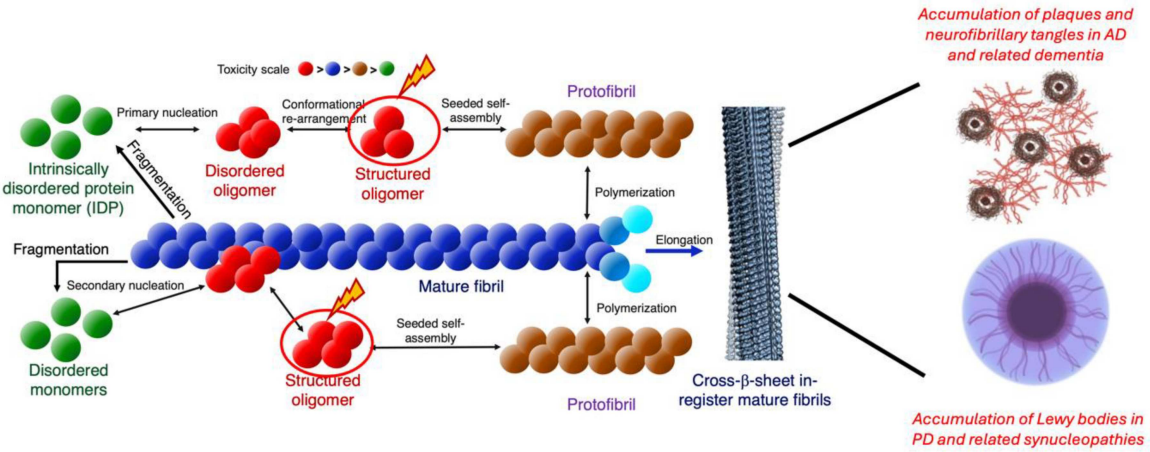


Fig. 1. Known molecular mechanisms underlying amyloid pathogenesis to self-assembled insoluble aggregates of mature fibrils further elongated to cross-β sheet fibrils and eventually plaques and neurofibrillary tangles in AD and related dementia, and Lewy bodies in PD and related synucleopathies. IDPs of Aβ, tau, and αS may undergo primary or predominantly secondary nucleation [16] to form oligomers, which are the critical nuclei for the growth of seeds into protofibrils and fibrils, which, through further polymerisation, lead to mature fibrils. Oligomers and mature fibrils may further undergo fragmentation or dissociate to form new seeds.

of reliable measurement techniques and predictive modelling [23–25]. Amyloid-β (Aβ) [26] and tau [27] form plaques and neurofibrillary tangles in AD and related dementia, and α-synuclein (αS) [28] form Lewy bodies in PD and other synucleopathies. The neurotoxicity in AD and PD is now generally attributed to low-molecular-weight aggregates or small oligomers rather than large amyloid fibrils and plaques [29, 30]. However, detection and isolation of these soluble oligomers has proven difficult due to their short lifetimes, low concentration, and structural heterogeneity [29] in solution, unless they are artificially engineered to be structurally stable [31].

The relationship between aggregation rates of amyloidogenic peptides and their morphological changes over time is important not only to comprehensively map the pathological protein self-assembly, but also to characterise the shapes and sizes of the small population of cytotoxic, short-lived, and structurally diverse misfolded oligomers to design new potential drug targets [32]. To detect and measure the metastable oligomers in physiological solution, several *in vitro* and *in vivo* techniques have been employed in the past, generating mainly only qualitative or semi-quantitative data [32], and therefore computational molecular modelling and computer simulations have played a major role in guiding experiments on protein self-assembly. The presence of biological and non-biological surfaces or surface–water interfaces is well known to mediate pathogenic protein self-assembly [33]. The interaction between biological lipid membranes and pathogenic peptides in the context of fibril formation has been extensively studied [34–36]. The interaction between non-biological surfaces and amyloid aggregation depends on the nature of the surfaces, which can play a crucial role in either catalysing or inhibiting the aggregation of amyloid proteins and is an interesting area that is only beginning to be explored [37, 38]. In this mini-review, we provide a perspective on the latest findings and recent advances made in mapping pathological protein self-assembly. We highlight the importance of computer-based molecular modelling and simulations supported by experimental investigations in bulk solution and on biological and non-biological surfaces to reveal molecular-level details of the assembly mechanisms, identifying potential early stages of self-assembly, the role of secondary structures, and routes to resist toxic aggregation with potential therapeutic intervening targets focussing on the proteins responsible for AD and PD pathologies, Aβ, tau, and αS.

2. Demystifying stages of pathogenic protein self-assembly from molecular simulations

As discussed above, it is experimentally very difficult or impossible to characterise thermally accessible states of Aβ, tau, and αS protein assemblies in AD and PD pathophysiology, so molecular modelling with appropriate benchmarking and experimental validation can help to identify the thermodynamic driving force behind specific protein morphologies formed in the self-assembly pathway, and also to estimate their kinetics (i.e., how fast these morphologies form) of formation and dissociation [39]. Computational modelling, in particular molecular dynamics (MD) simulations, can predict dynamic local and long-range interactions driving heterogeneous assemblies [40, 41]. Predictive models from MD may also provide mechanistic insights

into previously unknown self-assembly features at different stages of pathological protein aggregation, including those that could be validated, directly or indirectly, by experiments [42].

One of the earliest instances of very short MD simulations provided mechanistic insights into amyloidogenic misfolding that is involved in the multimerisation of PrP_{Sc} (pathogenic prion) [43]. The only information available from experiments was that the conformational conversion of PrP_C (cellular prion) to PrP_{Sc} occurred at low pH. Within ten nanoseconds (ns) of molecular dynamics, the simulations mapped conformational shifts in the N-terminal region that were mainly due to the breaking of charge-stabilised hydrogen bonded interactions at low pH, which was later confirmed by amide-proton exchange nuclear magnetic resonance (NMR) experiments [44]. The computational predictive power of molecular simulations is ever-increasing due to improvements in hardware and increased accessibility of high-performance computing platforms coupled with software developments in particular advanced sampling methods [45, 46] and the latest improved force fields [47, 48] and water models [7] for MD, allowing to reach extended timescales to map physically realistic, and biologically relevant, protein aggregation pathways.

3. Atomic models of pathological protein self-assembly

3.1. Modelling self-assembly of Aβ protein in AD

The mechanisms of initial misfolding and aggregation of Aβ in AD have been studied extensively by MD simulations supported by single-molecule experimental techniques [49–51]. As Aβ dimers were identified as the smallest toxic oligomers that could potentially assemble into neurotoxic protofibrils [52], recent long, multi-microsecond MD studies have investigated their detailed assembly to reveal differences in dimer morphologies from the U-shaped and S-shaped fibrillar morphologies indicating significant conformational re-arrangements in aggregation from small toxic oligomers to fibrils [53, 54]. These atomic scale insights are not readily available from high-resolution experimental techniques due to the broad distribution of short-lived dimeric shapes and their rapid self-assembly into higher-order structures. Similarly, recent transition path theory (TPT) network and Markov state models (MSM) based on MD-generated ensembles of dimers and higher-order oligomers have shown the predominance of Aβ oligomer shapes in directing self-assembly propensities; the compact metastable dimer matching to the oligomer distribution has been observed experimentally and may be more toxic than the extended dimers that self-assemble into larger fibrils [54, 55].

A comprehensive study involving microseconds scale MD simulations of ten different protein force fields and cross-correlation network analysis benchmarked by experimental NMR revealed the very nascent aggregation-favouring and aggregation-impeding propensities of fully folded, unfolded, and partially folded helical states of both $A\beta_{1-42}$ and αS_{1-140} (see Fig. 2a). The fully folded helical states optimise the direct intra-protein hydrophobic contacts between the termini and the central hydrophobic domain (CHD) of both proteins which resist aggregation (see Fig. 2b) [56], while the partially folded helical states may promote initial self-assembly by long-range allosteric coupling between the terminal residues and the CHD (Fig. 2c) [57].

In another recent work [58], extensive MD simulations predicted the molecular signatures of the difference in aggregation profiles visualised by atomic force microscopy (AFM) experiments on preformed oligomers in LS-shaped fibril fold (profibrillar 12-mers) between peptides $A\beta_{40}$ and $A\beta_{42}$ that may account for the higher pathogenicity of $A\beta_{42}$ in AD (Fig. 3). Modelling the orientation of both peptide assemblies on a single layer of graphene as the interface between graphene and water is an ideal platform to study peptide assemblies with AFM. From oligomer–graphene binding energies, the models predicted that amyloid beta undergoes chain elongation along the graphene sheet (orientation III, Fig. 3a, b). Predictions of oligomer model height profiles on top of graphene and hydrogen bond (H-bond) occupancies in three dimers of hexamer (dimer at one end of the oligomer, denoted as E1, dimer in the centre of oligomer, denoted as C, and dimer at the other end of oligomer, denoted as E2) forming two layers of the 12-mer and validated from AFM maps (Fig. 3c) revealed unidirectional growth profile for $A\beta_{40}$ and bidirectional growth for $A\beta_{42}$ at the graphene–water interface (Fig. 3d) that may explain the highly aggregation-prone nature and toxicity of $A\beta_{42}$.

3.2. Modelling self-assembly of tau protein in AD

The microtubule-associated protein tau (MAPT or simply τ) [59] is implicated in the pathogenesis of AD. Tau is an IDP responsible for the polymerisation and stabilisation of microtubules and has two major domains: (i) the projection domain, which includes the N-terminal and points away from the microtubule surface, and (ii) the C-terminal domain, which binds to microtubules [60]. The polymorphic nature of hTau40 [61] has precluded attempts to resolve its full atomic structure experimentally, and recent cryo-EM structures of tau filaments capture the structural polymorphism at fibrillar level with paired helical filaments (PHFs) [62–65], straight filaments (SF) [62, 63], narrow Pick's filament (NPF) [66] in frontotemporal

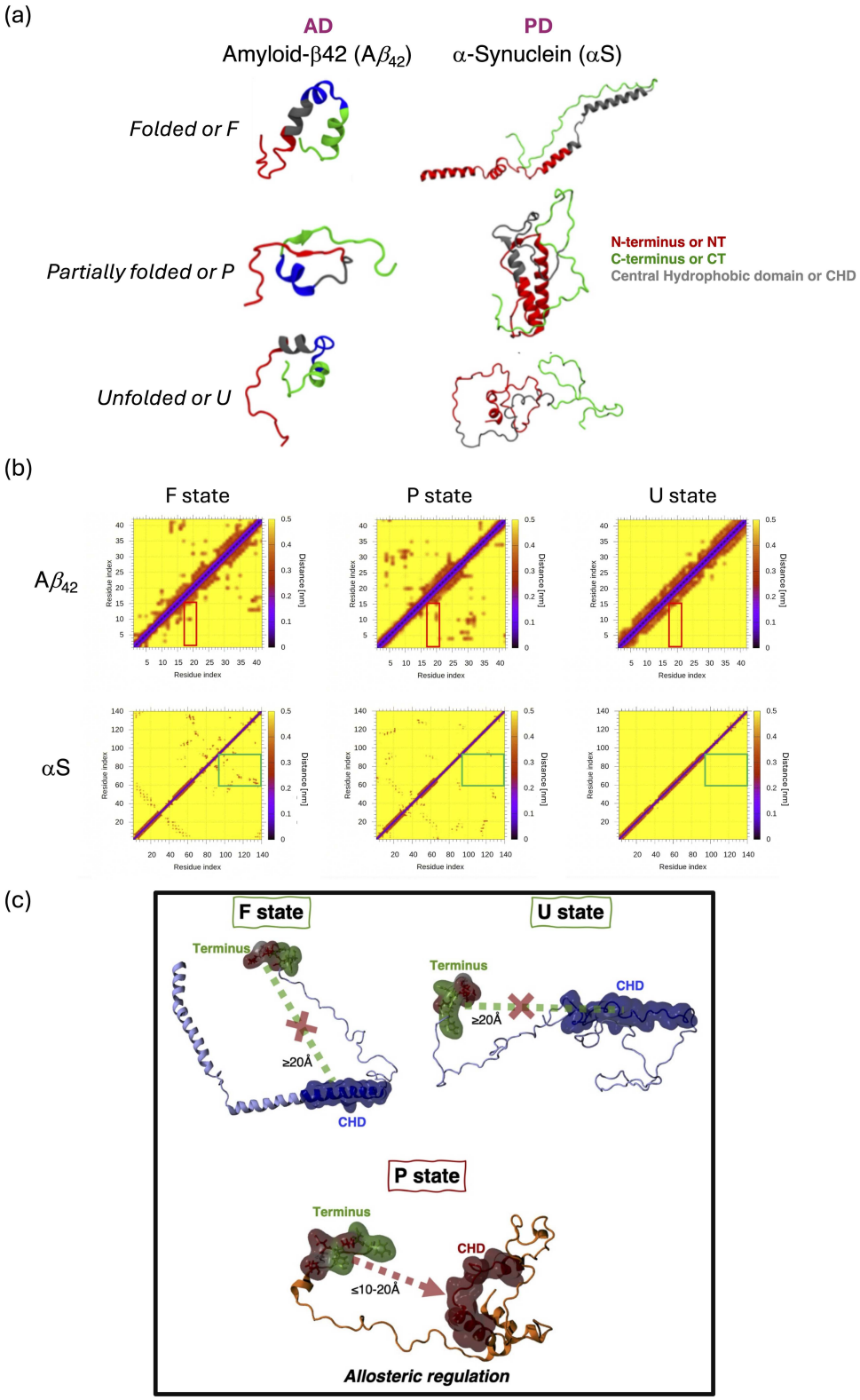
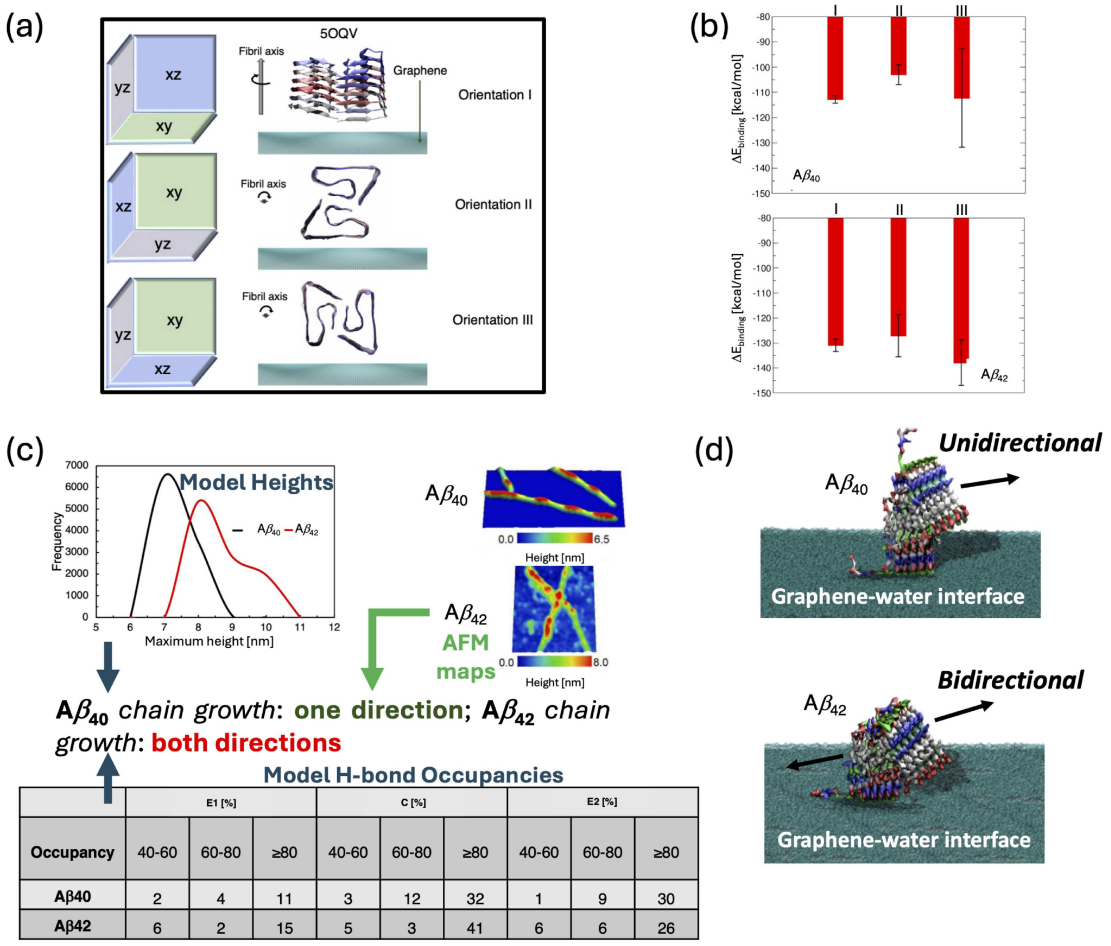


Fig. 2. (a) Representative folded (F), partially folded (P), and unfolded (U) helical state structures of Aβ_{42} in AD and αS in PD sampled from $\sim$ 36 microseconds MD simulations of helical structures employing ten alternative force fields and water models combinations. (b) Residue–residue contact maps showing the helical F state stabilising the interactions between the N-terminus and CHD of Aβ_{42}, and the C-terminus and CHD of αS that inhibit aggregation and missing in the P and the U states. (c) Long-range ($\leq$ 20 Å) allosteric regulation of the CHD by the termini of helical peptides Aβ_{42} and αS in their P state, which makes them aggregation-prone.



	E1 [%]			C [%]			E2 [%]		
Occupancy	40-60	60-80	≥80	40-60	60-80	≥80	40-60	60-80	≥80
Aβ40	2	4	11	3	12	32	1	9	30
Aβ42	6	2	15	5	3	41	6	6	26

Fig. 3. (a) Different starting orientations of Aβ preformed oligomers in fibril fold on top of a graphene sheet and at the graphene–water interface for running MD simulations. (b) Oligomer–graphene binding energies, showing that orientation III with a fibril axis along the graphene sheet is most favourable. (c) Model prediction of maximum oligomer heights on top of graphene and comparison of H-bond occupancies [%] between two ends of oligomers and the centre of oligomer supports visualisation of AFM maps to reveal Aβ_{40} chain elongation in one direction and Aβ_{42} chain elongation in both directions. (d) Depiction of unidirectional and bidirectional chain growth of Aβ_{40} and Aβ_{42}, respectively, at the graphene–water interface.

dementia, and very recently non-helical filaments [67] in AD, leaving out details of the morphological features of oligomer assemblies that could not be characterised experimentally. Post-translational modifications, including hyper-phosphorylation of tau, trigger its self-assembly by decreasing its microtubule-stabilising ability [60] and may act as an important target for disease-modifying therapies [68]. In addition to the formation of PHFs in AD, hyper-phosphorylated tau aggregates may form neurofibrillary or gliofibrillary tangles commonly known as "tauopathies" [69].

In this regard, MD simulations have helped reveal the dynamics of monomer misfolding and dimerisation of the four microtubule-binding (MTB) repeat domains (R1–R4) constituting the core of tau fibrils with R3 monomers forming β-sheets while both R2 and R3 repeats aggregate into metastable β-sheet-rich dimers, especially residues composed of the PHF6 hexapeptide of R3 [70]. A more recent MD simulation study predicted the aggregation propensity of the repeat domains of tau peptide where the R3–R4 (residues 306–378) monomer may form transient β-hairpins within the R3 repeat and between the R3 and R4 repeats in bulk solution, but spontaneous β-sheets insertion was not observed in modelling on the membrane surface [71]. MD simulations have also been coupled with MSM and TPT models to uncover the tau misfolding kinetics and structural features of the key R3 repeat domain at the atomic scale, where a critical intermediate state was noted for the formation of two target β-sheet structures [72].

A number of recent simulations have focused on investigating the morphologies of tau oligomers and fibrils. For instance, it was predicted through atomistic MD simulations that the C-shaped conformation of the fibril core is retained only by the R3–R4 repeat domains, while the R1–R2, first and second repeat domains, tend to have a linear

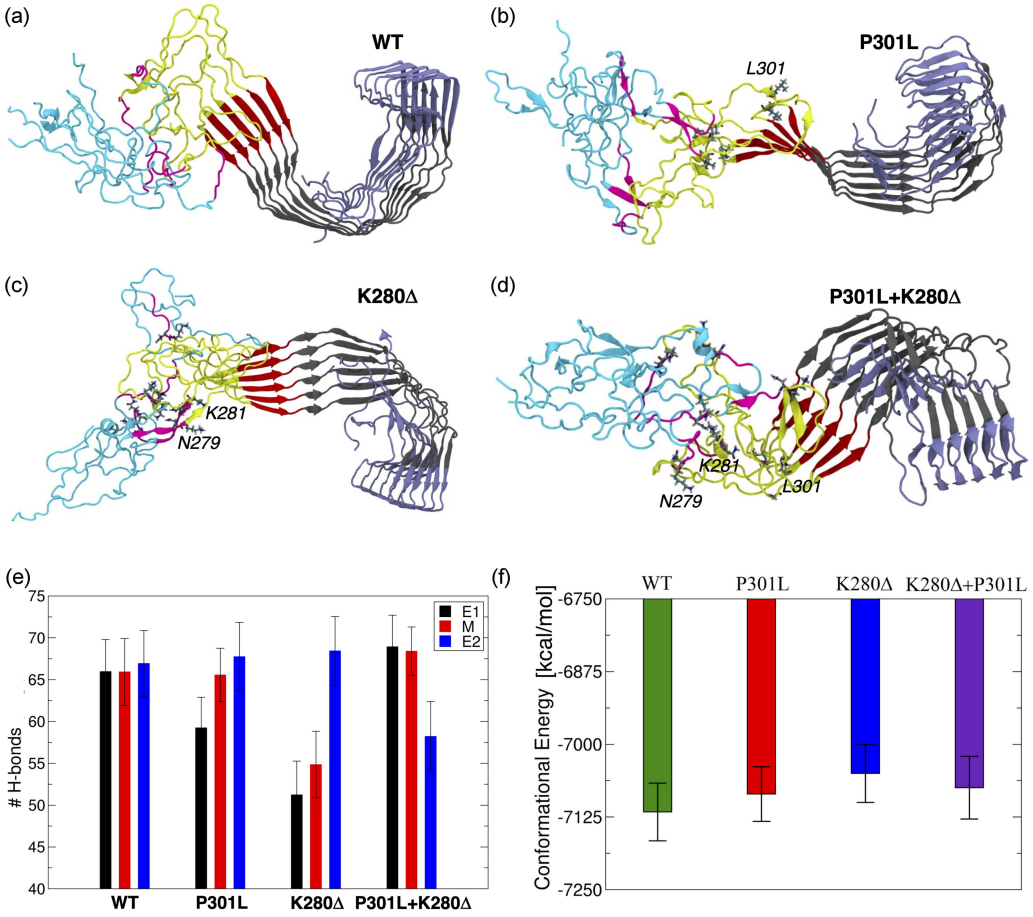


Fig. 4. (a–d) Final conformation of tau microtubule-binding domain (MBD) pro-fibrillar structure formed after 1 μs each of free molecular dynamics (MD) in water under physiological conditions for (a) WT, (b) P301L, (c) K280Δ, (d) and double mutant P301L + K280Δ. (e) Comparison of the average number of H-bonds (computed over the last 250 ns dynamics) between two monomer chains at one end (E1), central or middle dimer (M), and the other end dimer (E2) of each hexamer tau MBD model. (f) Comparison of conformational energies for the crystalline domain of the tau MBD hexamers.

shape [73]. This model finding of a C-shaped fibrillar core formed by R3 and R4 domains was later confirmed by experimental cryo-EM structures of tau protofibrillar straight filaments (SF) [62]. Multiscale MD simulations (both atomistic and coarse-grained MD) to explore the conformational features of hyper-phosphorylation on tau repeat domains (R1–R4) showed that hyper-phosphorylation exposes the repeats to bulk solution, which could further promote tau filament self-assembly [74]. The latest study also proposed MD models of the hyper-phosphorylated NPF fibril repeat domains at three experimentally observed phosphorylated Ser sites (S262, S324, and S356) in the MTB domain [75]. Mutations E264G and D358G were engineered on the wild-type (WT) narrow Pick's filament to understand the function of E264 and D358 residues on the local conformations and compare them with the fibrillar architecture of hyper-phosphorylated NPF from microseconds scale atomistic MD simulations. The models revealed that the mutant and hyper-phosphorylated NPF showed a major morphological departure from the WT narrow Pick's filament and that the repeat-specific sequence of the C-terminal hexapeptide strongly guides and influences the conformational properties of the PGGG motif that flanks the hexapeptide in tau [75].

Through four-microsecond atomistic MD simulations, we recently modelled the driving forces behind the assembly of pro-fibrillar hexameric oligomers of two familial mutations within the MTB repeat R1–R4 domains of tau, namely the P301L substitution and the deletion mutation K280Δ (both known to cause frontotemporal dementia). The models identify their pro-aggregation capability due to their innate core packing by R2 and R3 and the overall stability of hexamers in their fibrillar fold that facilitates pro-fibrillar elongation compared to the WT and a control double mutant, P301L + K280Δ of the MTB R1–R4 domain (Fig. 4a–d) [76]. Based on H-bond networks, the models predicted H-bond strengths with regard to three dimers in the hexamer: the dimer at one end of the fibril (denoted as E1), the dimer in the middle of

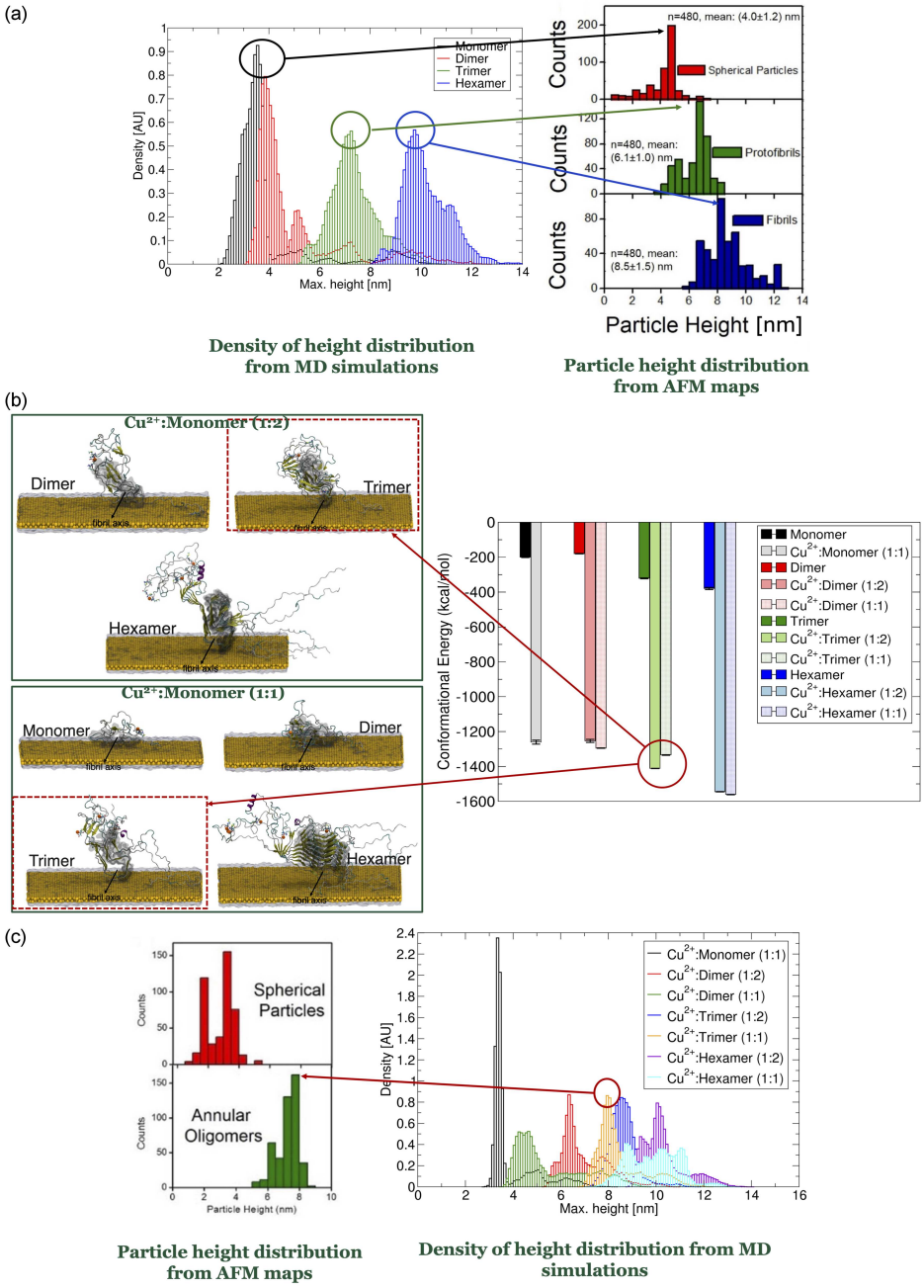


Fig. 5. (a) Density of height distribution profile of αS assemblies without Cu(II) at the gold–water interface from MD simulations matched up with particle height distributions from AFM maps on ultra-flat gold predicts αS trimer as the critical nucleus for fibril growth. (b) Representative snapshots of Cu (II)-bound monomer and oligomer constructs at different Cu(II) concentrations (low concentration — Cu^{2+}: monomer = 1 : 2 and high concentration — Cu^{2+} : monomer = 1 : 1). Increase in Cu(II) concentration do not sample more stable trimers, as confirmed by their conformational energies. (c) Model predicted trimer height distributions at higher Cu(II) concentrations on the gold–water interface confirmed to be annular oligomers from AFM particle height profile on ultra-flat gold.

the fibril (denoted as M), and the dimer at the other end of the fibril (denoted as E2). We noted almost no difference in H-bond strengths of E1 and E2 for WT tau, indicating that the WT hexamers may not undergo pro-fibrillar growth in one direction. The maximum difference in H-bond strength of E1 and E2 is observed for K280Δ followed by P301L + K280Δ, with P301L showing the least difference (Fig. 4e), which supported previous experimental observations that K280Δ mutation leads to enhanced overall aggregation kinetics with increased nucleation and elongation rates, while the P301L mutation leads to a stunted fibril growth rate compared to K280Δ. The models predicted that the

thermodynamic stabilities of hexamers in the fibrillar fold and core packing of the tau crystalline domain are significantly altered by the missense mutations, P301L and K280Δ, indicating they may be more oligomer-like (Fig. 4f).

3.3. Modelling self-assembly of αS protein in PD

Several lines of evidence have shown that β-sheet-rich αS oligomers may trigger neurotoxicity mainly by disrupting membrane integrity, including impairment of protein degradation and function of mitochondria and endoplasmic reticulum [77]. In a previous work [78], we used extensive MD simulations to model the location of several hotspots in the hydrophobic segment (residues 71–82) of non-amyloid β component (NAC region) fibrils that initiates αS self-assembly, which in both termini of NAC could change the populations of different fold morphologies adopted by NAC. The models predicted that at a lower temperature, both WT and mutant αS are sensitive to the solution environment, including the physiological salt concentration, which decreases the stability of WT NAC fibrils and may shift the relative stability of different NAC mutants. The models provide new insights into the polymorphic conformational states of αS fibrils to help predict the binding sites of new and existing protective inhibitors.

A recent MD study of full-length αS monomer misfolding and dimerisation revealed that both monomers and dimers mainly adopt disordered conformations with partial helices around the N-terminus, which is known to bind lipids and form α-helices [79]. β-sheets were mainly formed in the N-terminal tail and the NAC region, with the C-terminus remaining mostly unstructured. Further, dimerisation enhanced the β-sheet content with a subsequent decrease in disorder, with the interaction of the C-terminus with the N-terminal tail and NAC regions indicating the prevention of αS self-assembly [80]. Multi-scale MD was also recently used in conjunction with NMR and cross-linking mass spectrometry (XLMS) to probe the interactions of αS with anionic lipid cellular membrane [81]. The computational and experimental models reveal a break in the helical structure of the NAC region of αS that possibly promotes oligomer formation. Specifically, liposome-bound αS showed β-strand formation in the NAC region, and MSM models indicated a membrane-interacting αS mechanism *via* the dynamic helix break in the NAC region for pathogenesis in PD. To identify the cellular lipid membrane-mediated polymorphic folds of αS fibrils, six structures of αS fibril–lipid complexes were identified with cryo-EM, and the lipid–fibril interactions were revealed using MD simulations along with solid-state NMR (ss-NMR) spectroscopy [82]. The models revealed that phospholipids promote an unusual arrangement of protofilaments, which fill the fibril central cavities, identifying a potential mechanism for the neurotoxicity in PD by fibril-induced lipid extraction.

In recent work [83], we modelled different oligomeric assemblies (monomers, dimers, trimers, and hexamers) of the αS protein at low and high copper (Cu(II)) concentrations, because copper is one of the metals found in high concentration in the post-mortem PD patient brains. MD simulations were performed at the gold–water interface, as αS aggregates were visualised and quantified on ultra-flat gold in water through AFM (Fig. 5) [83]. Our model distribution of density of the heights of assemblies in a Cu(II)-free environment predicted proximity of αS monomer and dimer with spherical particles measured in AFM, trimer with the protofibrillar fold in AFM and hexamer having a fibrillar fold (see Fig. 5a). So, we propose that trimers are the minimal critical nucleus for elongation of αS protofibrils at the gold–water interface. Our simulations show that there are significantly tighter assemblies with increasing concentrations of Cu(II), as seen from the assembly height distributions at the gold–water interface and their conformational energy profiles, except for the trimers for which we noted that increased copper concentration does not make the trimers thermodynamically more stable (Fig. 5b). At low Cu(II) concentration, the trimer retains their assembly fold, but at higher Cu(II) concentration, the trimer shifts to a different conformation indicative of an atypical fold, which was confirmed from AFM particle heights to be annular-shaped oligomers corresponding to our model heights of trimers at high copper concentration (Fig. 5c). We propose such highly toxic annular oligomers as potential drug targets for treating PD.

3.4. Modelling the aggregation-resistant αS helical tetramer protein in PD

The latest experimental findings by a number of groups have proposed that αS may exist as α-helically folded tetramer that resists further aggregation under normal physiological conditions [84, 85]. There are many contradictory viewpoints on whether αS is a predominantly disordered cytosolic monomer that is aggregation-prone [86], as long understood [87], or a cytosolic α-helically folded tetramer that is aggregation-impeding, as recently discovered [88–90]. It is now believed that the unfolded monomeric and helically folded tetrameric states may be in dynamic equilibrium with each other [91, 92], as evident from the familial PD causing missense mutations which shifted the tetramers to pro-aggregating monomers precipitating neurotoxicity by decreasing αS solubility [93]. It was also shown that homologous E→K mutations destabilise αS multimers (including helical tetramers) and induce monomer aggregation at membranes to form vesicle-rich inclusions [89, 94].

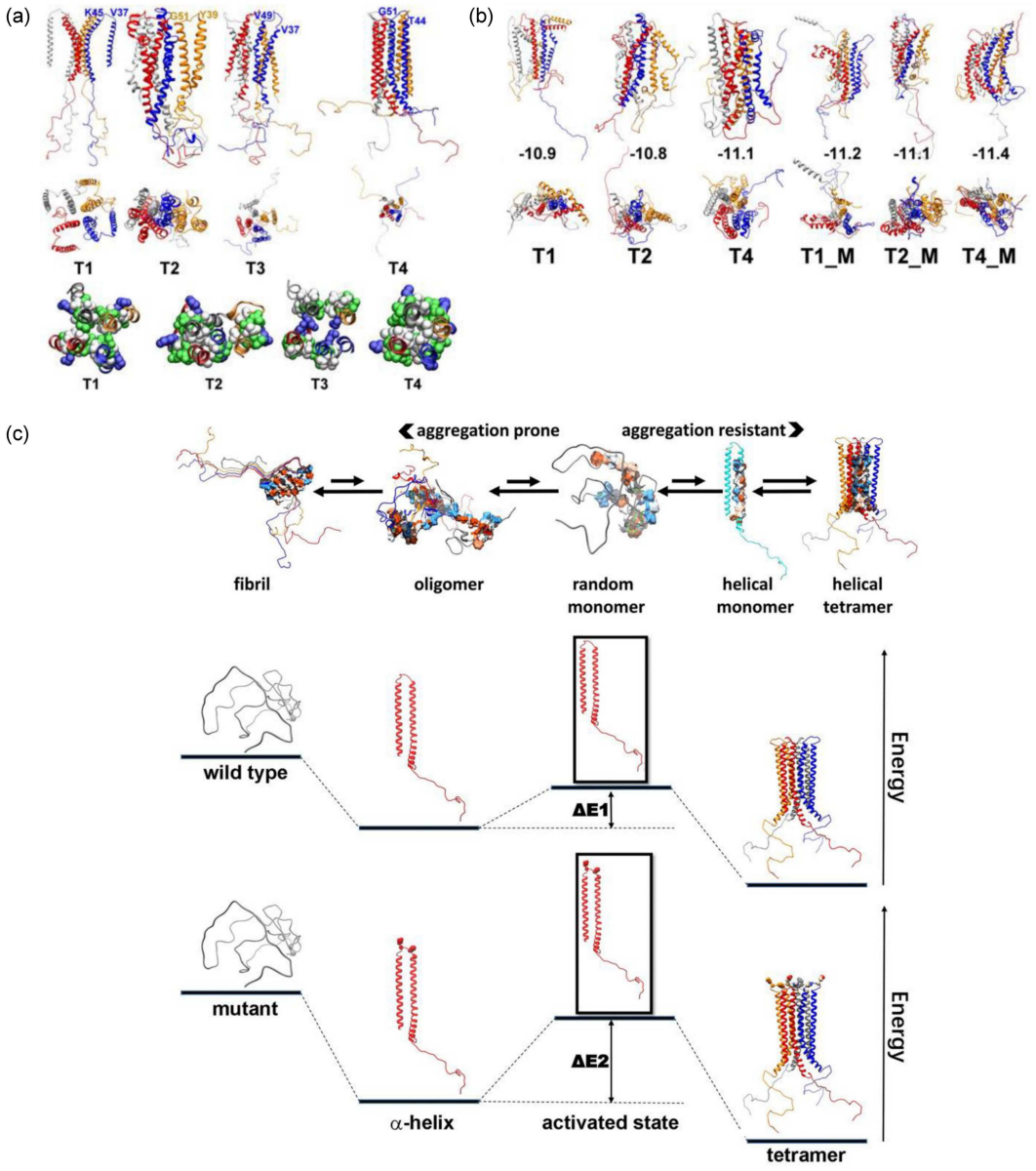


Fig. 6. (a) Initial conformations of four full-length α-helical αS tetramers (T1–T4) in both side and top views. The first and last residues of the loop that connects α1 and α2 segments are labelled. The heavy atoms (top view) of residues 71–82 are represented as van der Waals spheres with hydrophobic, acidic, basic, and polar residues coloured white, red, blue, and green, respectively. (b) Final structures of different helical αS tetramer conformations (T1, T2, and T4 for wild type, T1_M, T2_M, and T4_M for mutant) computed following 200 ns simulations, shown in side and top views. The tetramer T3 is not stable and is not included here. The corresponding conformational energy (in 10^3 kcal/mol) averaged over the final 20 ns of dynamics is shown for each tetramer. (c) Top: Role of helical tetramer in the pathological aggregation of αS. The hydrophobic NAC is also represented as the surface. Bottom: The proposed molecular pathways to decreased tetramer:monomer ratios in the mutant. Energy levels are estimated according to the calculated conformational energy of disordered αS monomers, helical αS monomers and tetramers.

However, there are very limited computationally verified models or MD simulation studies to date on the helical αS tetramers. One MD study that used a fragment-based approach to construct energetically favourable full-length αS suggested that the sampled structures with amphipathic helices can self-assemble via hydrophobic contacts to form tetramers [95]. In another study, a combination of replica exchange MD (REMD) and variational Bayesian weighting (VBW) methods was used to generate monomers, α-helical- and β-strand-rich αS trimers and tetramers in an attempt to resolve the controversy regarding experimentally observed αS native structure [96]. The authors noted that the ensemble is dominated by disordered monomers, with very few helical trimers and tetramers, although

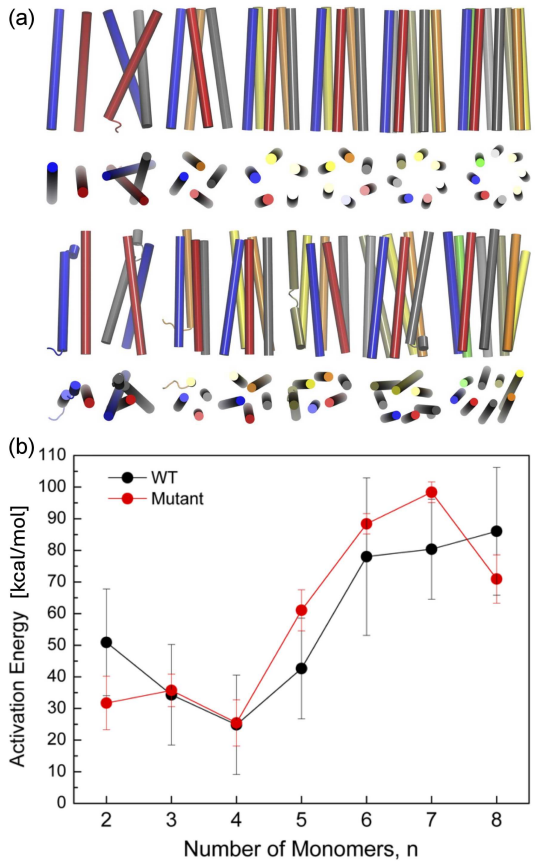


Fig. 7. (a) Designed αS NAC oligomers from dimer to octamer in side and top views at the start of the MD simulations (top panel, after minimization and equilibration) and after 200 ns of unconstrained dynamics (bottom panel) in water. (b) Activation energy of formation of αS multimers from monomer to octamer. Horizontal axis number n from 1 to 8 indicates growth from monomer to octamer.

the tetrameric states had significant helical content. Another simulation study observed that αS tetramer from a completely disordered state exhibited appreciably reduced stable β-sheets in comparison to dimers and a more stable helical content than either monomer or dimer [97]. Finally, by employing a steric parameter, correlations were obtained between the main and side chains of αS monomers and tetramers, revealing residues consisting mostly of parts of KTKEGV repeats that could potentially mediate the formation of helical tetramers [98].

Given that hydrophobic packing plays an important role in the folding and stabilisation of globular proteins [99], for IDPs like αS that display a broad distribution of conformational substates under physiological environment [7], hydrophobic interactions have been suggested to be a major driving force governing their self-assembly into oligomers, as "hydrophobes" also pack along the fibril growth axis [100]. We note that most homomers have high structural symmetry [101], and cyclic symmetries are thought to be the basic building blocks for the *de novo* design of self-assembling proteins such as water-soluble α-helical barrels [102] and helical bundles with high thermodynamic stability [103]. We used this preliminary knowledge to model and rationally design αS tetramers, which are nevertheless homomeric assemblies. We probe the free energy landscape of αS tetramerisation from four alternatively designed WT broken α-helical constructs (T1–T4) and their familial mutants T1_M, T2_M, and T4_M (Fig. 6a,b) and identify the active state corresponding to the conformation attained by a monomer when bound in a stable tetramer [104]. In the process, we designed the most thermodynamically stable (tetramer T4) *de novo* broken α-helical tetramer with a reconstructed loop motif using available experimental data [105]. Our results highlight that optimisation of inter-monomeric hydrophobic packing in NAC regions facilitates assembly to a stable broken water-soluble α-helical tetrameric construct, with secondary roles of the termini in regulating stability [104]. Moreover, we show that PD-causing familial mutations may create a much higher energy barrier for association of α-helical monomers into the aggregation-resistant α-helical tetramer, shifting tetramer–monomer equilibrium back towards aggregation-prone disordered monomers (Fig. 6c).

Following our designed *de novo* broken α-helical αS tetramer assembly in [104] with residues Val3–Val44 and Lys51–Thr92 forming two α-helices, we further designed a more stable αS broken α-helical tetramer construct using the same αS helical monomer as the building unit. Additionally, oligomers/multimers from dimers to octamers were modelled using the same designed broken α-helical monomer structure (Fig. 7a). The initial helical multimeric structures contained NAC regions with Cn symmetry. We characterize the thermodynamic and kinetic properties from MD simulations of both WT and quadruple mutated (E46K + H50Q + G51D + A53T) α-helical tetramers from the re-designed, more stable *de novo* α-helical tetramer assembly in order to elucidate the proposed hypothesis that tetramers may be ubiquitous in nature compared to α-helical αS oligomers from dimer to octamer [106]. Our models revealed that although the conformational stability of αS oligomers increases linearly with the number of monomers, the assembly of αS multimers proceeds *via* multiple energy barriers. The tetramer shows the lowest activation energy (Fig. 7b), which may explain its ubiquity.

3.5. Modelling co-assembly of pathological proteins in AD and PD

It is becoming increasingly clear that co-aggregation or cross-seeding assembly of amyloid proteins may be more neurotoxic than their self-assembly in AD and PD pathogenesis [107], and

there are several clinical overlaps in symptoms and pathologies between AD and PD. A previous MD study investigated the plausible early assembly pathways in PD and AD through cross-dimerisation of αS_{1-95} and $A\beta_{1-42}$ to reveal that the imperfect KTKEGV repeats in the N-terminus of αS may be responsible for forming inter-protein salt bridges with Aβ and NAC in αS may closely interact with Aβ hydrophobic core to form these hetero-assembled pathological protein complexes [108]. In a more recent study, MD simulations investigated the impact of αS–Aβ hetero dimerisation, showing that αS directly interacted with Aβ monomers and dimers, aggregating to potentially toxic β-barrel intermediates [109]. The αS–Aβ binding was mediated by the N-terminal end and NAC region in αS and the central hydrophobic cluster (CHC) C-terminus in Aβ.

Recent REMD simulations identified the conformational ensembles formed by the co-aggregation of CHC of Aβ ($A\beta_{16-22}$) and each of two core segments of tau (PHF6* and PHF6) [110]. The hetero-oligomers formed were found to be rich β-sheet, with PHF6 and $A\beta_{16-22}$ aggregate forming closed β-barrels, while PHF6* and $A\beta_{16-22}$ aggregate form open β-barrels. Hydrophobic and π–π stacking interactions were found to be crucial for the formation of toxic closed β-barrel between PHF6 and $A\beta_{16-22}$.

4. Conclusions

To identify new therapeutic targets for common proteinopathies such as Alzheimer's (AD) and Parkinson's diseases (PD), a comprehensive map of the molecular-level detailed pathway of early stages of self-assembly of pathological proteins is required, especially to structurally define the rare, polymorphic and short-lived toxic oligomeric intermediates. A number of *in vitro* and *in vivo* experimental techniques in the past have attempted to uncover the morphologies of pathogenic oligomers, but with little to no success, mainly due to a lack of a reliable quantification method. Computer-based molecular modelling and molecular dynamics (MD) simulations then have provided significant insights in guiding experiments on protein self-assembly. In this mini-review, we have focussed on the recent advances and latest findings from modelling and MD simulations of pathological protein self-assembly in bulk solution and on surfaces and interfaces that reveal some key structural and morphological details, thermodynamic driving forces, and kinetics of formation of several assembly constructs, including oligomers of proteins amyloid-β (Aβ) and tau in AD and α-synuclein in PD. We remain hopeful that fundamental simulation-guided research will uncover new therapeutic targets for these common proteinopathies and foster efforts to re-engineer functionality in pathogenic amyloids, like designer nanostructured materials [111–113].

Acknowledgments

This review article is dedicated with love to the memory of Professor Marek Cieplak, a dear mentor and friend of DT. To paraphrase a colleague's description of the generous and knowledgeable Marek, even when Google did not work, Marek worked. His passion and enthusiasm for modelling disordered linker regions in proteins informed much of our current work in IDPs as drug targets.

References

[1] T.P.J. Knowles, M. Vendruscolo, C.M. Dobson, *Nat. Rev. Mol. Cell Biol.* **15**, 384 (2014).

[2] C.A. Ross, M.A. Poirier, *Nat. Rev. Mol. Cell Biol.* **6**, 891 (2005).

[3] M.T. Heemels, *Nature* **539**, 179 (2016).

[4] V.L. Feigin, T. Vos, E. Nichols et al., *Lancet Neurol.* **19**, 255 (2020).

[5] World Health Organization (WHO), *Neurological Disorders. Public Health Challenges*, World Health Organization, Geneva 2006.

[6] À. Gómez-Sicilia, M. Sikora, M. Cieplak, M. Carrión-Vázquez, *PLoS Comput. Biol.* **11**, e1004541 (2015).

[7] S. Bhattacharya, L. Xu, D. Thompson, *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **8**, e1359 (2018).

[8] R.N.L. Lamptey, B. Chaulagain, R. Trivedi, A. Gothwal, B. Layek, J. Singh, *Int. J. Mol. Sci.* **23**, 1851 (2022).

[9] N. Bittner, C.S.M. Funk, A. Schmidt, F. Bermpohl, E.J. Brandl, E.E.A. Algharably, R. Kreutz, T.G. Riemer, *Drugs Aging* **40**, 953 (2023).

[10] S. Reardon, *Nature* **613**, 227 (2023).

[11] R. McShane, M.J. Westbya, E. Roberts, N. Minakaran, L. Schneider, L.E. Farrimond, N. Maayan, J. Ware, J. Debarros, *Cochrane Database Syst. Rev.* **2019**, CD003154 (2019).

[12] P. Chopade, N. Chopade, Z. Zhao, S. Mitragotri, R. Liao, V.C. Suja, *Bioeng. Transl. Med.* **8**, e10367 (2023).

[13] S. Srivastava, R. Ahmad, S.K. Khare, *Eur. J. Med. Chem.* **216**, 113320 (2021).

[14] J.A. Szász, V.A. Constantin, K. Orbán-Kis et al., *Brain Sci.* **11**, 826 (2021).

[15] T. Sinnige, *Chem. Sci.* **13**, 7080 (2022).

[16] J. Habchi, S. Chia, R. Limbocker et al., *Proc. Natl. Acad. Sci. USA* **114**, E200 (2017).

[17] Ł. Mioduszewski, M. Cieplak, *Phys. Chem. Chem. Phys.* **20**, 19057 (2018).

[18] M. Wojciechowski, À. Gómez-Sicilia, M. Carrión-Vázquez, M. Cieplak, *Mol. Biosyst.* **12**, 2700 (2016).

[19] A.B. Poma, H.V. Guzman, M.S. Li, P.E. Theodorakis, *Beilstein J. Nanotechnol.* **10**, 500 (2019).

[20] A. Kamada, A. Levin, Z. Toprakcioglu, Y. Shen, V. Lutz-Bueno, K.N. Baumann, P. Mohammadi, M.B. Linder, R. Mezzenga, T.P.J. Knowles, *Small* **16**, 1904190 (2020).

[21] L.R. Volpatti, T.P.J. Knowles, *J. Polym. Sci. B Polym. Phys.* **52**, 281 (2014).

[22] T.P.J. Knowles, M.J. Buehler, *Nat. Nanotechnol.* **6**, 469 (2011).

[23] S. Guerin, S.A.M. Tofail, D. Thompson, *Cryst. Growth Des.* **18**, 4844 (2018).

[24] K. Tao, J. O' Donnell, H. Yuan et al., *Energy Environ. Sci.* **13**, 96 (2020).

[25] R.B. Svensson, H. Mulder, V. Kovanen, S.P. Magnusson, *Biophys. J.* **104**, 2476 (2013).

[26] M.G. Iadanza, M.P. Jackson, E.W. Hewitt, N.A. Ranson, S.E. Radford, *Nat. Rev. Mol. Cell Biol.* **19**, 755 (2018).

[27] M. Goedert, D.S. Eisenberg, R.A. Crowther, *Annu. Rev. Neurosci.* **40**, 189 (2017).

[28] A.L. Mahul-Mellier, J. Burtscher, N. Maharjan, L. Weerens, M. Croisier, F. Kuttler, M. Leleu, G.W. Knott, H.A. Lashuel, *Proc. Natl. Acad. Sci. USA* **117**, 4971 (2020).

[29] A.J. Dear, T.C.T. Michaels, G. Meisl, D. Klenerman, S. Wu, S. Perrett, S. Linse, C.M. Dobson, T.P.J. Knowles, *Proc Natl Acad Sci USA* **117**, 12087 (2020).

[30] P.H. Nguyen, A. Ramamoorthy, B.R. Sahoo et al., *Chem. Rev.* **121**, 2545 (2021).

[31] R. Limbocker, N. Cremades, R. Cascella, P.M. Tessier, M. Vendruscolo, F. Chiti, *Acc. Chem. Res.* **56**, 1395 (2023).

[32] K. Kulenkampff, A.M. Wolf Perez, P. Sormanni, J. Habchi, M. Vendruscolo, *Nat. Rev. Chem.* **5**, 277 (2021).

[33] F. Grigolato, P. Arosio, *Biophys. Chem.* **270**, 106533 (2021).

[34] J.M. Kenyaga, Q. Cheng, W. Qiang, *J. Biol. Chem.* **298**, 102491 (2022).

[35] N. El Mammeri, O. Gampp, P. Duan, M. Hong, *Commun. Biol.* **6**, 467 (2023).

[36] A.S. Kurochka, D.A. Yushchenko, P. Bouř, V.V. Shvadchak, *ACS Chem. Neurosci.* **12**, 825 (2021).

[37] T. John, A. Gladytz, C. Kubeil, L.L. Martin, H.J. Risselada, B. Abel, *Nanoscale* **10**, 20894 (2018).

[38] A. Morriss-Andrews, J.E. Shea, *J. Chem. Phys.* **136**, 065103 (2012).

[39] P. Ricchiuto, A.V. Brukhno, S. Auer, *J. Phys. Chem. B* **116**, 5384 (2012).

[40] L. Xu, S. Bhattacharya, D. Thompson, *Methods Mol. Biol.* **2340**, 379 (2022).

[41] S. Bhattacharya, L. Xu, D. Thompson, *Methods Mol. Biol.* **2340**, 401 (2022).

[42] D. Matthes, V. Gapsys, J.T. Brennecke, B.L. De Groot, *Sci. Rep.* **6**, 33156 (2016).

[43] D.O.V. Alonso, S.J. DeArmond, F.E. Cohen, V. Daggett, *Proc. Natl. Acad. Sci. USA* **98**, 2985 (2001).

[44] L. Calzolai, R. Zahn, *J. Biol. Chem.* **278**, 35592 (2003).

[45] S. Samantray, W. Schumann, A.-M. Illig, M. Carballo-Pacheco, A. Paul, B. Barz, B. Strodel, *Computer Simulations of Aggregation of Proteins and Peptides*, Vol. 2340, 2022, p. 235.

[46] B. Strodel, *Curr. Opin. Struct. Biol.* **67**, 145 (2021).

[47] J. Huang, S. Rauscher, G. Nawrocki, T. Ran, M. Feig, B.L. de Groot, H. Grubmüller, A.D. MacKerell Jr., *Nat. Methods* **14**, 71 (2016).

[48] P. Robustelli, S. Piana, D.E. Shaw, *Proc. Natl. Acad. Sci. USA* **115**, E4758 (2018).

[49] Y. Zhang, Y.L. Lyubchenko, *Biophys. J.* **107**, 2903 (2014).

[50] W. Zheng, M.Y. Tsai, M. Chen, P.G. Wolynes, *Proc. Natl. Acad. Sci. USA* **113**, 11835 (2016).

[51] M. Hashemi, Y. Zhang, Z. Lv, Y.L. Lyubchenko, *Nanoscale Adv.* **1**, 3892 (2019).

[52] G.M. Shankar, S. Li, T.H. Mehta et al., *Nat. Med.* **14**, 837 (2008).

[53] V.H. Man, P.H. Nguyen, P. Derreumaux, *J. Phys. Chem. B* **121**, 2434 (2017).

[54] V.H. Man, P.H. Nguyen, P. Derreumaux, *J. Phys. Chem. B* **121**, 5977 (2017).

[55] A.M. Illig, B. Strodel, *J. Chem. Theory Comput.* **16**, 7825 (2020).

[56] S. Bhattacharya, L. Xu, D. Thompson, *ACS Chem. Neurosci.* **10**, 2830 (2019).

[57] S. Bhattacharya, L. Xu, D. Thompson, *Sci. Rep.* **10**, 7597 (2020).

[58] P.N. Nirmalraj J. List, S. Battacharya, G. Howe, L. Xu, D. Thompson, M. Mayer, *Sci. Adv.* **6**, eaaz6014 (2020).

[59] M. Morris, S. Maeda, K. Vossel, L. Mucke, *Neuron* **70**, 410 (2011).

[60] Y. Wang, E. Mandelkow, *Nat. Rev. Neurosci.* **17**, 22 (2016).

[61] M.D. Mukrasch, S. Bibow, J. Korukottu, S. Jeganathan, J. Biernat, C. Griesinger, E. Mandelkow, M. Zweckstetter, *PLoS Biol.* **7**, e1000034 (2009).

[62] A.W.P. Fitzpatrick, B. Falcon, S. He et al., *Nature* **547**, 185 (2017).

[63] Y. Shi, A.G. Murzin, B. Falcon et al., *Acta Neuropathol.* **141**, 697 (2021).

[64] G.I. Hallinan, M.R. Hoq, M. Ghosh, F.S. Vago, A. Fernandez, H.J. Garringer, R. Vidal, W. Jing, B. Ghetti, *Acta Neuropathol.* **142**, 227 (2021).

[65] B. Falcon, W. Zhang, M. Schweighauser, A.G. Murzin, R. Vidal, H.J. Garringer, B. Ghetti, S.H.W. Scheres, M. Goedert, *Acta Neuropathol.* **136**, 699 (2018).

[66] B. Falcon, W. Zhang, A.G. Murzin, G. Murshudov, H.J. Garringer, R. Vidal, R.A. Crowther, B. Ghetti, S.H.W. Scheres, M. Goedert, *Nature* **561**, 137 (2018).

[67] P. Duan, A.J. Dregni, N. El Mammeri, M. Hong, *Proc. Natl. Acad. Sci. USA* **120**, e2310067120 (2023).

[68] N. Basheer, T. Smolek, I. Hassan, F. Liu, K. Iqbal, N. Zilka, P Novak, *Mol. Psychiatry* **28**, 2197 (2023).

[69] Y. Zhang, K.M. Wu, L. Yang, Q. Dong, J.T. Yu, *Mol. Neurodegener.* **17**, 28 (2022).

[70] H. He, Y. Liu, Y. Sun, F. Ding, *J. Chem. Inf. Model.* **61**, 2916 (2021).

[71] P.H. Nguyen, P. Derreumaux, *J. Phys. Chem. B* **126**, 3431 (2022).

[72] H. Liu, H. Zhong, Z. Xu, Q. Zhang, S.J.A. Shah, H. Liu, X. Yao, *Phys. Chem. Chem. Phys.* **22**, 10968 (2020).

[73] X. Li, X. Dong, G. Wei, M. Margittai, R. Nussinov, B. Ma, *Chem. Commun.* **54**, 5700 (2018).

[74] L. Xu, J. Zheng, M. Margittai, R. Nussinov, B. Ma, *ACS Chem. Neurosci.* **7**, 565 (2016).

[75] A.E. Sahayaraj, R. Viswanathan, F. Pinhero, A. Abdul Vahid, V. Vijayan, *ACS Chem. Neurosci.* **14**, 136 (2023).

[76] O. Maraba, S. Bhattacharya, M. Conda-Sheridan, D. Thompson, *Nano Express* **3**, 044004 (2022).

[77] G. Fusco, S.W. Chen, P.T.F. Williamson et al., *Science* **358**, 1440 (2017).

[78] L. Xu, S. Bhattacharya, D. Thompson, *Phys. Chem. Chem. Phys.* **20**, 4502 (2018).

[79] H.A. Lashuel, C.R. Overk, A. Oueslati, E. Masliah, *Nat. Rev. Neurosci.* **14**, 38 (2012).

[80] Y. Zhang, Y. Wang, Y. Liu, G. Wei, F. Ding, Y. Sun, *ACS Chem. Neurosci.* **13**, 3126 (2022).

[81] S.B.T.A. Amos, T.C. Schwarz, J. Shi, B.P. Cossins, T.S. Baker, R.J. Taylor, R. Konrat, M.S.P. Sansom, *J. Phys. Chem. B* **125**, 2929 (2021).

[82] B. Frieg, L. Antonschmidt, C. Dienemann et al., *Nat. Commun.* **13**, 6810 (2022).

[83] O. Synhaivska, S. Bhattacharya, S. Campioni, D. Thompson, P.N. Nirmalraj, *ACS Chem. Neurosci.* **13**, 1410 (2022).

[84] W. Wang, I. Perovic, J. Chittuluru et al., *Proc. Natl. Acad. Sci. USA* **108**, 17797 (2011).

[85] T. Bartels, J.G. Choi, D.J. Selkoe, *Nature* **477**, 107 (2011).

[86] F.X. Theillet, A. Binolfi, B. Bekei et al., *Nature* **530**, 45 (2016).

[87] P.H. Weinreb, W. Zhen, A.W. Poon, K.A. Conway, P.T. Lansbury, *Biochemistry* **35**, 13709 (1996).

[88] B. Fauvet, M.K. Mbefo, M.-B. Fares et al., *J. Biol. Chem.* **287**, 15345 (2012).

[89] U. Dettmer, N. Ramalingam, V.E. von Saucken et al., *Hum. Mol. Genet.* **26**, 3466 (2017).

[90] E.S. Luth, T. Bartels, U. Dettmer, N.C. Kim, D.J. Selkoe, *Biochemistry* **54**, 279 (2015).

[91] D. Selkoe, U. Dettmer, E. Luth, N. Kim, A. Newman, T. Bartels, *Neurodegener. Dis.* **13**, 114 (2014).

[92] U. Dettmer, D. Selkoe, T. Bartels, *Curr. Opin. Neurobiol.* **36**, 15 (2016).

[93] U. Dettmer, A.J. Newman, F. Soldner et al., *Nat. Commun.* **6**, 7314 (2015).

[94] S. Nuber, M. Rajsombath, G. Minakaki, J. Winkler, C.P. Müller, M. Ericsson, B. Caldarone, U. Dettmer, D.J. Selkoe, *Neuron* **100**, 75 (2018).

[95] O. Ullman, C.K. Fisher, C.M. Stultz, *J. Am. Chem. Soc.* **133**, 19536 (2011).

[96] T. Gurry, O. Ullman, C.K. Fisher, I. Perovic, T. Pochapsky, C.M. Stultz, *J. Am. Chem. Soc.* **135**, 3865 (2013).

[97] J.Y. Mane, M. Stepanova, *FEBS Open Bio* **6**, 666 (2016).

[98] Y. Cote, P. Delarue, H.A. Scheraga, P. Senet, G.G. Maisuradze, *ACS Chem. Neurosci.* **9**, 1051 (2018).

[99] G.D. Rose, R. Wolfenden, *Annu. Rev. Biophys. Biomol. Struct.* **22**, 381 (1993).

[100] B. Li, P. Ge, K.A. Murray et al., *Nat. Commun.* **9**, 3609 (2018).

[101] C.H. Norn, I. André, *Curr. Opin. Struct. Biol.* **39**, 39 (2016).

[102] A.R. Thomson, C.W. Wood, A.J. Burton, G.J. Bartlett, R.B. Sessions, R.L. Brady, D.N. Woolfson, *Science* **346**, 485 (2014).

[103] P.S. Huang, G. Oberdorfer, C. Xu et al., *Science* **346**, 481 (2014).

[104] L. Xu, S. Bhattacharya, D. Thompson, *Chem. Commun.* **54**, 8080 (2018).

[105] E. Kara, P.A. Lewis, H. Ling, C. Proukakis, H. Houlden, J. Hardy, *Neurosci. Lett.* **546**, 67 (2013).

[106] L. Xu, S. Bhattacharya, D. Thompson, *Phys. Chem. Chem. Phys.* **21**, 12036 (2019).

[107] B. Ren, Y. Zhang, M. Zhang et al., *J. Mater. Chem. B* **7**, 7267 (2019).

[108] J.C. Jose, P. Chatterjee, N. Sengupta, *PLoS One* **9**, e106883 (2014).

[109] F. Huang, Y. Liu, Y. Wang, J. Xu, J. Lian, Y. Zou, C. Wang, F. Ding, Y. Sun, *Phys. Chem. Chem. Phys.* **25**, 31604 (2023).

[110] X. Li, Y. Chen, Z. Yang, S. Zhang, G. Wei, L. Zhang, *Int. J. Biol. Macromol.* **254**, 127841 (2024).

[111] M. Gunnoo, P.-A. Cazade, A. Orlowski, M. Chwastyk, H. Liu, D.T. Ta, M. Cieplak, M. Nashde, D. Thompson, *Phys. Chem. Chem. Phys.* **20**, 22674 (2018).

[112] M. Gunnoo, P.-A. Cazade, A. Galera-Prat et al., *Adv. Mater.* **28**, 5619 (2016).

[113] G. Nawrocki, P.A. Cazade, D. Thompson, M. Cieplak, *J. Phys. Chem. C* **119**, 24404 (2015).

In memory of Professor Marek Cieplak

REVIEW

The Role of Cavities in Biological Structures

Ł. Mioduszewski[a], K. Wołek[b] and M. Chwastyk[b,*]

[a] *Cardinal Stefan Wyszyński University, Dewajtis 5, 01-815 Warsaw, Poland*
[b] *Institute of Physics, Polish Academy of Sciences, al. Lotników 32/46, PL-02668 Warsaw, Poland*

Doi: 10.12693/APhysPolA.145.S51 *e-mail: chwastyk@ifpan.edu.pl

We investigate the significance of cavities within biological structures, ranging from single proteins to large complexes, such as viruses and even protein clusters composed of intrinsically disordered proteins. Utilizing our SPACEBALL algorithm, we detect empty spaces within these structures and quantify their volumes. This enables us to elucidate the impact of cavities on the properties of the given structures. Finally, we discuss how the presence of cavities in protein clusters facilitates the assessment of their hydration levels within a coarse-grained implicit solvent approach. Our discussion aims to demonstrate that the functions of various proteins originate from their specific tertiary structures containing cavities.

topics: cavities in biological structures, molecular dynamics simulations, viruses, simulations of gluten

1. Introduction

In the intricate world of biomolecular science, the diversity of shapes that proteins, their aggregates, and complexes can adopt is a captivating and fundamental phenomenon. These varied tertiary or quaternary structures play a pivotal role in determining the functional capabilities of biomolecules, as they dictate their ability to interact with other molecular entities [1]. Within this realm of structural exploration, a number of techniques are employed to unveil the hidden architectures of biomolecules, each offering a unique perspective. Notable among these methodologies are nuclear magnetic resonance (NMR) spectroscopy, X-ray crystallography, and electron microscopy, which empower researchers to construct precise atomic models [2]. These models, once derived, serve as the foundation for theoretical analyses, shedding light on the detailed mechanisms of biomolecular systems.

Intriguingly, these investigative techniques occasionally unveil enigmatic and fascinating topological features within biomolecular structures. Such findings include the knotting of the protein's main chain [3, 4], the entanglement of two chains within multi-chain proteins [5–7], or the hidden cavities within a molecule's core [8, 9]. The most recent of these discoveries is widely discussed in the context of pathogenesis-related proteins of class 10 (PR-10), a category of plant proteins that has long been a source of scientific interest. Despite their conspicuous presence and high expression levels, PR-10 proteins continue to confound researchers by defying easy categorization of their functions. Beyond their purported roles, these enigmatic proteins have been found to participate in various biological processes, including the regulation of development and symbiotic interactions with other organisms [10]. Furthermore, they feature a hollow cavity within their molecular core, formed by a relatively short polypeptide chain comprising 154–163 residues. This cavity is surrounded by a seven-stranded antiparallel β-sheet, intersected by an elongated C-terminal α-helix, as illustrated in Fig. 1 (see also [11]). The figure presents the tertiary structure of the yellow lupine LIPR-10.2B protein obtained from the Protein Data Bank (PDB). This structure was extracted from the LIPR-10.2B/zeatin complex, which involves the plant hormone, *trans*-zeatin [12].

These structural elements are supported by a V-shaped framework formed by two additional helices, H1 and H2, as demonstrated in previous studies [13, 14] and depicted in Fig. 1. This distinctive folding pattern, commonly referred to as the PR-10 fold or Bet v1 fold, owes its nomenclature to the elucidated crystal structure.

To gain insights into the roles of these proteins, it becomes imperative to precisely determine the position of the cavity within a protein and describe its unique characteristics. This necessity served as the impetus for research initiated by Professor Marek Cieplak, resulting in the development of an algorithm and the establishment of the public server known as SPACEBALL [15, 16]. This innovative program facilitates the objective identification of cavity positions and the detailed description of their geometrical and chemical attributes. Notably,

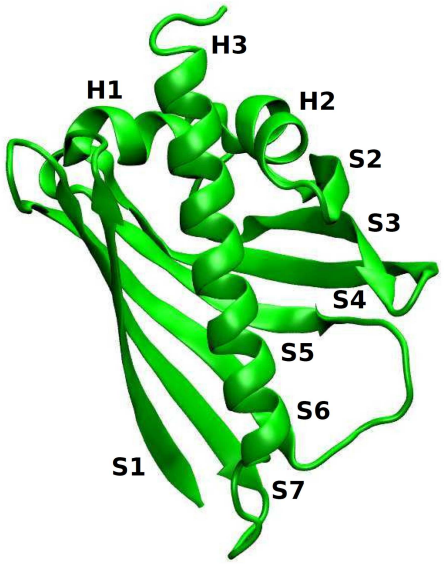


Fig. 1. The native structure of the LIPR-10.2B/zeatin PR-10 protein (PDB: 2QIM) [11]. The β-strands (S) and α-helices (H) are numbered consecutively from the N- to the C-terminus. The three zeatin molecules within the protein's hydrophobic cavity are omitted from the presentation as they were excluded from the analysis.

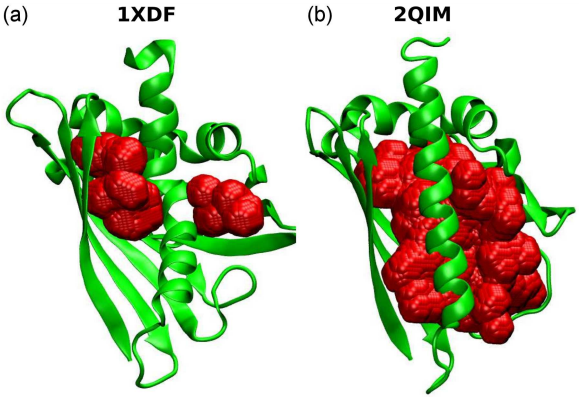


Fig. 2. The structures of two PR-10 protein representatives with PDB codes (a) 1XDF and (b) 2QIM. The protein structure is indicated in green, and the detected cavities are highlighted in red.

SPACEBALL not only enables the characterization of cavities within individual proteins, but also extends its utility to protein aggregates, such as gluten, and complex systems, like virus capsids.

2. Proteins with cavities

In 2013, the group led by Professor Cieplak calculated the volumes of cavities and described the surfaces of eighteen plant pathogenesis-related proteins of class 10. At that time, they characterized the cavities as large, given that the average calculated volume was 326 ± 162 Å^3. Three years later, an updated algorithm was published, and it yielded an average volume of 1309 ± 556 Å^3 for the same set of algorithm parameters. The new version of the algorithm accurately accounted for regions in the immediate proximity of the cavity walls. Since the surface of the cavity interior is often highly irregular, resulting in a large volume, the portion of the cavity volume near the cavity wall contributes significantly to its total volume [16]. Additionally, for calculations using the van der Waals radii proposed by Pauling instead of those proposed by Tsai et al. [17], the average volume was 1494 ± 609 Å^3. The largest cavity, with a volume of 2179 ± 16 Å^3, was detected in LIPR-10.2B/zeatin protein (PDB: 2QIM) [11], while the smallest, 273 ± 10 Å^3, in LIPR-10.2A protein (PDB: 1XDF) [18]. It is important to note that in the case of 1XDF, the protein's interior is composed of three smaller cavities, and the given value represents the volume of the largest one. Nevertheless, even when the volumes of all three are summed, this protein ranks at the lower end of the list of calculated volumes. The structures of these two proteins and the cavity positions are presented in Fig. 2.

The expression of PR-10 proteins increases after viral, bacterial, or fungal infection, as well as due to abiotic factors, such as cold, drought, oxidative stress, or UV radiation [19, 20]. Despite the wide range of factors impacting their expression, no unique function can be attributed to them [20], as mentioned in the introduction. These proteins exhibit considerable uniformity in their behavior, with notable disparities primarily observed in the internal cavity volumes and variations in the optimal folding time [15]. Despite these variations, they demonstrate mechanical robustness and display nearly identical structural rupture patterns when subjected to mechanical forces [15]. This suggests a high stability of the PR-10 fold. Interestingly, this stability is not immediately apparent, given the presence of a large cavity in their structures [15]. Consequently, it is suggested that this protein family may serve as versatile ligand binders, playing diverse roles in small-molecule signaling, transport, or storage. It is essential to highlight that, owing to variations in cavity volumes, shapes, topologies, and internal surface amino acid compositions, individual proteins within this family may offer distinct chemical environments for ligands. This suggests that different proteins possess the capability to host and potentially transport ligands with varying atomic compositions, in line with previous suggestions in the literature [12]. Furthermore, this suggests that such proteins can serve as selectors for ligands.

In 2020, another group led by Professor Cieplak extended the aforementioned analysis to calculate the volumes of cavities within each of the 24 280 single-chain protein structures from the CATH

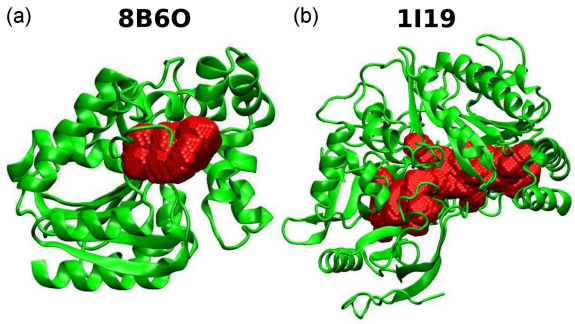


Fig. 3. (a) Tertiary structure (green) and cavity position (red) in haloalkane dehalogenase (PDB: 8B6O). (b) Tertiary structure (green) and cavity position (red) in cholesterol oxidase II (PDB: 1I19).

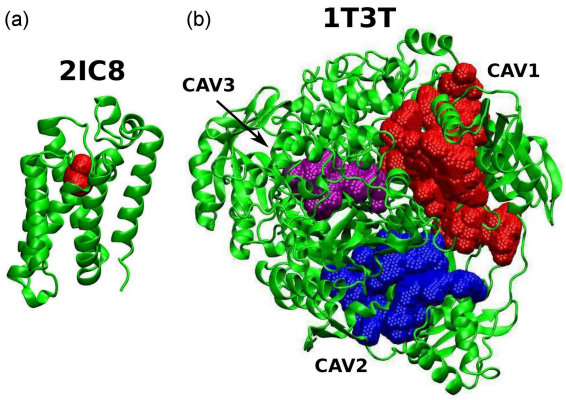


Fig. 4. (a) Structure (green) of the intramembrane protease GlpG from *Escherichia coli* (PDB: 2IC8) [25] with a single cavity (red) detected using the SPACEBALL algorithm. (b) Structure (green) of the StPurL protein (PDB: 1T3T) [26] with three cavities, labeled CAV1 (red), CAV2 (blue), and CAV3 (purple).

database [21]. Their findings demonstrated the existence of cavities with volumes of almost 40 nm^3 (PDB: 1KMP, 1KMP, 1PNZ) and a great number of smaller ones, showcasing the diverse range of cavity sizes within protein structures. It should be emphasized that the volume of the smallest considered cavity is of the order of 12 $Å^3$, which is sufficient to accommodate a single water molecule. This indicates that cavities initially considered to be large were found, upon examination of all available structures, to be relatively small compared to structures with much larger cavities. Moreover, very often the role of these large cavities is much better specified. Beyond ligand binding or small molecule transport, as mentioned in the case of PR-10 proteins, there are several other reasons for the presence of cavities within protein structures. Now, we will discuss the most interesting ones.

Some cavities act as active sites where enzymatic reactions take place. These folds provide a specific microenvironment for the binding and transformation of substrates. One example of structures with an active site buried in a cavity is haloalkane dehalogenases, where the active site is deeply embedded in the predominantly hydrophobic cavity at the interface of the α/β-hydrolase core domain and the helical cap domain [22]. The tertiary structure and the position of the cavity with a volume of 802 ± 34 $Å^3$ are presented in Fig. 3a. Another example of an active site deeply buried within a protein structure is cholesterol oxidase II. The active site in this case consists of a cavity (with a volume of 1977 ± 13 $Å^3$) bounded on one side by the β-pleated sheet in the substrate-binding domain and, on the opposite side, by the isoalloxazine ring of the flavin adenine dinucleotide cofactor covalently attached to the protein [23], as presented in Fig. 3b.

Moreover, cavities can play a role in regulating protein activity. Changes in cavity conformation may control the accessibility of substrates to the active site or modulate the protein's overall function. For example, the protease GlpG of *Escherichia coli* can be inactivated via the selective stabilization of the flexible C subdomain by cavity-filling mutations in this subdomain. On the other hand, cavity-creating mutations might enhance GlpG activity by providing even more flexibility [24]. The structure of GlpG with a cavity of volume 179 ± 19 $Å^3$, calculated using the SPACEBALL algorithm, is presented in Fig. 4a (see also [25]).

Cavities also play a crucial role in allosteric regulation, where binding at one site (allosteric site) influences the activity or conformation of another site in the protein. Surprisingly, the presence of empty spaces in the protein can trigger domain movements that facilitate the activation of the enzyme. In the investigation conducted by Tanwar et al. [26], focusing on the FGAR-AT protein derived from *Salmonella typhimurium* (StPurL), it was elucidated that this protein contains specific hydrophobic cavities that allow for breathing motions. The residues delineating these vacant regions establish a correlation network, interlinking them with the active centers, thereby constituting a functional communication conduit. Additionally, the protein's regions containing cavities, even if lacking a contiguous network with the active center, demonstrate inherent plasticity, rendering them capable of accommodating substantial structural perturbations, including those leading to direct steric conflicts with adjacent neighbors. Here, we show that the mentioned empty spaces are much larger and more extensive than previously detected [26]. Using the SPACEBALL algorithm, we identified three main cavities in the structure of the StPurL protein, as presented in Fig. 4b (see also [26]). Their volumes are: $V_{\mathrm{CAV1}} = 6674 \pm 249$ $Å^3$, $V_{\mathrm{CAV2}} = 3416 \pm 111$ $Å^3$, and $V_{\mathrm{CAV3}} = 1177 \pm 93$ $Å^3$. The positions of these cavities within the protein's structure are marked in red, blue, and purple, respectively.

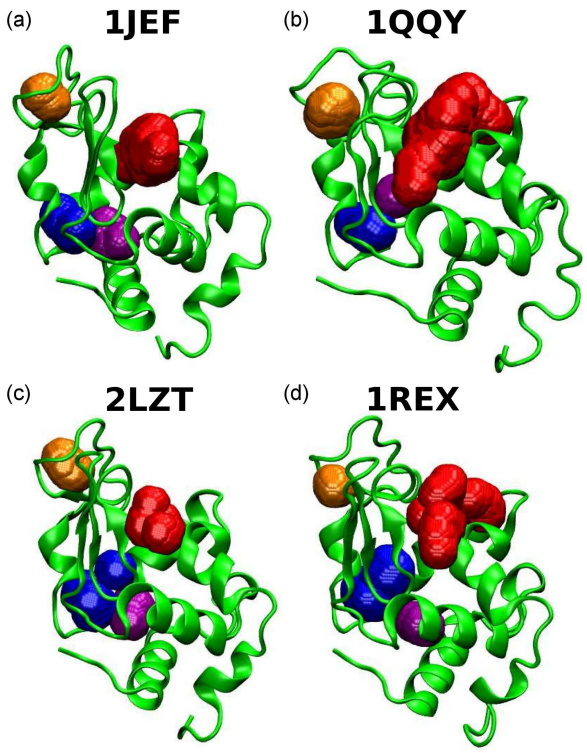


Fig. 5. Lysozyme structures from different biological species: turkey (PDB: 1JEF), dog (PDB: 1QQY), hen (PDB: 2LZT), and human (PDB: 1REX). Specific cavities, detected using the SPACEBALL algorithm, are distinguished by different colors. The largest cavity is marked in red, the smallest one in orange, the second-largest in purple, and the third-largest in blue.

Another intriguing example highlighting the impact of cavities on protein activity is found in the case of lysozymes. According to H. Li and Y.O. Kamatari [27], the location of cavities within structures from different biological species remains the same, despite variations in amino acid sequences. We further investigated whether the volumes of cavities at these specific locations are comparable. To do so, we utilized the SPACEBALL algorithm for structures from various biological species, including turkey (PDB: 1JEF) [28], dog (PDB: 1QQY) [29], hen (PDB: 2LZT) [30], and human (PDB: 1REX) [31]. The detected cavities align with the positions reported in the literature [27]. The volumes of these cavities, calculated using the SPACEBALL algorithm, are recorded in Table I; see also Fig. 5.

Based on the presented data, we observe that, despite a significant difference in the volume of the largest cavities, the volumes of the others, which are more precisely defined, are comparable. The substantial difference in the volumes of the largest cavities arises from their less precisely defined shapes, resembling pockets, whose volumes cannot be determined with high accuracy, yet their positions remain consistent. This observation underscores the

TABLE I

The volumes of cavities detected in lysozyme structures from different biological species. The first column specifies the particular species, the second column indicates its PDB code, and the third through sixth columns provide the volumes of specific cavities, as presented in Fig. 5.

Species	PDB	V_1 [Å^3]	V_2 [Å^3]	V_3 [Å^3]	V_4 [Å^3]
dog	1QQY	397 ± 42	51 ± 12	39 ± 11	29 ± 9
turkey	1JEF	194 ± 24	68 ± 11	49 ± 11	39 ± 12
human	1REX	154 ± 20	70 ± 10	40 ± 8	30 ± 5
hen	2LZT	87 ± 10	75 ± 9	67 ± 7	41 ± 8

potential significance of cavities in the functional attributes of lysozyme, supporting the perspective that these cavities play a pivotal role in the catalytic cycle of lysozymes. Their presence allows for a level of mobility within the active site, maintaining a constant volume available for water molecules. This arrangement is posited to contribute to the hydrolysis of substrate molecules. Furthermore, this outcome supports the notion that cavities are evolutionarily conserved elements essential for protein function [27].

The examples discussed so far illustrate the pivotal role of cavities in proteins, revealing their significance. Thus far, our focus has been on cavities within individual protein chains. Now, we turn our attention to a larger system — the protein complex known as the capsid, which serves as the protective protein coat of a virus.

3. Interior of viral capsid

A virus capsid is an assembly of proteins that shields viral genomes, possessing remarkable mechanical properties that have captured scientific interest. This fascination has led to extensive studies of various capsids to unveil their elastic behavior [32–34]. Computer simulations of nanoindentation experiments [35–38] have revealed that virus capsids, especially those protecting single-stranded RNA, exhibit significant elasticity. The study emphasizes that capsid sturdiness results from a combination of protein mechanical properties and interprotein binding [39]. Here, we explore whether the cavity within the capsid also affects the virus's stability.

Our theoretical research is based on molecular dynamics (MD) simulations, which can be conducted using either all-atom or coarse-grained models. While all-atom simulations offer valuable insights, they are limited by computation time. To address these challenges, we employed a coarse-grained molecular dynamics model to explore the mechanical response of a virus capsid, taking the

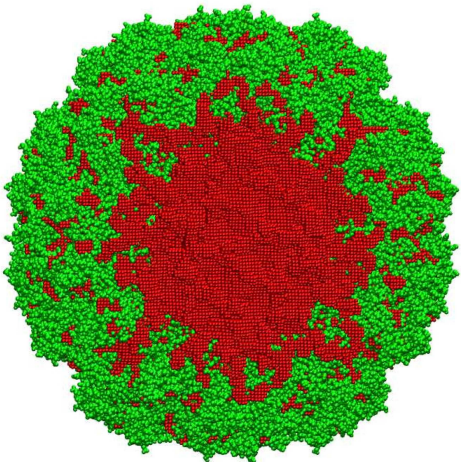


Fig. 6. The capsid of the CCMV virus in its basic native structure (PDB: 1CWP) [40]. The structure is highlighted in green, while the detected cavity is marked in red.

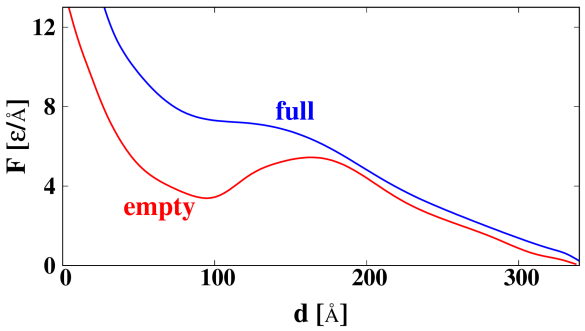


Fig. 7. The force F acting on the opposite walls of the simulation box during the squeezing of the virus capsid as a function of the distance d between the walls, considering both empty and full (with the inclusion of an RNA molecule) CCMV capsids.

cowpea chlorotic mottle virus (CCMV) [40] as an example. The structural analysis includes the identification of a cavity with a volume on the order of 5185 ± 2 nm^3, calculated using the SPACEBALL algorithm, as depicted in Fig. 6 (see also [40]).

The model used for the MD simulations has been developed by Professor Cieplak's group over many years. This is a Gō-like [41] model where each residue interacts with other residues via a pairwise Lennard–Jones potential, while residues in a single chain are connected by harmonic bonds. The model is based on representing each amino acid residue as a single pseudo-atom with an implicit solvent, and the temperature is controlled by a Langevin thermostat [34, 42]. The molecular dynamics within this approach is based on a contact map, i.e., a list of residues in contact, determined from the PDB structure through atomic overlaps [17, 43]. Finally, the native contacts between the Cα atoms i and j at distance r_{ij} are described by the Lennard–Jones potential

$$V(r_{ij}) = 4\epsilon \left[\left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{6} \right], \qquad (1)$$

where σ_{ij} is calculated as $\sigma_{ij} = 2^{-1/6}\, d_{ij}$ for each ij pair, so that the potential minimum coincides with the native distance d_{ij}, and the binding energy parameter ϵ is of the order 110 pN/Å. The interactions between residues not in the contact map are purely repulsive and are modeled using a truncated Lennard–Jones potential, with a cutoff at the minimum of 4 Å [41, 44–46]. The same criteria for atom dynamics were applied to describe interactions between protein chains forming the fully assembled virus capsid [39, 42].

Within the aforementioned model, we conducted a mechanostability analysis of the virus capsid through nanoindentation studies [39]. These studies involved measuring the force, F, acting on the opposite walls of the simulation box while squeezing the virus capsid as a function of the distance, d, between walls. We considered two structures of the capsid: the empty structure (PDB: 1CWP) and the same one but with the inclusion of an RNA molecule composed of 3171 bases. The RNA is modeled as a chain of beads connected by a harmonic potential with an equilibrium distance of 5.8 Å. RNA beads interact only via the excluded-volume effect, which imposes a repulsive interaction for other RNA beads within a distance closer than 8 Å or amino acids closer than 6 Å [39]. Both examined capsids exhibited a linear elastic response under small deformations, followed by a sudden force drop, signaling irreversible structural changes due to bond rupture within the capsid, as depicted in Fig. 7.

It must be noted that in our implicit solvent model, amino acids receive only random "kicks" from the thermostat and experience friction, but we do not consider the potentially stabilizing role of water within the capsid. However, since the capsid wall is semipermeable, allowing water to flow in and out of the capsid [47], we expect this effect to be negligible.

Figure 7 illustrates markedly different mechanical properties of the examined structures. Inter-protein bonds break much more easily in the empty capsid but are much more stable in the case of the one with RNA. This observation reveals another role of cavities within biological structures, i.e., they serve as activators of a cascade of inter-chain bond ruptures after the initial bond breaks, which can be considered the trigger of this process. Such a situation is not observed in structures filled with an RNA molecule that stabilizes the full structure. This may suggest that a larger ratio of empty space to the space occupied by the genetic material within the virus increases the opportunity for the virus to break and release the genetic material, but further research is needed to test this hypothesis [48].

4. Cavities in protein clusters

Cavities, as discussed thus far, have typically been described within proteins or their complexes with well-defined structures. Now, we will explore a slightly different system where cavities can emerge, namely, within intrinsically disordered protein (IDP) clusters. A notable example of such a system is storage proteins of grains, with gluten made from wheat [49] serving as an interesting case study due to its importance in the elasticity of dough in breadmaking [50]. Gluten can be categorized into two main fractions: shorter, water-soluble gliadins and longer, insoluble glutenins [49]. Glutenins are expected to contribute significantly to the elasticity of gluten [49, 50]. In contrast, storage proteins from maize and rice do not exhibit the same extraordinary elastic properties. Here, we present the results of our simulations involving various systems of interest: gluten, its gliadin and glutenin fractions, as well as proteins from rice and maize, revealing another function of empty regions within biological structures.

The research on gluten, similar to the virus capsid case, was conducted within a coarse-grained model. However, the model described in the previous section can be applied to simulate systems with well-defined structures. In this assembly, intrinsically disordered regions are included, making it challenging to use a classic Gō-like model [51]. To analyze this system, we modified the model so that the contact map is no longer based on the native structures from PDB. Instead, the contact map is constructed dynamically based on the geometry of the chain at any given moment. The geometrical criteria for a contact are derived statistically from a large database of contacts. The contacts are turned on and off quasi-adiabatically, reflecting changes in the chain conformation. The model was validated on a set of IDPs and partially ordered proteins [51], demonstrating its ability to simulate not only IDPs but also large clusters of IDPs and partially structured proteins as the Gō-like contact map can coexist with the dynamic one.

The viscoelastic properties of storage proteins depend on their hydration levels [52]. In a coarse-grained model with an implicit solvent [53], water molecules cannot be represented explicitly. Instead, their presence is visualized by cavities within the simulated system, large enough to accommodate at least one water molecule but small enough to remain part of a separate protein cluster. The analysis of the cavities in the simulated systems was conducted using our SPACEBALL algorithm [16], enabling the identification of the number and volume of the cavities.

Our simulations were conducted in several steps. Initially, the simulation box was compressed to achieve the desired protein concentration of 3.5 residues per cubic nanometer, as detailed in [52]. In the next step, the system was equilibrated in preparation for periodic box deformations. The simulation box was deformed by shearing, as illustrated in Fig. 8a, or by pulling, as shown in Fig. 8b. In both cases, the position of two opposing walls changed periodically, back and forth. The wall displacement as a function of time was a sinusoid with an amplitude of 1 nm and an oscillation period of 40 μs. The residues were attracted to the walls with the Lennard–Jones potential [54]. Following five full oscillation cycles, the system underwent the next equilibration. We also performed control simulations with no periodic deformation. Finally, the simulation box with a well-equilibrated system was stretched in one direction to induce the rupture of the protein network. The same pair of box walls continually attracted protein residues, causing them to adhere to these walls and enabling the stretching of the entire system [54].

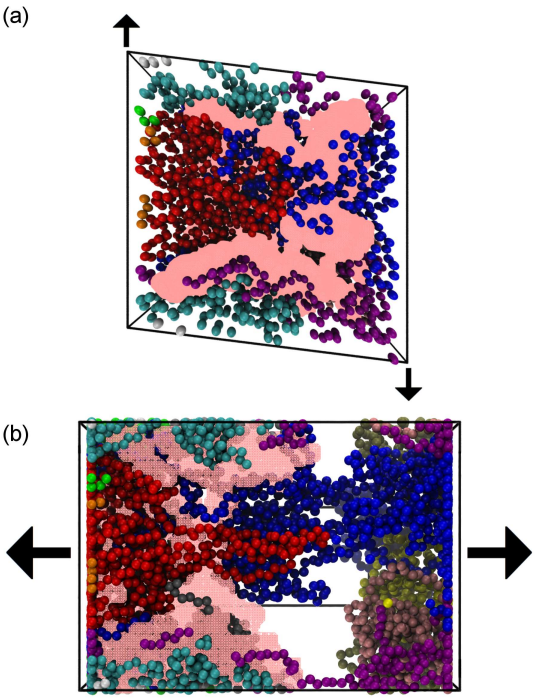


Fig. 8. Examples of gluten systems during shearing (a) and pulling (b) deformations. In both panels, the left and right walls attract residues, while all other walls have periodic boundary conditions. Residues are represented as balls, and cavities are shown in pink.

In our simulations, we observed a reduction in the average volume of the largest cavity during the stretching of all simulated storage proteins [53]. This indicates a decrease in the amount of solvent inside the protein cluster. The result is depicted in Fig. 9 (see also [53]) as the ratio of the volumes of the largest cavity, averaged over the second and first parts of the stretching trajectory. The observation is consistent with experimental findings [55] and may correspond to the stretching-induced release of water molecules [56].

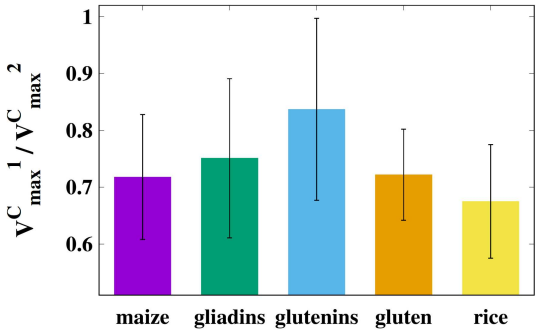


Fig. 9. The ratio of the largest cavity volumes, averaged over the second ($V^{C}_{max}1$) and first ($V^{C}_{max}2$) halves of the stretching trajectory for the five simulated storage protein systems. The data for this graph were obtained from [53]. Both halves of the stretching trajectory were simulated after the periodic mechanical deformation of the sample.

The high uncertainties in Fig. 9 result from volume fluctuations during stretching due to the dynamic nature of this process.

It is important to note that the final stretching is the last stage of the simulation, and the results presented in Fig. 9 only cover the first and second halves of that stage. Other figures display results from an earlier stage of the simulation, namely periodic deformation, where we periodically distorted the simulation box to investigate its effects on the system.

To explore this, we examined the total number of cavities in the systems resulting from either pull or shear periodic deformation. Figure 10a illustrates the ratio of the number of cavities in the systems after and before box deformation in the pulling mode. The decrease in the number of cavities suggests that the pulling mode facilitates the merging of smaller cavities into larger ones during the elongation step. On the other hand, in the shearing mode, as depicted in Fig. 10b, the number of cavities seems to slightly increase only for gluten. However, this change is still within the error bar, assessed at 20%, indicating that the shearing mode does not seem to lead to the merging of smaller cavities into larger ones.

This observation is confirmed by the results pertaining to the volume of the largest cavity. Its average value increases significantly in the pulling mode, while it remains relatively stable in the shearing mode, as presented in Fig. 11. On the other hand, when considering the total volume of cavities in the examined systems, a slight increase is observed in the pulling mode, as depicted in Fig. 12a. This increase is primarily due to the merging of small cavities, too small to contain a water molecule, into larger ones. Consequently, the average total cavity volume is higher. In contrast, the situation remains stable in the case of the shearing mode, as presented in Fig. 12b.

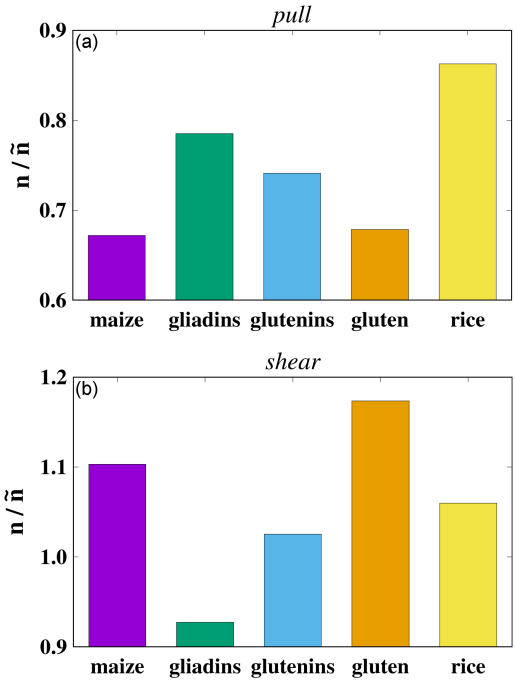


Fig. 10. The ratio of the number of cavities before ($\tilde{n}$) and after (n) periodic deformation for the 5 simulated systems in the pulling mode (a) and the shearing mode (b). The uncertainty is approximately ± 0.1 for pulling and ± 0.2 for shearing.

The presented results show more pronounced changes in cavity properties after periodic deformation in the pulling mode (normal stress) compared to the shearing mode. This aligns with the "loops and trains" theory [57], which predicts that after elongation, proteins form "trains" composed of parallel chains connected by hydrogen bonds, leading to the expulsion of water from the system. In the undeformed state, proteins form loops that are partially solvent-exposed and establish hydrogen bonds with water (all plant storage proteins contain high amounts of hydrophilic residues [49, 58, 59]). Closing these loops by elongating them in one direction acts as a kinetic trap, compelling the proteins to remain in the "train" state even when they are no longer deformed and no stress is applied. This phenomenon explains the lower number of cavities, the higher volume of the largest cavity (due to water expulsion), and minimal change in the total cavity volume — outcomes expected in an explicit solvent simulation but not necessarily in an implicit solvent one. In the shearing mode, "trains" are not formed, and the aforementioned process does not occur, resulting in much smaller changes.

While the differences between the studied systems were smaller than those arising from the deformation mode, it is noteworthy that the number of cavities before and after deformation changed the most in the gluten system (after pulling deformation, it was the lowest of all, and after shearing, it was the highest), making gluten much more

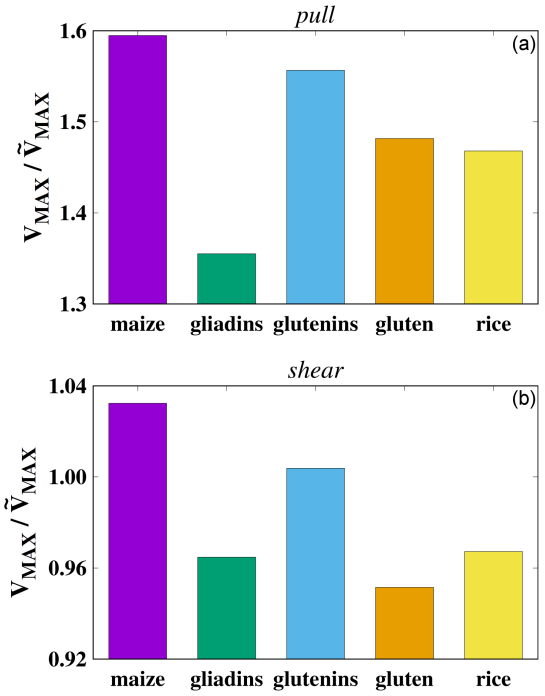


Fig. 11. The ratio of the volumes of the largest cavity after (V_{MAX}) and before ($\tilde{V}_{\mathrm{MAX}}$) periodic deformation for the five simulated storage protein systems. The error bars are approximately ± 0.12 for the pulling mode (a) and ± 0.07 for the shearing mode (b).

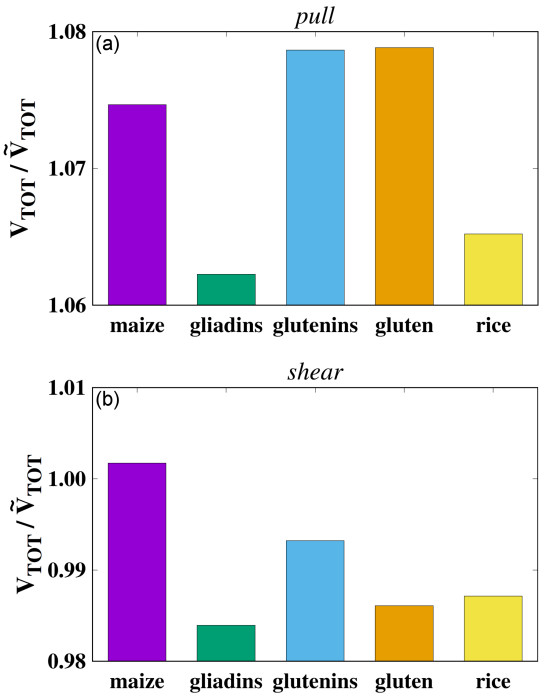


Fig. 12. The ratio of the total volumes of all cavities in the system after (V_{TOT}) and before ($\tilde{V}_{\mathrm{TOT}}$) periodic deformation for the five simulated storage protein systems. The error bars are approximately ± 0.11 for the pulling mode (a) and ± 0.18 for the shearing mode (b).

susceptible to changes than, for example, rice storage proteins. This susceptibility may stem from the fact that the gluten protein network solidifies after deformation, forming more hydrogen bonds [50, 57]. Cavities act as "holes" in that network, allowing adaptation to large deformations [57].

5. Conclusions

We have surveyed various structures containing cavities, detecting and calculating their volumes. With support from literature data, we described the functions of empty spaces within the considered structures. We concluded that, in the case of PR-10 proteins, their interior can serve as a space for ligands specific to a particular protein, consequently allowing such proteins to act as selectors for ligands. Using the examples of haloalkane dehalogenase and cholesterol oxidase II, we demonstrated that cavities provide an environment for enzymatic reactions. Through the example of GlpG, we discussed the role of cavities in regulating protein activity. Examining the StPurL protein, we illustrated that cavities are crucial in allosteric regulation. Finally, based on the analysis of lysozyme structures from different species, we asserted that cavities play a pivotal role in the catalytic cycle of lysozyme. Our calculations supported the notion that cavities are evolutionarily conserved elements in protein structure.

Analyzing the mechanostability and structure of the cowpea chlorotic mottle virus led us to the hypothesis that the large ratio of empty space to the space occupied by the genetic material within the virus increases the opportunity for the virus to break and release the genetic material.

For the simulated plant storage protein systems, we correlated changes in the size and number of cavities after periodic deformation with the "loops and trains" theory [57]. This enabled us to demonstrate that, even in an implicit solvent model, cavities can serve as a measure of the hydration level [52], and processes like solvent expulsion can be simulated using only geometric constraints [54]. The variations in the number of cavities also contributed to confirming the unique nature of the viscoelastic gluten protein network [53].

In all the discussed phenomena, cavities within protein systems play a pivotal role. The research initiated by Professor Cieplak in 2013 allowed us to uncover this role and may shed more light on many other systems in the future.

Acknowledgments

The authors acknowledge Professor Marek Cieplak for encouraging the investigation of structures containing cavities, knots, and various unstructured systems, explored using coarse-grained models.

This research has received support from the National Science Centre (NCN), Poland, under grant No. 2018/31/B/NZ1/00047 and the European H2020 FETOPEN-RIA-2019-01 grant PathoGelTrap No. 899616. The computer resources were supported by the PL-GRID infrastructure.

References

[1] J. Janin, R.P. Bahadur, P. Chakrabarti, *Q. Rev. Biophys.* **41**, 133 (2008).

[2] M. Schneider, X. Fu, A.E. Keating, *Proteins* **77**, 97 (2009).

[3] M. Jamroz, W. Niemyska, E.J. Rawdon, A. Stasiak, K.C. Millett, P. Sulkowski, J.I. Sulkowska, *Nucleic Acids Res.* **43**, D306 (2015).

[4] M. Chwastyk, M. Cieplak, *J. Phys. Chem. B* **124**, 11 (2020).

[5] P. Dabrowski-Tumanski, A.I. Jarmolinska, W. Niemyska, E.J. Rawdon, K.C. Millett, J.I. Sulkowska, *Nucleic Acids Res.* **45**, D243 (2017).

[6] M. Kroger, J.D. Dietz, R.S. Hoy, C. Luap, *Comput. Phys. Commun.* **283**, 108567 (2023).

[7] Y.I. Zhao, M. Chwastyk, M. Cieplak, *J. Chem. Phys.* **146**, 225102 (2017).

[8] M. Chwastyk, E.A. Panek, J. Malinowski, M. Jaskólski, M. Cieplak, *Front. Mol. Biosci.* **7**, (2020).

[9] M. Chwastyk, M. Cieplak, *Front. Mol. Biosci.* **8**, 692230 (2021).

[10] H. Fernandes, K. Michalska, M. Sikorski, M. Jaskolski, *FEBS J.* **280**, 1169 (2013).

[11] H. Fernandes, O. Pasternak, G. Bujacz, A. Bujacz, M.M. Sikorski, M. Jaskolski, *J. Mol. Biol.* **378**, 1040 (2008).

[12] H. Fernandes, A. Bujacz, G. Bujacz, F. Jelen, M. Jasinski, P. Kachlicki, J. Otlewski, M. Sikorski, M. Jaskolski, *FEBS J.* **276**, 1596 (2009).

[13] M. Gajhede, P. Osmark, F.M. Poulsen, H. Ipsen, J.N. Larsen, R.J.J. van Neerven, C. Schou, H. Lowenstein, M.D. Spangfort, *Nat. Struct. Biol.* **3**, 1040 (1996).

[14] J. Biesiadka, G. Bujacz, M.M. Sikorski, M. Jaskolski, *J. Mol. Biol.* **319**, 1223 (2002).

[15] M. Chwastyk, M. Jaskolski, M. Cieplak, *FEBS J.* **281**, 416 (2014).

[16] M. Chwastyk, M. Jaskolski, M. Cieplak, *Proteins* **84**, 1275 (2016).

[17] J. Tsai, R. Taylor, C. Chothia, M. Gerstein, *J. Mol. Biol.* **290**, 253 (1999).

[18] O. Pasternak, J. Biesiadka, R. Dolot, L. Handschuh, G. Bujacz, M.M. Sikorski, M. Jaskolski, *Acta Cryst.* **D61**, 99 (2005).

[19] S. Führer, J. Unterhauser, R. Zeindl, R. Eidelpes, M.L. Fernández-Quintero, K.R. Liedl, M. Tollinger, *Int. J. Mol. Sci.* **23**, 8252 (2022).

[20] H. Fernandes, K. Michalska, M. Sikorski, M. Jaskolski, *FEBS J.* **280**, 1169 (2013).

[21] N.L. Dawson, T.E. Lewis, S. Das, J.G. Lees, D. Lee, P. Ashford, C.A. Orengo, I. Sillitoe, *Nucl. Acids Res.* **45**, D289 (2016).

[22] K.H.G. Verschueren, F. Seljee, H.J. Rozeboom, K.H. Kalk, B.W. Dijkstra, *Nature* **363**, 693 (1993).

[23] R. Coulombe, K.Q. Yue, S. Ghisla, A. Vrielink, *J. Biol. Chem.* **276**, 30435 (2001).

[24] R. Guoa, Z. Cangb, J. Yaoa, M. Kima, E. Deansc, G. Weib, S. Kangd, H. Hong, *PNAS* **117**, 22146 (2020).

[25] Y. Wang, Y. Zhang, Y. Ha, *Nature* **444**, 179 (2006).

[26] A.S. Tanwar, V.D. Goyal, D. Choudhary, S. Panjikar, R. Anand, *PLoS ONE* **8** e77781 (2013).

[27] H. Li, Y.O. Kamatari, *Subcell Biochem.* **72**, 237 (2015).

[28] K. Harata, M. Muraki, *Acta Cryst.* **D53**, 650 (1997).

[29] T. Koshiba, M. Yao, Y. Kobashigawa, M. Demura, A. Nakagawa, I. Tanaka, K. Kuwajima, K. Nitta, *Biochemistry* **39**, 3248 (2000).

[30] M. Ramanadham, L.C. Sieker, L.H. Jensen, *Acta Cryst.* **B46**, 63 (1990).

[31] M. Muraki, K. Harata, N. Sugita, K. Sato, *Biochemistry* **35**, 13562 (1996).

[32] B. Kiss, D. Mudra, G. Török, Z. Martonfalvi, G. Csik, L. Herenyi, M. Kellermayer, *Biophys. Rev.* **12**, 1141 (2020).

[33] A. Zolochevsky, S. Parkhomenko, A. Martynenko, Quantum, molecular and continuum modeling in nonlinear mechanics of viruses. *V.N. Karazin Kharkiv Nat. Univ. Ser. Med.* **44**, (2022).

[34] M. Cieplak, M. Robbins, *J. Chem. Phys.* **132**, 015101 (2010).

[35] M. Cieplak, in: *Computational Methods to Study the Structure and Dynamics of Biomolecules and Biomolecular Processes*, Ed. A. Liwo, Springer Series on Bio- and Neurosystems, vol 8. Springer, Cham 2019.

[36] M. Bravo, L.J. Gomez, J. Hernandez-Rojas, *Soft Matter* **16**, 3443 (2020).

[37] M. Medrano, A. Valbuena, A. Huete, M. Mateu, *Nanoscale* **11**, 9369 (2019).

[38] M. Aznar, S. Roca-Bonet, D. Reguera, *J. Phys. Condens. Matter* **30**, 264001 (2018).

[39] K. Wolek, M. Cieplak, *J. Condens. Matter Phys.* **29**, 474003 (2017).

[40] R. Konecny, J. Trylska, F. Tama, D. Zhang, N. Baker, C. Brooks III, A. McCammon, *Biopolymers* **82**, 106 (2006).

[41] M. Cieplak, T.X. Hoang, M.O. Robbins, *Proteins* **49**, 114 (2002).

[42] M. Cieplak, M.O. Robbins, *PLOS ONE* **8**, e63640 (2013).

[43] M. Chwastyk, A.P. Bernaola, M. Cieplak, *Phys. Biol.* **12**, 046002 (2015).

[44] M. Chwastyk, A. Galera-Prat, M. Sikora, Á. Gómez-Sicilia, M. Carrión-Vázquez, M. Cieplak, *Proteins* **82**, 717 (2014).

[45] M. Gunnoo, P.-A. Cazade, A. Orlowski, M. Chwastyk, H. Liu, D.T. Ta, M. Cieplak, M. Nashde, D. Thompson, *Phys. Chem. Chem. Phys.* **20**, 22674 (2018).

[46] Y.N. Zhao, M. Chwastyk, M. Cieplak, *Sci. Rep.* **7**, 39851 (2017).

[47] E. Tarasova, I. Korotkin, V. Farafonov, S. Karabasov, D. Nerukh, *J. Mol. Liq.* **245**, 109 (2017).

[48] H.V. Chaudhari, M.M. Inamdar, K. Kondabagil, *iScience* **24**, 102452 (2021).

[49] H. Wieser, Food Microbiol. **24**, 115 (2007).

[50] D.N. Abang Zaidel, N.L. Chin, Y.A. Yusof, *J. Appl. Sci.* **10**, 2478 (2010).

[51] Ł. Mioduszewski, M. Cieplak, *Phys. Chem. Chem. Phys.* **20**, 19057 (2018).

[52] Ł. Mioduszewski, *Eur. Biophys. J.* **52**, 583 (2023).

[53] Ł. Mioduszewski, M. Cieplak, *PLOS Comp. Biol.* **17**, e1008840 (2021).

[54] Ł. Mioduszewski, M. Cieplak, *Tribol. Lett.* **69**, 60 (2021).

[55] P. Belton, I. Colquhoun, A. Grant, N. Wellner, J. Field, P. Shewry, A. Tatham, *Int. J. Biol. Macromol.* **17**, 74 (1995).

[56] S. Liese, M. Gensler, S. Krysiak, R. Schwarzl, A. Achazi, B. Paulus, T. Hugel, J.P. Rabe, R.R. Netz, *ACS Nano.* **11**, 702 (2017).

[57] H. Singh, F. MacRitchie, *J. Cereal Sci.* **33**, 231 (2001).

[58] Y. Wu, J. Messing, *Front. Plant Sci.* **5**, 240 (2014).

[59] P. Chen, Z. Shen, L. Ming et al., *Front. Plant Sci.* **9**, 612 (2018).

In memory of Professor Marek Cieplak

REVIEW

Protein Dynamics in Tight Tunnels

M. Wojciechowski and M. Chwastyk*

Institute of Physics, Polish Academy of Sciences, al. Lotników 32/46, PL-02668 Warsaw, Poland

Doi: 10.12693/APhysPolA.145.S61 *e-mail: chwastyk@ifpan.edu.pl

We investigate the impact of narrow tunnels, such as the ribosomal exit tunnel and the entrance of the proteasome channel, on the dynamics of proteins with and without knots. Our exploration delves into the potential driving forces behind protein chain movement and their individual significance. Furthermore, within the framework of protein degradation facilitated by the proteasome, we analyze how the presence of knots influences the protein's entry into the proteasome chamber through diverse approaches. This discussion illustrates how molecular dynamics simulations within a coarse-grained structure-based model provide valuable insights into these intricate molecular processes.

topics: molecular dynamics simulations, knots, ribosome, proteasome

1. Introduction

Proteins serve as the workhorses of biology, participating in virtually every aspect of life, from fundamental cellular processes to complex physiological functions. Their versatility and diversity make them integral to the functioning of living organisms. Within a cell, the creation and degradation of proteins are crucial processes. The former involves the extrusion of proteins from the ribosomal exit tunnel, while the latter is related to the translocation of proteins into the chamber within the proteasome, a complex structure containing ATPases associated with diverse cellular activities (AAA+ proteases). Protein degradation plays a pivotal role in maintaining cellular homeostasis by breaking down damaged or unwanted conformations [1–5]. The structure and function of these proteins are tightly regulated to ensure proper turnover and overall cellular function. The crystallographic structures of the ribosome (PDB: 5XY3) [6] and the proteasome (PDB: 7QO3) [7], shown in Fig. 1 (see [8–10]), are provided by the Protein Data Bank (PDB).

Among all proteins, there is a special category that contains knotted conformations in their native state. This topology is not very common, as less than 2% of known structures from the human proteome are knotted [11]. It is extremely important to conduct research on them because knots in proteins have functional significance, and understanding the knotting process can provide insights into various biological processes [12]. Moreover, knots can influence the stability and folding kinetics of proteins, so knowledge about them can be applied to design more stable structures or predict the folding behavior of novel protein sequences [13]. Finally, a very important aspect of examining knotted proteins is their role in neurodegenerative diseases such as Alzheimer's, Parkinson's, and Huntington's diseases, where misfolded proteins aggregate and form insoluble deposits in the brain [14]. Knots in proteins may affect their propensity to aggregate and the structure of the resulting aggregates, potentially influencing disease progression. Therefore, part of our research on protein degradation is devoted to this not very common but super important group of structures.

2. CG Model of ribosome and proteasome

Computer simulations of the processes mentioned above can be carried out using either all-atom or coarse-grained (CG) approaches. Due to the significant conformational changes involved in protein creation and degradation, simulating them at the all-atom level can be challenging. Therefore, we employ a model developed by Professor Marek Cieplak's group over many years [15–17]. In this model, the protein is represented using a structure-based approach, as described in references [15, 18–21], with a chirality potential responsible for maintaining backbone stiffness. Each amino acid is depicted by a single bead positioned at the C_α position, and interactions between beads are based on whether the residues belong to the contact map or not. If two residues are on the contact map, the Lennard-Jones potential is applied to them, with the well depth denoted as ϵ. The contact map is determined using the overlap criterion between all atoms of residues, as observed in the fully folded native state. The characteristic length of these interactions corresponds to distances from the same state. Interactions between the remaining residues consist only of the repulsive

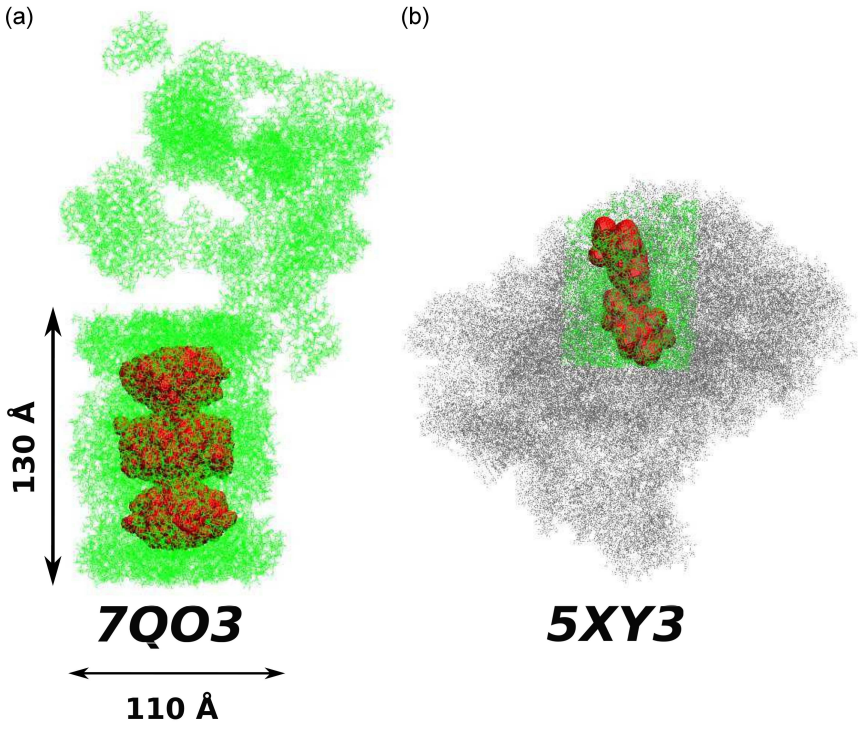


Fig. 1. Representation of the proteasome (panel (a)) and the ribosome (panel (b)) based on PDB crystallographic structures 7QO3 and 5XY3, respectively. In the proteasome, the green color denotes all heavy atoms of its structure. For the ribosome, heavy atoms are colored gray or green, with green representing only the atoms considered in our simulation. Red indicates cavities within both structures detected using the SPACEBALL server [8–10]. Both structures are presented at the same scale.

part of the Lennard-Jones potential, with a characteristic length of 4 Å, beyond which the potential is turned off. Temperature control and the influence of the solvent are introduced by the Langevin thermostat, with the room temperature set at $\approx 0.35\epsilon/k_{\mathrm{B}}$.

The ribosome is a macromolecular machine within a cell responsible for protein synthesis during the process of messenger RNA (mRNA) translation. Ribosomes facilitate the linkage of amino acids in a specific order as directed by the codons present in mRNA molecules. In eukaryotic cells, such as those found in humans, ribosomes are composed of two primary subunits: the small ribosomal subunit (40S) and the large ribosomal subunit (60S). These subunits consist of 2–6 RNA chains and approximately 50 proteins, totaling between 100 000 and 220 000 atoms [22–24]. Within both subunits, ribosomal RNA (rRNA) molecules provide the framework for the ribosome structure and play essential roles in catalyzing the chemical reactions involved in protein synthesis. Protein synthesis occurs at the peptidyl transferase center (PTC), and the newly created chain is directed toward the ribosomal exit tunnel. The size and geometry of this tunnel significantly depend on different domains of life, including bacteria, archaea, and eukarya [25], as well as the specific organism within each domain. The inner walls of the tunnel are rough and highly irregular, featuring several constriction sites. The narrowest constriction, with a radius of approximately 8 Å, is observed in eukaryotes, followed by a slightly wider one, of around 11 Å in radius, in archaeal ribosomes, and even wider in the bacterial case, with an average radius of approximately 15 Å. Due to the high complexity of the ribosome, we simplify its structure to only the cylinder with a radius of 70 Å containing the ribosomal exit tunnel aligned with its longitudinal axis, as illustrated in Fig. 1b, where all heavy atoms of the ribosome are colored gray. The bottom of the cylinder is positioned at PTC, and its length extends to encompass the farthest atoms from PTC. The tunnel under consideration is marked in green. Since the created cylinder is composed of 11 680 atoms, it is still a very complex system, and simulations of all its atoms would be highly inefficient. In this research, we focused on the impact of confinement inside the cylinder, making the interactions between its atoms less critical. Thus, we simplified our system by considering the ribosome structure as rigid. The interactions between the protein and the atoms of the tunnel are purely repulsive, as each ribosomal atom contributes to a soft repulsive potential, truncated at 4 Å, with an amplitude of ϵ. The bottom of the confinement cylinder is modeled as a repulsive wall characterized by the potential $\frac{3\sqrt{3}}{2}\epsilon\,(\sigma_0/z)^9$, where z signifies the distance from the plate and $\sigma_0 = 4 \times 2^{-1/6}$. This wall prevents any backward steps.

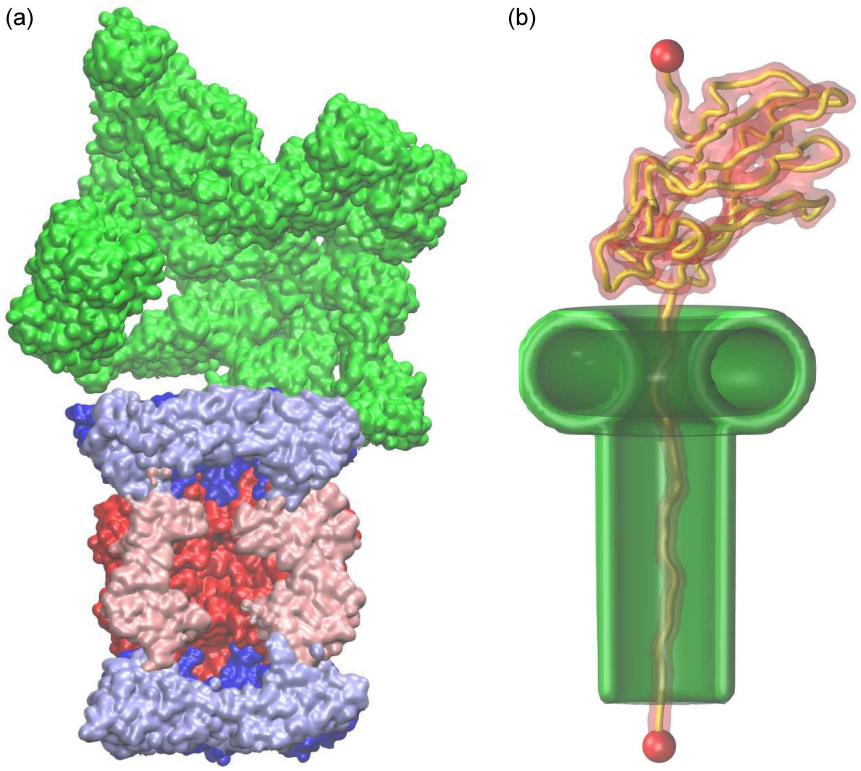


Fig. 2. The cross-section of the crystal structure of the 26S proteasome (PDB: 7QO3) and its theoretical model. The structures' parts closer to the reader are omitted to reveal inner chambers in both cases. In panel (a), the β-rings are colored red or pink, and the α-rings are blue or cyan, with the 19S cap in green. Panel (b) depicts a schematic representation of the entrance to the proteasome. A dragged protein must fit into a hole with a diameter of approximately 14 Å, represented by the torus. Below the torus, there is a straight pipe that keeps the chain unfolded. The entrance and pipe are depicted in green, while the protein is shown in yellow and red tube representation. The N- and C-ends of the chain are represented as red spheres. The pulling force is parallel to the pipe and is attached to the protein terminus (at the bottom of the picture).

On the other hand, we have the proteasome, a highly complicated and well-organized protein complex, as illustrated in Fig. 1a. It consists of multiple subunits arranged into a cylindrical structure resembling a barrel with an open entrance and exit. Protein degradation occurs within the inner chamber of this barrel-like structure. In eukaryotic cells, such as those in humans, the proteasome typically has a 26S structure, as depicted in Fig. 2. This 26S proteasome consists of two main components: the 20S core particle (CP) and the 19S regulatory particle (RP). The 19S regulatory particle recognizes proteins tagged for degradation with ubiquitin molecules, unfolds them, and translocates these target proteins into the central channel of the 20S core particle. Within the complex chamber, the target protein encounters proteolytic active sites that cleave it into smaller peptide fragments. These remaining peptide fragments are subsequently processed and released from the proteasome for recycling or presentation as antigens in immune responses. In our simulations, the presence of the proteasome structure is not directly modeled by the potentials of its individual atoms. Instead, we employ a highly simplified model consisting of a torus as the source of a continuous repulsive part of the Lennard-Jones potential, defined on its surface with a characteristic length of 6 Å. The major radius of the torus, $R_t = 13$ Å, and the minor radius, $r_t = 6$ Å, result in an entrance diameter of ≈ 14 Å, while the diameter of the entrance in a proteasomal crystal structure is 13 Å [26–28]. We enlarged it to accommodate the flexibility of the hole. Below the torus, we introduce a narrow tunnel, but its presence does not originate from the proteasome shape. The tunnel is used solely for technical reasons and prevents the elongated chain from refolding since we do not simulate the degradation process but only protein pulling through the entrance. In our simulations, the chain inside the tunnel is considered degraded, and it no longer impacts the simulation.

Despite the fundamentally different biophysics of protein creation and degradation, one aspect of these processes is shared — the protein must pass through a narrow tunnel or entrance with rugged walls. In the case of the ribosome, this occurs as the protein is pushed out through the ribosomal exit tunnel, while with the proteasome, the chain is pulled inside with the assistance of adenosine triphosphate (ATP) [1, 5].

In computer simulations, the creation of a protein in a ribosome is typically implemented by placing a fully synthesized chain in proximity to the peptidyl transferase center (PTC) and then monitoring the folding process [29–32]. Simulations often model the protein's exit from the ribosome using steered molecular dynamics with a constant pulling speed applied to the N-terminus [29]. Alternatively, some simulations apply a constant force to more accurately replicate the natural process [30]. In our simulations, we demonstrate that the protein chain can exit the ribosomal tunnel without external manipulation. We employ a sequential growth method, where each amino acid emerges at specific time intervals after the previous one is synthesized. As mRNA translation proceeds from the 5' to 3' ends, proteins are synthesized from the N-terminus to the C-terminus, causing the N-terminus to emerge first. Our approach resembles the one described by P.T. Bui and T.X. Hoang [33], with the distinction that our repulsive potential accounts for all heavy atoms of the ribosome part considered in simulations, and any backward motion is prevented by repulsion from the lower tunnel wall. Additionally, in our case, the growth process is implemented in a quasi-continuous manner.

Protein degradation, on the other hand, involves pulling the chain into the proteasome, and our focus is exclusively on this pulling process. Unlike the protein dynamics in the ribosomal exit tunnel, achieving this requires the application of a dragging force. The simplest approach is to drag the chain into the proteasome and through the torus hole at a constant speed. While this approach is not realistic, it serves to illustrate the difference between AFM-like protein stretching (AFM — atomic force microscopy) [18, 19] and proteasome-assisted unfolding. To implement a more realistic scheme, we start pulling with a constant force, which can be implemented either continuously or periodically. In the latter case, the force is applied for a specified time, approximately 4.5 μs, to either the C- or N-terminal chain ends, depending on the pulling scenario, and then turned off for an equal amount of time, allowing the protein to retract. This reflects the generation of force upon delivery of ATP (non-continuous). The final scheme represents a "ratchet-like mechanism" involving pulling for a maximum of 4.5 μs (or pulling 1 nm of chain), followed by blocking the protein's retraction during the absence of force. This mechanism is supported by biological proteasome action, where certain parts of the proteasome undergo bending upon ATP delivery, generating the force capable of dragging a small portion of the chain [34]. Subsequently, the protein chain is prevented from retracting and awaits the next ATP cycle, making periodic force application closely mimic biological evidence.

The aforementioned simplified description of on-ribosome protein creation and protein degradation by the proteasome could be improved by using more complex and detailed approaches. However, in our research, we deal with large conformational changes in proteins that occur within seconds, making simulations of such processes very expensive and time-consuming. The advantage of simulations within a coarse-grained model is that, thanks to their efficiency, a large number of different trajectories for a given process can be conducted, and the final result is obtained by averaging the considered quantities.

3. On-ribosome protein folding

We initiate simulations of protein chain creation by the ribosome with a single residue (the N-terminus) placed at the PTC. The next residue appears within a specified time, denoted as t_w, referred to as the waiting time. Determining its value is not straightforward, but in our previous work [35], we demonstrated that a waiting time of 5 000 τ is sufficient, and times longer than this value do not significantly affect the outcomes. Here, τ is equal to 1 ns, which is the characteristic time in our model. This means that in our simulations, protein synthesis occurs $\approx$ 4–5 orders of magnitude faster compared to biological systems. Based on the previous results [35], we assume that this waiting time value is sufficient, and thanks to the acceleration of adding new residues, our simulations are much more efficient, allowing us to examine large conformational changes in the newly created protein. As mentioned in Sect. 2, the process of the protein leaving the ribosomal exit tunnel can be implemented in various ways. In our model and simulations, our intention was to closely simulate the biological process. This means that there is no external force aiding the squeezing of the protein N-terminus through the rough channel. The rigid walls are also unhelpful in this process. The only forces acting on the N-terminus of the newly created protein facilitating its travel to the ribosome exit are the repulsive forces from neighboring protein residues, the repulsive forces from the bottom wall, and the atoms comprising the ribosome structure. In the case of smooth walls, the protein chain would easily slide toward the ribosomal exit tunnel. The walls of the real ribosomal exit tunnel are rough, and the movement of the first residue is not easy because it can become jammed in the alleys or nooks of the walls. Such a situation is presented in panel (b) in Fig. 3, which shows the distance of the N- and C-terminus from the PTC in function of time. The graph illustrates that the N-terminus became jammed $\approx$ 60 Å from the PTC, and all of the protein's residues became crowded into the space between this point and the PTC. This scenario is unrealistic because proteins do not typically jam during their creation in real situations, and it can be considered an artifact of the model. Nevertheless, this observation highlights the significant impact of irregularities in the exit tunnel walls on protein dynamics.

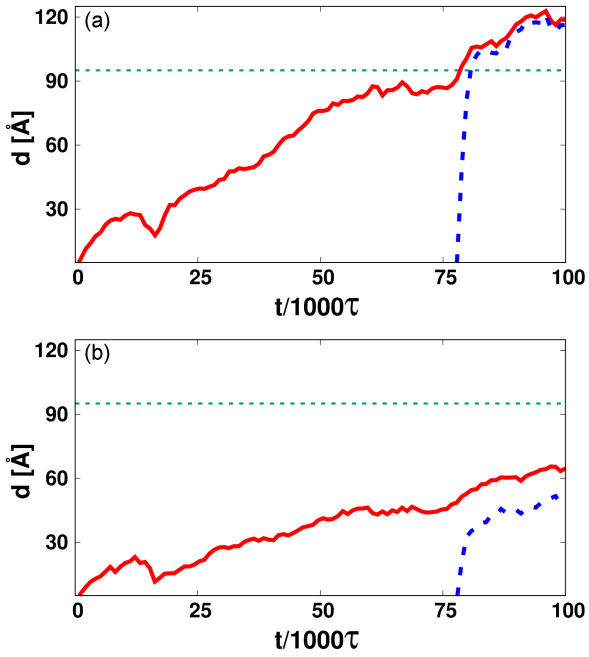


Fig. 3. The distances between the PTC and the N-terminus of the 1J85 protein chain (solid red line) or its C-terminus (dotted blue line) plotted as a function of time during the simulation within the ribosomal exit tunnel. Panel (a) illustrates the scenario where the N-terminus has reached the exit from the ribosomal structure, denoted by the green dotted line, while dragging the C-terminus outward. Panel (b) depicts the situation where the N-terminus was jammed within the tunnel. In both cases, the simulations were conducted at room temperature with a waiting time of $t_w = 100$ ns.

Most often, when the t_w is long enough, the N-terminus can find its way to reach the exit from the ribosome, as illustrated in Fig. 3a. Here arises a question regarding the mechanisms that play a crucial role in protein dynamics within the ribosomal exit tunnel. In our previous research [25], we proposed that the protein's movement toward the outside of the tunnel starts at the PTC and is primarily influenced by diffusion, interactions with the tunnel walls, and the increase in entropy associated with the escape. However, we did not investigate which of these mechanisms is crucial. By looking at Fig. 3a, we see that there is a rapid increase in the N-terminus' distance from the PTC when it reaches the top of the tunnel, denoted by the green dotted line. Above the outer surface of the ribosome, the protein began to fold, native contacts began to form, and there was a decrease in the protein's potential energy. As a consequence, the C-terminus of the created protein started to be pulled by the folded protein and traveled very quickly, without any jamming, outside the ribosome. This observation suggests that the change in potential energy is crucial for the protein's exit from the tunnel.

4. Protein degradation by the proteasome

Simulating protein degradation within the proteasome presents several challenges. As the proteasome employs ATP to pull proteins periodically and restricts backward motions, the most realistic simulation scenario involves periodic force pulling with a "ratchet-like mechanism" [34]. Here, we examine several pulling scenarios, ranging from the simplest to the most advanced, to assess pulling efficiency in each case.

In our study of protein behavior within the proteasome, we have considered three primary methods of protein stretching. The first method, referred to as "type AFM," involves standard AFM stretching and serves as a template and reference for proteasome-aided unfolding. In the second method, we pull the protein by its ends in the presence of a proteasome model, either at the C-end or N-end, while the other end remains free. This method of pulling is called "type I." Considering stretching from both termini is justified by the fact that the degradation mechanism is preceded by the protein marking with special tags, which can be attached to either the C- or N-terminal chain ends. The last type of pulling, referred to as "type II," involves pulling by one terminus while the other remains attached to its original position. All of these pulling types can be realized by different scenarios. The simplest one is constant speed pulling, in which we pull the protein end at a constant speed. This type of protein pulling generates a characteristic AFM-like force graph with peaks, as depicted in the top parts of each panel in Fig. 4. Each peak on the force-displacement graph corresponds to the breaking of contacts within a specific part of the protein structure and represents the protein's resistance to mechanical stress. The most resilient part of the structure yields the highest force peak, and its height, denoted as $F_{\max}$, can be used to characterize the protein mechanostability. Both types of proteasome-assisted pulling with constant speed are comparable to standard AFM stretching experiments and, as demonstrated in our previous research [36], they facilitate the pulling process. Lower forces were measured in most of the proteins considered. However, in the case of barnase (PDB: 1BNR) and titin (PDB: 1TIT), slightly higher force values were observed when pulling from the C- or N-termini compared to standard AFM stretching. This suggests that the presence of the torus facilitates the breaking of contacts and the unfolding of the protein chain, but in our simulations, we also observed the impact of the torus presence on knotted protein dynamics and their degradation.

Since the entrance to the proteasome is narrower than the average width of a knot in a knotted protein, the knot can block the entrance, making the proteasome useless [36]. However, in our simulations, we have demonstrated that the presence of the proteasome can untie deep knots, which are

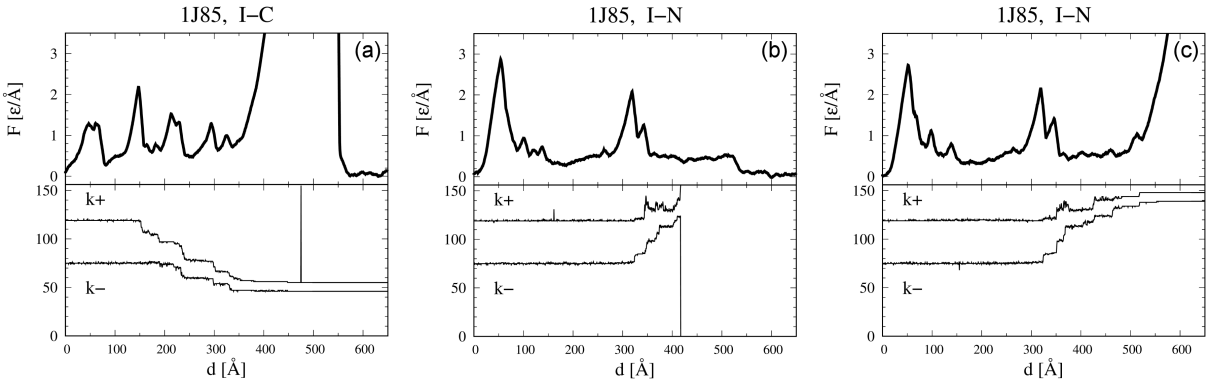


Fig. 4. The simulation results of the deeply knotted protein 1J85 stretching. In both AFM-like stretching and "type II" pulling, the knot almost always tightens. Panel (b) illustrates "type I" pulling, where the knot may slide along the chain and become untightened. Alternatively, the knot may slide down and tighten, similar to AFM stretching, as depicted in panels (a) and (c). The top part of each panel displays force versus distance, while the bottom ones show the position of the knot in the sequence. The knot is considered untied when its position along the protein chain is one or it equals the length of the chain.

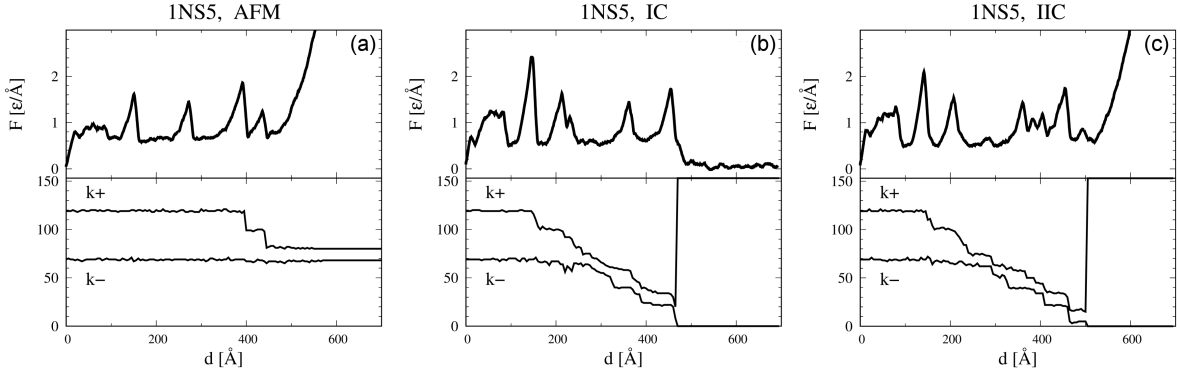


Fig. 5. This figure is similar to Fig. 4. All panels depict the simulation results of the deeply knotted protein 1NS5. Here, we observe knot untying only for C-end pulling (panels (b) and (c)). The N-end simulations only show knot tightening (panel (a)).

defined as knots where the termini are distant from the protein ends [37]. In our research, we considered two deeply knotted proteins: YibK methyltransferase from *Haemophilus influenzae* (PDB: 1J85) and YBEA from *E. coli* (PDB: 1NS5). In the case of the 1J85 protein, the ends of the knot, which consists of 156 residues, are located at residues 75 and 120 in the protein's native state. For 1NS5 (153 residues), the positions of the knot are at residues 69 and 119.

Figure 4 displays the results of our simulations of the degradation of the 1J85 protein with a knot. In this case, we observe that the AFM experiment leads to the tightening of the knot ends, as expected. Knot untying was observed only for "type I-N" simulations, and one trajectory resulting in this behavior is presented in Fig. 4b. Since the protein chain is not smooth, geometric constraints can arise, allowing knot tightening. We observed such trajectories as well, as shown in panels (a) and (c) in Fig. 4. Panel (a) illustrates how the knot blocks the entrance, but if the dragging force is very large, the knot may be pushed through the entrance. However, this is non-physical, as the value of the force is unrealistic ($F \sim 17\epsilon$ Å^{-1}). This occurrence was not observed in every trajectory; in some trajectories, the knot remained at the entrance even at larger forces ($F > 20\epsilon$ Å^{-1}). For other types of pulling, we observe only jamming.

The same observations were made in the case of another deeply knotted protein, 1NS5. The results of the AFM stretching of this protein are presented in Fig. 5a. The outcomes of the stretching using the method called "type I" are shown in Fig. 5b. Interestingly, when stretching the 1NS5 protein using the "type II" method, we also observed knot untying. In this method, we grasp the opposite end of the protein, making knot untying less straightforward. In the simplest scheme, the opposite end of the protein fluctuates during pulling and may accidentally pass out from the loop, thus untying the knot. However, in "type II," such a situation is not possible. Therefore, the only method to untie the knot is to move the entire loop to pass around the immobilized end.

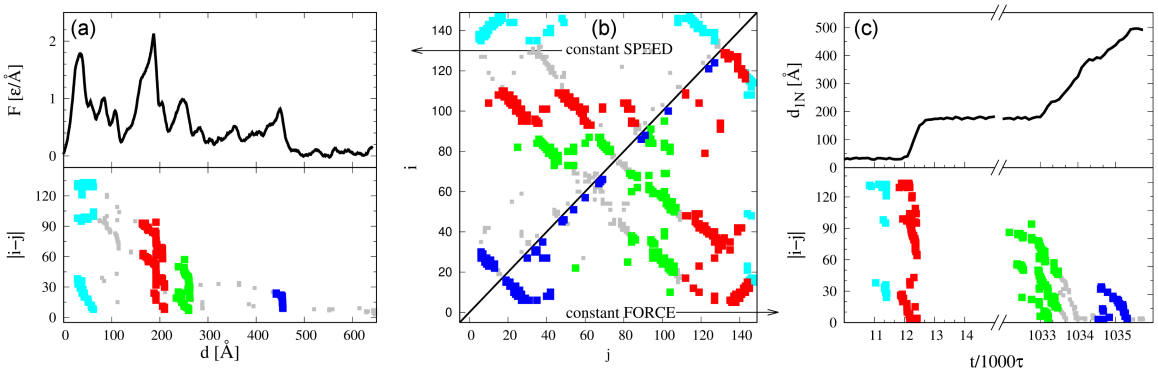


Fig. 6. The simulation results of protein 1AOH. Panel (a) presents an example dataset of constant speed pulling ("type I-C"). The protein is pulled into the proteasome by one end at a constant speed, resisting dragging with the depicted force. The largest force peak reaches approximately $\sim 2.3\epsilon$ Å^{-1}. The top part of panel (a) displays the force graph as a function of distance. Panel (c) illustrates an example of constant force pulling ($F = 1.6\epsilon$ Å^{-1}) ("type I-C"). The protein is pulled into the proteasome with a constant force while monitoring its distance from the proteasome entrance. The top part of this panel displays the distance graph as a function of time. Panel (b) presents the contact map (amino acid i versus j) for both simulation types. The breakings of specific contacts during the simulation of protein stretching with constant speed (panel (a)) and constant force (panel (c)) are indicated at the bottom of both panels by different colors corresponding to the colors of the contacts marked on the contact map presented in panel (b).

Another interesting aspect is the role of the pulling direction (C- versus N-end). When dragging from the C-end ("type I-C" and "type II-C"), we observed untying in 32% and 29% of cases, respectively. However, in the case of N-end dragging, untying is almost impossible (4% for "type I-N" and 0% for "type II-N"). Therefore, the direction of pulling (or knot sliding) may play an important role. Untying a knot in "type II" is more complicated since the other end is restrained, but it is still possible.

Interestingly, we observe in our simulations that the presence of the proteasome does not affect the measured mechanostability of knotted proteins, in contrast to similar simulations of proteins without any knots. This shows that the presence of knots on the protein chain affects the functionality of the proteasome, even leading to the proteasome becoming inactive when the entrance is blocked by the knotted protein.

According to our understanding of the mechanism of protein degradation by the proteasome, pulling the protein at a constant speed does not mimic the behavior of a protein in the presence of the proteasome. As explained in Sect. 2, a more realistic approach is pulling with a constant force. In this case, the protein chain can also be stretched using pulling types I or II. The comparison between the methods of protein pulling with constant speed and constant force is presented in Fig. 6.

Pulling with a constant force, apart from better mimicking the real mechanism, allows for the use of a lower force and, consequently, reduces the work needed to unravel the protein. This is possible thanks to the support from random forces provided by the Langevin thermostat. These forces facilitate protein unfolding or the untying of knots because they can resolve small steric clashes that may occur during protein pulling. In this case, the constant force only determines the direction of the chain. What is particularly interesting is how to compare the results obtained by pulling with a constant force and constant speed. One possible approach is well described in the literature [38] — a number of stretching simulations were conducted with different values of force, and the average pulling speed was calculated for each case. Protein mechanostability is defined as the extrapolation of the force values to a pulling speed of 80 residues/s, which is the pulling speed in the real proteasome [34]. Figure 6 shows different shapes of the graphs obtained during pulling with a constant force (panel (c)) or constant speed (panel (a)). The graph obtained during constant force pulling presents the distance L from the entrance to the proteasome as a function of time. The graph resembles multiple stair steps with different widths. Each step is associated with the breaking of a particular part of the structure. The contacts broken during the appearance of the particular steps are marked at the bottom part in panel (c). The time, $\tau_k \sim e^{\Delta E_k}$, necessary to break a particular group of contacts is related to the energy barrier, ΔE_k, that must be crossed [39]. The longest time is connected to the highest energy barrier.

Protein pulling with a constant force is also challenging. During such simulations, it is possible that steric clashes are so strong that thermal fluctuations cannot remove them. Additionally, knots in the protein chain also pose problems. In our simulations with continuous, constant force, we observed the proteasome becoming jammed by knots in proteins. For low pulling forces ($F \leq 1.6\epsilon$ Å^{-1}), only

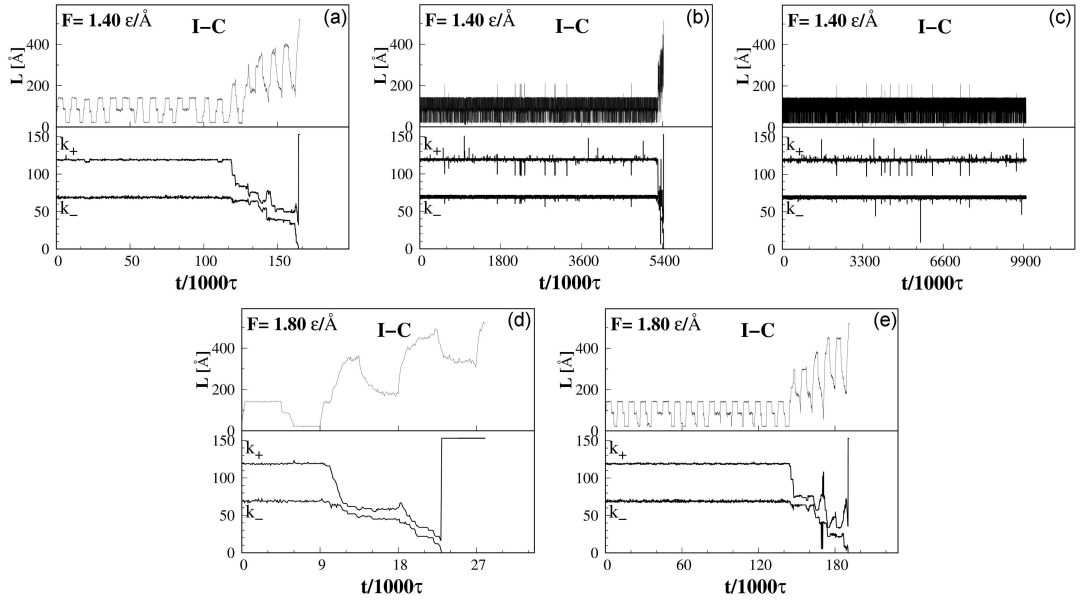


Fig. 7. The simulation results for protein 1NS5 within the periodic force model. Panels (a–c) present results obtained with a pulling force of $F = 1.40\epsilon$ Å^{-1}, while panels (d–e) present results obtained for $F = 1.80\epsilon$ $^{-1}$. Panels (a), (b), and (c) depict fast protein untying, a long simulation completed with knot untying, and a long unsuccessful simulation, respectively. Successful unknotting simulations are shown by a short simulation (panel (d)) and a long simulation (panel (e)).

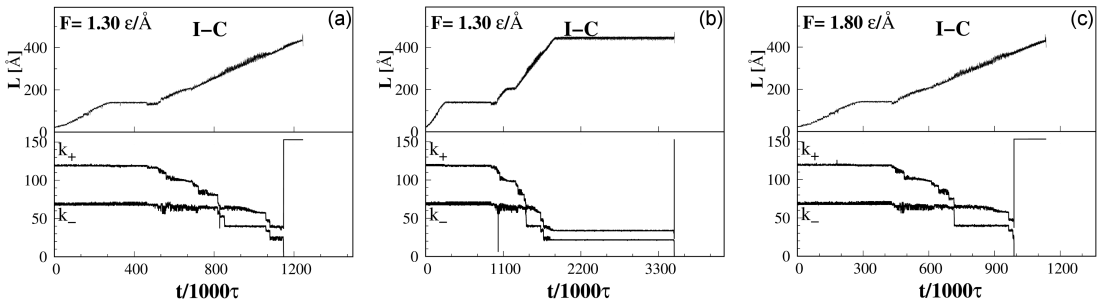


Fig. 8. The successful simulation results for 1NS5 within the ratchet force model. Panels (a), (b), and (c) depict fast pulling with a force of $F = 1.30$ ϵ/Å, slow pulling with the same force, and fast pulling again with a force of $F = 1.80$ ϵ $^{-1}$, respectively.

a portion of the protein without any knots was pulled into the chamber, and the pulled protein was halted at the position where the knot began. This indicates that the knot remained at its initial position. For higher forces, we observed the tightening of knots, but they did not move along the chain. This observation suggests that knotted proteins can block the proteasome, which cannot be true because in living cells, knotted proteins are usually degraded. To avoid this problem and achieve an even more realistic situation, we switch from the continuous application of constant force to a simulation with a constant force applied at given time intervals. Figure 7 shows five trajectories of protein pulling with different forces.

Applying the force in time intervals is also challenging because, in the absence of the force, we observe a backward movement of the chain, similar to the situation when part of the protein folds outside of the ribosome. However, this backward movement can be helpful in removing potential geometrical constraints. Repeatedly applying the force for a specific duration and then allowing the chain to relax can facilitate chain reorientation, increasing the possibility of the chain passing through the entrance. The number of times the force needs to be applied depends strongly on the protein, pulling force, and force application duration. This approach is also interesting because, as we showed in Fig. 7, the periodic force application can even enable the degradation of a knotted protein, in contrast to continuous force pulling. Figure 7 also depicts the varying times needed to untie knots in the knotted protein 1NS5, depending on the mentioned conditions. This process is very inefficient due to the return movements of the protein chain.

To prevent this behavior, we have implemented a "ratchet-like mechanism" that prevents the protein from moving backward when the force is switched off, mimicking the waiting for a new ATP portion in a real situation. Moreover, we do not specify the pulling period based on time, but rather consider the length of the pulled chain. We stop the pulling process when $d_{step} = 3.8$ Åof the chain is drawn into the proteasome. The size d_{step} was chosen to be comparable to the average size of a single amino acid. It is worth mentioning that if a geometrical clash occurs and no chain movement is possible, we pause the pulling process for 4.5 μs, allowing the Langevin thermostat to address the issue for another 4.5 μs before applying the force again. Importantly, no protein backward movement is allowed during this time. The advantage of this mechanism, compared to simple pulling with a constant force, is that we prevent pulling the same part of the chain many times, which can occur when it escapes during the absence of force. Avoiding multiple pulls of the same part of the protein chain using the "ratchet-like mechanism" is also favorable because periodically applying a small force has a lower likelihood of causing strong steric clashes. The trajectories from the simulation of the 1NS5 knotted protein, as presented in Fig. 8, exhibit this behavior. Moreover, we can see that this approach is much more stable, as the times necessary for protein unfolding are comparable regardless of the simulation parameters. The "ratchet-like mechanism" is considered the most realistic approach for pulling proteins into the proteasome.

5. Conclusions

In our simulations, we have demonstrated that the reduction in the protein's potential energy is crucial for the process of a protein exiting the ribosomal tunnel. As the partially folded protein emerges from the ribosome, it generates a pulling force on the remaining portion within the tunnel. While interactions between the protein chain and the ribosomal tunnel walls, along with diffusion, are important, their significance for the protein's movement outside the tunnel is secondary. This was confirmed by simulations of pulling the protein into the proteasome chamber without the "ratchet-like mechanism," resulting in the protein being retracted from the proteasome. This backward movement also correlates with a decrease in the protein's potential energy. It is worth noting that further investigation is needed, as these simplified models may not encompass all aspects of the processes under consideration.

Furthermore, our simulations lend support to the hypothesis regarding the directionality of knot tightening and untying. Given the critical role of the ribosome structure in simulations of knotting the 1J85 protein by pulling the C-terminus across a specially created loop, it suggests that knot degradation should be more feasible from the C-end. This implies that pulling the 1J85 protein from the N-terminus enables knot sliding and untying. This finding aligns with our simulation results using the constant-speed scheme for 1J85. The untying of the knot and successful engulfing of the protein were achievable only with the "type I-N" scheme.

Acknowledgments

The authors acknowledge Professor Marek Cieplak for encouraging the investigation of protein dynamics, explored using coarse-grained models. This research has received support from the National Science Centre (NCN), Poland, under grant No. 2018/31/B/NZ1/00047 and the European H2020 FETOPEN-RIA-2019-01 grant PathoGel-Trap No. 899616. The computer resources were supported by the PL-GRID infrastructure.

References

[1] F. Türker, E.K. Cook, S.S. Margolis, *Cell Chem. Biol.* **28**, 903 (2021).

[2] G.A. Collins, A.L. Goldberg, *Cell* **169**, 792 (2017).

[3] M. Bochtler, L. Ditzel, M. Groll, C. Hartmann, R. Huber, *Annu. Rev. Biophys. Biomol. Struct.* **28**, 295 (1999).

[4] S. Gottesman, *Annu. Rev. Genet.* **30**, 465 (1996).

[5] J.A.M. Bard, E.A. Goodall, E.R. Greene, E. Jonsson, K.C. Dong, A. Martin, *Annu. Rev. Biochem.* **87**, 697 (2018).

[6] Z. Li, Q. Guo, L. Zheng, Y. Ji, Y.T. Xie, D.H. Lai, Z.R. Lun, X. Suo, N. Gao, *Cell Res.* **27**, 1275 (2017).

[7] K.Y.S. Hung, S. Klumpe, M.R. Eisele et al., *Nat. Commun.* **13**, 838 (2022).

[8] M. Chwastyk, M. Jaskolski, M. Cieplak, *FEBS J.* **281**, 416 (2014).

[9] M. Chwastyk, M. Jaskolski, M. Cieplak, *Proteins* **84**, 1275 (2016).

[10] M. Chwastyk, E.A. Panek, J. Malinowski, M. Jaskólski, M. Cieplak, *Front. Mol. Biosci.* **7**, 591381 (2020).

[11] A.P. Perlinska, W.H. Niemyska, B.A. Gren, M. Bukowicki, S. Nowakowski, P. Rubach, J.I. Sulkowska, *Prot. Sci.* **32**, e4631 (2023).

[12] P. Virnau, L.A. Mirny, M. Kardar, *PLoS Comput. Biol.* **15**, e122 (2006).

[13] J.I. Sułkowska, P. Sułkowski, P. Szymczak, M. Cieplak, *Proc. Natl. Acad. Sci.* **105**, 19714 (2008).

[14] M. Cieplak, M. Chwastyk, Ł. Mioduszewski, B.R.H. de Aquino, *Prog. Mol. Biol. Transl. Sci.* **174**, 79 (2020).

[15] M. Cieplak, T.X. Hoang, M.O. Robbins, *Proteins* **49**, 114 (2002).

[16] M. Sikora, J.I. Sułkowska, M. Cieplak, *Plos Comput. Biol.* **5**, e1000547 (2009).

[17] J.I. Sulkowska, M. Cieplak, *Biophys. J.* **95**, 3174 (2008).

[18] M. Chwastyk, A. Galera-Prat, M. Sikora, Á. Gómez-Sicilia, M. Carrión-Vázquez, M. Cieplak, *Proteins* **82**, 717 (2014).

[19] M. Gunnoo, P.-A. Cazade, A. Orlowski, M. Chwastyk, H. Liu, D.T. Ta, M. Cieplak, M. Nashde, D. Thompson, *Phys. Chem. Chem. Phys.* **20**, 22674 (2018).

[20] Y.N. Zhao, M. Chwastyk, M. Cieplak, *Sci. Rep.* **7**, 39851 (2017).

[21] M. Chwastyk, A.P. Bernaola, M. Cieplak, *Phys. Biol.* **12**, 046002 (2015).

[22] R. Young, H. Bremer, *Biochem. J.* **160**, 185 (1976).

[23] K. Boström, M. Wettesten, J. Borén, G. Bondjers, O. Wiklund, S.O. Olofsson, *J. Biol. Chem.* **261**, 13800 (1986).

[24] N.T. Ingolia, L.F. Lareau, J.S. Weissman, *Cell* **147**, 789 (2011).

[25] M. Chwastyk, M. Cieplak, *Front Mol. Biosci.* **8**, 692230 (2021).

[26] A.M. Ruschak, T.L. Religa, S. Breuer, S. Witt, L.E. Kay, *Nature* **467**, 868 (2010).

[27] F. Zhang, M. Hu, G. Tian, P. Zhang, D. Finley, P.D. Jeffrey, Y. Shi, *Mol. Cell* **34**, 473 (2009).

[28] M. Groll, L. Ditzel, J. Löwe, D. Stock, M. Bochtler, H.D. Bartunik, R. Huber, *Nature* **386**, 463 (1997).

[29] D.A. Nissley, Q.V. Vu, F. Trovato, N. Ahmed, Y. Jiang, M.S. Li, E.P. O'Brien, *J. Am. Chem. Soc.* **142**, 6103 (2020).

[30] P. Dabrowski-Tumanski, M. Piejko, S. Niewieczerzal, A. Stasiak, J.I. Sulkowska, *J. Phys. Chem. B* **122**, 11616 (2018).

[31] J. Frank, R.L. Gonzalez Jr., *Annu. Rev. Biochem.* **79**, 381 (2010).

[32] A.H. Elcock, *Plos Comput. Biol.* **2**, e98 (2006).

[33] P.T. Bui, T.X. Hoang, *J. Chem. Phys.* **153**, 045105 (2020).

[34] R.A. Maillard, G. Chistol, M. Sen, M. Righini, J. Tan, C.M. Kaiser, C. Hodges, A. Martin, C. Bustamante, *Cell* **145**, 459 (2011).

[35] M. Chwastyk, M. Cieplak, *J. Phys. Chem. B* **124**, 11 (2020).

[36] M. Wojciechowski, P. Szymczak, M. Carrión-Vázquez, M. Cieplak, *Biophys. J.* **107**, 1661 (2014).

[37] Y.I. Zhao, M. Chwastyk, M. Cieplak, *J. Chem. Phys.* **146**, 225102 (2017).

[38] M. Wojciechowski, Á. Gómez-Sicilia, M. Carrión-Vázquez, M. Cieplak, *Mol. BioSyst.* **12**, 2700 (1016).

[39] J.I. Steinfeld, J.S. Francisco, W.L. Hase, *Chemical Kinetics and Dynamics*, 2nd ed., Prentice Hall, 1999.