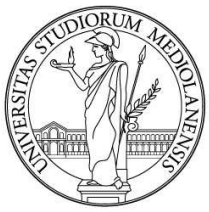


UNIVERSITÀ DEGLI STUDI DI MILANO

DIPARTIMENTO DI BIOSCIENZE

PhD Course in Molecular and Cellular Biology
XXXV Cycle
Academic Year 2021-2022

Building a model to simulate protein aggregation at atomistic detail

Emanuele Scalone
R12492

Scientific tutor:
Prof. Carlo Camilloni

SSD: FIS/07; BIO/10

Thesis performed at Dipartimento di Bioscienze, Università degli Studi di Milano.

I. Riassunto

Il progetto su cui mi sono dedicato nel mio dottorato è rappresentato dallo sviluppo di un metodo computazionale per simulare l'aggregazione di proteine in fibrille amiloidi. Come descritto più avanti, l'aggregazione proteica è un processo complesso che porta all'accumulo e alla deposizione di fibrille amiloidi, causando molte malattie tutt'ora non curabili. I dettagli atomistici di questo processo non sono ancora completamente compresi e la dinamica molecolare potrebbe integrare le tecniche sperimentali disponibili. Tuttavia, le grandi dimensioni dei sistemi in gioco e la lunghezza di simulazione necessaria rendono difficile l'applicazione di tali metodi. Seguendo gli studi di Camilloni e Sutto1–3, ho sviluppato un approccio che permette di simulare l'aggregazione proteica basato sulla conoscenza delle strutture iniziale e finale di tale processo. Durante il primo anno di dottorato, mi sono concentrato sullo sviluppo iniziale del metodo creando un singolo potenziale in grado di rappresentare due stati di una proteina a partire da due strutture. Per lo sviluppo della prima versione si è utilizzato SMOG, un software che crea modelli di structure-based (SB), cioè utilizzando una struttura di riferimento come minimo dell'energia potenziale. Contestualmente, ho sviluppato un codice in Python per unire diversi modelli

SMOG in uno ed eseguire simulazioni di dinamica molecolare utilizzando GROMACS. Per testare il metodo si è utilizzato un peptide derivato dalla Transtiretina. Il primo tentativo è stato quello di rappresentare a grana grossa il peptide TTR mappando su una singola sfera centrata su sul rispettivo Cα. Tuttavia, tale rappresentazione semplificata si è dimostrata inefficace nel rappresentare il processo di aggregazione. In seguito, sono passato ad una rappresentazione semi-atomistica parametrizzando tutti gli atomi tranne gli idrogeni utilizzando SMOG ed ho definito il modello come multi-GO. Con la nuova rappresentazione semi-atomistica, multi-GO ha dimostrato la capacità di simulare l'aggregazione delle proteine grazie al raffinamento del potenziale mediante diverse ottimizzazioni. La prima ottimizzazione è rappresentata dallo sviluppo di un codice Python per parametrizzare il potenziale Lennard-Jones (LJ) invece di SMOG. Di conseguenza, i parametri di legame (bonded) sono stati ottimizzati insieme a quelli non covalenti. Abbiamo basato i primi su GROMOS54a7 migliorando notevolmente il nostro modello. Inoltre, ho aggiunto la possibilità di imparare il potenziale LJ dalle simulazioni MD. Questa nuova versione è stata chiamata multi-eGO. I risultati ottenuti nel secondo anno sono stati pubblicati nel primo articolo su PNAS. Al terzo anno ho applicato questo metodo alla proteina Aβ42, poiché un training set basato su simulazioni

atomistiche statisticamente significativo è stato generato, permettendo la generazione di un potenziale multi-eGO. I test sulle proteine complete hanno evidenziato un problema importante nella parametrizzazione della geometria locale di multi-eGO, legato a uno sbilanciamento tra il potenziale LJ dei contatti locali e di quelli a lungo raggio. Durante il terzo anno sono riuscito a migliorare la descrizione di proteine lunghe inserendo un parametro che permette di raffinare la descrizione della geometria locale. Poiché la proteina Aβ42 è stata parametrizzata correttamente, ho potuto estendere le funzionalità di multi-eGO aggiungendo una piccola molecola alla simulazione. Ho svolto questo compito presso l'Università di Cambridge, ospite del prof. Michele Vendruscolo, dimostrando la possibilità di parametrizzare piccole molecole in questo framework. Complessivamente, in questi tre anni abbiamo sviluppato un nuovo modello in grado di simulare semi-quantitativamente l'aggregazione proteica. Sebbene siano necessarie diverse ottimizzazioni, multi-eGO potrebbe diventare un potente strumento in silico per descrivere ed investigare l'aggregazione proteica.

II. Summary

The project I have been focusing for the last three years was the development of a computational method to simulate protein aggregation into amyloids. As further described below, protein aggregation is a complex process leading to the accumulation and deposition of amyloid fibrils causing several and uncurable diseases. The atomistic detail of this process remains elusive and molecular dynamics could complement experimental techniques. However, the large size of the systems and long timescales of the process make difficult to apply such methods. Following the works from Camilloni and Sutto[1–3], I developed a structure-based (SB) approach to protein aggregation. During the first year of the PhD, I focused on the initial proof of concept of using two structures in a single SB model. The first version was based on a software called SMOG which creates SB models using a protein structure as input. In the first year I developed a python script to merge different SMOG models into one and perform MD simulations using GROMACS. The method was developed using a short peptide derived from the Transthyretin protein. The first attempt was to coarse grain the representation of the TTR peptide mapping a single bead per residue on the Cα. However, such simplified representation was not suitable to simulate protein

aggregation. Afterward, I used SMOG in a semi-atomistic representation to define our model called multi-GO, demonstrating the feasibility of our method, although several optimizations were required. The first optimization was to develop a python code to parametrize the LJ potential instead of SMOG. As a result, bonded parameters were optimized along with the non-bonded. We based bonded parameters on GROMOS54a7, and the improved geometry greatly improved our model. Moreover, I added the possibility to learn the LJ potential from MD simulations. This new version was named multi-eGO. The results obtained in the second year were published in the first article on PNAS. In the third year I applied this method on the full length Aβ42 protein, as the knowledge required to parametrize multi-eGO is available. Testing full-length proteins highlighted a major issue in multi-eGO related to an unbalance between LJ potential of local and long-range contacts. During the third year I was able improve the description of proteins by inserting a prior information describing the local geometry. As the Aβ42 protein was properly parametrized, I was able to extend multi-eGO functionalities by adding a small molecule in the simulation. I performed this task at Cambridge University hosted by prof. Michele Vendruscolo, proving the possibility of small molecules parametrization in multi-*e*GO. Altogether, during those three years we developed a new model able to

qualitatively simulate protein aggregation. Although several optimizations are necessary, multi-*c*GO could become an additional tool to simulate protein aggregation.

III. Aim of the project

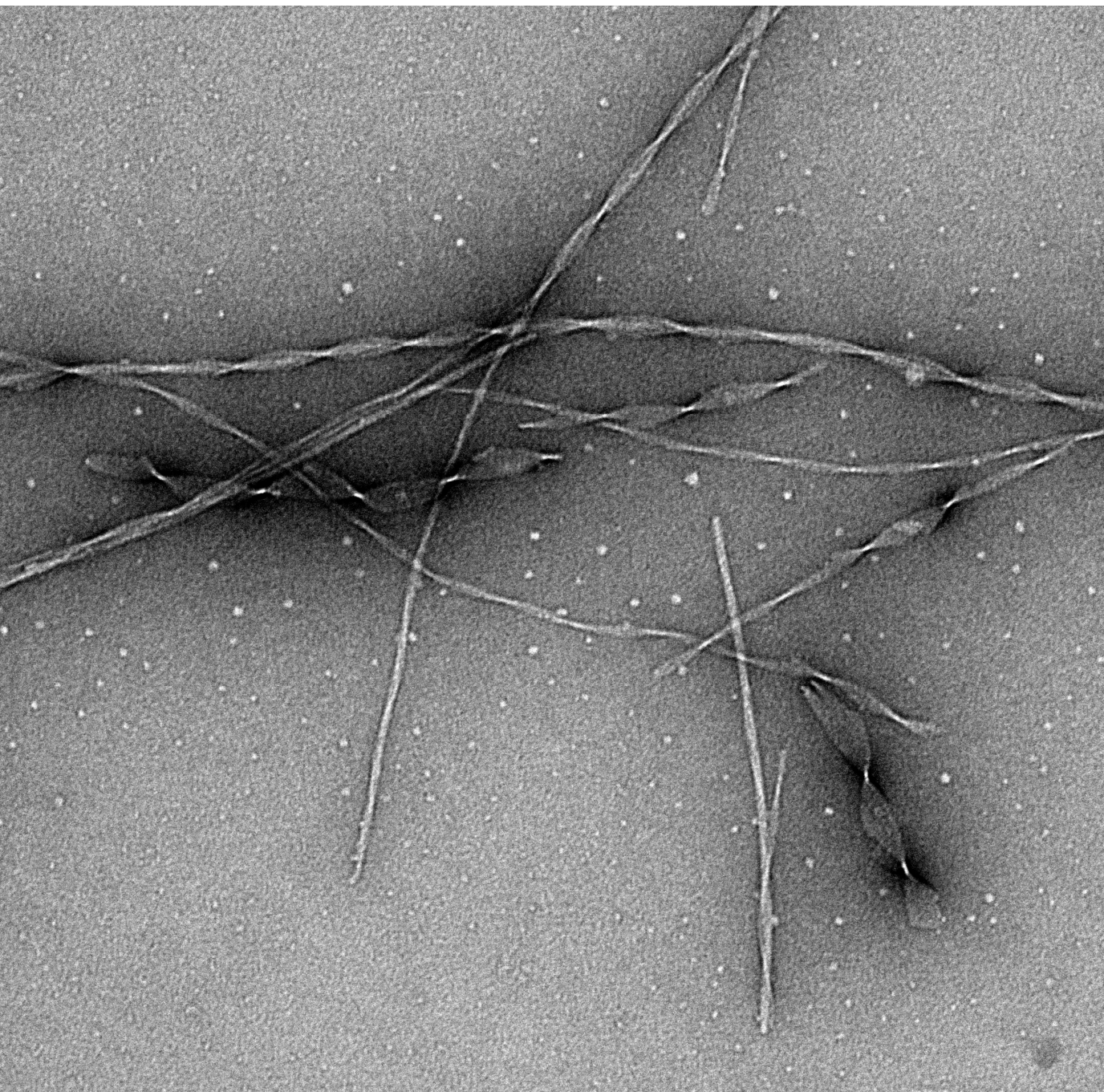
Amyloid fibrils are the result of the accumulation and deposition of proteins. All molecules composing the fibrils arrange in β-sheets repeated along the fibril axis and their sidechains closing in a dry steric zipper. The relevance of the amyloid state research is determined by several diseases linked to the formation of insoluble fibrils. Some of these diseases are epidemic globally and mostly uncurable. Amyloidosis is a complex process whose details remain elusive due to the considerable dynamics of the oligomeric intermediates. The latter have been reported to be toxic for the cell, increasing the interest to define the structure and the formation of oligomers and possibly develop a treatment targeting them. Given the dynamical properties of such process, performing high-resolution experimental observations has proven difficult. Yet, in such cases it is possible to perform Molecular Dynamics simulations to obtain some insights. However, the large size and the long time scales of the aggregation process prevents the application of common MD methodologies. To this aim, my PhD project is to develop a novel method to simulate the entire process of amyloidosis allowing the characterization of both structure and dynamics of oligomers and the consequent amyloid fibrils formation. The method is based on multiple Structure-Based

Models and thus called multi-*c*GO. I was able to perform concentration dependent MD simulations describing the aggregation kinetics of thousands of monomers at atomic resolution. I carried out the initial development of multi-*c*GO using an amyloidogenic peptide from Transthyretin, a protein involved in cardiac amyloidosis. The simulated kinetics and the structural features of the resulting fibrils are in qualitative agreement with in vitro experiments. Starting from monomers, multi-*c*GO simulations were able to describe the formation of primary nuclei from lower-order oligomers following their growth and the subsequent secondary nucleation events, until the maturation of the fibril is achieved. After the Transthyretin peptide, I have been improving multi-*c*GO to simulate the aggregation of full-length proteins with the possibility to perform drug discovery, study protein folding processes and, the assembly of protein complexes.

1. Introduction

“Dai, dai, dai.”

René Ferretti



The foundation of this project lays on the concept of protein folding. Proteins can perform their functions when properly folded, while several diseases derive because of misfolding. Among those, amyloidoses are one of the most representative as they cause Alzheimer's Diseases (AD) and Parkinson's Disease (PD). The current knowledge regarding protein folding allowed the development of the simplified computational model applied in this thesis.

1. Protein aggregation and amyloids

As defined by the symposium held in 2020 by the International Society of Amyloidosis[4]; an amyloid is *"a sub-microscopic fibrillary structure of a substance that histologically appeared amorphous"*. Its structure consists of *"a generic cross-β structure of amyloid fibrils of different origin […] derived from non-fibrillar precursors"*. The same symposium also provided the definition of amyloid fibril as *"the basic structure of all amyloids […] are built up by twisted protofilaments. An amyloid protofilament is a stack of protein layers in β-sheet structure, which when twisted about identical stacks, forms an amyloid fibril"*. These definitions are the result of observations that began in 1639[5] up to the present day, where approximately 50 proteins are defined as an etiologic agent in amyloidosis[4–7].

Although it is known the role of protein folding in aggregation, the detailed process by which proteins fails to achieve their native conformation and consequently aggregate, remains elusive[8,9]. The process of amyloid accumulation is slow and the incidence of developing amyloidoses increases with aging[10]. Specifically, more than 10 proteins are associated with age-related amyloid deposition such as amyloid-β protein (Aβ) and tau in AD[11,12], α-synuclein in Parkinson's Disease (PD)[13,14], transthyretin[15,16], islet amyloid polypeptide (IAPP) in type 2 diabetes mellitus[17] and AD[18], atrial natriuretic factor (ANF)[19], apolipoprotein AI (ApoAI)[20], and epidermal growth factor-containing fibulin-like extracellular matrix protein 1 (EFEMP1)[4,21]. These diseases are well known and have a heavy impact on the national health system of several states[22]. As example, AD is one of the leading causes of dementia with 24.4 million people affected with an increase of 4.6 million every year[11]. PD is the second most-common neurodegenerative disorder that affects 2-3% of the population over 65 years of age[14]. Indeed, other factors besides age can trigger amyloidosis besides aging such as genetic mutations[23–25], post-translational modifications, and high amyloidogenic precursor proteins concentration that affect protein homeostasis[10,26–30]. Aggregated fibrils can reach such concentrations as to impair the tissue function[31]. The accumulation of amyloid fibrils can impair the organ

functions in diverse ways, yet it is not understood why specific proteins target certain organs. As example, β2m targets the joints and guts[32]; the V30M transthyretin targets peripheral nerves[24]; light chain fibrils can deposit systemically[33]. The organ of deposition can depend on several factors such as the local pH, the presence of specific proteases, the presence of fibril seeds, specific interactions with glycosaminoglycans or cell receptors[28,34,35]. The deposition of insoluble proteins affects the tissue mechanically changing its architecture and preventing the physiological function[28,31]. The exchange of nutrients between tissues and blood stream can be reduced[28,31]. The accumulation of proteins can interact with membranes of mitochondria, lysosomes, endoplasmic reticulum (ER) and plasma membrane[5,36–38]. Hence, destruction of cell membrane, interactions with receptors and the formation of toxic ion channels[28,33,39]. The presence of high number of fibrils can trigger an inflammatory response which can further deteriorate the tissue, as in Alzheimer's disease[11,28]. Additionally, the presence of fibrils reduces proteasomal degradation[5,26,40], impair autophagy and mitochondrial function[5,41,42], produce reactive oxygen species (ROS)[43]. *In vitro* amyloids display cytotoxicity[5,36,44]. The amyloid structure is considered thermodynamically more stable than the native state[7]. However, the formation of fibrils requires a remarkably high protein concentration which proteins hardly

reach *in vivo*[7]. Moreover, the critical concentration required for amyloid formation is inversely proportional to the length of the protein sequence. Therefore, it is possible to theorize that biology has evolved to use proteins of at least 300-500 amino acid residues that reduce the probability of amyloid formation[7,45,46]. More precisely, the free energy barrier involved in the rearrangement of sequences of such length from native to amyloid is kinetically less favorable[7]. Moreover, during evolution proteins were selected to keep the aggregation prone sequences within their globular core[7,47–49] and promote alternate hydrophilic and hydrophobic sequences[50–52]. As further described below, as protein aggregation is the result of misfolding, several cell mechanisms evolved to prevent or correct such process. Such mechanisms involve molecular chaperones, the degradation processes to keep protein homeostasis and autophagy; and their impairment results in protein aggregation[5,7,53–55]. In this process, it has been observed the formation of cytotoxic oligomeric species prior to the amyloid deposition[5,36,56–65]. Oligomers can be defined as a variety of small soluble aggregates[66] and their toxicity can be related to the exposure of hydrophobic residues[5,67]. As further explained below, oligomers can include protofilaments which are considered the initial aggregate leading to the formation of fibrils[68]. It is important to mention that there are observations of organs

continuing to function despite the high number of fibrils[5,31,69]. This paradox, suggest that amyloid fibrils could have a protective role against the cytotoxic oligomers observed at the initial stages of aggregation[6,8,31].

It is worth mentioning that proteins able to assemble in amyloids can be part of functional and physiological pathways[70]. An amyloid fibril can be used to store proteins in a stable and insoluble state. The study of the functional roles of protein aggregation is emerging, paving the way to interesting applications. Functional amyloids can be found in a broad range of living organisms from bacteria to mammals suggesting a fundamental role in biology. Based on their properties, functional amyloids can be classified as: chemical storage, structure, information, loss-of-function, and signaling gain-of-function. An example of the first group could be represented by hormones[71] storage inside granules or yeast pyruvate kinase Cdc19. This can be related to membraneless granules as phase-separated agglomerate of amyloids. Curli[72] proteins have evolved as a structural amyloid in bacteria biofilm which could be exploited as a target for a new generation of antibiotics. Amyloids can handle memory[73] as information carrier. Cdc19 can be classified as a loss-of-function amyloid as it aggregates when not

useful to the cell[74]. Finally, an example of gain-of-function can be represented by RIP1/RIP3 which aggregation causes apoptosis cell signaling[70,75].

Structural biology of amyloids

Amyloid deposits in insoluble fibrils of about 10 nm diameter[31] providing a very peculiar X-ray diffraction pattern defined as cross-β[8,76–78]. The motif is characteristic of amyloids, and it is formed regardless the native fold of the protein, including the intrinsically disordered one and the rich in α-helices ones[8]. The cross-β structure is characterized by a continuous array of hydrogen bonds which provides a great strength and stability[7,77,79]. Fibrils can be either extracted from tissues or produced in laboratory. In such way, it is possible to ease the study and the descriptions of amyloid proteins and their related disease[8,80]. Typically, experimental techniques employed to study the structure of protein aggregates provide a low-resolution data, thus other methods must be complemented[68].

One of the most used techniques is the UV-visible absorbance or fluorescence[68,81]. Three dies are mostly used to bind amyloid aggregates: Congo red (CR), thioflavin T (ThT) and 1-anilino-8-naphthalene sulfonate (ANS)[82,83]. Conformation specific antibodies (Abs) can be used as another low-resolution

technique. Abs can detect the presence of aggregates and their location[84,85]. Protease digestion can provide indications about the compactness of the aggregate structure[86,87]. Atomic force microscopy (AFM) and electron microscopy (EM) can allow the visualization of macroscopic properties such as length, width, rigidity, and the degree of branching. AFM, transmission electron microscopy (TEM) and cryo-transmission electron microscopy (cryo-EM) shows that fibrils are composed of individual protofilaments, and different precursor proteins may lead to different number of protofilaments[8,88,89]. At the microscope, the fibril thickness is about 10 nm[31]. Also, AFM allows the visualization of the fibril growth as an additive elongation at the ends[90]. Fibrils measured using solid-state NMR (ssNMR) and X-ray crystallography shows the cross-β architecture as a result of backbone hydrogen bonding, which is accessible to all polypeptide chains resulting in extended β-sheets[7]. Also, pairs of β-sheets are interlocked by their amino acid sidechains[8] and defined steric zipper[7,91].

Such motif is not common in proteins as the steric zipper usually does not occur in two copies of the same sequence as in amyloids[77]. The atomistic details of fibrils also illustrate the main factors of their stability: hydrophobic effect; van der Waals stabilization, mutual polarization of the hydrogen-bonding groups of the

backbone and sidechains (polar zippers); and stacking interactions of the sidechains[8,77,91,92]. As suggested by the AFM elongation measures, the fibrils 3D structure is composed of thousands of copies of a protein or peptide[8]. Nowadays, due to the optimization of Cryo-EM protocols and hardware, several atomistic descriptions of amyloid fibrils are being published[88,89,93–105]. The comparison between the different fibrils proves the variety of polymorphism, which depends on the aggregating protein[8], and also on the aggregating conditions[80,99] and such polymorphism includes structural breaks within the fibrils[99]. Overall, it is clear that fibrils from the same precursor protein can aggregate in different structures involved in different diseases[103–105], different seeding characteristics[106], different rates of spread[107] and distinct pattern of neuropathology[77,108,109]. A common feature is a left-handed protofilament twist of the protofilament, although right-handed and straight fibrils were observed[8]. Such twist is attributed to the chirality of the L-amino acid residues and intrinsically related to the fibril topology[8]. The twist may change within the same fibril, different fibrils in the same sample and it is influenced by the aggregating condition used to produce the fibrils.

Aggregation kinetics

In addition to microscopy, aggregation kinetics are among the first measurements to be made when studying a new aggregating molecule. Several methods can be employed to measure aggregation kinetics such as circular dichroism spectroscopy (CD), NMR spectroscopy, infrared spectroscopy (FTIR), fluorescent spectroscopy, and scattering methods (using X-rays). In this project, wet lab kinetic measurements were based on fluorescent spectroscopy using Thioflavin T as dye. Usually, a sigmoidal kinetics is obtained by measuring the aggregate amount over time. Three distinct phases can be distinguished: the lag phase, growth phase, and final plateau[7,8,81]. The duration of the lag phase is called lag time (t_{lag}). The end of t_{lag} is generally defined as the point when the signal starts to increase. Different analytical definitions could refer to the signal increase and, in this thesis, we defined t_{lag} as the intercept of the pre-transition base line and the maximum derivative of the growth phase[81]. Another common parameter to study protein aggregation is the half time ($t_{1/2}$) as the time point related to the 50% signal. Many microscopic events occur throughout the aggregation process, and that the pattern of kinetics is a result of the predominant flow at that given time.

The most represented species during the entire process can be either monomers in the lag phase or fibrils once reached the plateau. Instead, the concentration of other small aggregates and oligomers remains low throughout the entire aggregation process[110]. During lag phase many microscopic processes occur and continue throughout the aggregation. At first there is the formation of small oligomers composed of a few molecules. Such oligomers display a low elongation or growth, suggesting a different structure compared to the amyloid one. Therefore, it is possible to define a primary nucleus as an oligomer with a stable conformation and because of that, able to elongate faster than an unstructured oligomer. It has also been demonstrated that oligomers often revert to the monomeric state, supporting the hypothesis that the energy barrier required to change from oligomer to primary nucleus is very large[110]. The energy barrier reduces based on the concentration increase[110,111] and once formed, the primary nucleus can be seen as an energetically favorable transition state onto which monomers can attach leading to amyloid formation[7,29,81]. During the lag phase millions of primary nuclei are constantly formed as the number of monomers is extremely high until the start of the growth phase. The sigmoidal behavior of aggregation kinetics suggests nucleation as the rate limiting step, further proven by the addition of amyloid seeds triggering the aggregation without the lag

phase[68,112]. The process of elongation can be described as the step where monomers attach to the ends of a nuclei or a fibril. The monomer interaction matches the cross-β conformation of molecules already present in the fibril functioning as a template[7,113,114]. Thus, the fibrils growth is triggered immediately after the first nucleation event or by templating or seeding from existing aggregates[81,115,116]. It has been hypothesized that fibrils have a catalytic surface[8] which can further trigger the formation of nuclei[29,111,117,118]. Therefore, secondary nucleation can overcome the primary nucleation after a certain concentration of fibrils is formed. Molecular details about the catalytic surface of a fibril are yet to be characterized.

As mentioned above, such processes occur throughout the entire aggregation process, and it also possible to claim that secondary nucleation overcome the primary one during the lag phase strongly affecting its duration. The difficulties of measuring of low concentration fibril, could mistakenly lead to the assumption that primary nucleation is the main driving force of the lag phase, but fibrils are already present triggering the secondary nucleation. Also, fibrils can break leading to the fragmentation process which raises the number of free ends for monomer to attach[81,111,119,120]. Together those events triggered by fibrils exponentially

increment the aggregation rate, thus defining a very steep growth phase. Several factors could change the behavior of the aggregation kinetics such as protein sequence changes, the presence of other molecules[121,122], variations in buffer or experimental conditions[123,124], mechanical factors like shaking[125]. Such variations could be exploited as a tool to study aggregation kinetics and small molecules aimed at the amelioration of amyloidoses.

As much information can be obtained by the kinetics, an atomistic structural description of oligomers is still missing. As mentioned above, oligomers play a critical role in protein aggregation and pathogenicity. Solving their structure could answer to several question regarding their pathogenicity[5]. Nowadays single molecules approaches are proving to be effective compared to bulk methods to properly follow the formation of oligomeric species[126]. Different techniques can be valuable to study oligomerization such as: single molecule fluorescence detection (smFRET)[110], fluorescence correlation spectroscopy (FCS), single molecule photobleaching[127] and super resolution optical imaging[126]. It is possible to detect the amount of such oligomers through the combination of different instruments and experiments such as size-exclusion chromatography (SEC), mass spectroscopy (MS) and liquid scintillation[110]. For some proteins, the oligomeric

size is determined such as hIAPP[128] and β2m[129] as in the latter it is also available a model of the hexameric oligomer of native-like monomers. However, it is difficult to precisely determine their structure as they are unstable, dynamic, and heterogeneous in mass and structure[5,128,130]. Another approach is based on photobleaching[127]. The latest experimental setup takes advantage of single molecule fluorescence and a deep neural network[131], confirming the different hypothesis made so far: the formation of oligomers and the aggregation pathways are highly sensible to the environment, suggesting differences between *in vitro* and *in vivo* aggregation. There is a wide heterogenicity in aggregation in terms of number of subunits, length, elongation speed and structure of fibrils. Specifically, there are observed four types of aggregation of Aβ42[131].

Type I pathway can be described as the formation of oligomers of varied sizes at the beginning of aggregation but most of them do not become amyloid fibrils. Type II pathway is characterized by fibril-like oligomers which are able to elongate. Type III pathway requires a conformational conversion from disordered to cross-β structures and then elongate. Finally, Type IV can be ascribed to the secondary nucleation process in which oligomers interacts with the fibril surface into disordered oligomers.

2. Protein Folding and misfolding

Proteins are probably the most versatile biopolymers as they handle almost all cellular functions[9,132,133]. Unlike any other polymers, they achieve 3D structure to perform their specific biological activity[9,132,133]. The central dogma concerning protein folding is that their structure depends on the sequence of amino acids[77]. However, a thorough understanding of the folding process is far from trivial. The complexity of protein folding is also such that there is evidence that the same sequence allows for alternative structures to the native one including the "amyloid fold"[77]. Small proteins below 100 amino acids long are usually folded in a single domain and thus folding in a two-state transition[9,134,135]. This is the simplest protein folding process starting from a completely unfolded conformation to a folded one without intermediates[134]. Because of that, it is possible to experimentally observe the protein folding process with NMR and to develop initial simplified Molecular Dynamics (MD) models[136]. However, two-states folding prevents a detailed study of the different interatomic contacts occurring during the folding process. Instead, proteins with more than 100 amino acids, corresponding to more than 90% of the proteins in a cell, have a tendency to collapse in water solution into compact non-native conformations, namely

intermediate states[135,137]. For small proteins it is possible to measure the structural intermediates which could be misfolded conformations to be reorganized to reach the native state[137]. Research on protein folding provided a massive amount of structural information related to several protein sequences linking different protein activities with a structure but also finding unstructured regions and intrinsically disordered proteins (IDPs)[138].

The study of protein folding started in the early '60s to these days and different theories were hypothesized trying to explain the experimental data. One early accepted theory is proposed by Cyrus Levinthal. He hypothesized a random search performed by a polypeptide chain of its native conformation[139,140]. Meaning that all the conformations are equally probable[139,140]. However, given the length of a protein, this process would have been required billions of years[140,141] to occur which does not correspond to the experimental observation, as proteins folds in milliseconds to seconds. Hence, this is defined as the Levinthal Paradox. At the same time, Levinthal himself expanded the hypothesis including a pathway leading to the native state[142]. After that, the research continued as an effort to find the pathway(s) leading to the final structure observed in crystallography. A new model is built as a folding tunnels and energy landscapes

where the latter is the representation of the free energy of each conformation as a function of the degrees of freedom[137,143]. Upon initiation of folding conditions at the top of the highest conformational energy, the protein tends to change its conformation and reducing its energy reaching its native structure[8,143].

Several forces are acting on the protein during this entire process pushing the exploration of different conformations. Such forces can be solute enthalpy, solvent entropy and enthalpy, steric and charge repulsions, hydrophobic interactions, and solvation effects[8,9,143]. It is also important to include the contribution of chaperones during the entire folding process[137]. Hence, a protein can fold from different random conformations and follow different pathways to its native state, but also can be trapped in misfolded conformations[6–8,27,143,144]. Since all those forces are simultaneously applied, a stable conformation will fold reaching a minimal frustration among all interactions[141,145]. Specifically, the minimal frustration principle is the basis of the theoretical studies made by Gō and co-workers further described below[146,147]. Such principle describes a folded structure as compact and secondary and tertiary structures as not in conflict[145]. Super secondary structure accommodates both local hydrogen bonding and hydrophobic packing[145]. This principle can stand true when the folding process

could be interpreted in sequential and cooperative events[9]. First, there is the formation of hydrogen bonding along the backbone, favoring the formation of α-helices[9,148,149]. Subsequently, there is a hydrophobic collapse determining the tertiary structure of proteins where sidechains cover an important role[9]. Thus, the solvation covers a fundamental role in protein folding[150–153]. As mentioned earlier, starting from a large conformational space, the probability of folding in a specific native state is very low. But the initial folding process narrows the probability of several conformations, hence increasing the chances to fold into the native conformation. [6,154]Mutagenesis experiments has proven the fold is maintained as long as the hydrophobic core is maintained[9,154,155].

Multiple folding pathways are possible and some of them do not directly end in the final native state or lead to topological traps[144,156–158]. As different alternative conformations can be reached during the protein folding process[77,137], or caused by pathological conditions; cells have different mechanisms to prevent misfolded molecules[40,65,159–161]. An example could be the ability of proteins to fold rapidly in a very ordered manner starting from secondary to tertiary structures. As an enhanced folding velocity allows cells to gain biological activity quickly, probably it also reduces the chances of getting stuck in misfolded conformations as less

competing molecular contacts can occur[27]. Supporting this hypothesis, secondary structure interactions like α-helices are formed as soon as the protein is expressed in vivo preventing misfolded conformations[162,163]. A second example could be the cell environment as being regulated to enhance the efficacy of protein folding and discouraging any off-pathway conformations. Changes of the cell environment could be related to pH, oxidation, limited proteolysis, metal ions and osmolytes[27,28]. One must remember that folding *in vitro* occurs in diluted fractions of protein and *in vivo* proteins are surrounded by a crowded environment of other macromolecules[137]. Additionally, during the translation process, protein chains are sequentially synthetized, thus long-range interactions between residues far in sequence does not occur. Considering the exposure of hydrophobic residues during translation, there is a high chance of aggregation sensitive interactions which must be addressed. To this aim, chaperones were evolved with the specific task of helping proteins to fold in their native conformation by optimizing the folding efficiency[160,164]. Specifically, chaperones are involved in *de novo* folding, refolding of stress-denatured proteins, oligomeric assembly, intracellular protein transport and assistance in proteolytic degradation. Finally, when a protein has failed to properly fold, cells have a quality control

system labelling misfolded proteins to degradation, removing dangerous mutants and conformational intermediates[27,28,165].

When those control systems fail, proteins can be trapped in misfolded conformations[137,143] exposing hydrophobic residues to the solvent[8]. Thus, non-native hydrophobic interactions will occur with other misfolded proteins instead of occurring in the protein core. Such uncontrolled and pathological interactions can cause aggregate and insoluble structures including the amyloid fibrils. An unstable conformation is required to cause protein aggregation. But not all misfolded proteins form fibrils as other factors may be involved. These include the role of charged residues providing a repulsive force preventing the aggregation; thus, neutral protein fragments are more aggregation prone[27,166,167]. As a misfolded conformation exposes residues that in the native conformation are in the core, a protease can target such residues. The resulting cleavage releases shorter protein fragments which are more likely to aggregate and further targeted by other classes of proteases[63,168–171].

3. Computational solutions

Considering the difficulties in measuring the entire protein aggregation process at atomistic resolution, Molecular Dynamics (MD) simulations can naturally complement the current experimental approaches. The first bottleneck in applying MD simulations to protein aggregation is the size of the system and the kinetics timescales. Ideally, one would simulate thousands of protein molecules, atomistical represented, including water and ions molecules. Nowadays, one of the biggest systems ever simulated is a detailed aerosol composed of a single SARS-CoV2 virus in a water droplet, containing 1,016,813,441 atoms for only 2.42 ns[172]. Despite the great advances in computational power and the standardization of MD simulations, a suitable system for protein aggregation is not yet feasible for the hardware available. One reason is the different timescales covered by experiments and simulations, as the first can span from seconds to months and the second can reach the microseconds order[173,174]. An estimation has been proposed, to simulate the addition of one molecule of Aβ42 to a fibril tip requires a system containing roughly 60000 atoms and 5 seconds simulation, which is not possible without enhanced sampling, umbrella sampling or metadynamics[173–175]. Those methods require a priori knowledge to bias the

simulation which are not available when studying oligomers formation. Nevertheless, some proteins were characterized such as Alzheimer's Aβ peptide[176–184], α-Synuclein[185–193], amyloid precursor protein (APP)[194,195]; and other proteins by simulating monomers, dimers, and oligomers[196].

Despite the difficulties in building large systems diverse types of MD were employed, from atomistic models of a few molecules in explicit water, to Coarse Grain (CG) simulations describing several molecules at low resolution and implicit water. Specifically on simulating protein aggregation, one must decide either to simulate a small number of molecules and atomistically describe their interactions or to simulate several molecules at low resolution and try to match macroscopic observations.

Beside the computational demand required by an atomistic MD simulation of protein aggregation, two additional problems were recently highlighted[173]. The first one is the inaccuracy of the current force-fields to represent IDPs, and protein-protein interactions as they were developed to simulate single folded proteins. As the interest for both IDPs and multi-protein is increasing, different force-field upgrades were proposed improving the description of folded proteins, disordered proteins, fast-folding proteins, and multidomain proteins[197–200].

Although the monomeric must be as accurate as possible, the aggregation process must be covered by force-fields. Recently, the top 10 most used force-fields were tested and none of them has proven to be suitable for protein aggregation[173]. The second issue regarding an atomistic simulation of protein aggregation is the protein concentration. A system containing a few molecules at low concentration requires a large amount of water reducing the simulation timescale. Thus, highly concentrated system is usually simulated, reducing the time between diffusion-controlled protein encounters and consequently proteins are not able to undergo a coil-to-β transition.

CG models are developed to capture the essential atomistic details with simplified interactions and geometries and reproduce macroscopic experimental properties under different condition[201,202]. The advantages of a simplified model are an increased efficiency, faster dynamics, and larger time steps, at the expense of resolution[201,202]. Different CG models can be build based on the macroscopic property to reproduce including protein folding and aggregation[203], lipid membranes[204,205], polymers[206], and soft matter systems[207]. In general, a CG model compared to a MD force-field has less degrees of freedom or simplified descriptions. Specifically, the Hamiltonian defining a system can have less

parameters (such as the electrostatic) and the protein chain description can be reduced to beads ranging from 1 bead per amino acid to 1 bead per atom (in this case is fully atomistic). An important example could be a lattice model developed to study protein folding. By representing the protein chain in a lattice model one can distinguish to type of interactions: hydrophobic (H) and polar (P). H sidechains can weakly interact with each other, and the native fold maximize such interactions based on the HP sequence. Namely, P interactions tends to interact most with the solvent and determining the maximum collapse of HH contacts. Two of the most popular CG force-field are Martini[208] and SIRAH[209] both reducing the description of molecules at different resolutions. The first is developed to parametrize proteins together with lipids and membranes[210]. The force-field is parametrized by five interactions sites: polar, nonpolar, apolar, charged, and water. Every interaction site is mapped to the molecule by beads which can be of three sizes: regular, small, and tiny. The interaction between each bead is parametrized based an interaction matrix[208,210]. SIRAH was developed to simulate CG models of protein and DNA with a focus on the solvation effect and long range electrostatics[211,212]. It is a physical-based force-field with six-terms Hamiltonian and each amino acid is represented by CG beads of different size corresponding to the Van der Walls (VdW) radii[209,212]. Other CG models are

UNRES[213], OPEP[214], PRIMO[215], MS-CG[216] and REM[217]. Considering the effort used in a CG model to represent macroscopic features, CGnet is emerging as it is developed using a machine learning approach[218]. Specifically, CGnet is trained to maintain all physically relevant invariances and to use a priori knowledge[218].

Among those, protein aggregation was simulated in different CG parametrization such as the aforementioned HP models[219], tube models[220], lattice models[221,222] and minimalistic models[223,224]. Different proteins were studied using a CG model including Aβ42[225], along with a fibril seed[226]; and α-synuclein[193,227].

Structure Based Models

One additional CG model is the structure-based (SB) model or Gō Model. The first approach was developed by Gō, Taketomi, and Ueda as a lattice model to study protein folding[147]. A lattice model can be represented as a 2D grid where each bead is connected by a bond making a linear representation of a protein[228]. Beads that are separated by the unit length and not connected by a bond have a non-local interaction and applying a Metropolis Monte Carlo is possible to simulate the folding process of the protein[147,229]. So, it is concluded that the specific long-range interactions are essential for highly cooperative stabilization

of the native conformation and short-range interactions accelerate the folding and unfolding transitions[228].

Other lattice models are developed after the Gō model where contact energies are not biased based on the knowledge of the protein structure but on the sequence[151,229,230]. Such bias is defined according to the consistency theory based on the observation of the first protein crystal structures[229] and anticipating the aforementioned protein folding process and the energy landscape perspective[229,231]. Proteins achieve their folding with optimal local interactions defining the secondary structure and non-local interactions defining the tertiary structure, both possible as result of evolutionary selection applied on the amino acid sequences[229]. Following this theory, it is possible to define that optimal local and non-local contacts define a minimal frustrated structure, defined minimal frustration principle[141,145]. The latter can be implemented in a Gō Model to perform protein folding simulations in a simplified model[229]. Such implementation can be defined as attractive interactions between amino acids close in space based on the reference native state, usually obtained from a crystal structure. As consequence, assuming the conformational energy minimum is populated by native conformation as the less frustrated one, and the Gō Model

an implementation of the latter; the crystal state defines the minimum energy configuration in a SB simulation. The advantages are typical of a CG model dramatically decreasing the computational cost of simulating large systems for long timescales. Those advantages allow the study of protein folding pathways of homolog proteins[229,232–235], large conformational changes[2,236,237], metamorphic proteins[2,3], the folding upon binding of disordered proteins with different partners[238,239], and nowadays a similar case can be represented by protein aggregation.

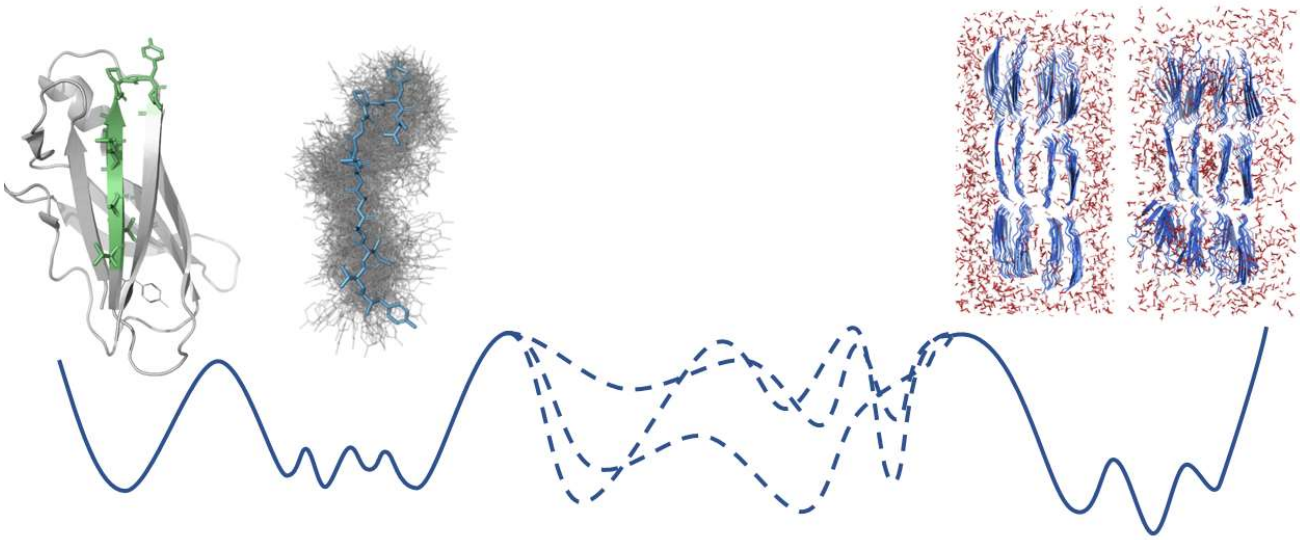
4. Multi-*e*GO as a novel tool

As mentioned above, a CG model can be simplified by reducing the Hamiltonian or the atomistic representation. In this thesis both approaches are applied. Considering the conformational energy minimum in a Gō model and that both the native state and the aggregate state are two energy minima, we followed the previous work of multiple-basin SB models[236] and metamorphic proteins[1,3]. During the PhD I developed multi-*e*GO, a hybrid SB model that includes nonbonded interactions derived from both the dynamics of the native state of the protein in solution and the structure of the amyloid fibril. As multiple structures are included in a single SB model, we implemented and optimized transferable bonded and all-atom interactions to remove any dependencies from the initial conformations[2]. In this thesis I show that multi-*e*GO can be used to follow the aggregation of thousands of molecules as a function of their initial concentration and directly compared to experimental macroscopic measurements.

2. Results and Discussion

"Perché un'altra televisione diversa, è (im)possibile!"

René Ferretti



In this chapter are described the development and issues of the early multi-eGO development followed by a first published version of the method. Following the previous studies on metamorphic proteins[1–3] we build the system as a CG model to minimize the number of simulated atoms. As a reference protein we chose the Transthyretin 105-115 amyloidogenic peptide ($TTR_{105-115}$)[240,241].

The promising results obtained with TTR led me to apply multi-*e*GO to Aβ42. However, the transition from a small peptide to a full-length IDP protein showed the inability of multi-*e*GO to properly simulate a single Aβ42 molecule. To improve the reliability of multi-*e*GO, we added a prior simulation, referred as random coil, during the parametrization of the LJ potential. This improvement, together with the optimization of the local geometry, allowed me to simulate a proper Aβ42 monomer and an initial aggregation using 1000 molecules. These results let me to visit the Department of Chemistry of the University of Cambridge for three months as a period abroad, hosted by prof. Michele Vendruscolo. During this time, I further developed multi-*e*GO by the addition of a small molecule. Despite the ligand parametrization must be optimized, I have demonstrated the possibility to simulate protein aggregation with small molecule and possibly pave the way to a potential drug discovery branch of multi-*e*GO.

Transthyretin

Transthyretin (TTR) has been known for a long time as the first experiments of thyroxine binding were conducted in 1952[242]. In 1969 it was found a second interaction with the retinol-binding protein (RBP)[243] and retinol (Vitamin A)[244] as it is involved in the transport of both retinol and thyroid hormone in plasma[245]. Recently, it has been associated to TTR a neuroprotective role[245]. Also, the diseases associated with TTR have been known for a long time as "A peculiar form of peripheral neuropathy" has been observed in 1939[246] and published in 1952. Soon after, the disease was referred as hereditary/familial amyloid polyneuropathy (FAP) and found to be endemic in northern Portugal[246], Japan[247] and Sweden[248] and reported in several world areas[64]. FAP is caused by the V30M mutation of TTR[64,249] (ATTRV30M) in endemic regions and it is the most common identified variant worldwide[24,25]. Later, up to 130 dominantly inherited mutations causing ATTR amyloidosis (ATTRv, v for variant[4]) were found[64,250,251]. Those mutations could be involved in the destabilization of the native structure of TTR, disrupting the tetrameric assembly and leading to aggregation[252–254]. ATTR amyloidoses can be classified in two different classes depending on whether the protein is wild-type (ATTRwt) or with mutations (ATTRv)[64]. ATTRwt causes the senile systemic amyloidosis (SSA)[30], an age-related disease

where amyloid deposition occurs in tissues with high mechanical stress[255]. Among those, the heart muscle is commonly affected, and the presence of amyloids cause cardiomegaly and death due to the organ failure[30,256,257]. ATTRv amyloidosis is a hereditary, autosomal dominant, and adult-onset systemic disease and thus they can be referred to the aforementioned FAP[64]. ATTRv aggregates deposit and accumulate mainly in peripheral nerves and heart but also in eyes, kidneys and systemically. In general, the clinical presentation is diverse, but overall ATTRv leads to system failure[16,24,25,64]. Until 2011 the only treatment for hereditary ATTRv was liver transplant with different survival rates depending on the type of mutations[258–260]. Nowadays there are two additional approaches to treat ATTRv[24,25,64,245]: the first method is the reduction of the amount of TTR produced through target gene silencing[261,262]; the second through the stabilization of the TTR protein preventing the aggregation[244,263,264] by small-molecules[265–267].

The structural knowledge of the TTR protein has a significant role in the development of new treatments. The first X-ray structure has been characterized in 1978[268] describing a homotetrameric protein of 55 kDa and composed of 127 amino acids. Specifically, the tetramer is the formation of dimers of dimers. The structure is determined by eight β-strands arranged in two antiparallel β-sheets

and one short α-helix[268]. The two dimers are assembled around the central channel of the protein determining two hydrophobic pockets were two thyroxine (T4) molecules binds[240] whilst two RBP bind on the external surface of TTR[269–272]. It is noteworthy that RBP bound to retinol stabilizes the TTR tetramer[244] and the stabilization of TTR is used as a strategy to ameliorate the related illnesses[273].

In 1987 two distinct types of ATTR fibrils were found from FAP and SAP patients[274]. The two fibrils were defined as type A, consists in C-terminal ATTR fragments and full-length ATTR (FL-ATTR); while Type B consists only in FL-ATTR[275–277]. Both fibril types are conserved in all organs and do not change over time. Those findings suggest the presence of at least two aggregation mechanisms. Fibrils containing C-terminal fragments are found mostly in ATTRwt fibrils and in the majority of ATTRv. While Type B fibrils are mostly composed of ATTRV30M and ATTRT114C.

To structurally validate the two ATTR fibrils type, early experiments of ATTR fibril formation in vitro were done in 1991 using both the FL-ATTR and different derived fragments[278]. The first low resolution description of the ATTRV30M fibril was made in 1995[279] using transmission electron microscopy (TEM). Fibrils displayed four protofilaments, a hollow fibril core, and a square-

like cross-section. Soon after, in 1996 and 1997 it was measured the X-ray diffraction confirming the cross-β structure and four left-hand twisted cross-β-sheets[76,78]. Nowadays, there are three high-resolution structures of ATTR. The first one describes a small fragment derived from the G strand of TTR[89]; the second one using an in vitro fibril obtained from the FL protein[280]; and finally, the Cryo-EM structure derived from brain tissue of a patient[100].

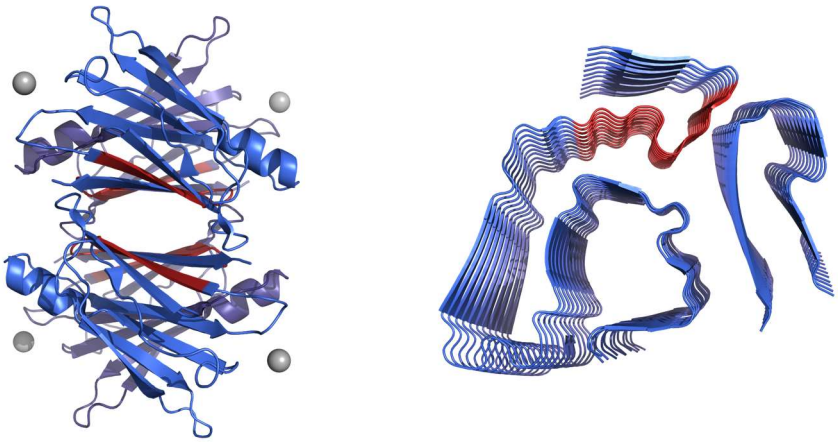


Figure 1 – Structure of the full-length Transthyretin protein. On the left the tetramer complex (PDB 4TLT)[167], on the right the Cryo-EM amyloid structure obtained from a patient with hereditary V30M amyloidosis[100]. The G strand corresponding to the $TTR_{105-115}$ fragment is colored in red.

As TTR has being studied for a long time, it is one of the reference models employed to study protein aggregation. The molecular mechanism of TTR aggregation is yet to be elucidated. However, two mechanisms were hypothesized:

the first one is based on tetramer dissociation as the rate-limiting step[281]; the second one involves a cleavage process of the CD loop, accelerating the tetramer dissociation[63,276,282].

The first mechanism can be explained as mutations in TTR affect its stability resulting in dissociation of the tetramer into unfolded monomers and dimers, promoting the formation of insoluble intracellular and potentially toxic aggregates[61,283–285]. The broad plethora of destabilizing mutations can lead to different structural deformation pathways[281]. Thus, it could explain the different tropism for specific organs, the age at onset and other disparities in the prognosis of the disease[281]. An unstable TTR protein can form dimer intermediates prone to further aggregation[286,287]. During aggregation, the formation of oligomers of distinct size has been observed[56,61,285,287]. Also, it has been observed that nonfibrillar TTR can deposit before amyloid fibril formation[15]. Eventually, those oligomers can form protofibrils and fibrils[15,257].

As type A fibrils contains C-terminal TTR fragments *in vivo*[171,257,275–277,288]; the second mechanism is supported by the observation of a proteolytic cleavage can accelerate the formation of cytotoxic oligomers and consequently protein aggregation [63,168,169,171]. Plasmin and trypsin could be the proteases involved in

TTR cleavage between residues 48-49[168]. The resulting aggregates of the C-terminal does not show significant differences from the full-length one, suggesting that the cleavage destabilize the tetrameric structure and accelerates the oligomer formation rather than defining a distinct second aggregation mechanism[63].

Being TTR one of the reference models in protein aggregation, studies led to the characterization of the aggregation propensity of different TTR segments[167,278,289]. Specifically, the fragment we employed in this PhD project is the $TTR_{105-115}$ fragment, corresponding to the aforementioned G β-strand of the FL-TTR. This peptide was among the first to be characterized in 1991[278] where the aggregation was described at neutral pH. Following this study, NMR analysis of the $TTR_{105-115}$ fragment highlighted the dominant random-coil structure with sporadic turn or helical elements[289]. Also, there is no propensity in forming β-strands in solution. It is noteworthy that this study highlights an absence of water solubility and the necessary use of acetonitrile/water solution[289]. As mentioned earlier in the paragraph, the $TTR_{105-115}$ fibril was characterized using solid state NMR (ssNMR)[89] and Cryo-EM and used as a reference in the current PhD project.

Three models of this fibril were proposed, as a doublet, triplet, and quadruplet of protofilaments. The arrangement of the cross-β sheets is antiparallel.

Aβ42

Aβ is key to the pathogenesis of Alzheimer's disease together with tau[290]. AD is the main cause of dementia[22,290,291] and an age-related disease[292]. Aβ peptide was first identified in 1984 as constituent of meningovascular amyloid and subsequently in amyloid neuritic plaques[293]. The pathobiology is complex and heterogeneous with two types of amyloid depositions as primary etiological agent: the extracellular one composed of Aβ molecules, and the neurofibrillary tangles (NFTs) composed of tau[290]. Tau deposits prevents cell to receive nutrients and other essential molecules inside the neuron and Aβ blocks synapses[292]. Aβ is derived from its precursor protein APP as a result of different sequential proteolytic cleavage by the γ-secretase complex[96,294,295]. Such complex can target different protein site resulting in different sizes of Aβ where the 40 and 42 residues long are the most represented (Aβ40 and Aβ42 respectively). Aβ fragments are produced mainly in endosomes and released from neurons based on synaptic activity[296,297]. Moreover, APP mutations can increase the ratio of Aβ42 to Aβ40[298], the overall concentration of Aβ42[299] and the consequent

assembly into filaments[300]. Such deposits in the brain can impair the physiological activity of neurons, including glucose metabolism, and impair the brain circuitries[301]. Aβ deposition can be detected up to 22 years before the symptom's onset[302]. Once the Aβ fibrils accumulates in the extracellular environment, there is a subsequent intracellular tau deposition following the AD onset[303,304]. Indeed, other factors are reported to participate to Aβ deposition and thus involved in AD development, including the e4 gene variant of apolipoprotein E (APOE)[290,305–308]. Amyloid deposits result in chronic inflammation activated by the microglia and the consequent atrophy[301] compromising the normal brain functions. The brain initially compensates the nerve cells damage until the latter is too severe, starting the cognitive decline[292]. Additionally, to the cognitive decline, different symptoms are consequent to the broad brain damage[290], memory loss, confusion as to time or place, basic body functions such as swallowing, and a set of behavioral symptoms. The latter can be depression, personality changes, and loss of interest in activities the patients used to enjoy[292,301]. Together, the development of these symptoms allows the classification of AD onset in three different stages: pre-clinical, mild cognitive impairment, and dementia due to AD[309,310]. Among the newest discoveries on the onset of AD, the role of Aβ against infections is being revived[290]. The first study

in 1991 demonstrated the presence of herpesvirus in brains of patients with AD, as well as within the amyloid plaques[311–314]. Recently a molecular network analysis shed light on the putative roles of viral-host interactions involving the innate immunity and APP processing[315]. Subsequently to a viral infection it is observed an increase of Aβ deposition[316]. This correlates with the presumed role of Aβ as an antimicrobial peptide against bacteria, fungus, and herpesvirus[317–320]. Thus, it is possible to hypothesize Aβ aggregation as an antimicrobial defense mechanism[321]. Additionally, it is observed that HSV-1 viral particles can catalyze Aβ42 aggregation in vitro[322]. Another related role of Aβ with the innate immunity is the interaction with the gut microbiome and AD pathogenesis[290]. The gut-brain axis links gut microbiome with the innate immune system of the central nervous system (CNS) and may modulate AD pathogenesis[290]. Difference abundance and diversity of microbial species is found between AD patients and control[323,324].

Currently, four FDA-approved medications are available to treat cognitive impairment and dysfunction to treat symptomatic AD[290]. Despite the effort by the pharmaceutical industry, today there is no treatment available to effectively modify AD onset[325]. Instead, there are symptomatic treatments acting on

cholinesterase inhibitors, NMDA receptor modulator[290]. Clearly, a need of a new strategy for drug development in AD is emerging. Also, there one must consider that phase 3 trials intervened on patients with symptomatic AD, meaning that the amyloid deposition in the brain started 15-20 years before the trials[302,326]. Considering the onset of clinical symptoms is due to irreversible synaptic and neural loss, the ability of a drug to reverse the effect is unlikely[327]. Crucial clinical interventions should be applied in the preclinical phase of AD to prevent irreversible brain damage[290]. Additionally, one must remember that there are other factors besides aggregation that can lead to AD and can be targets for therapeutic intervention[290]. Considering the broad diffusion of AD, several new drugs are under trials. At the beginning of 2022 there were 143 agents in 172 clinical trials for AD, as the majority designed to be disease-modifing[328]. It is important to mention the anti-amyloid monoclonal antibody aducanumab, considered the first disease-modifying therapy to be approved for AD released in the market in 2021[329]. Two more monoclonal antibodies are under review by the FDA[330–332].

Both Aβ40 and Aβ42 are intrinsically disordered, thus the characterization of the native conformation ensemble is determined by NMR in presence of

disaggregation agents. Overall, the NMR data shows the lack of secondary or tertiary structures. Thus, the conformation of Aβ peptides is largely random, with an extended chain structures and turns or bend-like structures at residues Asp7-Glu11 and Phe20-Ser26[333]. As IDPs lack of a stable native conformation, are described in terms of structural ensembles and free energy landscapes. MD is a particularly useful tool to describe such ensembles including the Aβ42 ones which have been recently published along with kinetic ensembles describing the transition between each transient structure[183]. Such ensembles allowed to simulate the interaction of a small molecule and potentially explain the protective role against aggregation[184].

Three major types of Aβ inclusions can be found in the brain of AD patients: diffuse and focal deposits in the parenchyma and vascular deposits. Diffuse deposits are characterized by lightly packed Aβ filaments. Focal deposits are in the form of dense core plaques of tightly packet filaments surrounded by loosely packet filaments. When the disease further reaches an advanced state, both diffuse and focal Aβ deposits are systemically spread. Amyloid fibrils in neuritic plaques are 8 nm in diameter with different putative oligomers surrounding the fibril[334]. The structure of a synthetically derived Aβ42 is characterized by Cryo-EM

displaying a diameter of 7 nm long with two twisted protofilaments consisting of Aβ42 monomers in an "LS" shape[335]. Cryo-EM structures of Aβ40 and Aβ42 aggregates from patients are available[96,336]. Interestingly, the fibrils are polymorphic and different compared to the one formed in vitro and the protofilament are similar[336]. In a first set of Cryo-EM fibrils, different peptides are found including Aβ40, Aβ38, Aβ(2-40), Aβ37, Aβ36, and Aβ39. Such data suggest that chain length variations are the most abundant modifications. Specifically, three morphologies were modelled. The first one is relatively abundant with a diameter of 7.4 nm and a crossover of 41.5 nm and contains Aβ40 molecules. Morphology II is characterized by a diameter of 12.3 nm and a crossover of 129 nm. Morphology III structure displays a diameter of 17.9 nm and a crossover of 142.8 nm. In a second set of Cryo-EM fibrils, two types of Aβ42 fibrils from brain of patients with different AD are characterized[96]. The first type is isolated from individuals with sporadic AD a predominance of twisted Aβ fibrils with two identical S-shaped protofilaments is observed. The second one was obtained from individuals with familial AD and other conditions and differs from the first between the orientation of a few peptide groups affecting the secondary structure assignment. Aβ40 and Aβ42 fibrils are structurally different[96].

Aβ42 is characterized by an increased hydrophobicity and thus aggregation propensity with respect to Aβ40 and plaques presents a less dense core compared to fibrils made of both Aβ40 and Aβ42[112,290,298,337]. The deposition of amyloid fibrils follows spatiotemporal spreading, suggesting a seeded aggregation similarly to prion disease[338–340]. Furthermore, different aggregate structural variation of Aβ can have different biochemical features and the ability to induce distinct pathological phenotypes[338].

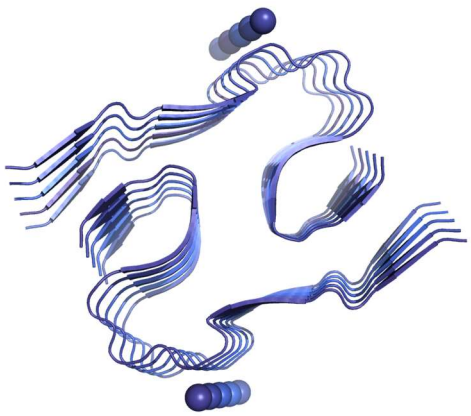


Figure 2 – Image of Aβ42 fibril from human brain from Yang et al.[96]. The structure is modelled from G9 to A42 residues. Each filament is made of two identical protofilaments arranged in the S-shape. The secondary structure is composed of 5 β-strands of at least three residues. The S-shape is determined by hydrophobic clusters.

Aβ peptides aggregates into high-order oligomers, protofibrils and fibrils[57,341]. Oligomers are reported to be cytotoxic, and the control of their dynamics is

covering an important role against AD onset[57,110]. According to the primary role in AD, the oligomerization process of Aβ42 has been recently described[110]. Most of the oligomers dissociate into their monomeric precursors without forming any fibrils. Considering that a fibril must originate from an oligomer, results shows that only a minority of oligomeric species are able to shift to an amyloid fibril. Specifically, to Aβ42, new steps of oligomerization are found and potentially exploited to develop new drugs against AD onset. Aβ42 aggregation occurs in two-step mechanism involving oligomers as necessary intermediates in the fibril formation. The first step is the generation of oligomers through the interaction of monomers with the catalytic surface of pre-formed fibrils. A small but critical concentration of amyloid fibrils can trigger the formation of oligomeric species creating a positive feedback loop[342]. Later, there is a competition between oligomer dissociation to monomers and the formation of fibrillar aggregates and the interconversion to fibrillar aggregates occurs in the same timescale of the overall aggregation. Thus, the fraction of oligomers converting from non-fibrillar forms to growth-competent fibrillar oligomers is determined be the rates of their conversion and dissociation. Oligomers are a heterogeneous population of different size and structurally different[110].

1. Multi-GO with a CG model

Following the previous studies on metamorphic proteins[1–3], we defined the SB models of the native TTR105-115 (PDB ID: 4TLT)[167] and the amyloid (PDB ID: 2M5K)[89] using SMOG 2 tool[343,344]. Conventional SB models are able to build a force field using only one input structure and simulate a single molecule. Ideally, merging the two models would be enough to allow the simulation to start from an initial -native- state to a final -aggregate- one. As consequence we stated the novelty in our method by the name multi-GO. Our first attempt was to use a Coarse Grain (CG) representation due to the necessity of simulating a large number of molecules. We chose to represent a single bead mapped to the Cα per residue. However, the very first simulations using only 16 molecules with multi-GO showed the inability of a proper representation using a CG model, resulting in a complete collapse of the fibril as shown in Figure 3. This result clearly indicates the inability of the Cα resolution to represent the complexity of the interactions occurring in an amyloid fibril. The antiparallel conformation of the two β-sheets is maintained, but as the beads were allowed to interact with each other in a self-contact, a collapse of the fibril is the optimal conformation for this

simple representation. Thus, we moved to a heavy atom representation of the molecules.

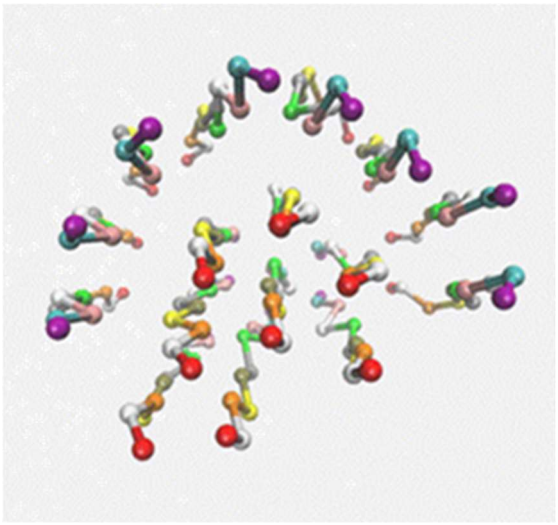


Figure 3 – CG model of $TTR_{105-115}$ peptide. Beads are mapped according the Cα of each residue. The N and C terms are colored in purple and red respectively. This arrangement covers all the parametrized LJ in a mistaken manner as the fibril is collapsed.

2. Multi-*e*GO: an *in-silico* lens to look into protein aggregation kinetics at atomic resolution

In the following paper I contributed to the development of the initial model multi-GO and the refined multi-eGO model. Wet-lab experiments were performed in collaboration with Stefano Ricagno and Sara Pellegrino groups.

As shown below, multi-GO results provided kinetics and fibrils of the aggregation of $TTR_{105-115}$ peptide and laid out a methodological way forward to protein aggregation. However, the poor representation of the dynamical behavior of the peptide delivered unreliable results. Thus, we implemented an explicit water MD of the $TTR_{105-115}$ peptide during the parametrization. Each contact was rescaled based on their probability, enhancing the description of our model referred as multi-eGO. This update provides comparable kinetics and fibril structures obtained with experimental data.

PNAS RESEARCH ARTICLE | BIOPHYSICS AND COMPUTATIONAL BIOLOGY

Multi-eGO: An in silico lens to look into protein aggregation kinetics at atomic resolution

Emanuele Scalone[a], Luca Broggini[a,b], Cristina Visentin[a,b], Davide Erba[a], Fran Bačić Toplek[a], Kaliroi Peqini[c], Sara Pellegrino[c], Stefano Ricagno[a,b,1], Cristina Paissoni[a,1], and Carlo Camilloni[a,1]

Edited by Susan Marqusee, University of California, Berkeley, CA; received February 25, 2022; accepted May 17, 2022

Protein aggregation into amyloid fibrils is the archetype of aberrant biomolecular self-assembly processes, with more than 50 associated diseases that are mostly uncurable. Understanding aggregation mechanisms is thus of fundamental importance and goes in parallel with the structural characterization of the transient oligomers formed during the process. Oligomers have been proven elusive to high-resolution structural techniques, while the large sizes and long time scales, typical of aggregation processes, have limited the use of computational methods to date. To surmount these limitations, we here present multi-*e*GO, an atomistic, hybrid structure-based model which, leveraging the knowledge of monomers conformational dynamics and of fibril structures, efficiently captures the essential structural and kinetics aspects of protein aggregation. Multi-*e*GO molecular dynamics simulations can describe the aggregation kinetics of thousands of monomers. The concentration dependence of the simulated kinetics, as well as the structural features of the resulting fibrils, are in qualitative agreement with in vitro experiments carried out on an amyloidogenic peptide from Transthyretin, a protein responsible for one of the most common cardiac amyloidoses. Multi-*e*GO simulations allow the formation of primary nuclei in a sea of transient lower-order oligomers to be observed over time and at atomic resolution, following their growth and the subsequent secondary nucleation events, until the maturation of multiple fibrils is achieved. Multi-*e*GO, combined with the many experimental techniques deployed to study protein aggregation, can provide the structural basis needed to advance the design of molecules targeting amyloidogenic diseases.

protein aggregation | molecular dynamics | aggregation kinetics | structure-based models | amyloids

Significance

Alzheimer's and Parkinson's diseases are incurable pathologies associated with the aberrant aggregation of specific proteins into amyloid fibrils. Understanding the mechanisms leading to protein aggregation, by characterizing the structures of the oligomeric species populated in the process, would have a tremendous impact on the design of therapeutic molecules. We propose that a structure-based molecular dynamics simulations approach can allow the aggregation kinetics of thousands of monomers to be followed at high resolution. Having shown that simulations can describe the aggregation of a Transthyretin amyloidogenic peptide, we demonstrate how they can provide a wealth of structural information. We foresee that integrating the latter with the many techniques developed to study protein aggregation will support the design of molecules to modulate amyloidogenesis.

Amyloid fibril formation is a highly specific self-assembly process, requiring a large degree of similarity between the interacting amino acid sequences (1). Amyloids, resulting from the uncontrolled transition of normally soluble proteins, were originally found to be associated with neurodegenerative diseases (2, 3). More recently, they have also been associated with several physiologic functions (4, 5). Amyloid fibrils share a cross-β architecture in which β-strands are oriented perpendicularly to the fibril axis, allowing the formation of a dense intermolecular hydrogen bond network with sidechains contributing to both intramolecular and intermolecular interactions (6, 7). In vitro, the amyloid fold seems to be accessible to a large number of, if not all, proteins (ordered or disordered) or even short sequences of amino acids (8, 9). Thermodynamic considerations, indeed, suggest that native proteins are metastable species under physiological conditions, with the global free-energy minimum corresponding to their amyloidogenic state (10).

Protein aggregation into amyloid fibrils is an inherently dynamic process. Many interconverting species of differing sizes and structures can be populated over multiple time scales (11). The description of amyloid fibril formation thus requires an understanding of the properties of the end states, that is, monomers and fibrils, and of the different oligomeric species that are transiently populated in between. Remarkably, in diseases like Parkinson's, Alzheimer's, type 2 diabetes mellitus, and cardiac amyloidosis, some oligomeric species may be the primary pathogenic agents (12–16). Furthermore, toxic oligomers have been found in model proteins and associated to specific physicochemical properties like size and hydrophobicity, although it is not yet clear whether these are relevant for all amyloidogenic diseases (17–19). Structural approaches based on solid-state NMR (ssNMR) and cryogenic electron microscopy (cryo-EM) have revealed the atomic structures of amyloid fibrils formed by different proteins in diverse conditions (6, 7, 20–22). The aggregation process itself can only be studied at very low resolution, by aggregation kinetics assays, where experimental conditions are tuned to induce the in-solution interconversion of protein monomers into amyloid fibrils. Seeds obtained by previously formed fibrils can also be employed to catalyze the interconversion (23).

Author contributions: E.S., C.P., and C.C. designed research; E.S., L.B., C.V., D.E., F.B.T., and K.P. performed research; E.S., L.B., and C.V. analyzed data; E.S., S.P., S.R., C.P., and C.C. wrote the paper; and S.P., S.R., C.P., and C.C. supervised the project and provided resources.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: stefano.ricagno@unimi.it, cristina.paissoni@unimi.it, or carlo.camilloni@unimi.it.

This article contains supporting information online at http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2203181119/-/DCSupplemental.

Published June 23, 2022.

Chemical kinetics analyses provide a framework to dissect the microscopic mechanisms at play in fibril formation (24). Aggregation is described by a network of microscopic processes including primary nucleation and elongation, as well as secondary nucleation processes, such as fragmentation and surface-induced nucleation. By globally fitting multiple, accurate, aggregation kinetic traces obtained for multiple initial monomer concentrations, it is possible to estimate the rates for the different microscopic processes and use these to interpret the macroscopic observations. Such analyses have highlighted the preference of proteins associated with amyloidogenic diseases to aggregate through secondary nucleation mechanisms, whereas physiological amyloids are proposed to be mainly controlled by primary nucleation (25, 26). Drug design strategies have been implemented, based on the kinetic modulation of such mechanisms (27). Nonetheless, despite its power, chemical kinetics fails to provide detailed structural information on the species at play during the process.

Given the inherently transient and dynamic nature of the species populated in an assembly process, molecular dynamics (MD) simulations naturally complement current experimental approaches (27, 28). Aggregation kinetics simulations have mainly been employed to characterize the early events in the oligomerization of a few peptides at high concentrations, because of the combination of challenges resulting from system sizes and relevant time scales. Implicit solvent models have been employed to mitigate these problems and have allowed the oligomerization of 20 monomers of Aβ40 and Aβ42 over hundreds of nanoseconds to be studied (29). Simulating larger systems over longer time scales requires coarse-grain (CG) models. Notably, fibrils may, in fact, be formed by tens of thousands of monomers.

CG simulations, with simplified interactions and geometries, including HP models (30), tube models (31), and lattice models (32, 33), as well as other minimalistic approaches (34, 35), have been used to make hypotheses on the general principles of protein aggregation, also informing chemical kinetics models (36). A recent overview can be found in ref. 28. In the field of protein folding simulations, the most adopted CG models are structure based (SB), also known as Gō models (37–41), recently reviewed in ref. 38. SB models are an implementation of the principle of minimal frustration [or the folding funnel (42)]: Attractive interactions are defined only between amino acids or atoms that are close in space in the native crystal state; consequently, the minimum energy configuration is the native crystal configuration. This allows folding and unfolding transitions to be efficiently studied, by dramatically decreasing the cost of evaluating interactions and accelerating the overall diffusion in conformational space; for example, the folding time of a protein can be rescaled from milliseconds to hundreds of nanoseconds.

In keeping with the observation that the amyloid structure is the global free-energy minimum of a protein at high concentration (10), we here describe multi-*ϵ*GO, a hybrid SB model that includes nonbonded interactions derived from both the dynamics of the soluble protein and the structure of the amyloid fibril, and transferable bonded interactions that are optimized to reproduce the results of state-of-the-art explicit solvent molecular force fields. While SB models, including more than one reference structure and/or hybrid terms, have been employed to study differences in protein folding pathways of homolog proteins (37–41), large conformational changes (43–45), metamorphic proteins (46, 47), and the folding upon binding of disordered proteins with different partners (48, 49), here we show that multi-*ϵ*GO can be used to follow the aggregation of thousands of monomers, as a function of their initial concentration, at high resolution. Our results are qualitatively in agreement with experiments and enable the structural investigation of the aggregation of proteins into amyloid fibrils.

Results

To develop multi-*ϵ*GO, we used the Transthyretin 105–115 amyloidogenic peptide ($TTR_{105–115}$) (50, 51). Transthyretin is a well-studied amyloidogenic protein responsible for both sporadic and genetic cardiac and systemic amyloidosis (52). $TTR_{105–115}$ has often been used as a model system to study aggregation, and three amyloid polymorphisms have been determined at atomic resolution by a combination of multiple techniques including ssNMR and cryo-EM (20). NMR analysis of monomeric $TTR_{105–115}$ in solution indicates that it primarily populates a random-coil structure with a low percentage of turns or helical elements (53). Multi-*ϵ*GO was built to include information from the structure or the dynamics of the end states and uses them to infer the properties of the intermediate oligomeric states (*SI Appendix*, Fig. S1). To have a realistic conformational ensemble reference of monomeric $TTR_{105–115}$, we performed an explicit solvent MD simulation using the a99SB-*disp* force field (54). This simulation well represented the behavior of $TTR_{105–115}$ in solution, showing a broad flexibility and sporadic turns, as reported by the radius of gyration distribution and the per-residue contact probability map in Fig. 1, as well as by a secondary structures populations analysis (*SI Appendix*, Fig. S2*A*), in agreement with previous NMR chemical shifts measurements (53).

Multi-*ϵ*GO Reproduces the Conformational Dynamics of $TTR_{105–115}$ in Solution. Following previous studies on metamorphic proteins (46, 47), we initially defined the multi-GO SB force field, at all heavy atom (nonhydrogens) resolution, as a combination of terms obtained from two reference structures (*Materials and Methods*), namely, the protein in its native monomeric state (extracting $TTR_{105–115}$ coordinates from Protein Data Bank [PDB] 4TLT, corresponding to the crystal structure of TTR) and the amyloid fibril (PDB 2M5K) (*SI Appendix*, Fig. S1). A multi-GO simulation of a $TTR_{105–115}$ monomer explored only extended configurations with an average radius of gyration of 1.05 nm, in comparison with 0.83 nm of the a99SB-*disp* one (Fig. 1*A*) and the conformational ensemble did not show long range contacts (*SI Appendix*, Fig. S3*A*). The multi-GO ensemble described above did not capture the conformational freedom of the monomeric state, and consequently may not capture that of early intermediate oligomeric states.

To increase the descriptive power of the model, we introduced multi-*ϵ*GO as a hybrid transferable/SB model (*Materials and Methods*). The most relevant differences are that all bonded interactions, and, in particular, proper dihedral angles, are transferable, while nonbonded interactions are learned from a reference simulation for the monomeric state, that is, the a99SB-*disp* force field simulation introduced above, and a reference amyloid fibril structure (PDB 2M5K). Remarkably, while the multi-GO simulation explored only extended configurations, the multi-*ϵ*GO model could better recapitulate $TTR_{105–115}$ dynamics in solution, with an average radius of gyration of 0.90 nm. The contact probability map for a99SB-*disp* and multi-*ϵ*GO, shown in Fig. 1*B*, as well as the secondary structures populations analysis in *SI Appendix*, Fig. S2*A*, indicate that multi-*ϵ*GO can also qualitatively describe the intramolecular transient interactions of the peptide.

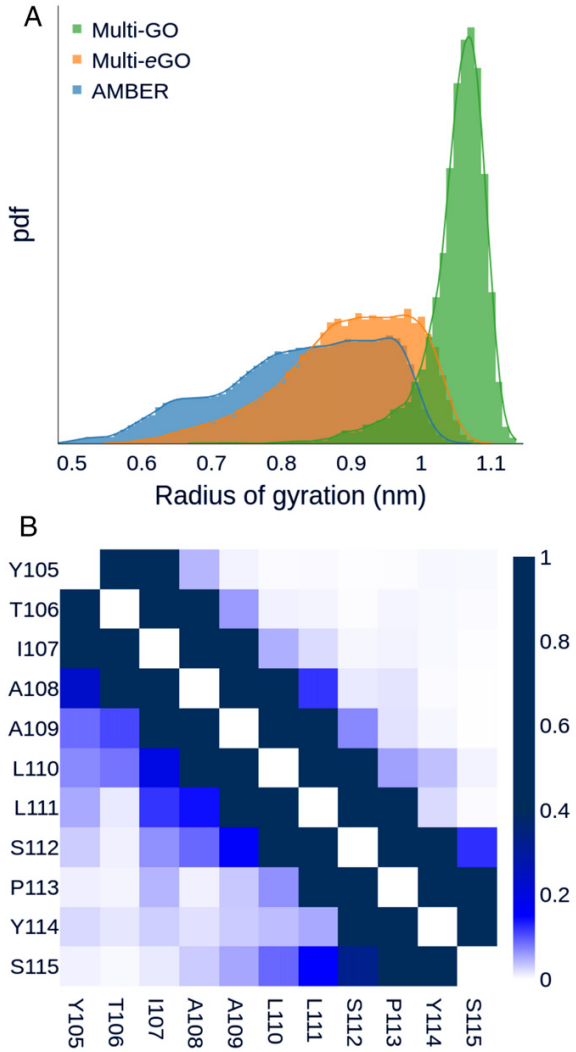


Fig. 1. $TTR_{105–115}$ peptide monomer dynamics. (*A*) Gyration radius distribution of $TTR_{105–115}$ peptide conformational ensemble according to multi-GO (green), multi-*e*GO (orange), and a99SB-*disp* (blue) simulations. The multi-GO distribution describes an open conformation with a single peak at 1.07 nm. The a99SB-*disp* simulation shows multiple peaks over a broad range of values. The multi-*e*GO distribution is shifted toward more extended conformations than a99SB-*disp* but still shows a broad range of values in qualitative agreement with the former. (*B*) Per-residue probability contact map for the a99SB-*disp* (*Lower Left*) and multi-*e*GO (*Upper Right*) simulations.

A qualitative comparison between the interaction energies of the multi-GO and multi-*e*GO $TTR_{105–115}$ models was obtained, summing all the native pairs strengths for all pairs of amino acids (*SI Appendix*, Fig. S3*B*). This showed limited differences in the interaction among residues between the two models. This is expected, as they are both trained on the same fibril structure. This analysis suggests that the transferable bonded interactions of multi-*e*GO are key to improve the agreement of the model with the a99SB-*disp* simulation.

Multi-*e*GO Can Simulate $TTR_{105–115}$ Aggregation. Using multi-*e*GO, a total of 15 simulations (each involving 4,000 $TTR_{105–115}$ monomers) as a triplicate of five different concentrations, between 7 and 13 mM, were produced (*SI Appendix*, legends for Datasets S1 and S2). The resulting aggregation kinetics are shown in Fig. 2*A* as the number of monomers forming assemblies, from decamers to larger ones, at a given time. It should be noted that the time scale of the simulations is only nominal, and comparisons with experimental data should consider a scaling factor. Simulations displayed sigmoidal concentration-dependent kinetics, where an increase in monomer concentration resulted in a reduction of the lag phase. We also observed that the variability of the curves increased inversely with the initial monomer concentration (55). From the resulting curves, we obtained the half-time, $\tau_{1/2}$, shown in Fig. 2*B*, and the growth rate, *r*, as the slope of the straight line fitting the region of the curve around $\tau_{1/2}$ (*SI Appendix*, Fig. S4). The double log plot of $\tau_{1/2}$ as a function of concentration (Fig. 2*B*) showed a bilinear trend with a change of slope, the scaling exponent γ, at concentrations lower than 8.5 mM, suggesting that, at high monomer concentration, a dominant aggregation mechanism becomes saturated (56).

To test the ability of multi-*e*GO to capture differences between seeded and unseeded aggregation kinetics, we also performed three seeded simulations at 7 mM by adding a 10-monomer oligomer seed, obtained from previous 13-mM simulations (compare *Multi-eGO Can Provide Structural Details for $TTR_{105–115}$ Aggregation Kinetics*, in the following). As shown in Fig. 2*C*, the addition of the seed led to a marked decrease of $\tau_{1/2}$ and a reduction of its variability. The observed growth rate *r*, instead, remained the same as for the unseeded simulations performed at the same concentration (i.e., 7 mM).

Multi-*e*GO $TTR_{105–115}$ Simulations Can Form Polymorphic Fibrils. The 18 simulations performed at five different concentrations yielded a total of 41 distinct fibrils. These fibrils grew in length

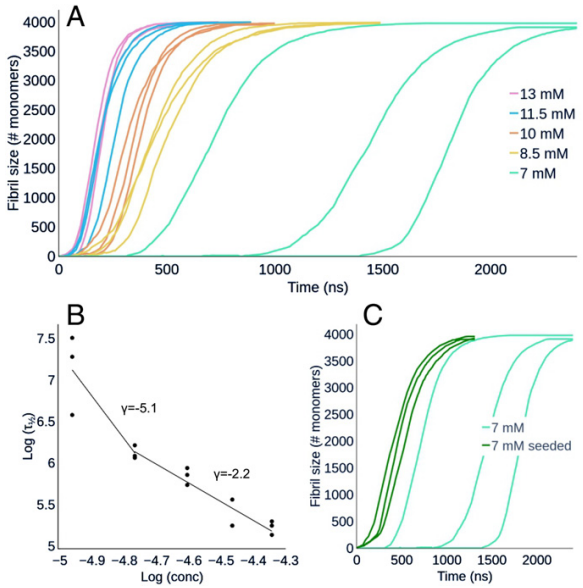


Fig. 2. (*A*) Simulated aggregation kinetics. Curves represent the number of monomers involved in an aggregate of at least 10 monomers as a function of nominal simulation time. See also *SI Appendix*, legends for Movies S1 and S2. (*B*) Log–log plot of the half-times, $\tau_{1/2}$, as a function of the initial monomer concentration. The points are fitted with two straight lines in the range 7 mM to 8.5 mM and 8.5 mM to 13 mM. (*C*) Aggregation kinetics of seeded and unseeded 7-mM simulations. Curves represent the number of monomers involved in an aggregate of at least 10 monomers as a function of nominal simulation time. The addition of the seeds reduces $\tau_{1/2}$ and makes it less variable compared to the unseeded simulations, leaving the slope of the growth is unaffected. See also *SI Appendix*, legend for Movie S3.

Table 1. Summary of the main structural features of the fibrils discussed in this work

Source	Number of fibrils	Length (Å)	Number of β-sheets in filaments	Number of filaments in fibril	Twist (deg)
13 mM	16	163 to 368	4 to 10	2 to 5	−0.18 to −0.73
11.5 mM	9	210 to 390	5 to 13	2 to 6	−0.23 to −0.85
10 mM	6	170 to 415	4 to 9	2 to 6	−0.2 to −0.79
8.5 mM	4	260 to 515	6 to 14	2 to 4	−0.15 to −0.54
7 mM	3	435 to 480	7 to 17	4 to 6	−0.11 to −0.48
7 mM seeded	3	443 to 485	13 to 14	5	−0.10 to −0.80
ssNMR model			4	2 to 4	−0.85
TEM*			4	3 to 8	−0.71 to −0.81

The first six rows indicate fibrils formed in silico in our multi-*e*GO simulations; ssNMR is for the fibrils corresponding to PDB codes 2M5K, 2M5M, and 3ZPK; and TEM are those observed in vitro in this work.
*Values for TEM are estimates; *SI Appendix*, Table S2.

from 163 Å to 515 Å, with an average length of 360 Å (Table 1 and *SI Appendix*, Table S1). All fibrils displayed the expected cross-β topology with a parallel and in-register stacking of chains in the same β-sheet as shown in Fig. 3. The average distance between β-strands in the cross β-sheet was 4.7 Å. Facing β-sheets were antiparallel and shifted by 2.5 Å, resulting in the even-numbered sidechains of one peptide interacting with the odd-numbered sidechains of the two opposite peptides. According to previous nomenclature (20), we define a protofilament as a structure made up of two antiparallel β-sheets; the further addition of two β-sheets in a protofilament determines a filament (Fig. 3*A*). The β-sheet content in a filament was shown to vary from 4 to 17, with an average of 10. Furthermore, filaments that grew in the peptide chain direction through interactions between the N- and C-terminal residues (Fig. 3*B*), could form a fibril. This head-to-tail interaction resulted from the Y105 sidechain interacting with both the S115 carboxyl group and the Y105 sidechain of the facing β-sheet. The number of filaments in a fibril varied from two to six, with an average of four. Mature fibrils displayed a twist per monomer between −0.1° and −0.85°, measured as the torsion angle between two vectors obtained from Y105-Cα and S115-Cα carbons of subsequent molecules in the same β-sheet. Single filaments displayed a more pronounced twist of −5° compared to mature fibrils (Fig. 3). At higher concentration, we saw the formation of more fibrils, indicating that more nuclei are produced than at lower concentration (*SI Appendix*, Fig. S5). Since the monomer number was fixed at 4,000, the fibrils grown at higher concentration were shorter in length than the ones obtained at lower concentration (Table 1 and *SI Appendix*, Table S1). At higher concentration, we also observed fibrils adhering together (Fig. 3*C*). Again, given the fixed and relatively small number of monomers, some protofilaments were not able to become fibrils, due to monomer depletion. We did not observe any specific differences in the fibrils formed at 7 mM in seeded and unseeded simulations.

Compared to the reference model determined by ssNMR and cryo-EM (20), the only remarkable difference is that our fibrils do not display any wet cavity within filaments. The cavity in the reference model accommodates structured water molecules that interact with exposed sidechains. In our model, all sidechains in a filament are tightly packed; therefore, we also observed a variable number of β-sheets in a filament, whereas the reference model always contained four. Indeed, there is evidence of such variations (57, 58). The reference model illustrates a structural polymorphism, based on the number of filaments, from doublet to quadruplet. In our simulations, we saw the same polymorphism but extended to six filaments in a single fibril.

Finally, we compared the multi-*e*GO and multi-GO aggregation kinetics by performing four multi-GO simulations of $TTR_{105–115}$ aggregation as function of the concentration (*SI Appendix*, Fig. S2*B*). Multi-GO quickly formed aggregates at concentrations, between 2 and 0.5 mM, that are lower than those used in multi-eGO simulations and experiments, that is, 10 mM to 20 mM (53). Furthermore, the multi-GO amyloid fibrils grew laterally instead of elongating (*SI Appendix*, Fig. S2*C*), indicating that multi-GO cannot correctly capture the $TTR_{105–115}$ aggregation process.

$TTR_{105–115}$ In Vitro Aggregation Experiments Recapitulate Multi-*e*GO Simulations. To validate the in silico aggregation kinetics, we performed aggregation assays monitored by Thioflavin T (ThT) fluorescence (Fig. 4). The $TTR_{105–115}$ peptide was incubated at 37 °C at different concentrations (i.e., 13, 10, and 7 mM), and ThT fluorescence was monitored over 150 h. ThT fluorescence increased over time, indicating (Fig. 4*A*) concentration-dependent aggregation kinetics. The lag phase at 13 mM was considerably shorter compared to at 10 mM or 7 mM. The fluorescence plateau was reached faster in the most concentrated samples, whereas, at the lowest concentration tested (i.e., 7mM), the plateau was not observed during the overall incubation time. The mean values of three independent experiments were subjected to nonlinear regression analysis, using a Boltzmann sigmoidal equation. From the regression, we derived the experimental $\tau_{1/2}$ of 33.7 ± 4.3 h, 62.0 ± 16.6 h, and 125.7 ± 10.2 h for 13, 10, and 7 mM, respectively. Fig. 4*B* shows a linear correlation between peptide concentrations and half-times in a double log plot, with the slope γ comparable to that obtained from simulations in the range 8.5 mM to 13 mM (namely, −2.0 and −2.2 for the experiments and simulations, respectively; compare Fig. 2*B*). This is of note given the relative simplicity of our model and the fact that it does not include any specific information about the kinetics of the process.

The aggregates of the $TTR_{105–115}$ peptide obtained by the aggregation kinetics experiments were negatively stained and analyzed by transmission electron microscopy (TEM). As reported in Fig. 4, we observed remarkable polymorphism in all conditions tested. Morphological analysis identified six main different types of structures (Fig. 4 *C–E* and *SI Appendix*, Table S2). The mean width at the cross-over is 37 ± 4 Å, as previously observed by Fitzpatrick et al. (20). The fibrils cross-over in the six polymorphs ranges from 1,041 ± 24 Å to 1,185 ± 43 Å; the diameter varies considerably, ranging from 115 ± 11 Å to 326 ± 21 Å. Representative pictures of each identified morphology are reported in Fig. 4*F*, and the main fibril parameters are summarized in *SI Appendix*, Table S2. Notably, the observed widths

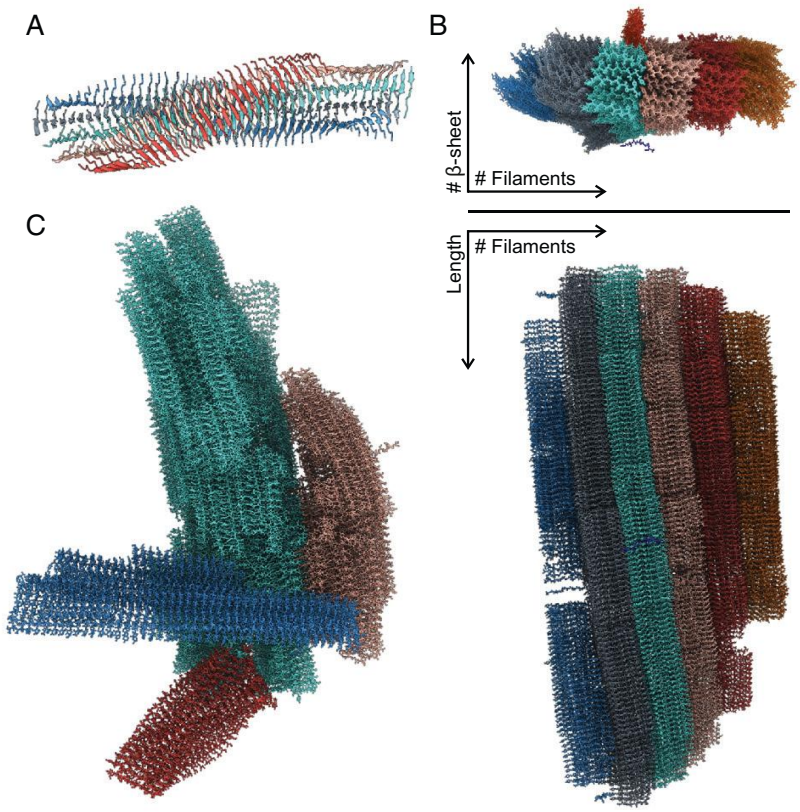


Fig. 3. (*A*) A filament model observed at the end of a simulation. Colors indicate the five different β-sheets composing the filament. (*B*) Top and side view of a mature fibril with colors indicating the different filaments. From the top view, it is possible to see a peptide which is about to attach to the fibril and a protofilament which is perpendicular to the main fibril. (*C*) Multiple mature fibrils (each represented with a different color) interacting with each other from one of the 13-mM simulations.

between cross-overs correspond to multiples of the peptide chain length, suggesting the presence of up to eight aligned filaments (*SI Appendix*, Table S2).

Multi-*e*GO Can Provide Structural Details for $TTR_{105–115}$ Aggregation Kinetics. Having shown that multi-*e*GO could simulate the aggregation of $TTR_{105–115}$ from monomers to fibrils, with structural and kinetic features compatible with experimental data, it was possible to observe the structures populated along the self-assembly process in detail. In Fig. 5*A* and *SI Appendix*, Fig. S6, the number of monomers, dimers, and trimers in our aggregation kinetics are shown as a function of time. We observed that the number of free monomers displayed a sigmoidal behavior, symmetric with respect to that of the fibril size. The number of dimers and trimers showed, instead, a noisy but relatively constant trend until the end of the lag phase (t_{lag}), defined as the intersection between a straight line, tangent to the aggregation kinetic curve at $\tau_{1/2}$ with slope r, and the time axis, and quickly dropped after this time. This suggests that, once fibrils start to grow, most monomers contributed to the fibril growth instead of forming new oligomers, and oligomers formed before t_{lag} dissolve over time.

The time-resolved distribution of oligomer sizes of the first 13-mM simulation before t_{lag} (*SI Appendix*, Fig. S7 for all simulations) is shown in Fig. 5*B*. This analysis allows the emergence and growth of primary nuclei to be followed and suggests that fibrils stem from primary nuclei composed of around 10 monomers. Assuming the simulations prior to t_{lag} at equilibrium, and thus averaging over this time window, we observed how, at all concentrations, dimers and trimers were the most represented oligomeric species, with populations in the 5 to 10% and 0.5 to 2% range, respectively. Higher-order oligomers were scarcely populated, stressing the need to simulate large numbers of monomers to study aggregation (compare Fig. 5*C*). Interestingly, apart from their populations, oligomers do not show other concentration-dependent properties (*SI Appendix*, legend for Dataset S3). In *SI Appendix*, Fig. S8 are shown the distributions of the radius of gyration, and the average β-populations per residue, of oligomers, ranging from dimers to decamers, populated in multi-*e*GO aggregation kinetics before t_{lag}. The analyses indicate that dimers, trimers, and tetramers are extended and disordered; around 20% of pentamers can form intermolecular β-sheets, and this fraction increases with higher oligomer orders. The increase in β-structure is also reflected by the distributions of the radius of gyrations that become less and less broadly distributed (*SI Appendix*, Fig. S8).

The structures of oligomers involved in primary nucleation are shown in Fig. 5*D*. All primary nuclei displayed two antiparallel β-sheets. Observing the trajectory, we were able to describe their formation. Free monomers spontaneously assemble into small oligomers, forming the first β-sheet. Once the β-sheet reaches a size of five to six monomers, other monomers interact with the β-sheet surface, triggering the formation of a second

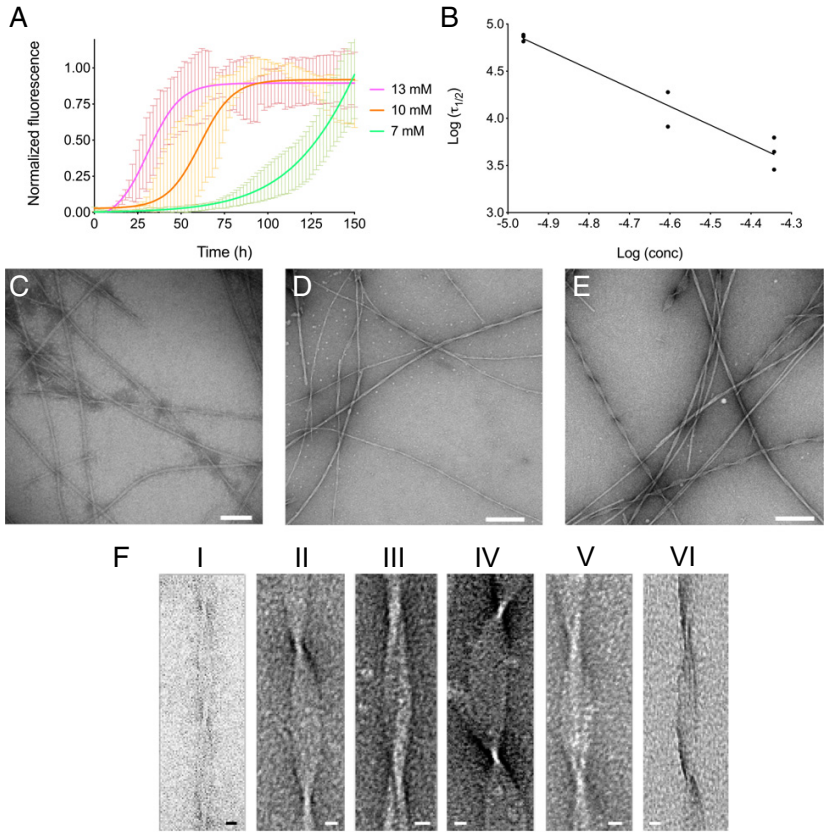


Fig. 4. TTR peptide aggregation kinetics in vitro. (*A*) Aggregation kinetics of the $TTR_{105–115}$ peptide at 13, 10, and 7 mM are shown in magenta, orange, and green, respectively. TTR peptide at 37 °C were obtained by monitoring ThT fluorescence. The mean value of three independent experiments analyzed by linear regression using Boltzmann sigmoidal equation is reported. (*B*) Log–log plot of the in vitro half-times, $\tau_{1/2}$, as a function of the initial monomer concentration. (*C–E*) Electron micrographs of fibrils formed by $TTR_{105–115}$ peptide incubated at 13 mM (*C*), 10 mM (*D*), or 7 mM (*E*) at 37 °C for 150 h. (Scale bars, 100 nm, *C;* 200 nm, *D* and *E*.) (*F*) Representative TEM images of the six main fibrillar morphologies. The detailed structural parameters of each morphology are reported in *SI Appendix*, Table S2. (Scale bars: 10 nm.)

β-sheet docked by sidechain/sidechain interactions. Once a primary nucleus is formed, its growth can be followed. Each β-sheet provides two ends for elongation, so primary nuclei have four ends. With regards to elongation, peptides generally dock at the N terminus toward the C terminus as shown in Fig. 6*A*. A generally highly twisted, cross-β protofilament exposes sidechains and termini for secondary nucleation, but we observed that only the sidechain faces could trigger the formation of further β-sheets, thus forming a filament (Fig. 6*B*). The addition of each β-sheet decreased the twist. Filaments, made up of at least four β-sheets, can further growth, both through their sidechains faces (Fig. 6*B*) and through their termini, as exemplified in Fig. 6*C*. Growth can occur by N to N terminus (head-to-head) as well as N to C terminus (head-to-tail) interactions. Importantly, a newly N to C β-sheet can grow into a new protofilament, while a newly formed N to N β-sheet must firstly convert into an N to C sheet before further growth can occur. A fibril is thus formed when two filaments are linked head to tail. Remarkably, the formation of new β-sheets always occurs with monomers sliding on the surface before eventually docking. At high monomer concentration, we subsequently observed interactions between fibrils. Fragmentation events were not observed in any simulations.

To test whether the described mechanism is consistent with the macroscopic kinetics shown in Fig. 2, we performed a chemical kinetics analysis of our simulated data using Amylofit (59–61). Simulations could only be globally fitted using a "multistep secondary nucleation, unseeded model" as shown in *SI Appendix*, Fig. S9. This is compatible with the positive curvature with an increased slope at lower concentration displayed by $\tau_{1/2}$ in a double log plot (compare Fig. 2*B*) that can be interpreted as the saturation of secondary nucleation (56) at high monomer concentration (e.g., all the catalytic fibril surface is occupied by monomers). Furthermore, our seeded simulations did not show variations in the rate constant r (compare Fig. 2*C*), supporting the hypothesis that secondary nucleation is a multistep process with a first step (monomer attachment on the surface) that is concentration dependent and a second step (monomers rearrangement on the surface) that is concentration independent (56). Amylofit analysis correlates with our observations where the addition of molecules on the cross-β surface implies the exploration of different conformations prior to latching on to and starting to form a new oligomer. Globally, the $TTR_{105–115}$ aggregation process described by our simulations is consistent with the hierarchy proposed by Fitzpatrick et al. (20).

Discussion

Amyloidogenesis is the result of an out-of-equilibrium, concentration-dependent process; thus, it cannot be easily followed by high-resolution structural biology techniques (6). Indeed, while NMR,

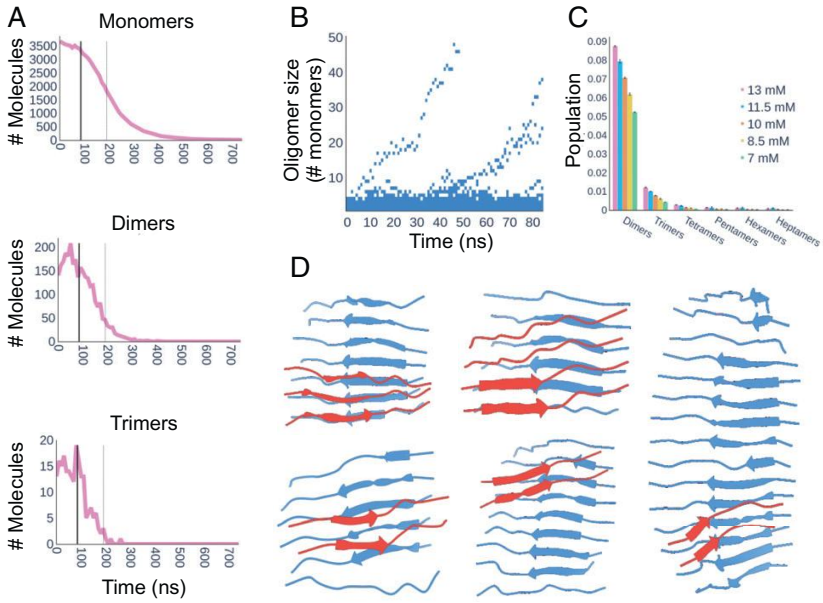


Fig. 5. (*A*) Number of monomers, dimers, and trimers over time of the first 13-mM simulation. The t_{lag} (solid thick line) and $\tau_{1/2}$ (dashed thin line) are reported in each plot. The monomer decrease is symmetrical compared to fibril growth (compare Fig. 2*A*). The number of dimers and trimers is relatively constant until t_{lag} and quickly drops after this time. (*B*) Time-resolved evolution of oligomer order of the first 13-mM simulation before t_{lag}. (*C*) Oligomer order populations in the time window from zero to t_{lag} and averaged over the three replicates. The different colors represent the different concentrations, while the error bars represent the SD over the three simulations. In this time window, only monomers and dimers are significantly populated (*C*); nonetheless, one can observe the emergence of two fibrils (*B*). (*D*) Five representative structures of primary nuclei extracted from five different simulations. The β-sheets colored in blue, representing the initial oligomer, are stabilized as primary nuclei by interacting with the second β-sheet, colored in red.

X-ray crystallography, and cryo-EM have been instrumental to investigating, at high resolution, the early steps associated with protein misfolding, and the structures of resulting amyloid fibrils, they have only provided low-resolution information about the transient oligomers populated along the process (62, 63). In this respect, most of the fundamental mechanical understanding of protein aggregation is based on the combination of multiple low-resolution techniques, in particular, aggregation kinetics studied by ThT fluorescence, and chemical kinetics analysis (24).

MD simulations could provide the resolution, in time and space, to observe the emergence of oligomers, nuclei, and fibrils from a solution of monomeric proteins (64). This would greatly assist the understanding of the determinants of the different aggregation mechanisms at play, the observation of the effect of mutations, and, subsequently, the SB design of drugs targeting specific oligomeric species. Unfortunately, computer power is far from being able to enable such simulations, using conventional classical mechanics transferable force fields (64).

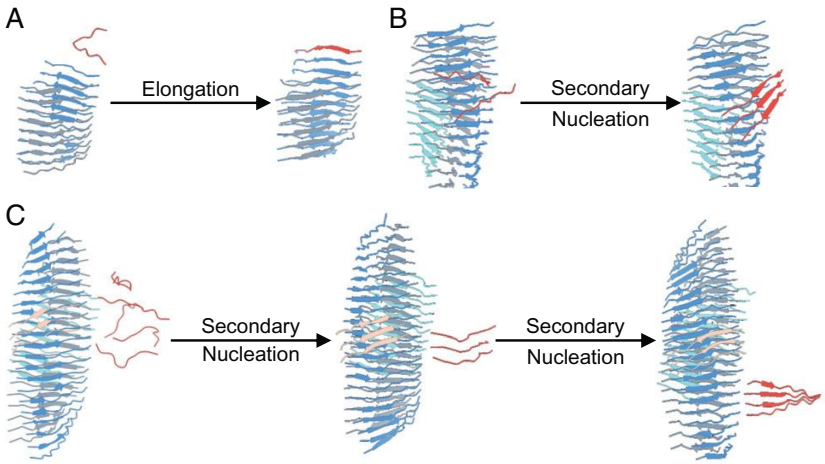


Fig. 6. Fibril growth secondary processes. (*A*) Elongation: the docking of a single monomer in red at one end of a β-sheet in blue. (*B* and *C*) Surface-induced secondary nucleation can occur both on the front surface of a β-sheet (*B*) and on the side surface (*C*). Peptides can slide on the surface before locking. At least three peptides are required to form a secondary nucleus.

Simulations have thus been employed to study the oligomerization of a few peptides (29, 65) and the interactions of peptides with preformed fibrils (66), and, in most cases, to understand the determinants of protein aggregation by only studying protein monomers (67–72). These studies, while valuable, fall short in providing indications about oligomers that are, by their nature, extremely unlikely and short lived. Alternatively, CG simulations have been used to investigate the general principles of protein aggregation, but without providing system specific information (28, 30–36).

In this study, we set out to develop a simplified force field that can allow protein aggregation, comprising thousands of monomers, to be studied with current state-of-the-art computing resources. Our force field builds on the success of SB (Gō) models to study protein folding and binding (38, 41, 48, 73, 74). The multi-ϵGO force field introduced here is 1) at atomic resolution (excluding hydrogens); 2) locally transferable, with bonded and excluded volume interactions derived from a transferable force field or optimized consequently; and 3) structure (or ensemble) based using multiple references and symmetrized so that, once an interacting pair is defined, this can be formed both intramolecularly as well as intermolecularly. We have shown, in Fig. 1, that this combination can describe the conformational ensemble of a disordered peptide, in qualitative agreement with more-accurate conventional explicit solvent MD simulations. Most importantly, multi-ϵGO can describe the aggregation kinetics of thousands of monomers, showing concentration-dependent features that are compatible with experimental data (Figs. 2–4). Indeed, the comparison of the scaling exponents γ derived from experimental and simulated $\tau_{1/2}$ data (Fig. 2*B* and Fig. 4*B*) showed comparable values, confirming the robustness of the model, with a difference observed only at the lowest concentration. TEM morphological analysis of the fibrils highlighted a remarkable degree of polymorphism. Specifically, we classified six main fibril morphologies with highly variable cross-over and width (Fig. 4*F* and *SI Appendix*, Table S2). These data indicate that each fibril is formed by three to eight filaments, in agreement with in silico observed fibrils (Fig. 3, Table 1, and *SI Appendix*, Tables S1 and S2). On the contrary, all morphologies share a mean width at the cross-over compatible with previous measurements and associated with a hydrated cavity (20) that is not observed in silico, indicating that future improvements should try to better account for solvation effects that may only be indirectly captured by our native pair interactions (41).

Simulations can be subsequently used to formulate hypotheses on the oligomeric species populated along the process and provide a structural model for the mechanisms of primary and secondary nucleation. Interestingly, we can show how primary nuclei are on the order of 10 monomers and organized into two opposed β-sheets (Fig. 5*D*). These nuclei are populated for less than 0.1% in the lag phase of the kinetics, in comparison to dimers and trimers that are populated around 10 and 1%, respectively (Figs. 5 *A*–*C* and *SI Appendix*, Figs. S6–S8). This indicates how relevant it is to simulate aggregation using large monomer numbers. Oligomer populations drop immediately after the formation of the first nuclei. Following the growth of primary nuclei, our model also shows that elongation occurs through the preferential binding of the N-terminal region of the peptide (Fig. 6). Once a protofilament is formed, secondary nucleation can be observed. Secondary nucleation arises, initially, by the formation of nuclei on the exposed sidechains of the filament surface; then monomers can slide over the surface and eventually dock into it, triggering the formation of an additional β-sheet layer (Fig. 6). Once at least four β-sheet layers are formed, we observe additional secondary nucleation events, catalyzed by interactions with the N and C termini (Fig. 6). Secondary nucleation contributes to overall fibril growth, but it does not lead to independent oligomer formations that detach to form new protofibrils. We hypothesize that this is a size effect, resulting from the relatively small number of monomers, that are shortly depleted by fibril formation. Nonetheless, using 4,000 monomers allowed us to observe multiple primary and secondary nucleation events, with secondary nucleation remaining the main effect that contributes to the exponential growth of the fibril mass, suggesting that this number may be large enough to recapitulate the key processes occurring during the aggregation process. Notably, the mechanisms inferred by applying a chemical kinetics analysis on the simulated aggregation kinetics agrees with what was observed in the simulation (*SI Appendix*, Fig. S9), suggesting that multi-ϵGO may complement and support experimental chemical kinetics models to provide a high-resolution time-resolved description of the microscopic processes at play during aggregation (24, 36).

In conclusion, we present the development of multi-ϵGO, an SB model tailored to study amyloid-type protein aggregation. The model is promising in describing, at least qualitatively, the spontaneous aggregation of monomers into amyloid fibrils, as a function of the initial monomer concentration, thus providing a structural picture of the populated oligomeric species and of the associated aggregation mechanisms. We anticipate that our model can be combined with methodologies that allow the integration of simulations with the many complementary experimental techniques deployed to study protein aggregation (75–78). Eventually, the computational efficiency of multi-ϵGO, combined with the availability of amyloid fibrils structures (79, 80), and of already run—and publicly available—long MD simulations for disease relevant proteins (81, 82), will improve our understanding of the mechanisms and the associated oligomeric structures at play in different pathogenic and nonpathogenic self-assembly processes.

Materials and Methods

MD simulations in this work were performed with GROMACS (83). Models' parameterization and preparation was developed in python. All scripts, including ad hoc analysis tools, are freely available on GitHub (https://github.com/emalacs/multi-eGO/tree/ITR_paper). Simulations are available on Zenodo (*SI Appendix*, legends for Datasets S1–S3).

Multi-GO: A Multireference Gō-Like Model for Protein Aggregation. Multi-GO is a multireference SB force field, at all heavy atom (nonhydrogens) resolution, defined as a combination of terms obtained from two reference structures, namely, the protein in its native monomeric state and the amyloid fibril. This was originally developed to study metamorphic proteins (46, 47) using SMOG software (84). In this model, distances between covalently bonded atoms, as well as angles formed by three subsequent covalently bonded atoms, are derived only from the monomeric structure, because they describe the local geometry that is generally independent from the specific configuration. Dihedral angles are defined as in SMOG but obtained from both structures and halving the force constant to account for the double counting. Native pairs are obtained from both reference structures following SMOG rules; if two atoms are in contact in both structures, then the distance is defined as the minimum distance. All native pairs are symmetrized so that, if atom *i* and atom *j* are in contact in one reference structure, they can interact irrespectively of whether the two atoms belong to the same monomer or to two different monomers; such an approach has been successfully employed to describe domain swapped dimers (85, 86), and is needed to make intramolecular and intermolecular interactions indistinguishable.

Multi-*e*GO: An Enhanced Gō-Like Model for Protein Aggregation. In contrast to multi-GO, the multi-*e*GO force field is partitioned so that, while nonbonded interactions are SB, bonded interactions are, instead, transferable. The multi-*e*GO Hamiltonian, given a reference monomer simulation X_m and a reference amyloid structure X_a, is defined as

$$H(X;X_m,X_a) = \sum_{bonds} K_r(r-r_0)^2 + \sum_{angles} K_q(\theta-\theta_0)^2 + \sum_{improper} K_\phi[1+\cos(n\phi-\phi_0)] + \sum_{dihedrals} K_\psi[1+\cos(n\psi-\psi_0)] + \sum_{native} \varepsilon_n\left[\left(\frac{r_{ij,m,a}}{r_{ij}}\right)^{12} - 2\left(\frac{r_{ij,m,a}}{r_{ij}}\right)^{6}\right] + \sum_{others} \frac{C_{ij}^{(12)}}{r_{ij}^{12}},$$

where the parameters for bonds, angles, and improper and proper dihedrals are obtained from a transferable force field, specifically, GROMOS54A7 (87), that is a united atom force field already optimized without nonpolar hydrogens. Thus, the local geometry no longer depends on the monomeric structure as in multi-GO. Proper dihedrals terms describing the ϕ and ψ backbone angles were reoptimized as describe in the next section. Interactions between native pairs are defined for pairs of atoms farther than one residue along the sequence (if they belong to the same molecule) and closer than 5.5 Å in either the native monomeric simulation or in the amyloid structure. The interactions strength for native pairs is either $\varepsilon_n = \varepsilon$ when a pair is derived from the amyloid structure or heuristically scaled as

$$\varepsilon_n = \varepsilon\left(1 - \frac{\ln P}{\ln P_{\text{threshold}}}\right) \quad [1]$$

when derived from a reference MD simulation. Here P is the fraction of frames in the simulation where the pair of atoms is closer than 5.5 Å, and $P_{threshold}$ is a minimum fraction that should be considered, set to 0.09 in this work. This approach has the merit of giving an interaction strength equal to ε if the contact population is one, as for a single structure, and to smoothly go to zero when reaching the chosen threshold. The interaction length $r_{ij,m,a}$ is defined as either the average pair distance calculated over the P frames or the pair distance in the amyloid structure. If a native pair is defined multiple times, then only the one associated with the shorter $r_{ij,m,a}$ distance is retained. Care must be taken so that, if a native interaction is defined between two atoms belonging to the same or the neighbor residue, this interaction applies only intermolecularly; otherwise, once a pair interaction is defined in its length and strength, these parameters are the same irrespective of whether the interaction is formed intermolecularly or intramolecularly. Finally, excluded volume interactions for all other pairs, that is, all pairs of atoms that are not native and that are separated by more than three consecutive covalent bonds (*i* to *i*+4), are defined as in the GROMOS54A7 force field. Among the excluded volume interactions, we include all *i to i*+4 interactions involving a backbone nitrogen; in this case, the $C^{(12)}$ Lennard-Jones parameter is scaled down by a factor of 0.15. This is needed to effectively account for the missing amide hydrogen, and it is critical to avoid nonphysical configurations. Of note is that, in multi-*e*GO, masses are correctly set to include hydrogens; also, since force constants obtained from GROMOS54A7 are tuned to work at room temperature, that is, ∼300 K, ε, the reference interaction strength between all native pairs, is the only free parameter to be set in a system-dependent manner.

Multi-*e*GO Backbone Dihedrals Optimization. In the multi-*e*GO Hamiltonian, the intramolecular interactions between atoms belonging to consecutive amino acids are described only by transferable terms. Therefore, a dipeptide simulated with multi-*e*GO should closely mimic the conformational freedom of a dipeptide simulated at room temperature using a transferable force field in explicit solvent. Due to the SB nonbonded addition, parameters for the proper dihedral angles have been optimized building on the former hypothesis. Alanine, glycine, and proline dipeptides were simulated using CHARMM22* (88) and TIP3P (89) at 300 K for 1 μs each, and the resulting Ramachandran distribution was set as our target.

The same dipeptides were simulated using multi-*e*GO, initially setting the force constant K of the potential V_D describing proper dihedrals for the backbone ϕ and ψ angles to zero. V_D is defined as $V_D(\vartheta) = K(1+\cos(n\vartheta-\phi_0))$, with force constant K, multiplicity n, and phase ϕ_0.

Target and multi-*e*GO Ramachandran distributions are then compared calculating the following scoring function S (i.e., the cross-entropy):

$$\mathcal{S} = P(T)\log\frac{P(T)}{P(E)},$$

where $P(E)$ and $P(T)$ are the multi-*e*GO and target Ramachandran distributions. To optimize the parameters K, n, and ϕ_0 for the proper dihedral angles, we then followed an iterative procedure combining a Monte Carlo (MC) optimization followed by MD (90, 91). In detail, the effect of a given choice of the parameters is estimated by analytically reweighting the last multi-*e*GO MD simulations as

$$w(\varphi,\psi) = \exp\left[-\frac{(V_D^i(\varphi)+V_D^i(\psi)) - (V_D^{i-1}(\varphi)+V_D^{i-1}(\psi))}{k_BT}\right],$$

where V^i_D is the potential energy from the ith iteration written as the sum of multiple proper dihedral terms, k_B is the Boltzmann constant, and T is the temperature of the MD simulation. Optimal parameters are searched by MC under the constraint that the effective information N_{eff}, calculated over the N configurations generated by the last MD, as

$$N_{eff} = \frac{[\sum_{i=1}^N w_i]^2}{\sum_{i=1}^N w_i^2},$$

is greater than 0.6. This allows choosing parameters that do not dramatically alter the starting distribution. A new MD simulation is then performed with the chosen parameters, and the procedure is repeated until convergence of the scoring function (90, 91).

Data Availability. MD simulation trajectories data have been deposited in Zenodo [DOI: 10.5281/zenodo.6125955 (92), DOI: 10.5281/zenodo.6125424 (93), and DOI: 10.5281/zenodo.6414572 (94)].

ACKNOWLEDGMENTS. Funding was provided to C.P. and C.C. by Fondazione Cariplo (Grant CoronAld) and by University of Milano - Linea 1; C.C. is also supported by Fondazione Telethon (Grant GGP19134). This work was partially supported by Fondazione ARISLA (Project TDP-43-STRUCT) and Italian Ministry of Research (PRIN 20207XLJB2) (S.R.). E.S., C.P., and C.C. acknowledge CINECA for an award under the ISCRA initiative, for the availability of high-performance computing resources and support. E.S., C.P., and C.C. acknowledge PRACE for awarding them access to Piz Daint at CSCS, Switzerland. We acknowledge Louise Gourlay (University of Milano), Pietro Sormanni (University of Cambridge), Guido Tiana (University of Milano), and Michele Vendruscolo (University of Cambridge) for their suggestions and insights.

Author affiliations: [a]Dipartimento di Bioscienze, Università degli Studi di Milano, 20133 Milano, Italy; [b]Institute of Molecular and Translational Cardiology, IRCCS Policlinico San Donato, 20097 San Donato Milanese, Italy; and [c]Dipartimento di Scienze Farmaceutiche, Sezione Chimica Generale e Organica, Università degli Studi di Milano, 20133 Milano, Italy

1. F. Chiti, C. M. Dobson, Protein misfolding, amyloid formation, and human disease: A summary of progress over the last decade. *Annu. Rev. Biochem.* **86**, 27–68 (2017).
2. M. Brunori, From Kuru to Alzheimer: A personal outlook. *Protein Sci.* **30**, 1776–1792 (2021).
3. T. P. J. Knowles, M. Vendruscolo, C. M. Dobson, The amyloid state and its association with protein misfolding diseases. *Nat. Rev. Mol. Cell Biol.* **15**, 384–396 (2014).
4. D. Otzen, R. Riek, Functional amyloids. *Cold Spring Harb. Perspect. Biol.* **11**, a033860 (2019).
5. S. K. Maji *et al.*, Functional amyloids as natural storage of peptide hormones in pituitary secretory granules. *Science* **325**, 328–332 (2009).
6. D. S. Eisenberg, M. R. Sawaya, Structural studies of amyloid proteins at the molecular level. *Annu. Rev. Biochem.* **86**, 69–95 (2017).
7. R. Gallardo, N. A. Ranson, S. E. Radford, Amyloid structures: Much more than just a cross-β fold. *Curr. Opin. Struct. Biol.* **60**, 7–16 (2020).
8. M. Fändrich, C. M. Dobson, The behaviour of polyamino acids reveals an inverse side chain effect in amyloid structure formation. *EMBO J.* **21**, 5682–5690 (2002).
9. F. Chiti *et al.*, Designing conditions for in vitro formation of amyloid protofilaments and fibrils. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 3590–3594 (1999).
10. A. J. Baldwin *et al.*, Metastability of native proteins and the phenomenon of amyloid formation. *J. Am. Chem. Soc.* **133**, 14160–14163 (2011).
11. D. Willbold, B. Strodel, G. F. Schröder, W. Hoyer, H. Heise, Amyloid-type protein aggregation and prion-like properties of amyloids. *Chem. Rev.* **121**, 8285–8307 (2021).
12. D. M. Hartley *et al.*, Protofibrillar intermediates of amyloid beta-protein induce acute electrophysiological changes and progressive neurotoxicity in cortical neurons. *J. Neurosci.* **19**, 8876–8884 (1999).
13. J. Janson, R. H. Ashley, D. Harrison, S. McIntyre, P. C. Butler, The mechanism of islet amyloid polypeptide toxicity is membrane disruption by intermediate-sized toxic amyloid particles. *Diabetes* **48**, 491–498 (1999).
14. J. D. Knight, A. D. Miranker, Phospholipid catalysis of diabetic amyloid assembly. *J. Mol. Biol.* **341**, 1175–1187 (2004).
15. T. Tokuda *et al.*, Detection of elevated levels of α-synuclein oligomers in CSF from patients with Parkinson disease. *Neurology* **75**, 1766–1772 (2010).
16. M. Maritan *et al.*, Inherent biophysical properties modulate the toxicity of soluble amyloidogenic light chains. *J. Mol. Biol.* **432**, 845–860 (2020).

17. M. Bucciantini *et al.*, Inherent toxicity of aggregates implies a common mechanism for protein misfolding diseases. *Nature* **416**, 507–511 (2002).
18. S. Campioni *et al.*, A causative link between the structure of aberrant protein oligomers and their toxicity. *Nat. Chem. Biol.* **6**, 140–147 (2010).
19. F. Bemporad, F. Chiti, Protein misfolded oligomers: Experimental approaches, mechanism of formation, and structure-toxicity relationships. *Chem. Biol.* **19**, 315–327 (2012).
20. A. W. P. Fitzpatrick *et al.*, Atomic structure and hierarchical assembly of a cross-β amyloid fibril. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 5468–5473 (2013).
21. A. W. P. Fitzpatrick *et al.*, Cryo-EM structures of tau filaments from Alzheimer's disease. *Nature* **547**, 185–190 (2017).
22. P. Swuec *et al.*, Cryo-EM structure of cardiac amyloid fibrils from an immunoglobulin light chain AL amyloidosis patient. *Nat. Commun.* **10**, 1269 (2019).
23. F. Chiti, C. M. Dobson, Protein misfolding, amyloid Formation, and human disease: A summary of progress over the last decade. *Annu. Rev. Biochem.* **86**, 27–68 (2017).
24. T. C. T. Michaels *et al.*, Chemical kinetics for bridging molecular mechanisms and macroscopic measurements of amyloid fibril formation. *Annu. Rev. Phys. Chem.* **69**, 273–298 (2018).
25. M. Törnquist *et al.*, Secondary nucleation in amyloid formation. *Chem. Commun. (Camb.)* **54**, 8667–8684 (2018).
26. M. Andreasen *et al.*, Physical determinants of amyloid assembly in biofilm formation. *MBio* **10**, e02279-18 (2019).
27. M. Carballo-Pacheco, B. Strodel, Advances in the simulation of protein aggregation at the atomistic scale. *J. Phys. Chem. B* **120**, 2991–2999 (2016).
28. I. M. Ilie, A. Caflisch, Simulation studies of amyloidogenic polypeptides and their aggregates. *Chem. Rev.* **119**, 6956–6993 (2019).
29. B. Barz, Q. Liao, B. Strodel, Pathways of Amyloid-β aggregation depend on oligomer shape. *J. Am. Chem. Soc.* **140**, 319–327 (2018).
30. H. D. Nguyen, C. K. Hall, Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16180–16185 (2004).
31. S. Auer, F. Meersman, C. M. Dobson, M. Vendruscolo, A generic mechanism of emergence of amyloid protofilaments from disordered oligomeric aggregates. *PLOS Comput. Biol.* **4**, e1000222 (2008).
32. R. I. Dima, D. Thirumalai, Exploring protein aggregation and self-propagation using lattice models: Phase diagram and kinetics. *Protein Sci.* **11**, 1036–1049 (2002).
33. S. Abeln, M. Vendruscolo, C. M. Dobson, D. Frenkel, A simple lattice model that captures protein folding, aggregation and amyloid formation. *PLoS One* **9**, e85185 (2014).
34. A. Šarić *et al.*, Physical determinants of the self-replication of protein fibrils. *Nat. Phys.* **12**, 874–880 (2016).
35. M. Cheon *et al.*, Structural reorganisation and potential toxicity of oligomeric species formed during the assembly of amyloid fibrils. *PLOS Comput. Biol.* **3**, 1727–1738 (2007).
36. T. C. T. Michaels *et al.*, Dynamics of oligomer populations formed during the aggregation of Alzheimer's Aβ42 peptide. *Nat. Chem.* **12**, 445–451 (2020).
37. S. G. Estácio, C. S. Fernandes, H. Krobath, P. F. N. Faísca, E. I. Shakhnovich, Robustness of atomistic Gō models in predicting native-like folding intermediates. *J. Chem. Phys.* **137**, 085102 (2012).
38. S. Takada, Gō model revisited. *Biophys. Physicobiol.* **16**, 248–255 (2019).
39. C. Clementi, H. Nymeyer, J. N. Onuchic, Topological and energetic factors: What determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* **298**, 937–953 (2000).
40. J. Karanicolas, C. L. Brooks 3rd, The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Sci.* **11**, 2351–2361 (2002).
41. J. Karanicolas, C. L. Brooks 3rd, Improved Gō-like models demonstrate the robustness of protein folding mechanisms towards non-native interactions. *J. Mol. Biol.* **334**, 309–325 (2003).
42. J. N. Onuchic, Z. Luthey-Schulten, P. G. Wolynes, Theory of protein folding: The energy landscape perspective. *Annu. Rev. Phys. Chem.* **48**, 545–600 (1997).
43. K. Okazaki, N. Koga, S. Takada, J. N. Onuchic, P. G. Wolynes, Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 11844–11849 (2006).
44. R. B. Best, Y. G. Chen, G. Hummer, Slow protein conformational dynamics from multiple experimental structures: The helix/sheet transition of arc repressor. *Structure* **13**, 1755–1763 (2005).
45. L. Sutto, I. Mereu, F. L. Gervasio, A hybrid all-atom structure-based model for protein folding and large scale conformational transitions. *J. Chem. Theory Comput.* **7**, 4208–4217 (2011).
46. C. Camilloni, L. Sutto, Lymphotactin: How a protein can adopt two folds. *J. Chem. Phys.* **131**, 245105 (2009).
47. L. Sutto, C. Camilloni, From A to B: A ride in the free energy surfaces of protein G domains suggests how new folds arise. *J. Chem. Phys.* **136**, 185101 (2012).
48. D. Ganguly, J. Chen, Topology-based modeling of intrinsically disordered proteins: Balancing intrinsic folding and intermolecular interactions. *Proteins* **79**, 1251–1266 (2011).
49. M. Knott, R. B. Best, Discriminating binding mechanisms of an intrinsically disordered protein via a multi-state coarse-grained model. *J. Chem. Phys.* **140**, 175102 (2014).
50. A. Wojtczak, V. Cody, J. R. Luft, W. Pangborn, Structures of human transthyretin complexed with thyroxine at 2.0 Å resolution and 3′,5′-dinitro-*N*-acetyl-L-thyronine at 2.2 Å resolution. *Acta Crystallogr. D Biol. Crystallogr.* **52**, 758–765 (1996).
51. Y. Liang, M. O. Ore, S. Morin, D. J. Wilson, Specific disruption of transthyretin(105-115) fibrilization using "stabilizing" inhibitors of transthyretin amyloidogenesis. *Biochemistry* **51**, 3523–3530 (2012).
52. G. Merlini, V. Bellotti, Molecular mechanisms of amyloidosis. *N. Engl. J. Med.* **349**, 583–596 (2003).
53. J. A. Jarvis, A. Kirkpatrick, D. J. Craik, 1H NMR analysis of fibril-forming peptide fragments of transthyretin. *Int. J. Pept. Protein Res.* **44**, 388–398 (1994).
54. P. Robustelli, S. Piana, D. E. Shaw, Developing a molecular dynamics force field for both folded and disordered protein states. *Proc. Natl. Acad. Sci. U.S.A.* **115**, E4758–E4766 (2018).
55. F. Grigolato, P. Arosio, Sensitivity analysis of the variability of amyloid aggregation profiles. *Phys. Chem. Chem. Phys.* **21**, 1435–1442 (2019).
56. G. Meisl *et al.*, Differences in nucleation behavior underlie the contrasting aggregation kinetics of the Aβ40 and Aβ42 peptides. *Proc. Natl. Acad. Sci. U.S.A.* **111**, 9384–9389 (2014).
57. H. Komatsu, E. Feingold-Link, K. A. Sharp, T. Rastogi, P. H. Axelsen, Intrinsic linear heterogeneity of amyloid β protein fibrils revealed by higher resolution mass-per-length determinations. *J. Biol. Chem.* **285**, 41843–41851 (2010).
58. L. Radamaker *et al.*, Cryo-EM reveals structural breaks in a patient-derived amyloid fibril from systemic AL amyloidosis. *Nat. Commun.* **12**, 875 (2021).
59. G. Meisl *et al.*, Scaling behaviour and rate-determining steps in filamentous self-assembly. *Chem. Sci. (Camb.)* **8**, 7087–7097 (2017).
60. P. Arosio, T. P. J. Knowles, S. Linse, On the lag phase in amyloid fibril formation. *Phys. Chem. Chem. Phys.* **17**, 7606–7618 (2015).
61. G. Meisl *et al.*, Molecular mechanisms of protein aggregation from global fitting of kinetic models. *Nat. Protoc.* **11**, 252–272 (2016).
62. G. Fusco *et al.*, Structural basis of membrane disruption and cellular toxicity by α-synuclein oligomers. *Science* **358**, 1440–1443 (2017).
63. D. C. Rodriguez Camargo *et al.*, Stabilization and structural analysis of a membrane-associated hIAPP aggregation intermediate. *eLife* **6**, e31226 (2017).
64. B. Strodel, Amyloid aggregation simulations: Challenges, advances and perspectives. *Curr. Opin. Struct. Biol.* **67**, 145–152 (2021).
65. S. Samantray, F. Yin, B. Kav, B. Strodel, Different force fields give rise to different amyloid aggregation pathways in molecular dynamics simulations. *J. Chem. Inf. Model.* **60**, 6462–6475 (2020).
66. Z. Jia, J. D. Schmit, J. Chen, Amyloid assembly is dominated by misregistered kinetic traps on an unbiased energy landscape. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 10322–10328 (2020).
67. D. J. Rosenman, C. R. Connors, W. Chen, C. Wang, A. E. García, Aβ monomers transiently sample oligomer and fibril-like configurations: Ensemble characterization using a combined MD/NMR approach. *J. Mol. Biol.* **425**, 3338–3359 (2013).
68. J. Nasica-Labouze *et al.*, Amyloid β protein and Alzheimer's disease: When computer simulations complement experimental studies. *Chem. Rev.* **115**, 3518–3563 (2015).
69. M. W. van der Kamp, V. Daggett, Pathogenic mutations in the hydrophobic core of the human prion protein can promote structural instability and misfolding. *J. Mol. Biol.* **404**, 732–748 (2010).
70. M. Sanz-Hernández *et al.*, Mechanism of misfolding of the human prion protein revealed by a pathological mutation. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2019631118 (2021).
71. A. R. Walker, N. Baddam, G. A. Cisneros, Unfolding pathways of hen egg-white lysozyme in ethanol. *J. Phys. Chem. B* **123**, 3267–3271 (2019).
72. T. Le Marchand *et al.*, Conformational dynamics in crystals reveal the molecular bases for D76N beta-2 microglobulin aggregation propensity. *Nat. Commun.* **9**, 1658 (2018).
73. R. B. Best, G. Hummer, W. A. Eaton, Native contacts determine protein folding mechanisms in atomistic simulations. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 17874–17879 (2013).
74. Y. Levy, P. G. Wolynes, J. N. Onuchic, Protein topology determines binding mechanism. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 511–516 (2004).
75. S. Orioli, A. H. Larsen, S. Bottaro, K. Lindorff-Larsen, How to learn from inconsistencies: Integrating molecular simulations with experimental data. *Prog. Mol. Biol. Transl. Sci.* **170**, 123–176 (2020).
76. P. G. Bolhuis, Z. F. Brotzakis, M. Vendruscolo, A maximum caliber approach for continuum path ensembles. *Eur. Phys. J.* **94**, 188 (2021).
77. R. Capelli, G. Tiana, C. Camilloni, An implementation of the maximum-caliber principle by replica-averaged time-resolved restrained simulations. *J. Chem. Phys.* **148**, 184114 (2018).
78. A. Jussupow *et al.*, The dynamics of linear polyubiquitin. *Sci. Adv.* **6**, 3786–3800 (2020).
79. Y. Shi *et al.*, Structure-based classification of tauopathies. *Nature* **598**, 359–363 (2021).
80. Y. Yang *et al.*, Cryo-EM structures of amyloid-b 42 filaments from human brains. *Science* **375**, 167–172 (2022).
81. T. Löhr, K. Kohlhoff, G. T. Heller, C. Camilloni, M. Vendruscolo, A kinetic ensemble of the Alzheimer's Aβ peptide. *Nat. Comput. Sci.* **1**, 71–78 (2021).
82. P. Robustelli *et al.*, Molecular basis of small-molecule binding to α-synuclein. *J. Am. Chem. Soc.* **144**, 2501–2510 (2022).
83. M. J. Abraham *et al.*, GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).
84. J. K. Noel *et al.*, SMOG 2: A versatile software package for generating structure-based models. *PLOS Comput. Biol.* **12**, e1004794 (2016).
85. M. J. Bennett, M. P. Schlunegger, D. Eisenberg, 3D domain swapping: A mechanism for oligomer assembly. *Protein Sci.* **4**, 2455–2468 (1995).
86. M. J. Bennett, S. Choe, D. Eisenberg, Domain swapping: Entangling alliances between proteins. *Proc. Natl. Acad. Sci. U.S.A.* **91**, 3127–3131 (1994).
87. N. Schmid *et al.*, Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *Eur. Biophys. J.* **40**, 843–856 (2011).
88. K. Lindorff-Larsen, S. Piana, R. O. Dror, D. E. Shaw, How fast-folding proteins fold. *Science* **334**, 517–520 (2011).
89. W. L. Jorgensen, C. Jenson, Temperature dependence of TIP3P, SPC, and TIP4P water from NPT Monte Carlo simulations: Seeking temperatures of maximum density. *J. Comput. Chem.* **19**, 11791186 (1998).
90. R. Capelli, C. Paissoni, P. Sormanni, G. Tiana, Iterative derivation of effective potentials to sample the conformational space of proteins at atomistic scale. *J. Chem. Phys.* **140**, 195101 (2014).
91. A. B. Norgaard, J. Ferkinghoff-Borg, K. Lindorff-Larsen, Experimental parameterization of an energy function for the simulation of unfolded proteins. *Biophys. J.* **94**, 182–192 (2008).
92. E. Scalone, C. Camilloni, Dataset 1 for: Multi-eGO: An in-silico lens to look into protein aggregation kinetics at atomic resolution. Zenodo. 10.5281/zenodo.6125995. Deposited 17 February 2022.
93. E. Scalone, C. Camilloni, Dataset 2 for: Multi-eGO: An in-silico lens to look into protein aggregation kinetics at atomic resolution. Zenodo. 10.5281/zenodo.6125424. Deposited 17 February 2022.
94. E. Scalone, C. Camilloni, Dataset 3 for: Multi-eGO: An in-silico lens to look into protein aggregation kinetics at atomic resolution. Zenodo. 10.5281/zenodo.6414572. Deposited 5 April 2022.

Assessing the initial TTR fibril model

After the publication of the previous paper, I further tested the quality of the TTR fibril (PDB 2M5M) by an explicit water MD simulation. The simulation was performed using GROMACS[345] and Charmm22*[200,346,347]. Short range interactions were cut-off at 1 nm, with long range electrostatic handled using the particle mesh Ewald scheme[348]. The fibril was elongated by duplicating and aligning TTR monomers, 14880 water molecules were added and no ions were necessary. Following energy minimization, a two positional restraint steps were performed under NVT condition at 300K temperature for 500 ns using the velocity-rescale thermostat[349]. For the first one all the atoms were restrained, while in the second one the restraint was only on the backbone atoms. After, the production simulation was performed under NVT conditions at 300K for 335 ns. Considering the difficulties in simulating an amyloid fibril using a transferable force-field the fibril ends were positionally restrained during the production. The result was a collapse of the fibril. Additionally, a further test was performed by restraining three residues at the N and C terminals of the external chains. Although the additional positional restraint, internal chains could not keep a proper conformation. Considering the experimental setup[89], I subsequently added 10% v/v of acetonitrile as shown in Figure 4 by adding 1164 acetonitrile

molecules and a total of 14991 water molecules. The acetonitrile parametrization was performed by using AmberTools and adapted for Charmm22*. The production simulation was performed under NVT conditions at 300K for 367 ns in the same conditions of the previous one. Interestingly the acetonitrile inserted within the fibril cavity displacing the water molecules Figure 4B. Despite the addition of acetonitrile, the fibril was not stable as the acetonitrile stabilized the cavity it also inserted between the protofilaments of TTR fibril breaking the amyloid structure as shown in Figure 4. Thus, I have interpreted this result as a clear indication of the importance of acetonitrile in TTR aggregation and the reasonable inaccuracy of multi-*e*GO along with the difficulties of simulating fibrils in complicated experimental conditions.

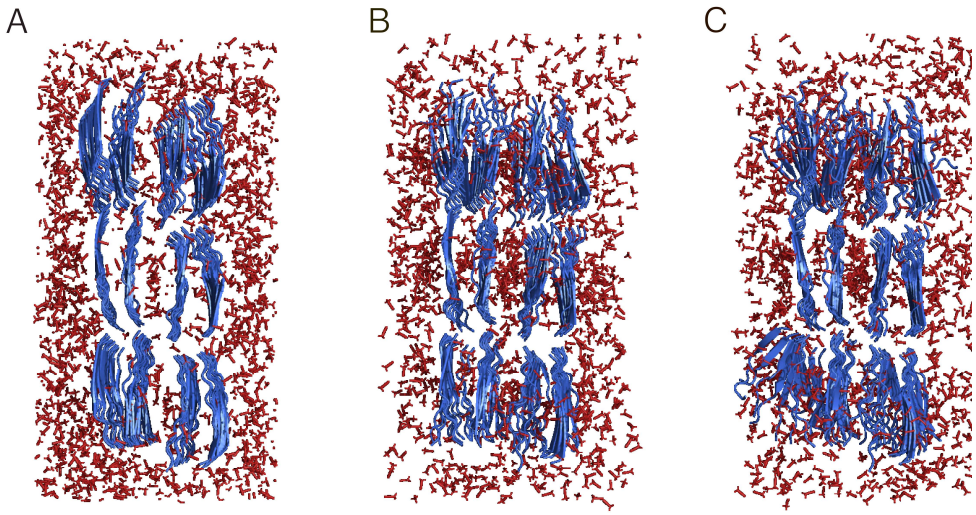


Figure 4 – Simulation of the $TTR_{105-115}$ with acetonitrile. A) The system contained 10% v/v of acetonitrile randomly placed in the box and subsequently water molecules were added. B) Acetonitrile molecules concentrating within the fibril cavity and protofilaments after 7 ns. C) Collapsed fibril at 367 ns (end of the simulation). The interactions between the N and C terminals were broken because of the acetonitrile.

3. Multi-*e*GO with prior information

The methodological basis of multi-*e*GO was published and reported above. Afterwards, several methodological updates, improvements and applications were done. In this chapter I will describe the new LJ parametrization along with initial testing and results of Aβ42 protein.

As explained in detail in the paper above, multi-eGO is a hybrid and transferable structure-based model defined from an ensemble of simulations and structures. Only the heavy atoms (non-Hydrogens) were included. Bonds, angles, and improper dihedrals were based on GROMOS54a7 force field, being already optimized without non-polar hydrogens. We defined excluded volume interactions parameters that is all pairs of atoms that are separated by more than 3 consecutive covalent bonds (i-i+4) using a custom $C^{(12)}$. In multi-*e*GO the structural information could be obtained from either a PDB structure or an MD simulation as an ensemble and parametrized as a LJ potential. The structure-based ensemble was defined by pairs of atoms farther than one residue along the sequence within a 5.5 Å cutoff. Using the TTR peptide we demonstrated the advantage of rescaling the LJ potential using an MD probability for each native

contact when simulating a short and unstructured peptide. The LJ potential was defined as the following equation:

$$\varepsilon_{i,j} = \varepsilon_0 \left(1 - \frac{\ln P_{i,j}^{MD}}{\ln P_{threshold}^{MD}} \right) = -\frac{\varepsilon_0}{\ln P_{threshold}^{MD}} \bullet \ln \frac{P_{i,j}^{MD}}{P_{threshold}^{MD}} \qquad \text{Eq 1}$$

Where ε_0 is the initial interaction energy supplied in input, $P_{i,j}^{MD}$ is the fraction of frames with a native contact obtained from an implicit solvent in an MD simulation using a transferable force-field and $P_{threshold}^{MD}$ is a minimum fraction that should be considered. Such LJ parametrization allowed an increase of protein flexibility.

Encouraged by our previous results, we built a multi-*e*GO model to simulate Aβ42 using the published MD trajectories from Löhr et al.[183] as reference. As can be noted from Figure 5, multi-*e*GO did not allow to capture the dynamics of the reference conformational ensemble of Aβ42 by using the same setup from $TTR_{105\text{-}115}$ nor the maximum value for ε. Our interpretation was on the unbalance between the probability of local contacts and the contacts residues far in sequence, where in the case of an IDP like Aβ42 is particularly emphasized. Further analysis of contact probabilities showed native contacts between residues close in sequence had a higher probability compared to native contacts far in sequence. Such

unbalance between the local and non-local contact probabilities caused an elongated protein conformation since the non-local contacts are too weak to provide a proper protein conformation.

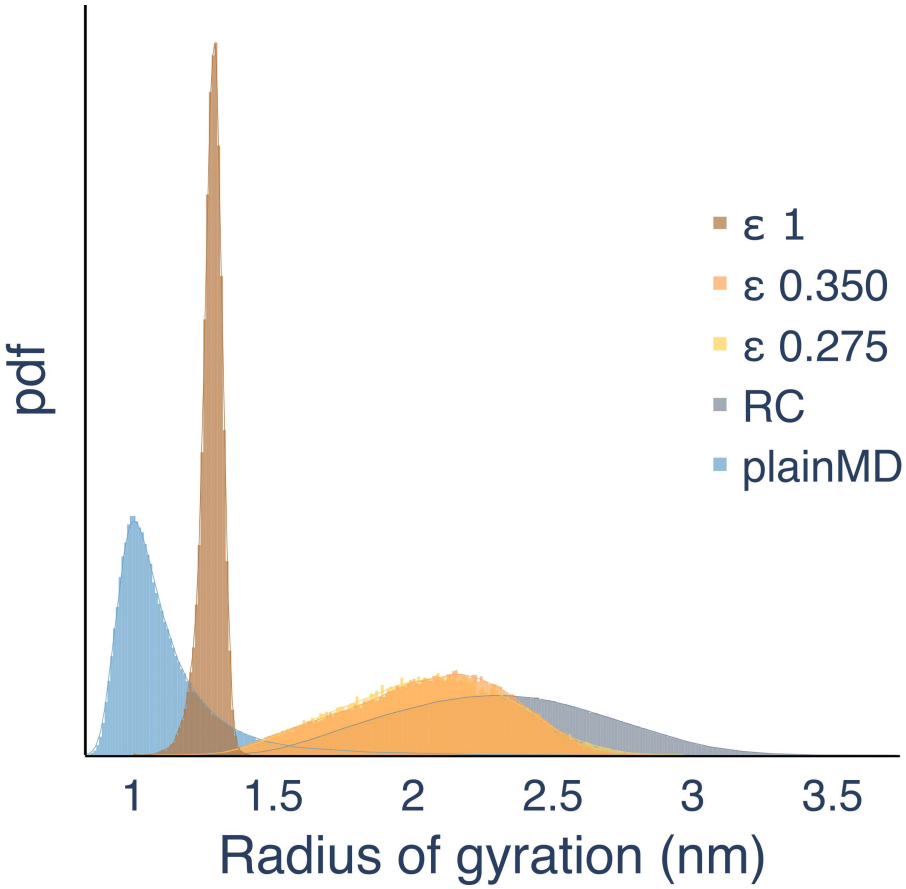


Figure 5 – Aβ42 monomer dynamics. Radius of gyration distribution at ε=0.275 (ochre), slightly higher ε value (orange), extreme increase of ε (brown), a random coil (RC) simulation (gray), Charmm22 (blue) used as a reference model from Löhr et al.[183]. The initial multi-eGO distributions describe an open conformation which does not overlap with the reference. Usually, that could suggest an increase of ε which did not resolve the issue. The extreme test of using the maximum value of ε describes a permanent dynamic with a single peak around 1.35 nm, suggesting an erroneous parametrization of multi-eGO. Observing the RC distribution, suggested an improper parametrization of the local contacts.*

Our hypothesis was supported by the similarity of the radius of gyration between the two simulations at ε=0.275, 0.350, and a simulation parametrized with bonded parameters exclusively. We defined the latter simulation as random coil (RC) where the radius of gyration depends exclusively on the local geometry as only exclusions terms are acting as non-bonded parameters. As the RC simulation provided information about the local contacts probability required to fine tune the LJ potential of local contacts, we improved the parametrization of the LJ potential by adding the RC probabilities. As the RC reweight can be applied on intramolecular contacts, we differentiated the parametrization by separating intramolecular and intermolecular contacts, and updating Eq1 as it follows:

$$\varepsilon_{i,j} = -\frac{\varepsilon_0^{intra}}{\ln P_{threshold}^{RC}} \bullet \ln \frac{P_{i,j}^{MD}}{P_{i,j}^{RC}} \qquad \text{Eq 2}$$

Here, $P_{i,j}^{MD}$ and $P_{i,j}^{RC}$ are the fraction of frames having a native contact in the MD and RC simulation respectively. In this new parametrization, the minimum fraction $P_{threshold}^{RC}$ is defined by the RC and not by the MD. This new equation allowed to balance the definitions of local and non-local contacts permitting the proper protein conformation derived from the non-local contacts.

Intermolecular contacts derived from PDB structures or MD simulations containing more than one molecule interacting were parametrized with the reformulated Eq1 as RC cannot be applied on intermolecular contacts:

$$\varepsilon_{i,j}^{inter} = -\frac{\varepsilon_0^{inter}}{\ln P_{threshold}^{MD}} \bullet \ln \frac{P_{i,j}^{MD}}{P_{threshold}^{MD}} \qquad \text{Eq 3}$$

The implementation of the LJ potential allowed intramolecular contacts to be applied also intermolecularly but not the opposite, as intermolecular contacts could cause distortions of the protein chain. The intermolecular contacts were applied also in cases where different type of molecules were inserted in a multi-*e*GO simulation such as a protein-ligand interactions. The separation of intramolecular and intermolecular contacts greatly improved multi-*e*GO as it eases the addition of different molecule types by keeping a proper description of the local geometry. There were cases where RC probabilities were higher than MD ones and it was our interpretation that such contacts should have a repulsive term on the local geometry. For this type of contacts only the $C^{(12)}$ was parametrized for the LJ potential and rescaled based on the ratio between MD probability and RC probability as it follows:

$$C_{i,j}^{(12)} = \sqrt{C_i^{(12)} C_j^{(12)}} - \ln \frac{P_{i,j}^{MD}}{P_{i,j}^{RC}} \langle d_{i,j}^{MD} \rangle^{12} \qquad \text{Eq 4}$$

Together, such optimizations allowed a proper representation of structured and unstructured full-length proteins as both cases are present during protein aggregation.

This new setup drastically improved the dynamics of Aβ42 molecule as shown in Figure 6. The radius of gyration of multi-eGO overlapped with the Charmm22* reference (Figure 6A). Comparison of the contacts between multi-eGO and the reference, both local and long-distance contacts, showed a high similarity between the two models with a score of 2.832 (Figure 6B). A further analysis of the reference and multi-*e*GO was performed by checking RMSD clustering shown in Figure 6C. The reference and multi-*e*GO trajectories were merged and a RMSD clustering was performed using a cutoff of 0.78 nm. Multi-*e*GO was able to quantitively reproduce the correct clusters of conformations and their relative populations.

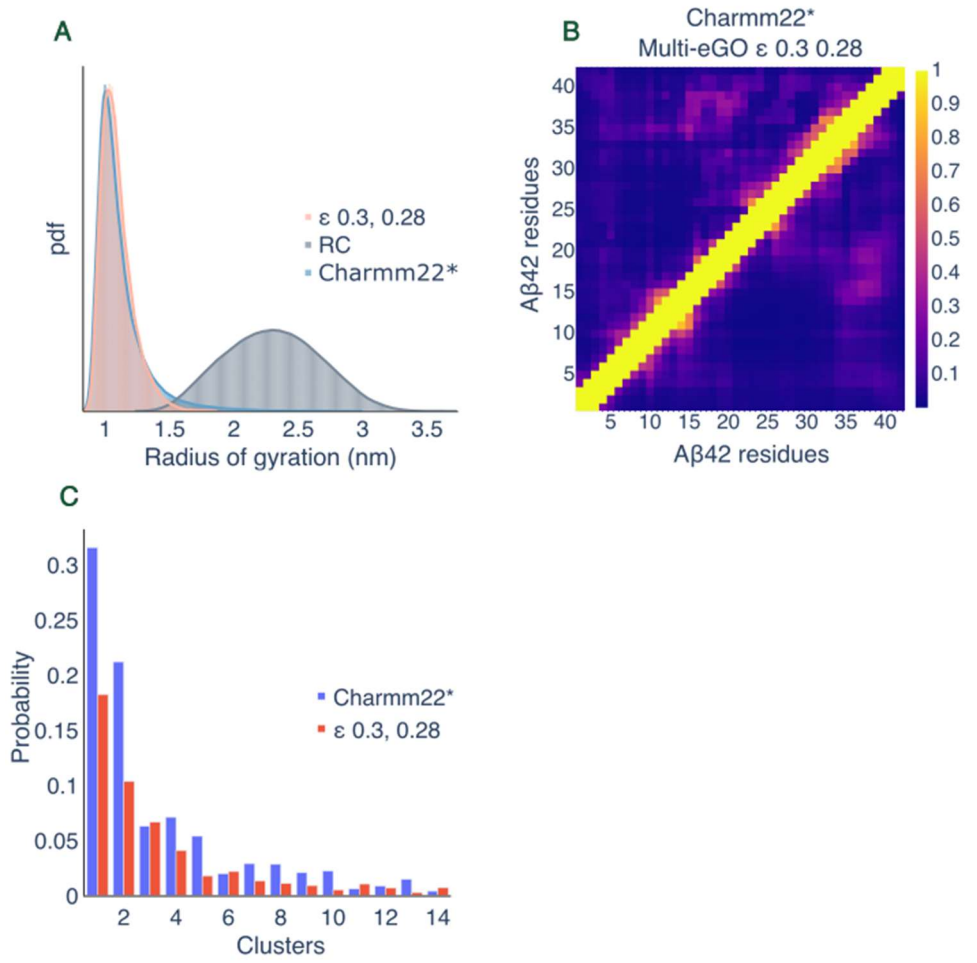


Figure 6 – Comparisons between the reference and multi-eGO. The LJ potential is parametrized by using ε=0.300 for intramolecular contacts and ε=0.280 for intermolecular contacts. (A) Radius of gyration distribution of the improved setup of multi-eGO (pink), reference MD (blue) and RC (grey). By the addition of RC to balance the LJ potential of local contacts, the radius of gyration distribution of multi-eGO and reference overlaps. (B) Contact maps of reference (upper-left) and multi-eGO (bottom-right). By comparing the two maps it is possible to appreciate the accuracy on both local and long-range contacts, suggesting a proper dynamics is achieved by multi-eGO.(C) Clustering analysis of multi-eGO and reference simulation. Qualitatively, the most represented classes are comparable.

The monomer of Aβ42 is also characterized in presence of a ligand preventing aggregation by Löhr et al.[184]. The small molecule stabilizes the native state of Aβ42 by a disordered binding mechanism, in which both the small molecule and the protein remain disordered. I tested the ability of multi-*e*GO to reproduce the ligand interactions following Eq3 for intermolecular contacts. In Figure 7 it is possible to appreciate the ability of multi-eGO to reproduce the ligand interactions and potentially simulate the aggregation process with several protein monomers and small molecules.

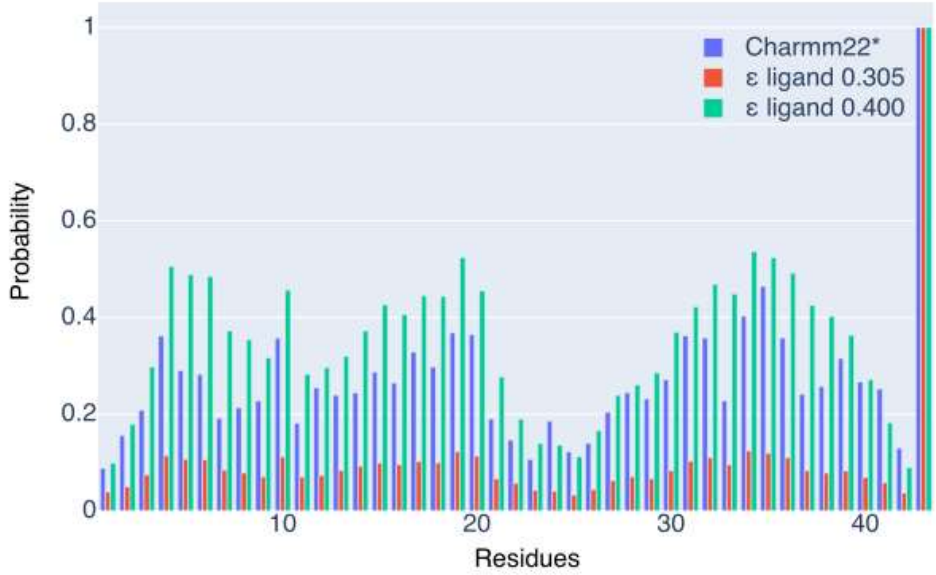


Figure 7 – Contact probability between the ligand from the reference simulation in blue, two different multi-eGO LJ potentials: low ε in red and high ε in green. Although a careful optimization of the LJ potential applied on protein-ligand contacts, the key probabilities were reproduced.

With the new multi-*e*GO, a simulation involving 1000 Aβ42 monomers at 100 μM was produced with the aim to simulate protein aggregation. Additionally, a second simulation was produced with the addition of 2 mM ligand. Although the simulations were short and did not describe a complete sigmoidal aggregation kinetics, we were able to observe Aβ42 oligomerization and the transition to a fibril-like conformation as shown in Figure 8A. Additionally, we were able to observe the interaction between several ligand molecules with Aβ42 oligomers in a non-stoichiometry manner displayed in Figure 8B.

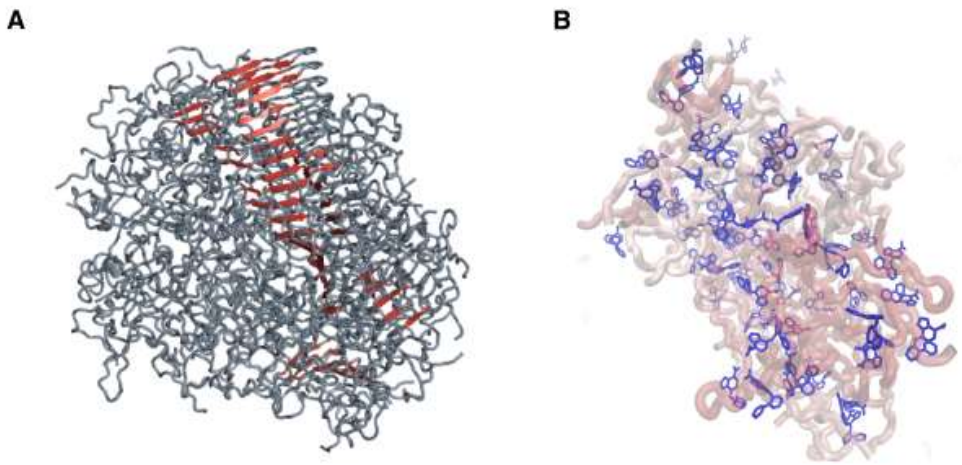
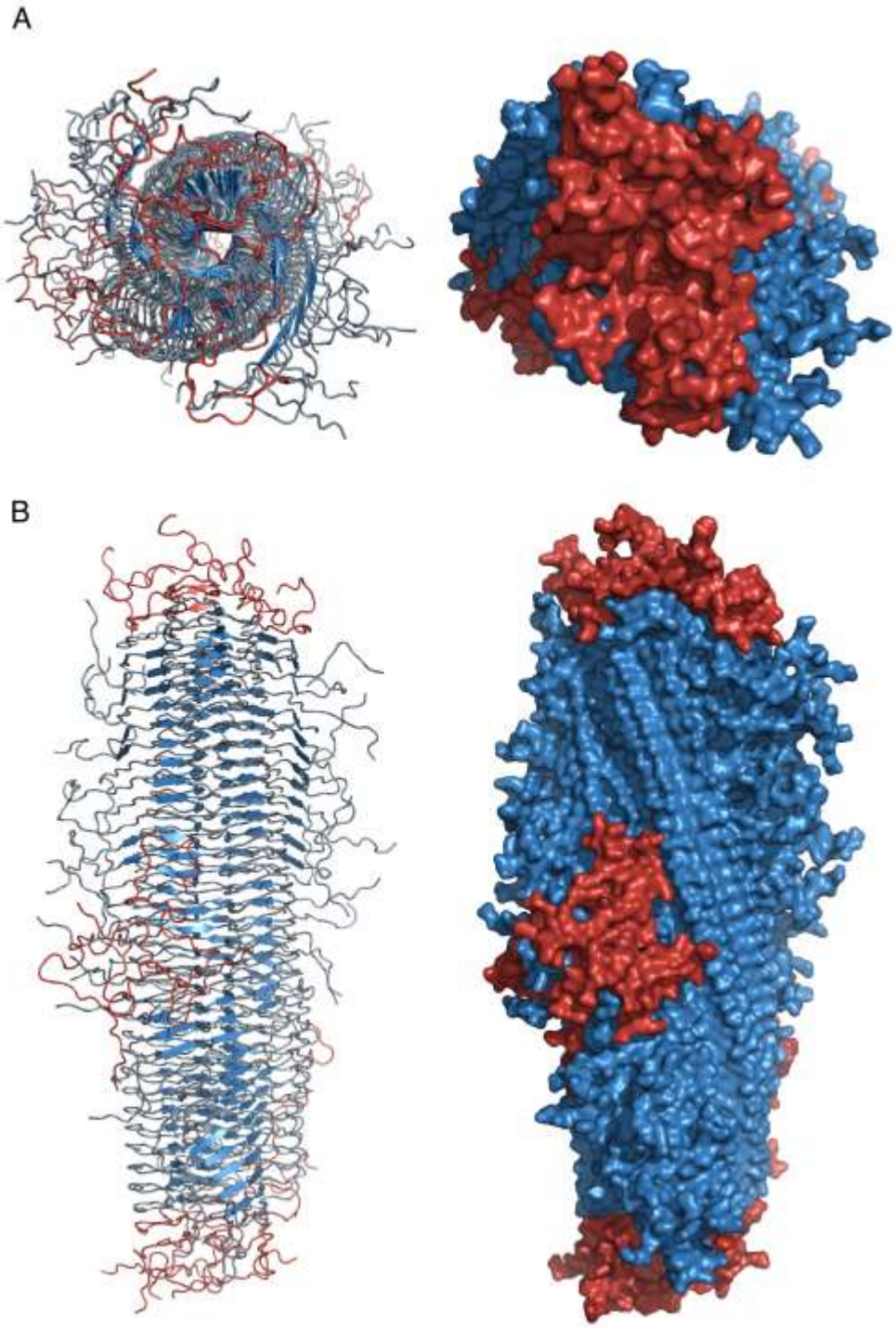


Figure 8 – A) Oligomer of 68 Aβ42 monomers. In red the cross-β conformation obtained during the rearrangement of molecules during the simulation. B) Aβ42 oligomer interacting with several small molecules. The interaction is non-specific and non-stoichiometric.

A third test simulation was produced with a fibril seed which showed the elongation of the two ends and an initial secondary nucleation process. In Figure 9A, Aβ42 monomers interacted with the two ends with the proper conformation of the fibril, defining an S-shape. Interestingly, the monomers interacting with the seed surface were retained by the N-terminal of the molecules within the seeds as shown in Figure 9B. Such observation could potentially explain the secondary nucleation process of Aβ42 where sticky N-terminals on the surface ease the interaction of monomers. Different monomers interacted with each other until an oligomer is formed, determining a catalytic property of the fibril surface.

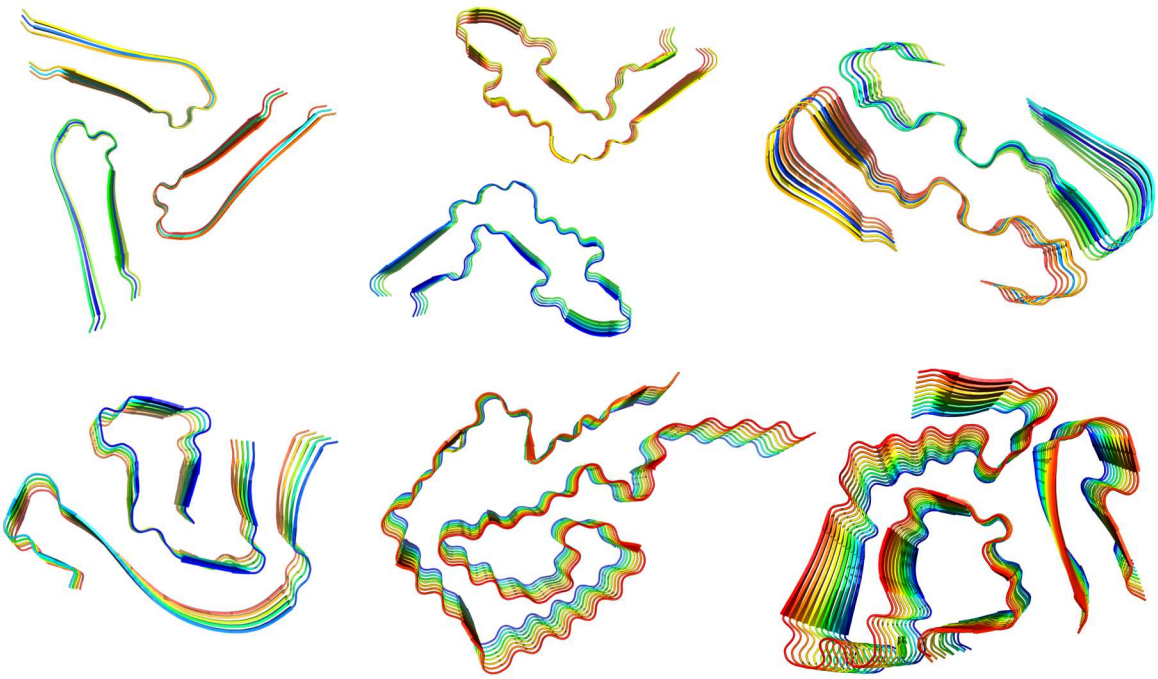
Figure 9 – A) Top view of the last frame of Aβ42 seeded aggregation with free monomers in red and seed in blue. It is possible to observe the free monomers arranging at the seed end in the proper S-shaped conformation. B) Side view of the fibril seed with a monomer interacting with the N-terminal of the fibril molecules. The N-terminal is not modelled in Cryo-EM maps, thus are disordered, and not involved in the cross-β.


A
B

3. Conclusions and Future Perspectives

"Nessuno è più basito"

Arianna Dell'Arti



In this thesis, I developed a simplified force-field to simulate protein aggregation. Using state-of-the-art computing resources, it allows to simulate thousands of monomers at atomic resolution (excluding hydrogens) and qualitatively provide aggregation models. It can be applied to any protein of interest as the bonded and excluded volume interactions are derived from a transferable force-field and consequently optimized. Multi-*e*GO is ensemble-based, learning the structural information from multiple references and symmetrized, so interacting pairs can be applied both intramolecularly and intermolecularly. In our first work we demonstrated the multi-*e*GO potential by simulating the aggregation of $TTR_{105-115}$ resulting in compatible kinetics and structures with reference experimental results. The difficulties encountered in the topology description of the TTR peptide fibrils using multi-*e*GO could be partially explained by the results of the fibril simulation using acetonitrile. The cavity is mainly filled by acetonitrile instead of water, thus multi-*e*GO might require additional details and acetonitrile. The application of multi-*e*GO using Aβ42 protein emphasized major issues in local geometry as shown in Figure 5 and addressed by updating the method by the insertion of a prior information of the protein chain. In Figure 6 the striking improvement is shown by a comparison of radius of gyration, contact map and RMSD clustering. Such promising results led to the simulation

of Aβ42 aggregation and the possibility to insert small molecules in multi-*e*GO. The aggregation kinetics was not fully described as the simulation was not long enough. However, multi-*e*GO managed to describe the oligomerization of Aβ42 and the transition from an unstructured oligomer to a fibril-like one. This encouraging result proves the potentiality of multi-*e*GO in simulating the atomistic details of protein aggregation. Seeded aggregation showed the possibility to study the elongation process but more important, the secondary nucleation process, as it is a drug discovery target to treat AD. The addition of a small molecule in an aggregation simulation can potentially be exploited in drug discovery processes as several thousand of protein and ligands can be simulated. The addition of a small molecule was a promising proof of concept. Contact probabilities were compatible with the reference simulation, although a proper parametrization of bonded interaction is mandatary as the ligand geometry was too rigid. A systematic approach to improve local geometry would be necessary as it was applied on proteins.

Future developments of multi-*e*GO include use of data-driven simulations (e.g., SAXS data, chemical shift, Cryo-EM density maps) for the training of the multi-eGO potential, or in case of observables relative to large assemblies, integration

of experimental data in multi-*c*GO simulation. As misfolded protein and oligomers expose hydrophobic residues, a future addition would be lipids parametrization as an aim to simulate protein aggregates in presence of a cellular membrane. Moreover, a future parametrization of glycans and nucleic acids can further expand the possibilities of multi-*c*GO as it would be possible to simulate the self-assembly of macromolecular complexes.

4. Published Articles

Scalone, E., Broggini, L., Visentin, C., Erba, D., Bacic Toplek, F., Peqini, K., Pellegrino, S., Ricagno, S., Paissoni, C., & Camilloni, C. (2022). Multi- *e* GO: An in silico lens to look into protein aggregation kinetics at atomic resolution. *Proceedings of the National Academy of Sciences*, *119*(26). https://doi.org/10.1073/PNAS.2203181119

The contribution to this paper was described in Results as it is the main manuscript of my PhD project. Here the Supplementary Materials are attached.

PNAS
www.pnas.org

Supplementary Information for

Multi-*e*GO: an *in-silico* lens to look into protein aggregation kinetics at atomic resolution

Emanuele Scalone[1], Luca Broggini[1,2], Cristina Visentin[1,2], Davide Erba[1], Fran Bačić Toplek[1], Kaliroi Peqini[3], Sara Pellegrino[3], Stefano Ricagno[1,2,*], Cristina Paissoni[1,*], and Carlo Camilloni[1,*]

[1]Dipartimento di Bioscienze, Università degli Studi di Milano, Milano, Italy

[2]Institute of Molecular and Translational Cardiology, IRCCS Policlinico San Donato, San Donato Milanese, Italy

[3]DISFARM, Dipartimento di Scienze Farmaceutiche, Sezione Chimica Generale e Organica, Università degli Studi di Milano, Milano, Italy

*Stefano Ricagno, Cristina Paissoni, Carlo Camilloni

Email: stefano.ricagno@unimi.it, cristina.paissoni@unimi.it, carlo.camilloni@unimi.it

This PDF file includes:

Other supplementary materials for this manuscript include the following:

Supplementary methods

Simulations details: $TTR_{105-115}$ monomer in explicit solvent

$TTR_{105-115}$ peptide in solution does not possess a unique well-defined structure being too short, consequently we performed a reference simulation using the a99SB-*disp* (1) force field. The starting structure was obtained from pdb 4TLT and the simulation was set to resemble the experimental conditions of $TTR_{105-115}$ aggregation, therefore, to be at pH 2.0 we protonated the N- and C- termini. The molecule was solvated with 3408 water molecules in a dodecahedron box initially 0.9 nm larger than the chain in each direction. One chlorine ion was added to neutralize the charges. Short range interactions were cut-off at 1 nm, with long range electrostatic handled using the particle mesh Ewald scheme (2). Lincs constraints were applied only to bonded hydrogens (3).

Following energy minimization, a positional restraint step was performed under NVT conditions at 300K temperature for 500 ps using the velocity-rescale thermostat (4); then the cell-rescale barostat (5) was used to equilibrate the system in the NPT ensemble to the target pressure of 1 atm for 1000 ps. A production MD simulation under NPT ensemble was run for 1.5 μs. This reference simulation was then analyzed to obtain the monomer native pairs and their corresponding interaction strength. These were defined as all the pairs of atoms, more than one residue apart, being closer than 5.5 Å with a population P larger than a threshold $P_{threshold}$, 0.09 in this case.

Simulations details: Multi-eGO $TTR_{105-115}$ fibril

The strength of the native pair interactions ε is the only parameter to be set empirically, this value is then used as it is for all native pairs obtained from the fibril structure, while it is scaled as in equation (1) for native pairs obtained from the simulation of the monomer. To find an optimal value for $TTR_{105-115}$, we performed multiple simulations at fixed temperature using a pre-formed fibril. The fibril model was obtained extracting 64 chains from pdb 2M5K (6). Multiple simulations were performed testing a range of ε values between 0.265 and 0.295 kJ/mol at 315 K using a Langevin dynamics with an inverse friction constant of 50 ps and a timestep of 5 fs. The structure was placed in a cubic box 10 nm larger than the fibril in each direction. After the energy minimization and a positional restraint simulation of 1 ns, production simulations were performed for 200 ns. The optimal ε value was chosen to be the highest at which the structure of the fibril was stable with fibril ends monomers showing some flexibility and ability to partially detach from the fibril. An ε value of 0.275 kJ/mol has then been used for all subsequent simulations.

Simulations details: Multi-eGO $TTR_{105-115}$ aggregation kinetics

To setup the simulations for aggregation kinetics a monomer model was extracted from Y105 to S115 using pdb 4TLT (7) as reference and then, the C-terminal oxygen was added. Four thousand molecules were randomly placed in a cubic box whose side length depends on the desired concentration as $len\,[nm] = 10^{8^3}\sqrt{\frac{nmol}{N_A[C]}}$, where *nmol* is the number of molecules to add in the system, N_A is the Avogadro Number, and [C] is the concentration. The resulting volumes are in the range of 1 attoliter. The system energy was minimized and then thermalized at a constant temperature of 310 K using positions restraint for 1 ns. Simulations were then performed at 5 different concentration, 13 mM, 11.5 mM, 10 mM, 8.5 mM, and 7 mM in triplicate and evolved long enough to form stable fibrils including most of the monomers. Furthermore, three seeded simulations were performed at 7 mM using a structure of a 10 molecules protofilament obtained at 13 mM as a seed. To analyze the simulations, we used a modified version of the GROMACS *clustsize* tool and homemade python scripts based on MDTraj (8) and MDAnalysis (9, 10). *clustsize* provides a matrix of clusters-sizes as a function of time. The kinetic curves were built by multiplying the number of clusters at every frame by the cluster dimension. Four multi-GO aggregation kinetics were initialized similarly, but using 2000 monomers and concentrations of 2, 1, 0.7 and 0.5 mM, respectively.

$TTR_{105-115}$ peptide synthesis

$TTR_{105-115}$ peptide (YTIAALLSPYS) was prepared by microwave assisted Fmoc solid-phase peptide synthesis on Wang resin (0.4 mmol/g loading) using the CEM Liberty Blue synthesizer. The

coupling reaction was performed with 5 eq excess of the amino acid (0.2 M in DMF), DIC (0.5 M in DMF) and Oxyma Pure (1M in DMF) as coupling reagents. The MW cycle was as following: 15s at 75°C, 170W followed by 110s at 90°C, 40W. N-Fmoc deprotection was performed using 20% piperidine in DMF with a MW cycle of 15s at 75°C, 155 W, followed by 60 s at 90°C, 50 W. Full-cleavage from the resin was performed by shaking the resin for 3 hours in a mixture of TFA/TIPS/H2O/phenol 90:2.5:2.5:5. After cleavage, the peptide was precipitated and washed using ice-cold diethyl ether. $TTR_{105-115}$ was purified in reverse-phase by using RP-HPLC with a ADAMAS C-18 column from Sepachrom (10 µm, 250 × 21.2 mm, phase A 0.1% TFA in water, phase B: 1% TFA in ACN with a gradient of 20-100% phase B over 40 min at a flow rate of 20 mL/min.).

$TTR_{105-115}$ aggregation assays

Lyophilized $TTR_{105-115}$ peptide was dissolved at 13 mM in 10% acetonitrile/water solution and pH adjusted to 2.5 with HCl. The solution was sonicated on ice for 15 min and then centrifugated at 4°C for 15 min at 20800 *x g*. The stock solution was eventually diluted to the final concentrations for the experiments (13 mM, 10 mM, and 7 mM) in 10% acetonitrile/water solution. Freshly prepared ThT was added to a final concentration of 20 µM. 50 µL of each condition was then pipetted into black polystyrene 96-well half-area plates with clear bottoms and polyethylene glycol coating (Corning). Each condition was performed in duplicate in each experiment. Plates were sealed to prevent evaporation and incubated at 37 °C under quiescent conditions in a Varioskan Lux plate reader (Thermo Fisher Scientific). Upon excitation at 450 nm, fluorescence at 480 nm was recorded through the bottom of the plate every 120 min. All the experiments were performed in triplicates, except for the 10 mM condition that was done in duplicate. The mean ThT fluorescence values from the independent experiments were normalized and subjected to nonlinear regression analysis, using Boltzmann sigmoidal equation.

TEM analysis on $TTR_{105-115}$ fibrils

Freshly prepared $TTR_{105-115}$ fibrils were analyzed by TEM. 4-µl droplet of sample was applied onto a 400-mesh copper carbon-coated grids (Agar Scientific) glow discharged for 30 s at 30 mA using a GloQube system (Quorum Technologies). After 1-min incubation, excess of sample was removed and the grid was stained with 2% (wt/v) uranyl acetate solution, blotted dry, and imaged on a Talos L120C transmission electron microscope (Thermo Fisher Scientific) operating at 120 kV. Morphological characterization of the fibrils was performed using the software ImageJ (11).

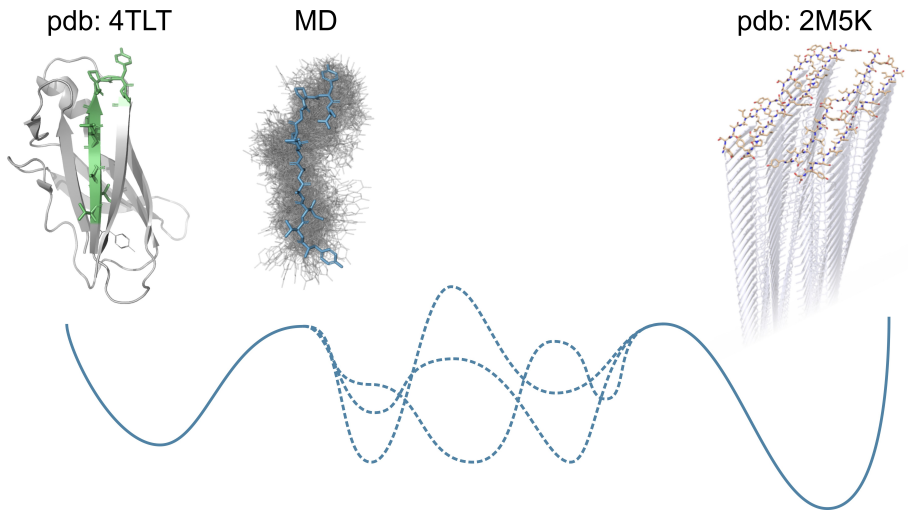


Figure S1. Schematic representation of multi-GO and multi-*e*GO models. Multi-GO model is based completely on two structures, one representing the native structure and one the fibril structure. Multi-*e*GO model instead includes information about the conformational dynamics of the native state and the fibril structure, furthermore multi-*e*GO local geometry is described using a transferable potential.

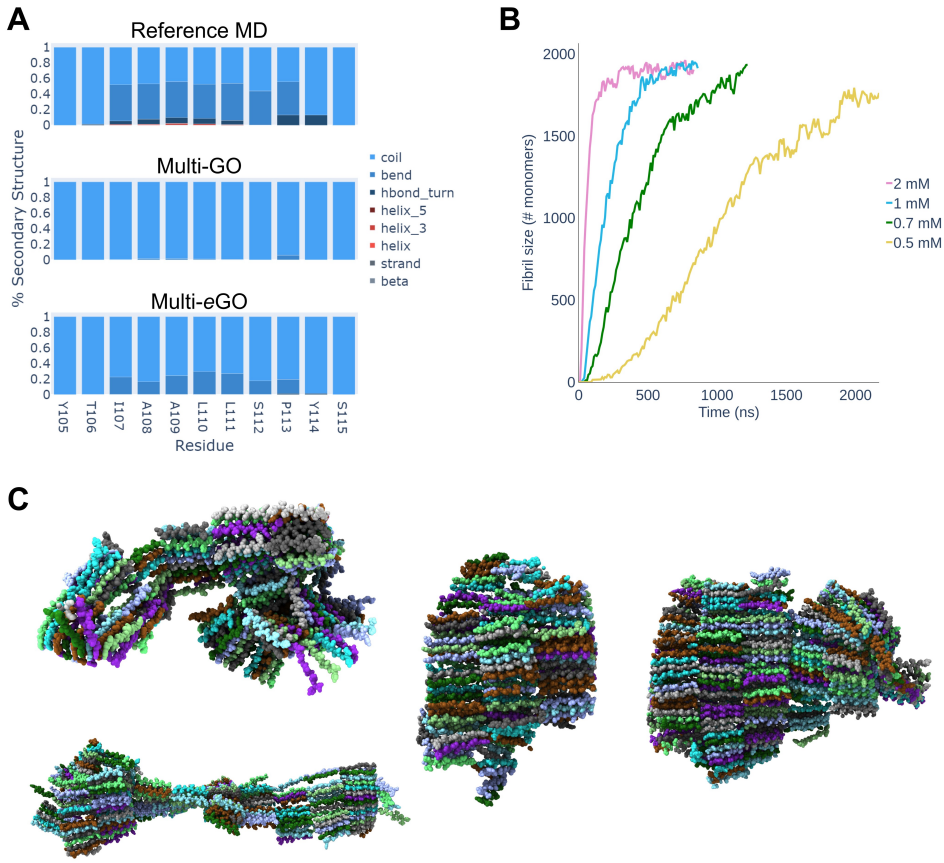


Figure S2. (A) Secondary structures populations for the $TTR_{105-115}$ monomer simulated using the a99SB-*disp* (top), multi-GO (middle) and multi-*e*GO (bottom) force fields. (B) Simulated aggregation kinetics using multi-GO. Curves represent the number of monomers involved in an aggregate of at least 10 monomers as a function of nominal simulation time. (C) Representative structure of amyloid fibrils formed at the end of the multi-GO aggregation kinetics.

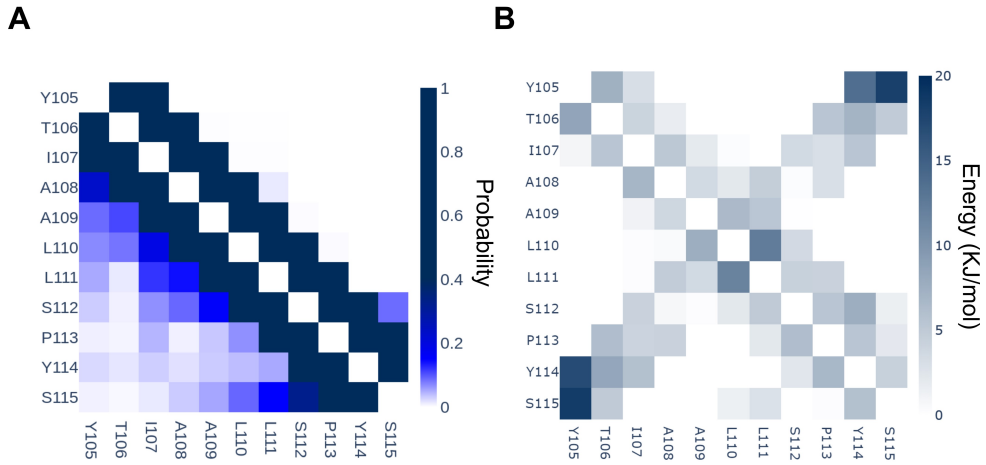


Figure S3. (A) Per residue probability contact map for the $TTR_{105-115}$ peptide conformational ensemble according to a MD simulation with a99SB-*disp* (lower left) and multi-GO (upper right). (B) Lower left multi-GO, Upper right multi-*e*GO.

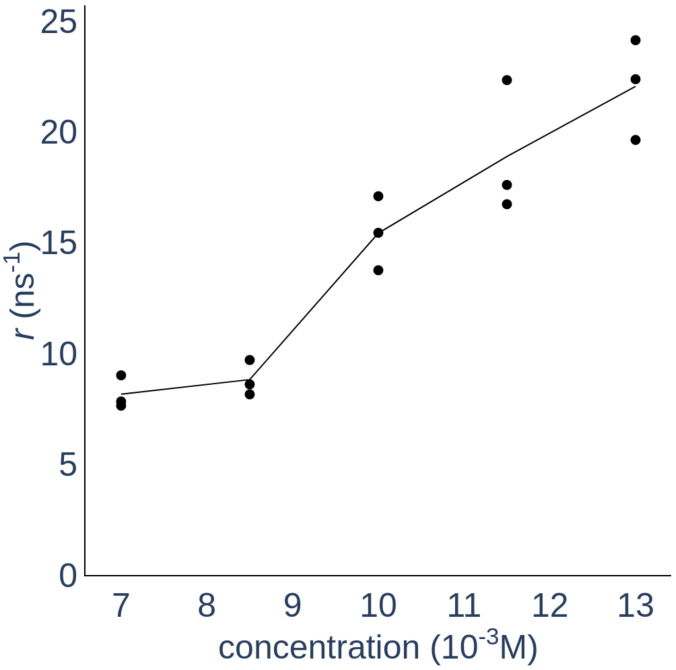


Figure S4. Growth rates, *r*, as a function of the initial monomer concentration, for the 15 simulated aggregation kinetics.

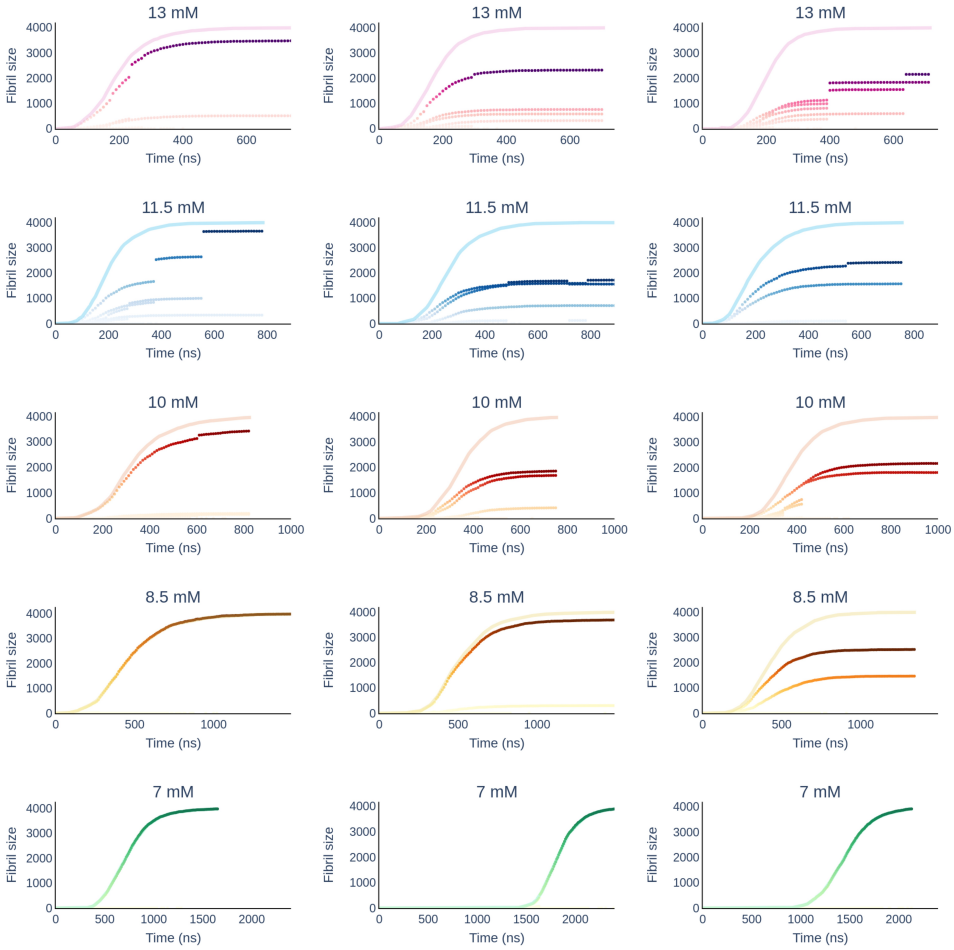
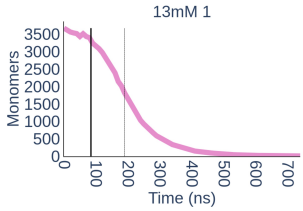
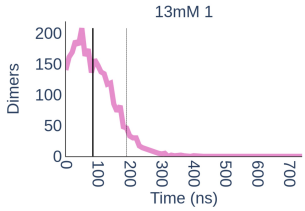
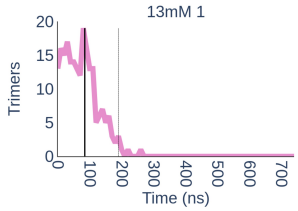
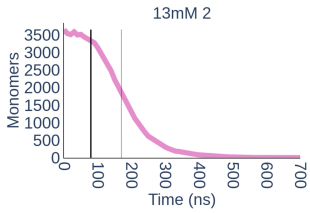


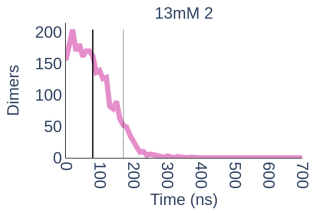
Figure S5. Fibrils grow for all simulations. The size of every fibril formed in the simulation is defined by the number of monomers. At high concentration different fibrils are formed and during the simulation they merge into a single fibril as pointed by the gaps.

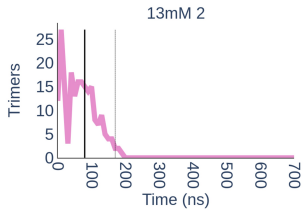

13mM 1
Monomers
Time (ns)

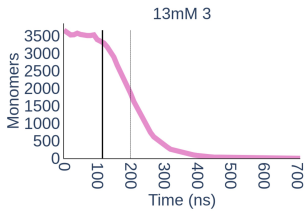

13mM 1
Dimers
Time (ns)

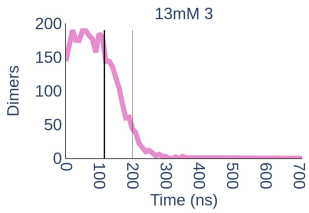

13mM 1
Trimers
Time (ns)

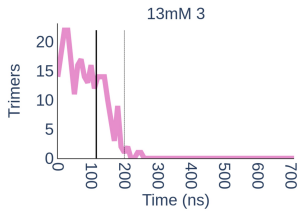

13mM 2
Monomers
Time (ns)

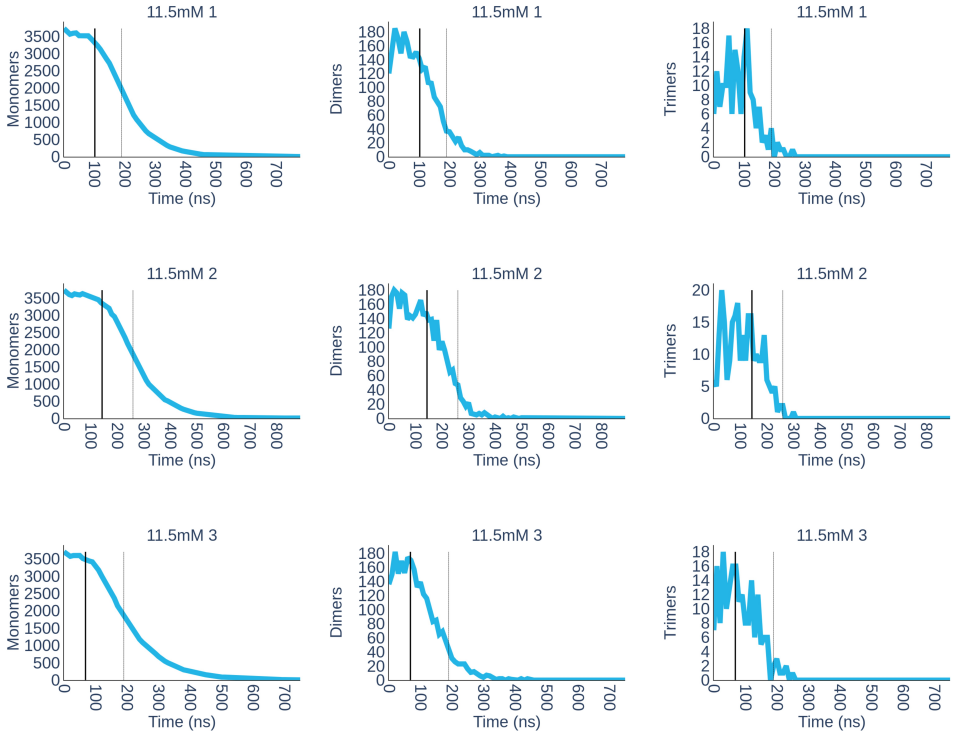

13mM 2
Dimers
Time (ns)

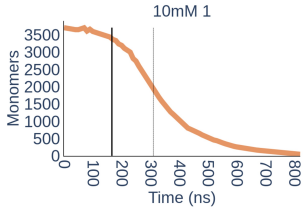

13mM 2
Trimers
Time (ns)

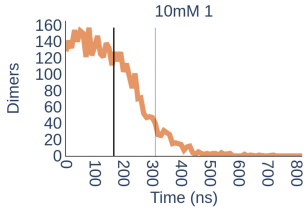

13mM 3
Monomers
Time (ns)

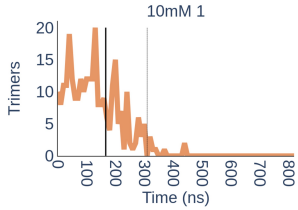

13mM 3
Dimers
Time (ns)

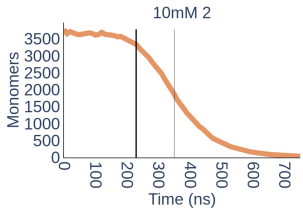

13mM 3
Trimers
Time (ns)

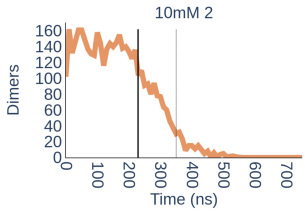

11.5mM 1
Monomers
Time (ns)
11.5mM 1
Dimers
Time (ns)
11.5mM 1
Trimers
Time (ns)
11.5mM 2
Monomers
Time (ns)
11.5mM 2
Dimers
Time (ns)
11.5mM 2
Trimers
Time (ns)
11.5mM 3
Monomers
Time (ns)
11.5mM 3
Dimers
Time (ns)
11.5mM 3
Trimers
Time (ns)

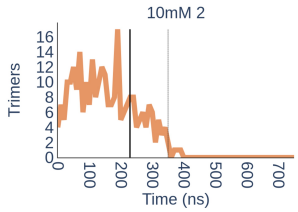

10mM 1
Monomers
Time (ns)

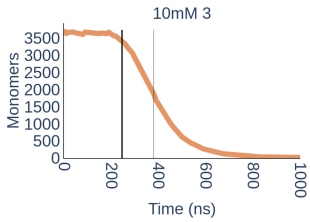

10mM 1
Dimers
Time (ns)

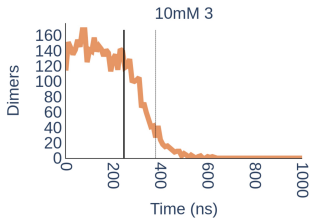

10mM 1
Trimers
Time (ns)

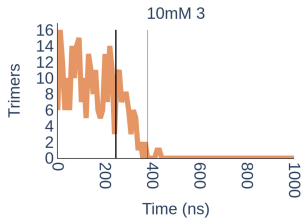

10mM 2
Monomers
Time (ns)

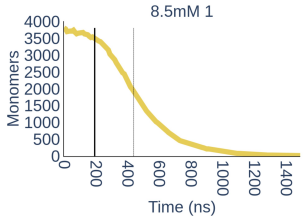

10mM 2
Dimers
Time (ns)

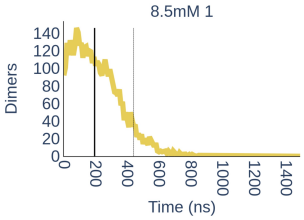

10mM 2
Trimers
Time (ns)

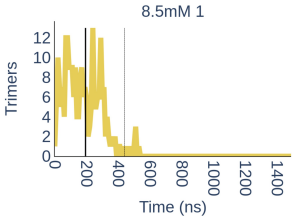

10mM 3
Monomers
Time (ns)

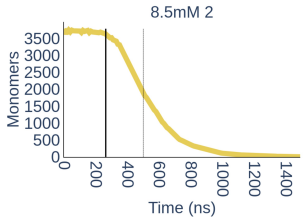

10mM 3
Dimers
Time (ns)

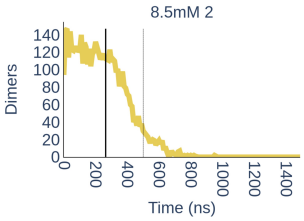

10mM 3
Trimers
Time (ns)

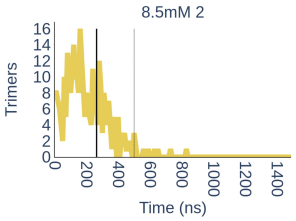

8.5mM 1
Monomers
Time (ns)

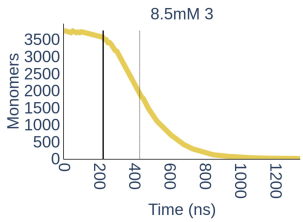

8.5mM 1
Dimers
Time (ns)

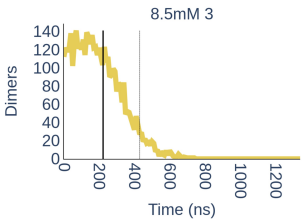

8.5mM 1
Trimers
Time (ns)

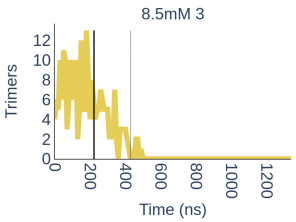

8.5mM 2
Monomers
Time (ns)


8.5mM 2
Dimers
Time (ns)


8.5mM 2
Trimers
Time (ns)


8.5mM 3
Monomers
Time (ns)


8.5mM 3
Dimers
Time (ns)


8.5mM 3
Trimers
Time (ns)

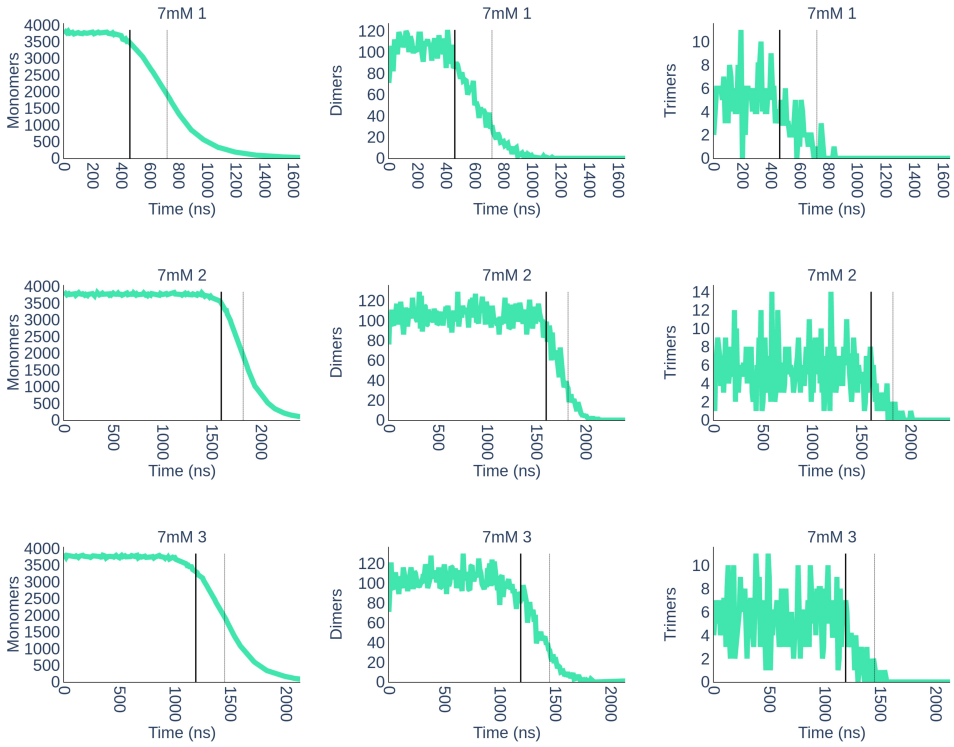


Figure S6. Number of monomers, dimers and trimers formed over time for each simulation.

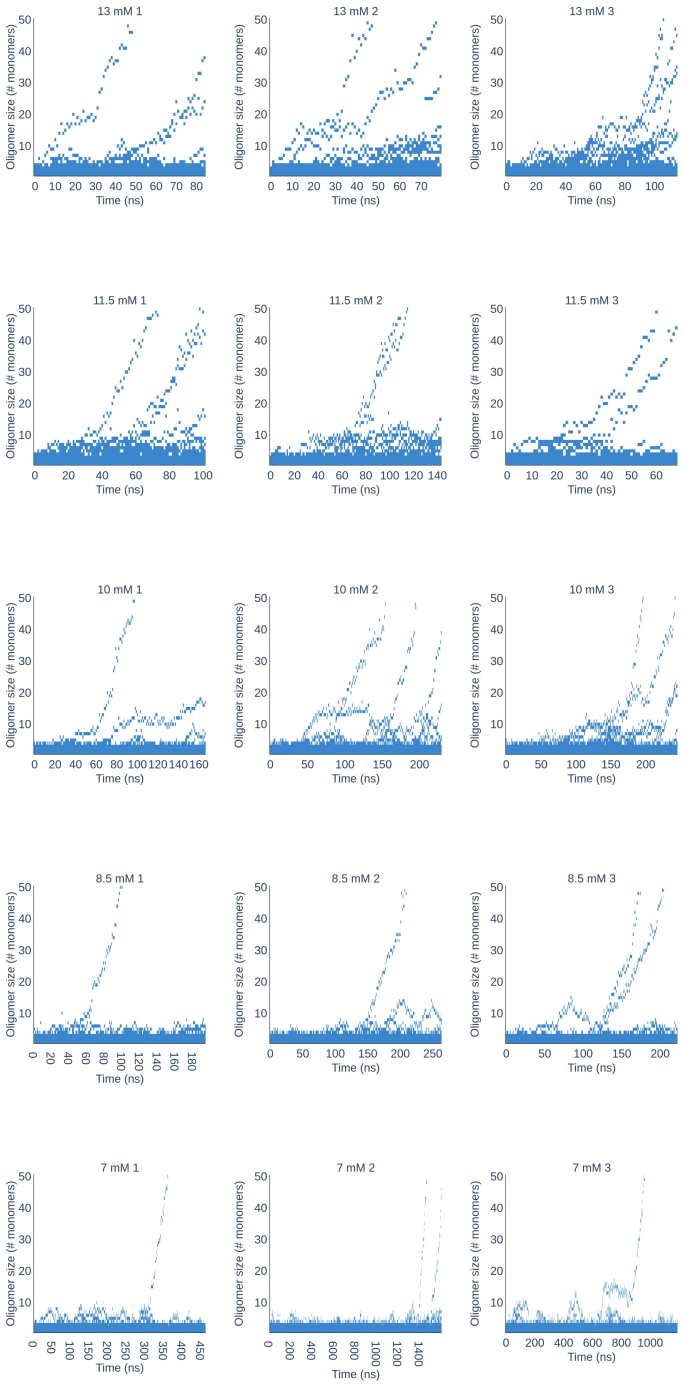
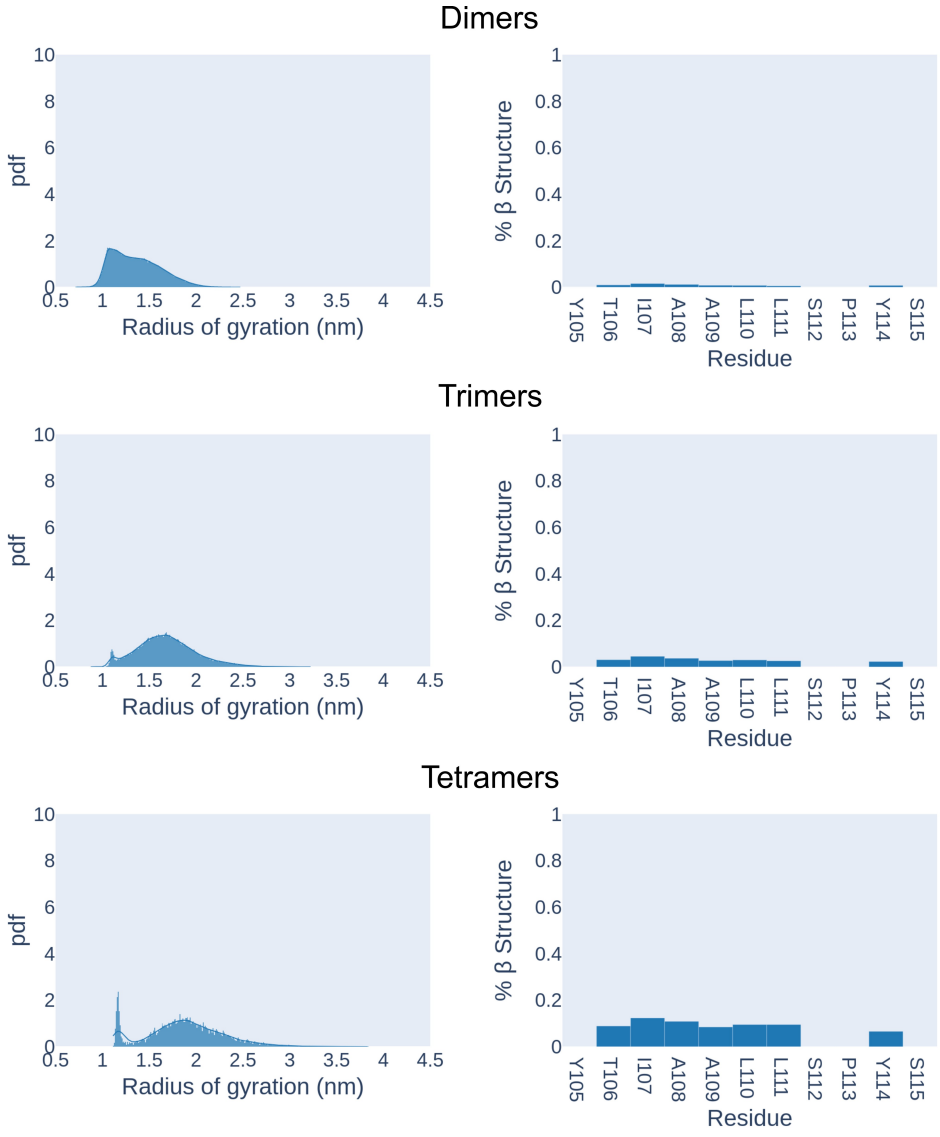
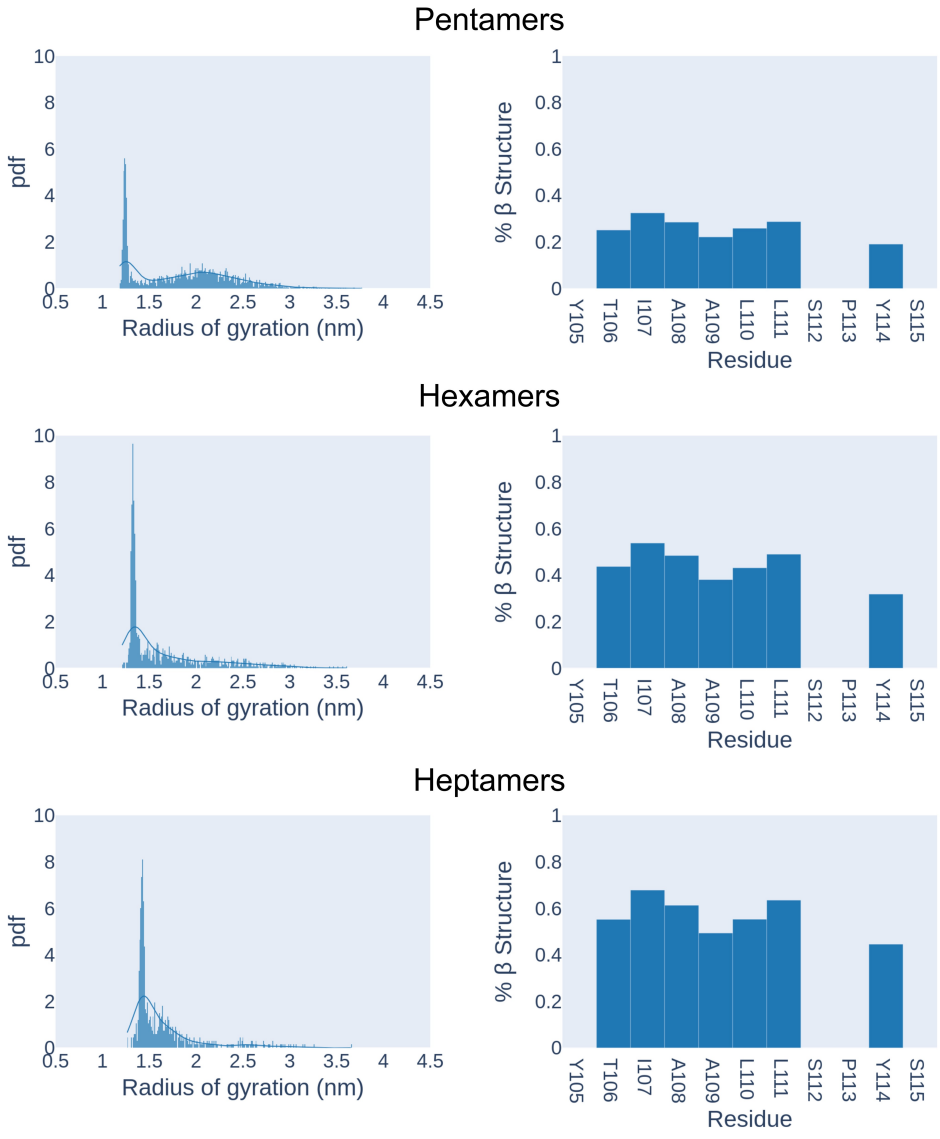


Figure S7. Time resolved evolution of oligomers order for all simulations before t_{lag}.


Dimers
pdf
Radius of gyration (nm)
% β Structure
Residue
Y105
T106
I107
A108
A109
L110
L111
S112
P113
Y114
S115
Trimers
pdf
Radius of gyration (nm)
% β Structure
Residue
Tetramers
pdf
Radius of gyration (nm)
% β Structure
Residue


Pentamers
pdf
Radius of gyration (nm)
% β Structure
Residue
Y105
T106
I107
A108
A109
L110
L111
S112
P113
Y114
S115
Hexamers
pdf
Radius of gyration (nm)
% β Structure
Residue
Heptamers
pdf
Radius of gyration (nm)
% β Structure
Residue

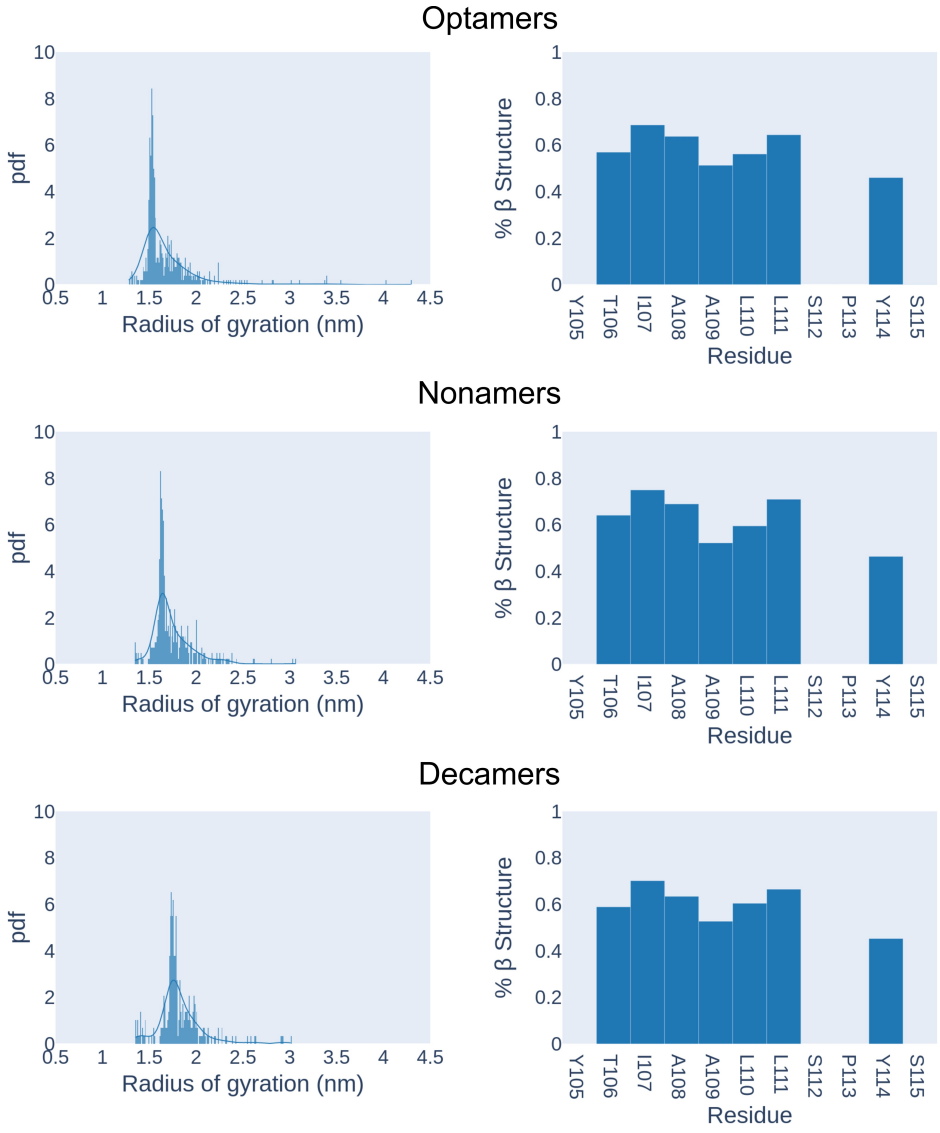


Figure S8. Distribution of the radius of gyration (left) and per residue average beta structure (right) for the oligomers, from dimers to decamers. populated during the multi-*e*GO aggregation kinetics until t_{lag}.

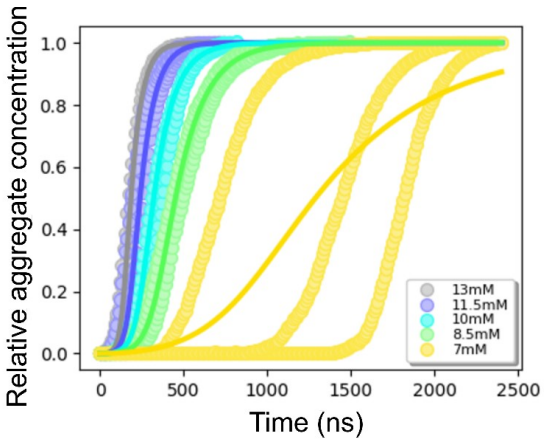


Figure S9. Chemical kinetics analysis by Amylofit of the 15 simulated aggregation kinetics traces. The data are fitted using the multi-step secondary nucleation, unseeded model.

Table S1. Detailed description of fibrils morphologies obtained *in-silico*.

	Replica	# Fibrils	Length (Å)	# β-sheets in filaments	# Filaments	Twist (°)	Number of monomers
13 mM	1	4	368	9	5	-0.22	2100
			232	9	2	-0.29	505
			211	5	3	-0.43	465
			300	7	2	-0.54	512
	2	7	320	10	5	-0.59	1513
			198	4	2	-0.58	206
			175	6	3	-0.73	336
			100	5	1	-4.97	86
			246	10	2	-0.50	324
			221	9	2	-0.32	584
			163	4	3	-0.56	324
	3	5	380	7	4	-0.47	1143
			190	8	2	-0.38	377
			190	8	3	-0.18	602
			255	9	4	-0.33	959
			215	7	4	-0.5	862
11.5 mM	1	4	390	6	4	-0.46	961
			280	12	4	-0.50	1717
			220	8	3	-0.23	909
			210	5	2	-0.85	336
	2	3	330	11	4	-0.24	1717
			300	11	3	-0.24	1560
			250	5	4	-0.30	713
	3	2	260	13	3	-0.33	1574
			310	12	6	-0.33	2420

10 mM	1	1	415	9	5	-0.70	3425
	2	3	366	5	6	-0.79	1690
			270	8	4	-0.23	1860
			214	5	3	-0.31	417
	3	2	372	9	4	-0.37	1811
			315	7	6	-0.24	2168
8.5 mM	1	1	515	9	3	-0.54	3980
	2	1	500	9	4	-0.15	3688
	3	2	355	14	3	-0.19	2524
			370	10	4	-0.34	1469
7 mM	1	1	435	17	5	-0.11	3980
	2	1	480	7	6	-0.48	3359
	3	1	490	13	4	-0.19	3313
7 mM seeded	1	1	443	13	5	-0.35	3959
	2	1	460	14	5	-0.10	3907
	3	1	485	13	5	-0.80	3929

Table S2. Detailed description of *in vitro* fibrils morphologies.

	Crossover (Å)*	Estimated twist (°)	Maximum width (Å)$^{\Delta}$	Estimated number of filaments in a fibril$^{\bullet}$	Width at crossover (Å)
I	1041 ± 24	-0.81	115 ± 11	3	39 ± 6
II	1112 ± 57	-0.76	237 ± 28	6	36 ± 3
III	1121 ± 33	-0.75	171 ± 9	5	36 ± 2
IV	1185 ± 43	-0.71	326 ± 21	8	36 ± 4
V	1060 ± 20	-0.80	247 ± 1	6	35 ± 1
VI	1115 ± 39	-0.76	191 ± 1	5	ND

* Crossover is defined as the length between two consecutive twists.
˜ Estimated twist is calculated by dividing 180° by the estimated number of molecules in a crossover (calculated considering an average distance between molecules of 4.7 Å).
$^{\Delta}$ Width is defined as the maximum diameter at the largest point within a crossover.
$^{\bullet}$ Calculated considering a peptide length of 38 Å

Movie S1 (separate file).
Multi-*e*GO MD trajectory for the aggregation kinetics of the first replica at 13 mM.
Available in Dataset S1, https://dx.doi.org/10.5281/zenodo.6125995

Movie S2 (separate file).
Multi-*e*GO MD trajectory for the aggregation kinetics of the second replica at 7 mM.
Available in Dataset S1, https://dx.doi.org/10.5281/zenodo.6125995

Movie S3 (separate file).
Multi-*e*GO MD trajectory for the aggregation kinetics of the first seeded replica at 7 mM.
Available in Dataset S1, https://dx.doi.org/10.5281/zenodo.6125995

Dataset S1 (separate file).

DOI: 10.5281/zenodo.6125995

https://dx.doi.org/10.5281/zenodo.6125995

Molecular dynamics simulation trajectories of TTR peptide monomers and aggregation kinetics:

- Amber99sb_disp: TTR monomer in explicit solvent using amber99disp force field.
- multi-GO-monomer: TTR monomer simulation using the multi-GO force field
- multi-eGO-monomer: TTR monomer simulation using the multi-*e*GO ensemble force field.
- multi-eGO-XXmM-Y: aggregation kinetics simulations of TTR using the multi-*e*GO force field at 13 mM, 11.5 mM and 10 mM concentration; replicate Y
- .ipynb files: analysis scripts employed for the aggregation kinetics simulations.

Dataset S2 (separate file).

DOI: 10.5281/zenodo.6125424

https://dx.doi.org/10.5281/zenodo.6125424

Molecular dynamics simulation trajectories of TTR peptide aggregation kinetics:

- multi-eGO-XXmM-Y: aggregation kinetics simulations of TTR using the multi-*e*GO force field at 8.5 mM, 7 mM and 7 mM seeded; replicate Y.

Dataset S3 (separate file).

DOI: 10.5281/zenodo.6414572

https://dx.doi.org/10.5281/zenodo.6414572

Multi-GO Molecular dynamics simulation trajectories of TTR peptide and aggregation kinetics; multi-*e*GO oligomer structures and trajectories:

- multi-GO-XXmM: aggregation kinetics simulations of TTR using the multi-*e*GO force field at XXmM concentration.
- Ensembles of all the oligomers, from dimers to decamers, sampled in each multi-*e*GO aggregation kinetics simulation.

SI References

1. P. Robustelli, S. Piana, D. E. Shaw, Developing a molecular dynamics force field for both folded and disordered protein states. *Proc Natl Acad Sci U S A* **115**, E4758–E4766 (2018).
2. U. Essmann, *et al.*, A smooth particle mesh Ewald method. *The Journal of Chemical Physics* **103**, 8577 (1998).
3. B. Hess, P-LINCS: A Parallel Linear Constraint Solver for Molecular Simulation. *Journal of Chemical Theory and Computation* **4**, 116–122 (2007).
4. G. Bussi, D. Donadio, M. Parrinello, Canonical sampling through velocity rescaling. *The Journal of Chemical Physics* **126**, 014101 (2007).
5. M. Bernetti, G. Bussi, Pressure control using stochastic cell rescaling. *The Journal of Chemical Physics* **153**, 114107 (2020).
6. A. W. P. Fitzpatrick, *et al.*, Atomic structure and hierarchical assembly of a cross-β amyloid fibril. *Proc Natl Acad Sci U S A* **110**, 5468–5473 (2013).
7. L. Saelices, *et al.*, Uncovering the mechanism of aggregation of human transthyretin. *Journal of Biological Chemistry* **290**, 28932–28943 (2015).
8. R. T. McGibbon, *et al.*, MDTraj: A Modern Open Library for the Analysis of Molecular Dynamics Trajectories. *Biophysical Journal* **109**, 1528–1532 (2015).
9. N. Michaud-Agrawal, E. J. Denning, T. B. Woolf, O. Beckstein, MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. *Journal of Computational Chemistry* **32**, 2319–2327 (2011).
10. R. Gowers, *et al.*, MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations in *Proceedings of the 15th Python in Science Conference*, (SciPy, 2016), pp. 98–105.
11. C. A. Schneider, W. S. Rasband, K. W. Eliceiri, NIH Image to ImageJ: 25 years of image analysis. *Nature Methods 2012 9:7* **9**, 671–675 (2012).

Rossi, Elio, Gabriella Leccese, Valerio Baldelli, Alessia Bibi, Emanuele Scalone, Carlo Camilloni, Moira Paroni, and Paolo Landini. 2022. "Inactivation of the Pyrimidine Biosynthesis *pyrD* Gene Negatively Affects Biofilm Formation and Virulence Determinants in the Crohn's Disease-Associated Adherent Invasive *Escherichia coli* LF82 Strain" *Microorganisms* 10, no. 3: 537. https://doi.org/10.3390/microorganisms10030537

The following paper was an aim to validate a small molecule interacting with the pyrD Gene. I contributed to the paper by setting up a docking experiment using Autodock Vina and Maestro Schrodinger.

microorganisms

MDPI

Article

Inactivation of the Pyrimidine Biosynthesis *pyrD* Gene Negatively Affects Biofilm Formation and Virulence Determinants in the Crohn's Disease-Associated Adherent Invasive *Escherichia coli* LF82 Strain

Elio Rossi †, Gabriella Leccese †, Valerio Baldelli, Alessia Bibi, Emanuele Scalone, Carlo Camilloni, Moira Paroni * and Paolo Landini *

Department of Biosciences, Università degli Studi di Milano, 20133 Milano, Italy; elio.rossi@unimi.it (E.R.); gabriella.leccese@unimi.it (G.L.); valerio.baldelli@unimi.it (V.B.); alessia.bibi@unimi.it (A.B.); emanuele.scalone@unimi.it (E.S.); carlo.camilloni@unimi.it (C.C.)

* Correspondence: moira.paroni@unimi.it (M.P.); paolo.landini@unimi.it (P.L.)

† These authors contributed equally to this work.

Abstract: In Crohn's disease (CD) patients, the adherent-invasive *Escherichia coli* (AIEC) pathovar contributes to the chronic inflammation typical of the disease via its ability to invade gut epithelial cells and to survive in macrophages. We show that, in the AIEC strain LF82, inactivation of the *pyrD* gene, encoding dihydroorotate dehydrogenase (DHOD), an enzyme of the de novo pyrimidine biosynthetic pathway, completely abolished its ability of to grow in a macrophage environment-mimicking culture medium. In addition, *pyrD* inactivation reduced flagellar motility and strongly affected biofilm formation by downregulating transcription of both type 1 fimbriae and curli subunit genes. Thus, the *pyrD* gene appears to be essential for several cellular processes involved in AIEC virulence. Interestingly, vidofludimus (VF), a DHOD inhibitor, has been proposed as an effective drug in CD treatment. Despite displaying a potentially similar binding mode for both human and *E. coli* DHOD in computational molecular docking experiments, VF showed no activity on either growth or virulence-related processes in LF82. Altogether, our results suggest that the crucial role played by the *pyrD* gene in AIEC virulence, and the presence of structural differences between *E. coli* and human DHOD allowing for the design of specific inhibitors, make *E. coli* DHOD a promising target for therapeutical strategies aiming at counteracting chronic inflammation in CD by acting selectively on its bacterial triggers.

Keywords: adherent-invasive *E. coli* (AIEC); Crohn's disease; dihydroorotate dehydrogenase (DHOD); curli fibers; stress response; virulence; adhesion factors

Citation: Rossi, E.; Leccese, G.; Baldelli, V.; Bibi, A.; Scalone, E.; Camilloni, C.; Paroni, M.; Landini, P. Inactivation of the Pyrimidine Biosynthesis *pyrD* Gene Negatively Affects Biofilm Formation and Virulence Determinants in the Crohn's Disease-Associated Adherent Invasive *Escherichia coli* LF82 Strain. *Microorganisms* **2022**, *10*, 537. https://doi.org/10.3390/microorganisms10030537

Academic Editor: Giovanni Di Bonaventura

Received: 24 January 2022
Accepted: 26 February 2022
Published: 28 February 2022

1. Introduction

Crohn's disease (CD) is characterized by chronic intestinal inflammation resulting from inappropriate and persistent activation of the intestinal mucosal immune system [1]. The pathophysiology of CD is multifactorial, including genetic and environmental factors resulting in an aberrant immune response [2–6]. Arguably, however, one of the main factors in CD pathogenesis is gut microbiota dysbiosis [7,8]. Indeed, metagenomic analysis of human gut microbiota in CD patients has outlined clear pattern changes in microbial abundances in comparison to healthy individuals, such as a depletion of symbionts belonging to the Firmicutes phylum, like Bifidobacteria and Clostridia [9,10]. It is thought that these bacterial species might exert a protective effect by production of short chain fatty acids, such as butyrate, with anti-inflammatory effects [11]. In contrast, adherent-invasive *Escherichia coli* (AIEC) have consistently been found to be enriched in ileal specimens from CD patients in comparison to healthy subjects [12]. Although early reports suggested a more predominant role of AIEC in CD than in ulcerative colitis (UC), the other main

inflammatory bowel disease, other works point to a role of the pathogen even in the latter disease [13,14]. The important role played by AIEC in IBD pathogenicity depends on their ability to invade intestinal epithelial cells and increase their permeability, triggering secretion of pro-inflammatory cytokines, thus ultimately resulting in chronic inflammation [15,16]. In addition, AIEC strains are able to survive and even replicate within human macrophages [17], leading to over-induction of the innate immune response. An extensive genetic and functional analysis, mostly carried out on the AIEC strain LF82, has identified several virulence factors and regulatory pathways involved in cell invasion and the induction of chronic inflammation in the human host, namely adhesion factors such as type 1 pili [18,19], specific invasive genes such as *ibeA* [20], regulatory genes such as the σ^E network [21], and efflux pumps [22].

Due to the direct role played by microorganisms in the disease, antibiotic treatments are routinely used in Crohn's disease treatment: for instance, an association of metronidazole and of the anti-inflammatory drug azathioprine has been shown to be effective in preventive post-operative Crohn's recurrence [23]. Interestingly, azathioprine is itself an antimicrobial drug, able to inhibit growth of *Mycobacterium avium* ssp. *paratuberculosis*, another bacterium associated with Crohn's disease [24], and to impair virulence factors' production in AIEC [25]. However, the long-term effectiveness of antibiotic treatments is questionable, as they seem to further promote dysbiosis, thus not tackling, or sometimes even exacerbating, one of the main triggers for chronic inflammation [26]. To ameliorate dysbiosis, probiotics, especially bacteria of the Lactobacillus and Bifidobacterium genera, have been utilised on CD patients, but clinical studies on probiotic efficacy are inconclusive [27–29]. Recently, we have performed an in vitro study comparing the immunomodulatory effects of Lactobacilli and Bifidobacteria probiotic strains, showing that production of pro-inflammatory cytokines and the activation of the IL-23/Th17 axis in response to AIEC is effectively counteracted by probiotics in cells from healthy subjects and from individuals suffering with ulcerative colitis, but not in cells from CD patients [30], thus suggesting that probiotics might have a limited impact in CD.

In this manuscript, we report that a mutation inactivating the *pyrD* gene of the AIEC LF82 strain results in the loss of its ability to grow in a macrophage environment-mimicking medium and inhibits biofilm and virulence determinants, such as curli fibers, type 1 fimbriae, and flagellar motility. The *pyrD* gene encodes dihydroorotate dehydrogenase (DHOD), an enzyme involved in the de novo biosynthesis of pyrimidines. Thus, our results point to a role of intracellular pyrimidine concentrations as a regulatory signal for genes involved in virulence and in host interaction, as already identified for other pathogenic bacteria such as *Pseudomonas aeruginosa* and *Shigella flexneri* [31,32]. Based on the observation that *E. coli* DHOD not only differs from the human enzyme, but also has a low sequence identity to DHOD from probiotic species like *Lactococcus lactis*, we propose that the development of specific inhibitors of *E. coli* DHOD might be a potentially interesting therapeutic strategy for CD remission via selective inhibition of AIEC growth and virulence.

2. Materials and Methods

2.1. Transposon Mutagenesis, Mutant Identification, and Screening on Acidic and Nutrient- Poor Medium

Transposon insertion mutagenesis was carried out using the EZ-Tn5<R6Kγori/KAN-2> transposome (Lucigen, Middleton, WI, USA) on LF82, an AIEC strain originally isolated from an ileal biopsy specimen from a CD patient [33]. Transposon mutagenesis and the determination of a transposon insertion site by rescue cloning were carried out according to the manufacturer's instructions. Through multiple transformation rounds, we obtained a library of 10,058 kanamycin resistant mutants, i.e., a 2.2× coverage of the *E. coli* LF82 genome (4534 genes). Growth in acidic and nutrient-poor medium (Acid Medium: 100 mM bis-Tris, 0.1% Casamino Acids, 0.16% glycerol, and 10 μM $MgCl_2$, and the pH was adjusted to 5.8 with 10M HCl) [34] was determined as follows: overnight bacterial cultures grown

in Yeast extract/Casamino acid (YESCA) medium (10 g/L casamino acids, 1.5 g/L Yeast extract) at 37 °C were diluted to OD_{600} = 0.02 in Acid Medium and overnight growth was constantly monitored in a microplate reader (SAFAS MP96).

2.2. Adhesion Factor Detection, Biofilm Quantification, Motility Assay, MIC Determination and LPS Integrity Evaluation

For phenotypical assays, bacteria were grown in YESCA medium, either at 30 °C or at 37 °C. When necessary, uracil was added at 0.25 mM from a 50 mM uracil solution in 50% dimethyl sulfoxide (DMSO); 0.25% DMSO was always added to control cultures. For adhesion factor detection, bacteria were grown on YESCA agar medium supplemented with either 0.004% Congo red and 0.002% Coomassie blue (CR) or 0.005% calcofluor (CF); dyes were always added to the medium after autoclaving. Bacteria were grown for 24 h at either 30 °C or 37 °C; phenotypes were better detected after a further 24–48 h incubation at 4 °C.

Biofilm formation was determined using the crystal violet (CV) assay as described previously [25]. For flagellar motility assays, bacterial cells were grown overnight in YESCA and normalized to an OD_{600} = 1. For each culture, 3 µL were spotted at the centre of a motility agar plate in the same growth medium used for overnight cultures, supplemented with 0.3% agar. Motility was determined by the diameter of the area colonized by the bacteria after 15 h of growth at 37 °C.

Determination of minimal inhibitory concentrations (MIC) of vidofludimus for LF82 was performed using standard 1:2 dilution methods in liquid YESCA medium, using an inoculum of 2×10^5 cfu/mL. To assess vidofludimus antimicrobial activity on Gram positive bacteria, we performed MIC determination on *Bacillus cereus* strain 971 and *Staphylococcus epidermidis* ATCC 155.

To evaluate LPS integrity, overnight cultures of LF82 and LF82*pyrD::*Tn5 mutants grown in YESCA medium were normalized to OD_{600} = 1.0 and serially diluted 1:10 six times. 3 µL of each dilution were spotted on LB and MacConkey agar in triplicates and plates were incubated overnight at 37 °C.

2.3. Gene Expression Determination by Quantitative Real-Time PCR

For RNA isolation, strains were grown either to mid-log phase or to the onset of the stationary phase in YESCA medium at 30 °C. Bacterial cells were harvested by centrifugation at 10,000× *g* for 5 min at 4 °C and cell pellets resuspended in 300 µL of DNA/RNA Shield (Zymo Research, Irvine, CA, USA). Total RNA was extracted using the Quick-RNA Miniprep Kit (Zymo Research, Irvine, CA, USA) after the addition of 1 mg/mL Lysozyme and 400 µg/mL Proteinase K to the ruspended pellet. RNA samples were checked by agarose gel electrophoresis to assess lack of degradation and quantified spectrophotometrically. Genomic DNA removal and reverse transcription were performed on 1 µg of total RNA, along with negative control samples incubated without reverse transcriptase using the QuantiTect Reverse Transcription Kit (Qiagen, Germantown, MD, USA). cDNA synthesis efficiency was verified by electrophoresis on agarose gel in comparison to negative controls. Real-time PCR was performed using the SYBR Green PCR master mixture, and the results were determined with a Rotor-Gene 3000 detection system (Corbett Research, Saffron Walden, UK). Reaction mixtures (15 µL) included 0.1 µg cDNA and 300 nM primers in the reaction buffer and enzyme supplied by the manufacturer. Primer sequences are listed in Table S1. A minimum of three biologically independent experiments were considered for analysis; negative control samples (i.e., non-retrotranscribed RNA) never showed significant threshold cycles. The relative transcript amounts were determined using 16S rRNA as the reference gene ([CtGene of interest-Ct16S] = ΔCt value).

2.4. Computational Models for Vidofludimus Binding to Dihydroorotate Dehydrogenase (DHOD) from Different Organisms

Docking calculations were performed using Glide version 2021.3. Receptor structures (hDHOD, PDB code: 2PRL and EcDHOD, PDB code: 1F76) were prepared using the Protein

Preparation Wizard of the Maestro graphical user interface (Schrödinger suite https://www.schrodinger.com/, Schrödinger, Mannheim, Germany) by removing all crystallographic waters and additives and optimizing the orientation of hydrogen bonds and the protonation state of histidine, aspartic acid and glutamic acid. The protein preparation was followed by a restrained minimization of the whole system. Target grids were built on receptor structures. Ligand sampling was set to 'Flexible' and we included the Epik state penalties in the docking score. For each compound, 5 poses were saved after a post-minimization of the ligand structure within the binding site. The docking protocol was initially tested for its ability to reproduce the binding mode of the native R2C in the crystal structure (2PRL). The program was successful in reproducing the experimentally determined binding mode of the compound as it corresponds to the best-scored pose. Then, docking with VF was followed by a 50 ns long Molecular Dynamics (MD) simulation using the first predicted pose. MD simulations were performed using the Desmond package of the Schrödinger suite. The chosen force field was OPLS4. The systems were solvated in a dodecahedron hexagon box with a 10 Å buffer and using TIP3P as water and 0.15 M of NaCl. MD simulation was preceded by energy minimization and equilibration steps.

2.5. Statistical Analysis

Statistical analysis was performed with Prism 9 software (GraphPad Software, San Diego, CA, USA). Student's *t*-tests for unpaired or paired samples were used to evaluate differences between means. Statistical significance between the means of more than two groups were performed using one-way ANOVA computing the Tukey's multiple comparisons test to evaluate differences between groups.

3. Results

3.1. Mutant Selection in an Acidic and Nutrient Stress Medium Mimicking the Macrophage Vacuole Environment

In order to identify genes that might be involved in AIEC virulence, we created a transposon mutagenesis library in the LF82 strain, using the EZ-Tn5<R6Kγori/KAN-2> transposome (Lucigen, WI, USA). Transposon insertion mutants were screened for their ability to grow in acid medium, recreating the harsh environment of the macrophage vacuole, characterized by low pH and limited nutrient availability [34]. Indeed, in this medium, even the LF82 parental strain was only able to carry out roughly two replications, with a maximal growth rate of 0.24 h^{-1} (Figure 1A and Figure S1). Out of the 10058 transposon mutants screened, 141 showed either a complete loss or a strong reduction in their ability to grow in acid medium, which was, however, often accompanied by growth defects also in the YESCA medium. Thus, we focused our attention on one mutant whose growth was not affected in YESCA, but totally abolished in acid medium. This mutant was compared to its parental LF82 strain in a more standardized experiment in which fresh overnight cultures of either strain grown in YESCA were resuspended to an OD_{600} = 0.085 and incubated overnight in acid medium, confirming total lack of growth by the mutant (Figure 1A).

The transposon insertion was mapped within the *pyrD* gene, encoding DHOD, the enzyme catalyzing the fourth step of the de novo pyrimidine biosynthesis, and the mutant strain will be referred to as LF82*pyrD*::Tn5 from now on. The transposon insertion site lies at nucleotide 368 of the 1011-bp long *pyrD* gene, immediately downstream of a portion of the gene coding for a domain involved in substrate binding, thus suggesting functional inactivation of the *pyrD* gene (Figure 1B). Loss of *pyrD* function was confirmed by restoration of LF82*pyrD*::Tn5 growth in acid medium by addition of 0.25 mM uracil (Figure 1A), which would also suggest that absence of exogenous pyrimidines, rather than acid sensitivity, is the reason for LF82*pyrD*::Tn5 inability to grow in this medium. Indeed, AIEC mutants deficient in pyrimidine biosynthesis have already been shown to be unable to survive in macrophages [35]. Although exogenous uracil rescued its defective phenotype in acid medium, the LF82*pyrD*::Tn5 mutant still showed a longer lag phase (120 vs. 75 min)

and a slightly, albeit statistically significant, slower growth rate (0.21 vs. 0.24 h^{-1}) than its parental strain even when exogenous uracil was provided (Figure 1A and Figure S1), suggesting that the *pyrD* mutation might affect LF82 fitness in acid medium even in the presence of excess pyrimidine availability.

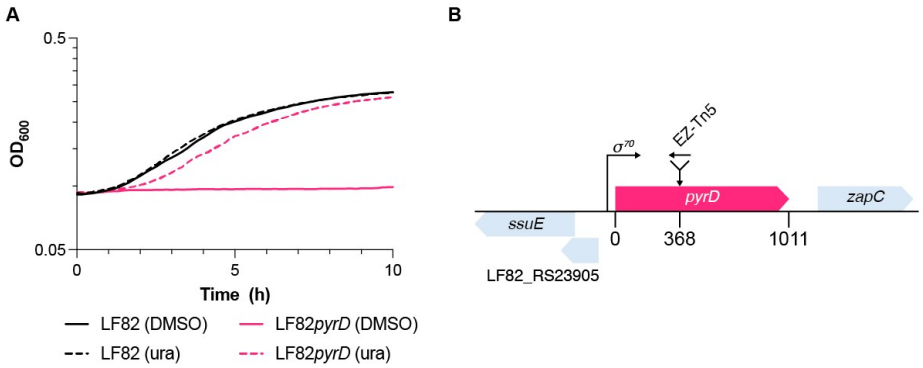


Figure 1. (**A**) Growth curves for LF82 (black) and LF82*pyrD*::Tn5 (LF82*pyrD* magenta) in acid medium mimicking the macrophage environment, either in the presence of 0.25 mM uracil (dashed lines) or 0.25% dimethyl sulfoxide DMSO (solid lines); (**B**) Localization of the EZ-Tn5<R6Kγori/KAN-2> transposon insertion site in the *pyrD* gene.

3.2. The LF82pyrD::Tn5 Mutant Is Impaired in Biofilm Formation and Adhesion Factors' Production and Displays a Slightly Reduced Flagellar Motility

In previous works, we showed that perturbations of intracellular nucleotide pools strongly affect biofilm formation and adhesion factors' production in *E. coli*: indeed, in the *E. coli* MG1655 laboratory strain, mutations in de novo pyrimidine biosynthesis genes strongly downregulate curli production, a major adhesion factor [36]. Likewise, the purine analogue azathioprine, and other 6-mercaptopurine drugs, can impair both curli fibers' production and cell motility in LF82 [25].

To assess whether the *pyrD* mutation could also impair bacterial adhesion and cell motility, both involved in AIEC virulence, we performed crystal violet biofilm staining and swimming motility assays comparing LF82 to its LF82*pyrD*::Tn5 derivative (Figure 2A). Since curli fibers production in most *E. coli* strains, including LF82 [25], is strongly inhibited at 37 °C in liquid media, we also performed biofilm and adhesion factor assays at 30 °C. Indeed, the LF82 wild type strain was more proficient in biofilm formation at 30 °C than at 37 °C, consistent with the production of curli fibers at lower growth temperatures (Figure 2A); however, the LF82*pyrD*::Tn5 mutant strain displayed a significant reduction in biofilm formation at both 30 °C and 37 °C, possibly suggesting the inhibition of multiple adhesion factors by the *pyrD* mutation (Figure 2A). The impaired ability to form biofilm by the LF82*pyrD*::Tn5 mutant is not due to any reduction in overall growth in the YESCA medium (Figure S2).

To further confirm the effects of the LF82*pyrD*::Tn5 mutation on curli fibers production, we performed phenotypic assays on solid medium supplemented with either Congo red or Calcofluor, dyes that can bind both curli fibers and the extracellular polysaccharide cellulose, often co-produced with curli, as well as other cell surface structures. While LF82 showed, respectively, red and fluorescent phenotypes on either Congo red (CR)- or calcofluor (CF)-supplemented media, binding to either dye was affected in the LF82*pyrD*::Tn5 mutant, particularly at 37 °C (Figure 2B), at which the mutant strain totally loses fluorescence on CF-supplemented medium and the pinkish coloration on CR displayed by its parental strain. Supplementation of 0.25 mM uracil fully overcame the effects of the *pyrD* mutations, restoring the ability of LF82*pyrD*::Tn5 to bind either dye (Figure 2B), thus strongly suggesting that LF82*pyrD*::Tn5 phenotypes are indeed due to a reduction of the pyrimidine availability.

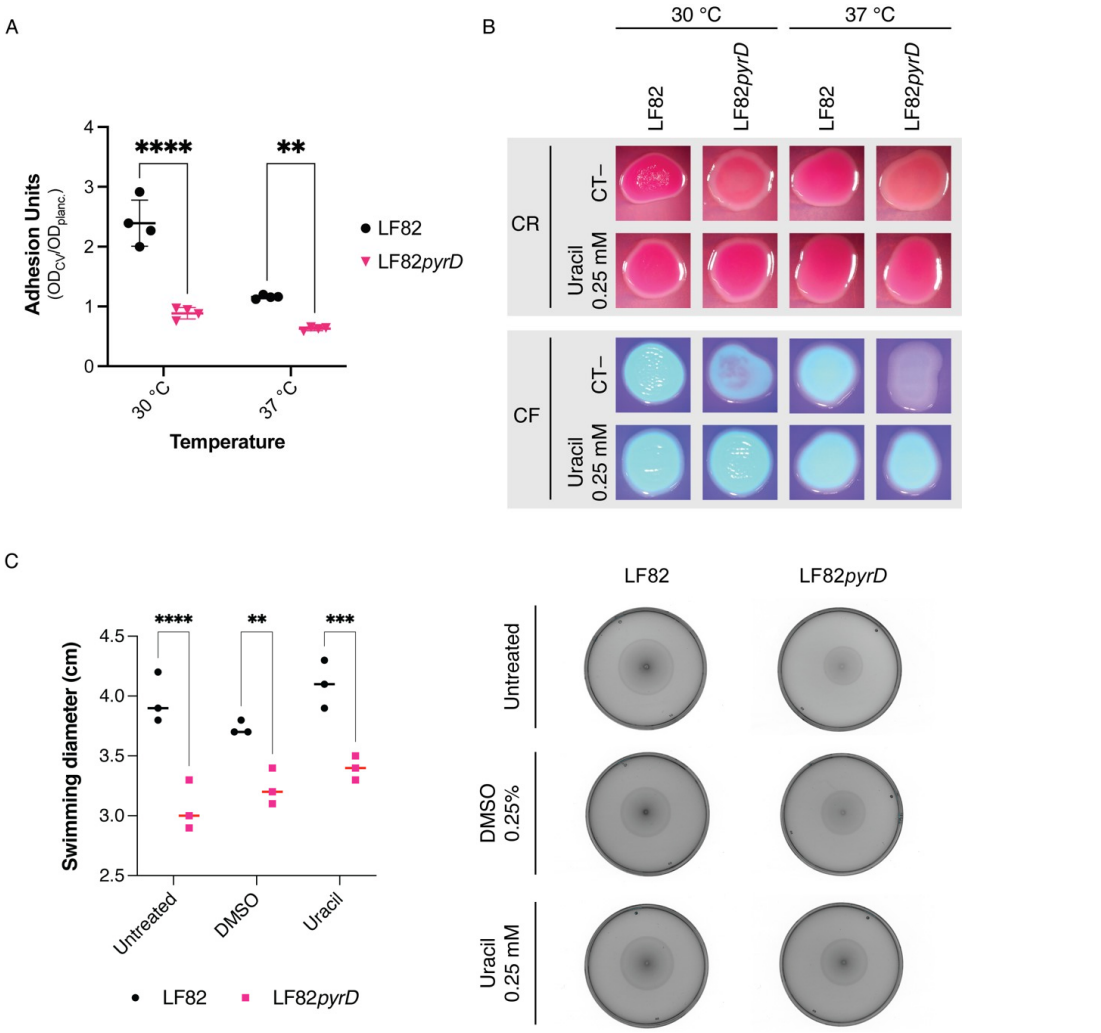


Figure 2. (**A**) Biofilm formation, measured as surface adhesion to polystyrene microtiter plates, using the Crystal violet method. Cultures were grown overnight in Yeast extract/Casamino acid (YESCA) medium at either 30 °C or 37 °C. Results of four independent values and median are shown. **, *p*-value < 0.01; ****, *p*-value < 0.0001, One-way ANOVA with Tukey's test for multiple comparisons; (**B**) LF82 and LF82*pyrD*::Tn5 (LF82*pyrD*) phenotypes grown either at 30 °C or 37 °C on YESCA medium, either with or without uracil supplementation, in the presence of either Congo red (CR) or Calcofluor (CF); (**C**) Swimming motility on YESCA soft agar plates (typical experiment, right) and measurement of halo diameters (three independent experiments and median are shown, left). **, *p*-value < 0.01, ***, *p*-value < 0.001, ****, *p*-value < 0.0001, Two-way ANOVA with Šidák's test for multiple comparisons.

Finally, LF82*pyrD*::Tn5 is also impaired in cellular motility (Figure 2C), suggesting that perturbation of intracellular pyrimidine nucleotide pools also affects this important virulence-related cell process in AIEC [21]. Interestingly, however, unlike for CR and CF phenotypes (Figure 2B), uracil supplementation did not restore full cellular motility to the mutant strain, thus suggesting that the modulation of flagellar motility might depend on the relative concentrations of some de novo pyrimidine biosynthesis pathway intermediates rather than the overall pyrimidine intracellular concentration.

Results in Figure 2 show that the LF82*pyrD::*Tn5 strain is hindered in production of adhesion factors, in biofilm formation and in cell motility. We reasoned that mutations in de novo pyrimidine biosynthesis might reduce the pool of UTP available for the activation of sugar precursors, thus affecting the production of extracellular polysaccharides such as peptidoglycan and lipopolysaccharide (LPS). In particular, mutations affecting LPS integrity have been shown to affect both adhesion to solid surfaces and cell motility in *E. coli* [37]. Thus, in order to verify whether the LPS structure might be impaired in the LF82*pyrD::*Tn5 strain, we determined its viability on MacConkey medium, as production of incomplete or aberrant LPS makes *E. coli* sensitive to bile salts present in this medium [38]. As shown in Figure S3A, we could not detect any reduction in viability comparing LF82*pyrD::*Tn5 strain growth in LB vs. MacConkey medium, thus suggesting that inactivation of the *pyrD* gene does not result in extensive perturbation of the LPS structure. As further assessment of the integrity of the cell envelope in the LF82*pyrD::*Tn5 mutant strain, we tested the activation of the σ^E-dependent *lptD* gene [39], as the σ^E regulon is activated in response to changes in LPS structure [40]: again, no induction of *lptD* transcription was observed in the *pyrD* mutant strain (Figure S3B).

3.3. The pyrD::Tn5 Mutation Results in Transcription Downregulation of Genes Encoding Curli Fibers and Type 1 Fimbriae

Inhibition of curli production in the LF82*pyrD::*Tn5 mutant would be consistent with our previous observations that inactivation of genes of the de novo pyrimidine biosynthetic pathway negatively impacts curli and cellulose production in the *E. coli* MG1655 laboratory strain via the downregulation of the *csgDEFG* operon [36]. The first gene of the *csgDEFG* operon codes for the CsgD regulatory protein that activates transcription of the *csgBAC* operon, encoding curli structural subunits, and of the *adrA* (*dgcC*) gene, which in turn promotes cellulose production by acting on the product of the cellulose biosynthetic *bcs* operon at the enzymatic level [41,42]. To verify this hypothesis, we compared expression of curli- and cellulose-related genes in LF82*pyrD::*Tn5 versus its parental strain. Gene expression was determined at 30 °C, both during exponential growth and at the onset of the stationary phase (Figure 3).

Inactivation of the *pyrD* gene strongly affected *csgB* expression, downregulating it by more than 10-fold, particularly at the onset of the stationary phase, in which *csgB* expression increases by 26-fold compared to the exponential phase (Figure 3). In contrast, it did not significantly affect *csgD* transcription levels, which, in contrast, were higher during the exponential phase, showing a ca. 33-fold decrease at the onset of the stationary phase (Figure 3). Our results suggest that in AIEC, unlike the *E. coli* MG1655 laboratory strain, perturbation of intracellular pyrimidine pools due to *pyrD* inactivation specifically targets *csgBAC*, encoding curli subunits, without affecting the CsgD regulon at large. Indeed, transcription of the cellulose-related *adrA* and *bscA* genes were not significantly different in LF82*pyrD::*Tn5 compared to its parental strain (Figure 3, ca. 1.3 fold WT vs. mutant). However, it must be pointed out that *adrA* transcription levels were almost undetectable also in the LF82 parental strain, possibly suggesting that *adrA* might not be transcribed at significant levels in AIEC, at least in our experimental conditions.

In addition to the curli structural operon, the *pyrD::*Tn5 mutation results in strong downregulation (ca. 20-fold in exponential phase, Figure 3) of the *fimA* gene, encoding the major subunit of type 1 fimbriae, an important virulence factor in AIEC, promoting its adhesion to epithelial cells [18,19], thus suggestted that perturbation of pyrimidine nucleotide pools negatively affects multiple adhesion factors in LF82. In contrast, *fliC*, encoding the main flagellar subunit, was not differently expressed in the LF82*pyrD::*Tn5 mutant, suggesting that reduced motility in this strain (Figure 2C) might be mediated at flagellar motility rather than at the flagellar gene transcription level.

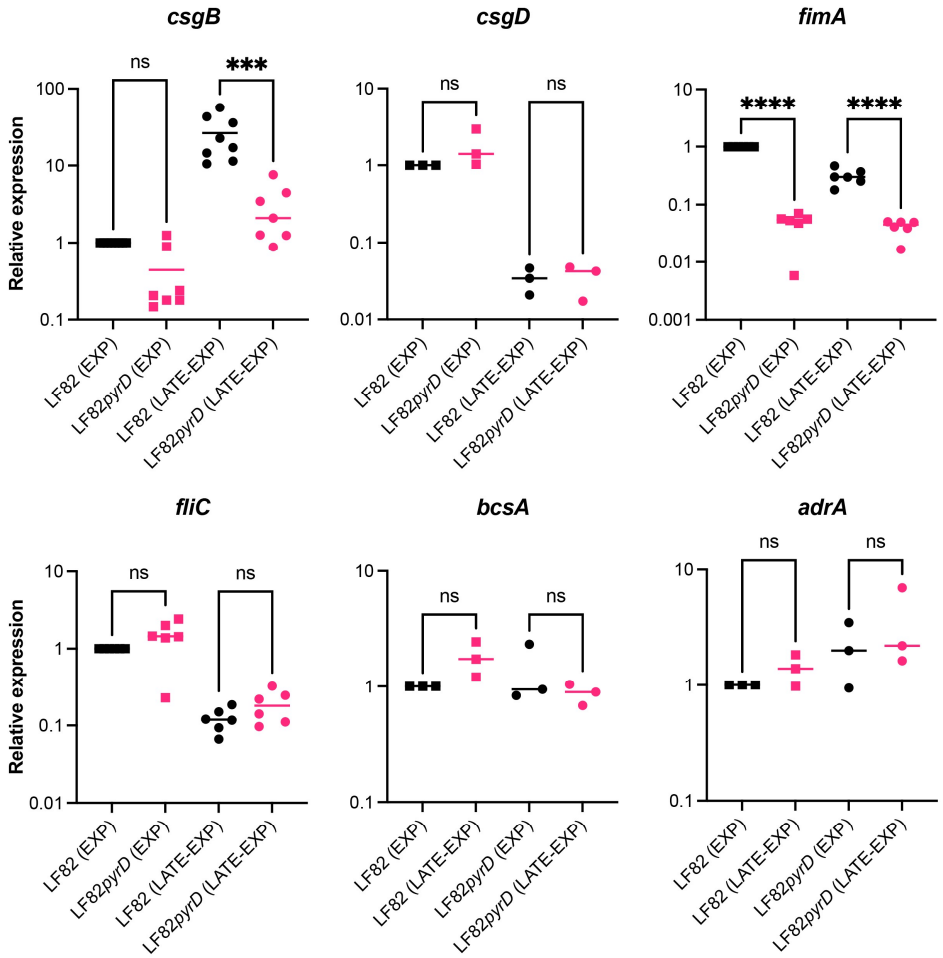


Figure 3. Determination of gene expression levels in the LF82 vs. the LF82*pyrD::*Tn5 (LF82*pyrD*) strain by qRT-PCR. RNA was extracted from bacterial cultures grown in YESCA medium at 30 °C either during exponential growth (EXP, OD_{600nm} = 0.4) or transition to stationary phase (LATE-EXP, OD_{600nm} = 1.0). Values are expressed as arbitrary units; transcription levels in the LF82 strain during exponential growth are set to 1. Data are from at least three independent experiments and median are shown. ns, not significant; ***, *p*-value < 0.001, ****, *p*-value < 0.0001 One-way ANOVA with Tukey's test for multiple comparisons.

3.4. Dihydroorotate Dehydrogenase (DHOD) as Potential Drug Target

Results presented so far suggest that inactivation of the *pyrD* gene, resulting in loss of dihydroorotate dehydrogenase (DHOD) activity, and consequent perturbation of intracellular pyrimidine pools, affects a variety of cellular processes involved in AIEC host colonization and virulence, such as cell adhesion and motility (Figure 2), and ability to grow in environments devoid of exogenous pyrimidines (Figure 1), such as the vacuoles formed during macrophage infection by AIEC [35]. Interestingly, inhibitors of the human DHOD protein, like vidofludimus (VF), have been widely studied as anti-inflammatory drugs in several pathologies, including CD [43]. We recently showed that purine synthesis inhibitors 6-mercaptopurines, a class of widely used anti-inflammatory drugs, possess antimicrobial activity against LF82 and can inhibit biofilm formation and motility in this bacterium at subinhibitory concentrations for growth [25], thus suggesting that their antimicrobial activity might contribute to their effectiveness in Crohn's disease treatment. We

hypothesized that VF might also inhibit *E. coli* DHOD, thus mimicking the effects of the *pyrD* mutation and hampering AIEC virulence, and that such inhibition might contribute to its anti-inflammatory action.

Unlike 6-mercaptopurines, however, VF showed no antibacterial activity on LF82 up to 256 µg/mL (Figure 4B), and it failed to inhibit biofilm formation at either 30 °C or 37 °C (Figure 4C), while only promoting a very slight phenotypic change of LF82 on CR-supplemented media at 37 °C (Figure 4A). VF even induced a slight increase in biofilm formation that was only statistically significant at 37 °C and did not show a clear dose-dependence, suggesting that these effects are not mediated by inhibition of DHOD activity. The lack of significant biological effects on LF82 by VF might be due either to inability to enter the bacterial cells of Gram negative bacteria, despite its small molecular size (molecular mass = 355.12), or to low binding affinity to bacterial DHOD. VF showed very poor antimicrobial activity on the Gram positive bacteria *Bacillus cereus* and *Staphylococcus epidermidis* (MIC = 128 µg/mL for either species, Figure 4B), which are typically more sensitive to antimicrobial agents not able to cross the outer membrane of Gram negative bacteria, possibly suggesting poor inhibition of bacterial DHOD. Even in M9Glucose minimal medium, in which pyrimidine nucleotides are exclusively synthesized by the de novo biosynthetic pathway, and thus growth is totally dependent on the DHOD activity, VF showed no antimicrobial activity against *E. coli* LF82 (Figure 4B).

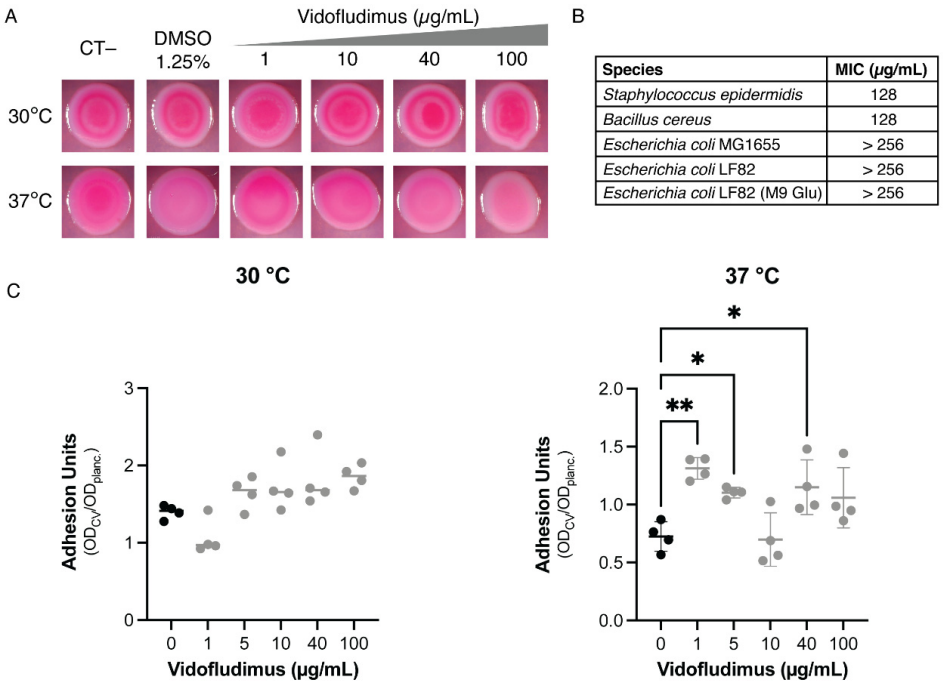


Figure 4. Effects of vidofludimus (VF) on LF82 phenotypes and on microbial growth. (**A**) Phenotypes on Congo red-supplemented YESCA agar medium, either at 30 °C or 37 °C, in the presence of growing VF concentrations. The final DMSO concentration was kept at 1.25% for all VF concentrations tested. (**B**) Minimal inhibitory concentrations (MIC) on *E. coli* (LF82, in YESCA and M9 Glu 0.2% media, and the laboratory strain MG1655, in YESCA medium) and the Gram positive bacteria *Bacillus cereus* and *Staphylococcus epidermidis* (in YESCA medium). (**C**) Biofilm formation, measured as surface adhesion to polystyrene microtiter plates, using the Crystal violet method, in the presence of increasing VF concentrations. Cultures were grown overnight in YESCA medium at either 30 °C or 37 °C. *, p-value < 0.05, **, p-value < 0.01 One-way ANOVA with Tukey's test for multiple comparisons. Only significant values are shown.

3.5. In Silico Analysis of Vidofludimus/DHOD Interaction

To further evaluate VF-DHOD interactions, we carried out in silico molecular docking experiments using GLIDE (see Materials and Methods). First we redocked the ligand 5-methoxy-2-[(4-phenoxyphenyl)amino]benzoic acid (R2C) to the corresponding experimentally-determined structure (PDB ID: 2PRL). The ligand shows a GLIDE score of −9.6, with a binding pose identical to the experimental one, where the main interaction is between the carboxylic acid of R2C and R136 and Q47 along with many other polar and non-polar interactions. Then, we docked VF to both human and *E. coli* DHOD (hDHOD and EcDHOD, respectively). Docking on hDHOD showed a GLIDE score of −9.4 with a binding pose comparable to R2C. In particular, the interaction between VF carboxylic acid and R136 and Q47 is maintained (Figure 5A).

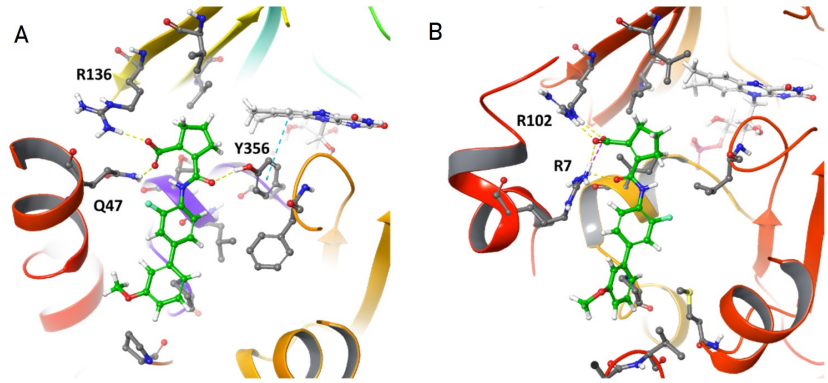


Figure 5. Model of VF-DHOD interaction. (**A**) VF (in green) docked to hDHOD showed an interaction between the VF carboxylic acid moiety and both R136 and Q47. A third interaction is observed between the amide group of VF and the hydroxyl group of Y356. (**B**) When VF is docked to EcDHOD we observed that the VF carboxylic acid moiety form interactions with R102 (equivalent to R136 in human) and R7 (in place of Q47 in human). The VF amide group does not form any interaction.

Furthermore, a third interaction is formed between the same carboxylic moiety with the hydroxyl group of Y356. Docking to EcDHOD again showed a comparable binding pose (Figure 5B). The carboxylic moiety of VF can interact with a conserved R102 (R136 in human) and with R7 (P44 in human) while lacking a third interaction present in the former. In addition, polar and non-polar interactions are also less optimised, resulting in a reduced GLIDE score of −8. Of note, EcDHOD and hDHOD are structurally similar with an RMSD of less than 1 Å and a sequence identity of 41%. Based on these results, we expect VF to be able to interact with both hDHOD and EcDHOD with similar binding affinities, which would be consistent with the hypothesis that the lack of VF antimicrobial activity might be due to its inability to enter bacterial cells. In contrast, a structural comparison between hDHOD, EcDHOD and the structures of the two DHOD isoenzymes in the Gram positive bacterium *Lactococcus lactis* (the *pyrD* homodimer and the *pyrD-pyrK* heterotetramer, 2DOR and 1EP2, respectively) show that in either DHOD form from this bacterium, the predicted binding site for VF is missing due to the lack of the two N-terminal helices necessary for DHOD-VF interaction (Figure S4).

4. Discussion

The ability of intracellular pathogenic bacteria such as AIEC to invade the gut epithelium and to survive in macrophages is a major trigger for chronic gut inflammation in CD. Thus, specific inhibition of AIEC virulence could be a promising strategy in CD treatment, as it might counteract inflammation without exacerbating gut dysbiosis, which can instead be observed in response to antibiotic therapy [26].

In order to identify potential targets for anti-virulence agents, we have generated a transposon insertion library in the AIEC strain LF82 and looked for mutants unable to survive in a medium mimicking the macrophage environment. Through this approach, we selected an insertion mutation in the *pyrD* gene (LF82*pyrD*::Tn5 mutant), encoding for DHOD, an enzyme part of the de novo pyrimidine biosynthetic pathway. The selection of the *pyrD* mutant is likely due to the lack of exogenous pyrimidines in the medium used in our screening (Figure 1); however, it confirms previous results showing that mutants in AIEC, as well as in *Salmonella*, impaired in their de novo pyrimidine biosynthetic pathway, are indeed unable to survive in macrophages [35,44], which is in line with the more general notion that, in several intracellular pathogens, auxotrophic mutants, either for nucleotides or for some amino acids, are avirulent, mostly because of their inability to grow or survive [45–47].

The pleiotropic effects of mutations in the pyrimidine biosynthetic pathway, such as the inability to form biofilm or to produce virulence factors, have already been described in the MG1655 laboratory strain of *E. coli* [36] as well as in other bacteria, such as *P. aeruginosa* [32], possibly via sensing of intracellular pyrimidine nucleotide pools by global regulators [48]. In this work, we showed that, in LF82, inactivation of the *pyrD* gene resulted in the inhibition of surface attachment (Figure 2) via downregulation of genes encoding curli fibers (Figures 2 and 3) and type 1 fimbriae (Figure 3) which are, arguably, the two main proteinaceous adhesion factors in *E. coli*. We had already described that, in *E coli* MG1655, mutations in the pyrimidine biosynthetic genes strongly affected transcription of the *csgDEFG* operon [36], which includes the CsgD regulatory protein presiding to curli and cellulose production; in contrast, in the LF82 background, gene expression downregulation by *pyrD* inactivation was only observed for the *csgBAC* operon, encoding curli structural subunits. This observation would suggest that, in LF82, sensing the lack of intracellular pyrimidines might be more selectively relayed towards turning off curli production, rather than to a more general effect on the CsgD regulon (Figure 3). Inhibition of curli production might be clinically relevant as, although curli production is turned off at 37 °C in laboratory conditions, antibodies against these structures have been isolated in the context of various diseases, suggesting their production would indeed take place during infection. Indeed, curli binding to the TLR1/2 Toll-like receptors and activation of the NOD-like receptor protein 3 inflammasome are thought to contribute to overall inflammation in CD [49,50].

While the role of curli fibers in AIEC infection has not yet been fully characterised, type 1 fimbriae are considered *bona fide* AIEC virulence factors that are able to promote bacterial adhesion to epithelial cells [18]. Interestingly, the expression of both curli fibers- and type 1 pili-encoding genes are negatively regulated in the presence of glucose in LF82, suggesting positive control by the CAP master regulator [25], whose activity appears to respond also to pyrimidine intracellular concentrations, as recently proposed [51]. At any rate, the inhibition of pyrimidine biosynthesis appears to have a very extensive impact on AIEC adhesion factors, reiterating its potential as a target for antimicrobial drugs.

Finally, in addition to survival in macrophages, biofilm formation and curli production, as well as cellular motility, another virulence factor in AIEC, was also affected by *pyrD* inactivation, probably via regulation of flagellar motility, as transcription of the *fliC* gene, part of the main flagellar operon, was unaffected in the LF82*pyrD*::Tn5 mutant strain (Figure 3). Unlike survival in the macrophage-mimicking medium and adhesion factors' production, flagellar motility was not fully restored by uracil supplementation (Figure 2C), possibly suggesting that accumulation of early intermediates of the de novo pyrimidine biosynthesis, rather than pyrimidine availability, might be involved in this process. Indeed, pyrimidine biosynthesis intermediates such as N-carbamoyl-aspartate can modulate production of c-di-GMP, a signal molecule involved in several processes, including flagellar motility [52].

The pleiotropic effects of the *pyrD* mutation on several virulence-related processes would make DHOD, the product of the *pyrD* gene, a suitable target for novel inhibitors of AIEC growth and/or virulence. This led us to hypothesize that VF, a known inhibitor of

human DHOD studied in clinical trials as an anti-inflammatory drug in the treatment of CD [43,53], could possess antimicrobial and antivirulence activity against AIEC via bacterial DHOD inhibition. Indeed, antimicrobial activity against microorganisms able to trigger chronic inflammation in CD has been proposed to be at least partly responsible for the anti-inflammatory activity of mercaptopurines, which, like VF, can inhibit de novo nucleotide biosynthesis in bacteria [24,25,54]. However, unlike mercaptopurines, VF possesses little or no activity either against AIEC or Gram positive microorganisms (Figure 4), although molecular docking analysis (Figure 5) predicts that VF can bind human and *E. coli* DHOD with similar affinity. The lack of antimicrobial activity by VF against AIEC (Figure 4) might thus depend on VF poor penetration into the cell, despite its low molecular mass (355.12), a behaviour observed also for several antimicrobial agents, such as clindamycin (MW 424.98) [55]. In contrast, comparison of the *E. coli* DHOD structure to both DHOD isoenzymes of the Gram positive, probiotic bacterium *Lactococcus lactis* (Figure S4) shows the absence of any potential binding site for VF in either DHOD isoform. Our results reiterate previous observations on the functional diversity of DHOD among bacteria [56] which, in addition to known structural differences between human and *E. coli* DHOD [57], would suggest the possibility to design DHOD inhibitors that might specifically target only a subset of selected bacterial species. In the context of a chronic inflammatory disease such as CD, targeting *E. coli* DHOD without affecting beneficial bacteria in the gut microbiota could eradicate an important trigger for chronic inflammation and help restore gut microbiota eubiosis, thus representing a promising therapeutic strategy.

Supplementary Materials: The following are available online at http://www.mdpi.com/article/10.3390/microorganisms10030537/s1, Table S1: List of oligonucleotides used for qRT-PCR experiments. Figure S1: Growth rates and lag phases of LF82 wild type strain and of the LF82*pyrD*::Tn5 mutant in Acid medium supplemented either with DMSO or uracil. Figure S2: Optical density of overnight cultures of LF82 wild type strain and of the LF82*pyrD*::Tn5 mutant in YESCA medium in microtiter plates used for biofilm determination experiments. Figure S3: (A) Viability of the LF82*pyrD*::Tn5 mutant compared to its parental strain on LB Agar and MacConkey media. (B) Relative expression levels of the *lptD* gene. Figure S4: structures of *Lactococcus lactis* PyrD protein.

Author Contributions: Conceptualization, E.R., P.L., M.P.; Funding acquisition, M.P.; Investigation, E.R., G.L., V.B., A.B., E.S., C.C.; Resources, M.P., P.L.; Supervision, E.R., M.P., P.L.; Writing—original draft, P.L.; Writing—review and editing, E.R., M.P., P.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was funded by Cariplo Foundation grant n. #2017-0816 to MP and by a Fondazione Telethon grant (GP191334) to CC.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: All data generated or analyzed during this study are included in this published article (and its supplementary information files).

Acknowledgments: We would like to thank Sara Biella for her assistance in carrying out the experimental work.

Conflicts of Interest: The authors declare that they have no conflict of interest.

References

1. de Souza, H.S.P.; Fiocchi, C. Immunopathogenesis of IBD: Current State of the Art. *Nat. Rev. Gastroenterol.* **2016**, *13*, 13–27. [CrossRef] [PubMed]
2. Duerr, R.H.; Taylor, K.D.; Brant, S.R.; Rioux, J.D.; Silverberg, M.S.; Daly, M.J.; Steinhart, A.H.; Abraham, C.; Regueiro, M.; Griffiths, A.; et al. A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. *Science* **2006**, *314*, 1461–1463. [CrossRef] [PubMed]
3. Villani, A.-C.; Lemire, M.; Fortin, G.; Louis, E.; Silverberg, M.S.; Collette, C.; Baba, N.; Libioulle, C.; Belaiche, J.; Bitton, A.; et al. Common Variants in the NLRP3 Region Contribute to Crohn's Disease Susceptibility. *Nat. Genet.* **2009**, *41*, 71–76. [CrossRef] [PubMed]

4. Ng, S.C.; Bernstein, C.N.; Vatn, M.H.; Lakatos, P.L.; Loftus, E.V.; Tysk, C.; O'Morain, C.; Moum, B.; Colombel, J.-F.; on behalf of the Epidemiology and Natural History Task Force of the International Organization of Inflammatory Bowel Disease (IOIBD). Geographical Variability and Environmental Risk Factors in Inflammatory Bowel Disease. *Gut* **2013**, *62*, 630. [CrossRef]
5. Wang, K.; Wu, L.; Dou, C.; Guan, X.; Wu, H.; Liu, H. Research Advance in Intestinal Mucosal Barrier and Pathogenesis of Crohn's Disease. *Gastroenterol. Res. Pract.* **2016**, *2016*, 9686238. [CrossRef]
6. Geremia, A.; Biancheri, P.; Allan, P.; Corazza, G.R.; Sabatino, A.D. Innate and Adaptive Immunity in Inflammatory Bowel Disease. *Autoimmun. Rev.* **2014**, *13*, 3–10. [CrossRef]
7. Ahmed, I.; Roy, B.C.; Khan, S.A.; Septer, S.; Umar, S. Microbiome, Metabolome and Inflammatory Bowel Disease. *Microorganisms* **2016**, *4*, 20. [CrossRef]
8. Nishida, A.; Inoue, R.; Inatomi, O.; Bamba, S.; Naito, Y.; Andoh, A. Gut Microbiota in the Pathogenesis of Inflammatory Bowel Disease. *Clin. J. Gastroenterol.* **2018**, *11*, 1–10. [CrossRef]
9. Walters, W.A.; Xu, Z.; Knight, R. Meta-analyses of Human Gut Microbes Associated with Obesity and IBD. *FEBS Lett.* **2014**, *588*, 4223–4233. [CrossRef]
10. Gevers, D.; Kugathasan, S.; Knights, D.; Kostic, A.D.; Knight, R.; Xavier, R.J. A Microbiome Foundation for the Study of Crohn's Disease. *Cell Host Microbe* **2017**, *21*, 301–304. [CrossRef]
11. Mentella, M.C.; Scaldaferri, F.; Pizzoferrato, M.; Gasbarrini, A.; Miggiano, G.A.D. Nutrition, IBD and Gut Microbiota: A Review. *Nutrients* **2020**, *12*, 944. [CrossRef] [PubMed]
12. Darfeuille-Michaud, A.; Boudeau, J.; Bulois, P.; Neut, C.; Glasser, A.-L.; Barnich, N.; Bringer, M.-A.; Swidsinski, A.; Beaugerie, L.; Colombel, J.-F. High Prevalence of Adherent-Invasive *Escherichia coli* Associated with Ileal Mucosa in Crohn's Disease. *Gastroenterology* **2004**, *127*, 412–421. [CrossRef] [PubMed]
13. Schippa, S.; Conte, M.P.; Borrelli, O.; Iebba, V.; Aleandri, M.; Seganti, L.; Longhi, C.; Chiarini, F.; Osborn, J.; Cucchiara, S. Dominant Genotypes in Mucosa-Associated *Escherichia coli* Strains from Pediatric Patients with Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* **2009**, *15*, 661–672. [CrossRef] [PubMed]
14. Thomazini, C.M.; Samegima, D.A.G.; Rodrigues, M.A.M.; Victoria, C.R.; Rodrigues, J. High Prevalence of Aggregative Adherent *Escherichia coli* Strains in the Mucosa-Associated Microbiota of Patients with Inflammatory Bowel Diseases. *Int. J. Med. Microbiol.* **2011**, *301*, 475–479. [CrossRef] [PubMed]
15. Glasser, A.-L.; Boudeau, J.; Barnich, N.; Perruchot, M.-H.; Colombel, J.-F.; Darfeuille-Michaud, A. Adherent Invasive *Escherichia coli* Strains from Patients with Crohn's Disease Survive and Replicate within Macrophages without Inducing Host Cell Death. *Infect. Immun.* **2001**, *69*, 5529–5537. [CrossRef] [PubMed]
16. Eaves-Pyles, T.; Allen, C.A.; Taormina, J.; Swidsinski, A.; Tutt, C.B.; Jezek, G.E.; Islas-Islas, M.; Torres, A.G. *Escherichia coli* Isolated from a Crohn's Disease Patient Adheres, Invades, and Induces Inflammatory Responses in Polarized Intestinal Epithelial Cells. *Int. J. Med. Microbiol.* **2008**, *298*, 397–409. [CrossRef] [PubMed]
17. Rolhion, N.; Darfeuille-Michaud, A. Adherent-Invasive *Escherichia coli* in Inflammatory Bowel Disease. *Inflamm. Bowel Dis.* **2007**, *13*, 1277–1283. [CrossRef]
18. Boudeau, J.; Barnich, N.; Darfeuille-Michaud, A. Type 1 Pili-mediated Adherence of *Escherichia coli* Strain LF82 Isolated from Crohn's Disease Is Involved in Bacterial Invasion of Intestinal Epithelial Cells. *Mol. Microbiol.* **2001**, *39*, 1272–1284. [CrossRef]
19. Barnich, N.; Boudeau, J.; Claret, L.; Darfeuille-Michaud, A. Regulatory and Functional Co-operation of Flagella and Type 1 Pili in Adhesive and Invasive Abilities of AIEC Strain LF82 Isolated from a Patient with Crohn's Disease. *Mol. Microbiol.* **2003**, *48*, 781–794. [CrossRef]
20. Cieza, R.J.; Hu, J.; Ross, B.N.; Sbrana, E.; Torres, A.G. The IbeA Invasin of Adherent-Invasive *Escherichia coli* Mediates Interaction with Intestinal Epithelia and Macrophages. *Infect. Immun.* **2015**, *83*, 1904–1918. [CrossRef]
21. Chassaing, B.; Darfeuille-Michaud, A. The σ^E Pathway Is Involved in Biofilm Formation by Crohn's Disease-Associated Adherent-Invasive *Escherichia coli*. *J. Bacteriol.* **2013**, *195*, 76–84. [CrossRef] [PubMed]
22. Fanelli, G.; Pasqua, M.; Colonna, B.; Prosseda, G.; Grossi, M. Expression Profile of Multidrug Resistance Efflux Pumps During Intracellular Life of Adherent-Invasive *Escherichia coli* Strain LF82. *Front. Microbiol.* **2020**, *11*, 1935. [CrossRef] [PubMed]
23. D'Haens, G.R.; Vermeire, S.; Assche, G.V.; Noman, M.; Aerden, I.; Olmen, G.V.; Rutgeerts, P. Therapy of Metronidazole With Azathioprine to Prevent Postoperative Recurrence of Crohn's Disease: A Controlled Randomized Trial. *Gastroenterology* **2008**, *135*, 1123–1129. [CrossRef] [PubMed]
24. Shin, S.J.; Collins, M.T. Thiopurine Drugs Azathioprine and 6-Mercaptopurine Inhibit *Mycobacterium paratuberculosis* Growth In Vitro. *Antimicrob. Agents Chemother.* **2008**, *52*, 418–426. [CrossRef]
25. Migliore, F.; Macchi, R.; Landini, P.; Paroni, M. Phagocytosis and Epithelial Cell Invasion by Crohn's Disease-Associated Adherent-Invasive *Escherichia coli* Are Inhibited by the Anti-Inflammatory Drug 6-Mercaptopurine. *Front. Microbiol.* **2018**, *9*, 964. [CrossRef]
26. Nitzan, O.; Elias, M.; Peretz, A.; Saliba, W. Role of Antibiotics for Treatment of Inflammatory Bowel Disease. *World J. Gastroenterol.* **2016**, *22*, 1078. [CrossRef]
27. Ganji-Arjenaki, M.; Rafieian-Kopaei, M. Probiotics Are a Good Choice in Remission of Inflammatory Bowel Diseases: A Meta Analysis and Systematic Review. *J. Cell Physiol.* **2018**, *233*, 2091–2103. [CrossRef]
28. Ghouri, Y.A.; Richards, D.M.; Rahimi, E.F.; Krill, J.T.; Jelinek, K.A.; DuPont, A.W. Systematic Review of Randomized Controlled Trials of Probiotics, Prebiotics, and Synbiotics in Inflammatory Bowel Disease. *Clin. Exp. Gastroenterol.* **2014**, *7*, 473–487. [CrossRef]

29. Butterworth, A.D.; Thomas, A.G.; Akobeng, A.K. Probiotics for Induction of Remission in Crohn's Disease. *Cochrane Database Syst. Rev.* **2008**, CD006634. [CrossRef]
30. Leccese, G.; Bibi, A.; Mazza, S.; Facciotti, F.; Caprioli, F.; Landini, P.; Paroni, M. Probiotic *Lactobacillus* and *Bifidobacterium* Strains Counteract Adherent-Invasive *Escherichia coli* (AIEC) Virulence and Hamper IL-23/Th17 Axis in Ulcerative Colitis, but Not in Crohn's Disease. *Cells* **2020**, *9*, 1824. [CrossRef]
31. Durand, J.M.B.; Björk, G.R. Metabolic Control through Ornithine and Uracil of Epithelial Cell Invasion by *Shigella flexneri*. *Microbiology* **2009**, *155*, 2498–2508. [CrossRef]
32. Ueda, A.; Attila, C.; Whiteley, M.; Wood, T.K. Uracil Influences Quorum Sensing and Biofilm Formation in *Pseudomonas aeruginosa* and Fluorouracil Is an Antagonist. *Microb. Biotechnol.* **2009**, *2*, 62–74. [CrossRef] [PubMed]
33. Boudeau, J.; Glasser, A.-L.; Masseret, E.; Joly, B.; Darfeuille-Michaud, A. Invasive Ability of an *Escherichia coli* Strain Isolated from the Ileal Mucosa of a Patient with Crohn's Disease. *Infect. Immun.* **1999**, *67*, 4499–4509. [CrossRef] [PubMed]
34. Bringer, M.-A.; Rolhion, N.; Glasser, A.-L.; Darfeuille-Michaud, A. The Oxidoreductase DsbA Plays a Key Role in the Ability of the Crohn's Disease-Associated Adherent-Invasive *Escherichia coli* Strain LF82 To Resist Macrophage Killing. *J. Bacteriol.* **2007**, *189*, 4860–4871. [CrossRef] [PubMed]
35. Thompson, A.P.; O'Neill, I.; Smith, E.J.; Catchpole, J.; Fagan, A.; Burgess, K.E.V.; Carmody, R.J.; Clarke, D.J. Glycolysis and Pyrimidine Biosynthesis Are Required for Replication of Adherent–Invasive *Escherichia coli* in Macrophages. *Microbiology* **2016**, *162*, 954–965. [CrossRef] [PubMed]
36. Garavaglia, M.; Rossi, E.; Landini, P. The Pyrimidine Nucleotide Biosynthetic Pathway Modulates Production of Biofilm Determinants in *Escherichia coli*. *PLoS ONE* **2012**, *7*, e31252. [CrossRef]
37. Genevaux, P.; Bauda, P.; DuBow, M.S.; Oudega, B. Identification of Tn 10 Insertions in the RfaG, RfaP, and GalU Genes Involved in Lipopolysaccharide Core Biosynthesis That Affect *Escherichia coli* Adhesion. *Arch. Microbiol.* **1999**, *172*, 1–8. [CrossRef]
38. Møller, A.K.; Leatham, M.P.; Conway, T.; Nuijten, P.J.M.; de Haan, L.A.; Krogfelt, K.A.; Cohen, P.S. An *Escherichia coli* MG1655 Lipopolysaccharide Deep-Rough Core Mutant Grows and Survives in Mouse Cecal Mucus but Fails To Colonize the Mouse Large Intestine. *Infect. Immun.* **2003**, *71*, 2142–2152. [CrossRef]
39. Dartigalongue, C.; Missiakas, D.; Raina, S. Characterization of the *Escherichia coli* σ^E Regulon. *J. Biol. Chem.* **2001**, *276*, 20866–20875. [CrossRef]
40. Tam, C.; Missiakas, D. Changes in Lipopolysaccharide Structure Induce the σ^E-dependent Response of *Escherichia coli*. *Mol. Microbiol.* **2005**, *55*, 1403–1412. [CrossRef]
41. Römling, U.; Rohde, M.; Olsén, A.; Normark, S.; Reinköster, J. AgfD, the Checkpoint of Multicellular and Aggregative Behaviour in *Salmonella typhimurium* Regulates at Least Two Independent Pathways. *Mol. Microbiol.* **2000**, *36*, 10–23. [CrossRef] [PubMed]
42. Zogaj, X.; Nimtz, M.; Rohde, M.; Bokranz, W.; Römling, U. The Multicellular Morphotypes of *Salmonella typhimurium* and *Escherichia coli* Produce Cellulose as the Second Component of the Extracellular Matrix. *Mol. Microbiol.* **2001**, *39*, 1452–1463. [CrossRef] [PubMed]
43. Herrlinger, K.R.; Diculescu, M.; Fellermann, K.; Hartmann, H.; Howaldt, S.; Nikolov, R.; Petrov, A.; Reindl, W.; Otte, J.M.; Stoynov, S.; et al. Efficacy, Safety and Tolerability of Vidofludimus in Patients with Inflammatory Bowel Disease: The ENTRANCE Study. *J. Crohn's Colitis* **2013**, *7*, 636–643. [CrossRef]
44. Ellis, M.J.; Tsai, C.N.; Johnson, J.W.; French, S.; Elhenawy, W.; Porwollik, S.; Andrews-Polymenis, H.; McClelland, M.; Magolan, J.; Coombes, B.K.; et al. A Macrophage-Based Screen Identifies Antibacterial Compounds Selective for Intracellular *Salmonella typhimurium*. *Nat. Commun.* **2019**, *10*, 197. [CrossRef]
45. Bange, F.C.; Brown, A.M.; Jacobs, W.R. Leucine Auxotrophy Restricts Growth of *Mycobacterium bovis* BCG in Macrophages. *Infect. Immun.* **1996**, *64*, 1794–1799. [CrossRef]
46. Pilatz, S.; Breitbach, K.; Hein, N.; Fehlhaber, B.; Schulze, J.; Brenneke, B.; Eberl, L.; Steinmetz, I. Identification of *Burkholderia pseudomallei* Genes Required for the Intracellular Life Cycle and In Vivo Virulence. *Infect. Immun.* **2006**, *74*, 3576–3586. [CrossRef]
47. Smith, D.A.; Parish, T.; Stoker, N.G.; Bancroft, G.J. Characterization of Auxotrophic Mutants of *Mycobacterium tuberculosis* and Their Potential as Vaccine Candidates. *Infect. Immun.* **2001**, *69*, 1142–1150. [CrossRef] [PubMed]
48. Beaumont, H.J.E.; Gallie, J.; Kost, C.; Ferguson, G.C.; Rainey, P.B. Experimental Evolution of Bet Hedging. *Nature* **2009**, *462*, 90. [CrossRef]
49. Tükel, Ç.; Nishimori, J.H.; Wilson, R.P.; Winter, M.G.; Keestra, A.M.; Putten, J.P.M.V.; Bäumler, A.J. Toll-like Receptors 1 and 2 Cooperatively Mediate Immune Responses to Curli, a Common Amyloid from Enterobacterial Biofilms. *Cell Microbiol.* **2010**, *12*, 1495–1505. [CrossRef]
50. Rapsinski, G.J.; Wynosky-Dolfi, M.A.; Oppong, G.O.; Tursi, S.A.; Wilson, R.P.; Brodsky, I.E.; Tükel, Ç. Toll-like Receptor 2 and NLRP3 Cooperate to Recognize a Functional Bacterial Amyloid, Curli. *Infect. Immun.* **2014**, *83*, 693–701. [CrossRef]
51. Lauritsen, I.; Frendorf, P.O.; Capucci, S.; Heyde, S.A.H.; Blomquist, S.D.; Wendel, S.; Fischer, E.C.; Sekowska, A.; Danchin, A.; Nørholm, M.H.H. Temporal Evolution of Master Regulator Crp Identifies Pyrimidines as Catabolite Modulator Factors. *Nat. Commun.* **2021**, *12*, 5880. [CrossRef] [PubMed]
52. Rossi, E.; Motta, S.; Aliverti, A.; Cossu, F.; Gourlay, L.; Mauri, P.; Landini, P. Cellulose Production Is Coupled to Sensing of the Pyrimidine Biosynthetic Pathway via c-di-GMP Production by the DgcQ Protein of *Escherichia coli*. *Environ. Microbiol.* **2017**, *19*, 4551–4563. [CrossRef]

53. Muehler, A.; Kohlhof, H.; Groeppel, M.; Vitt, D. Safety, Tolerability and Pharmacokinetics of Vidofludimus Calcium (IMU-838) After Single and Multiple Ascending Oral Doses in Healthy Male Subjects. *Eur. J. Drug Metab. Pharmacokinet.* **2020**, *45*, 557–573. [CrossRef] [PubMed]
54. Antoniani, D.; Rossi, E.; Rinaldo, S.; Bocci, P.; Lolicato, M.; Paiardini, A.; Raffaelli, N.; Cutruzzolà, F.; Landini, P. The Immunosuppressive Drug Azathioprine Inhibits Biosynthesis of the Bacterial Signal Molecule Cyclic-Di-GMP by Interfering with Intracellular Nucleotide Pool Availability. *Appl. Microbiol. Biot.* **2013**, *97*, 7325–7336. [CrossRef] [PubMed]
55. Leclercq, R.; Courvalin, P. Intrinsic and Unusual Resistance to Macrolide, Lincosamide, and Streptogramin Antibiotics in Bacteria. *Antimicrob. Agents Chemother.* **1991**, *35*, 1273–1276. [CrossRef] [PubMed]
56. Nørager, S.; Jensen, K.F.; Björnberg, O.; Larsen, S. *E. Coli* Dihydroorotate Dehydrogenase Reveals Structural and Functional Distinctions between Different Classes of Dihydroorotate Dehydrogenases. *Structure* **2002**, *10*, 1211–1223. [CrossRef]
57. Marcinkeviciene, J.; Rogers, M.J.; Kopcho, L.; Jiang, W.; Wang, K.; Murphy, D.J.; Lippy, J.; Link, S.; Chung, T.D.Y.; Hobbs, F.; et al. Selective Inhibition of Bacterial Dihydroorotate Dehydrogenases by Thiadiazolidinediones. *Biochem. Pharmacol.* **2000**, *60*, 339–342. [CrossRef]

Supplementary Materials

Table S1. Oligonucleotides used for RT-PCR experiments.

Name	Sequence (5′ → 3′)
16s_RT_for	TGTCGTCAGCTCGTGTCGTGA
16s_RT_rev	ATCCCCACCTTCCTCCGGT
csgB_RT_for	CATAATTGGTCAAGCTGGGACTAA
csgB_RT_rev	GCAACAACCGCCAAAAGTTT
csgD_RT_for	CCCGTACCGCGACATTG
csgD_RT_rev	CGTTCTTGATCCTCCATGG
bcsA_RT_for	TCGCGATTATCGTCGTCACG
bcsA_RT_rev	GGGTGCTCCAGCGGAATAAA
adrA_RT_for	GGCTGGGTCAGCTACCAG
adrA_RT_rev	CGTCGGTTATACACGCCCG
fimA_RT_for	CGCTTGCGCAGTTGATGCAG
fimA_RT_rev	CCGTCCCCAAGAAGGCAACA
fliC_RT_for	CAACTTACAGCGTATCCGTG
fliC_RT_rev	CGTTCACGCCGTTGAACTG
lptD_RT_for	GATGCGCTCGGTAATGTCC
lptD_RT_rev	CACCTTCCCAGACGTTGGT

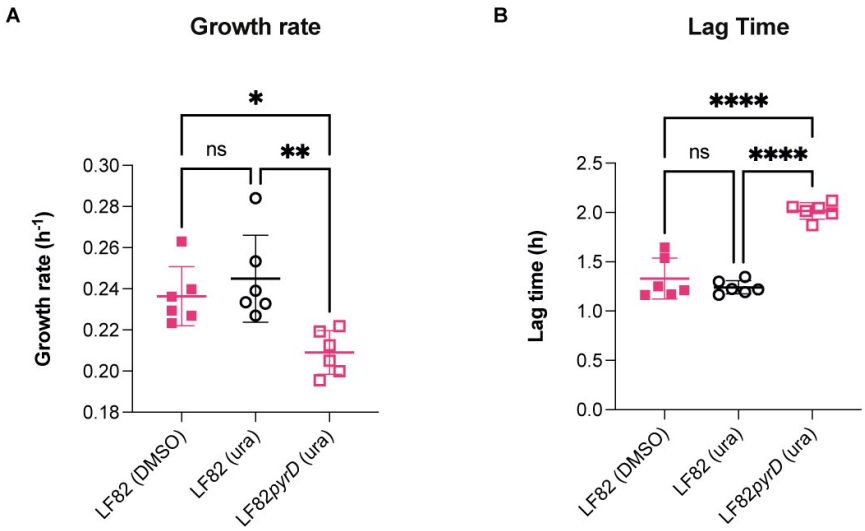


Figure S1. Analysis of growth rates (**A**) and lag time after inoculation (**B**) in Acid Medium for LF82 grown either with 0.25% DMSO, or 0.25mM uracil in DMSO, and for LF82*pyrD*::Tn5 in the presence of 0.25mM uracil. Results of six independent experiments are shown. *, *p*-value < 0.05; **, *p*-value < 0.01; ****, *p*-value < 0.0001, one-way ANOVA with Tukey's test for multiple comparisons.

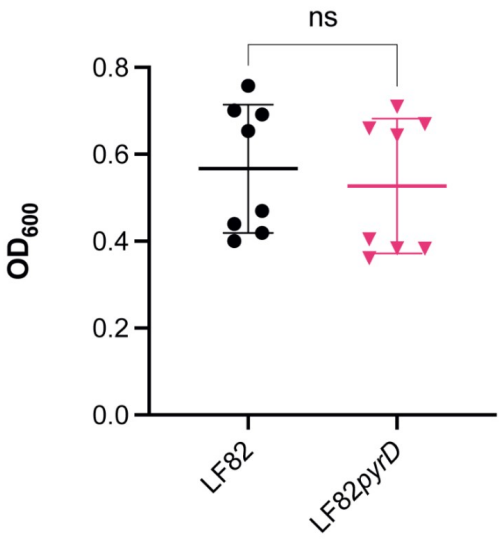


Figure S2. Optical density of overnight cultures of LF82 and LF82*pyrD*::Tn5 grown in YESCA medium in microtiter plates at 30 °C (conditions used for biofilm determination with crystal violet shown in Figure 2). Each point represents an independent biological replicate. ns, not significant (*p*-value = 0.89), Student t test.

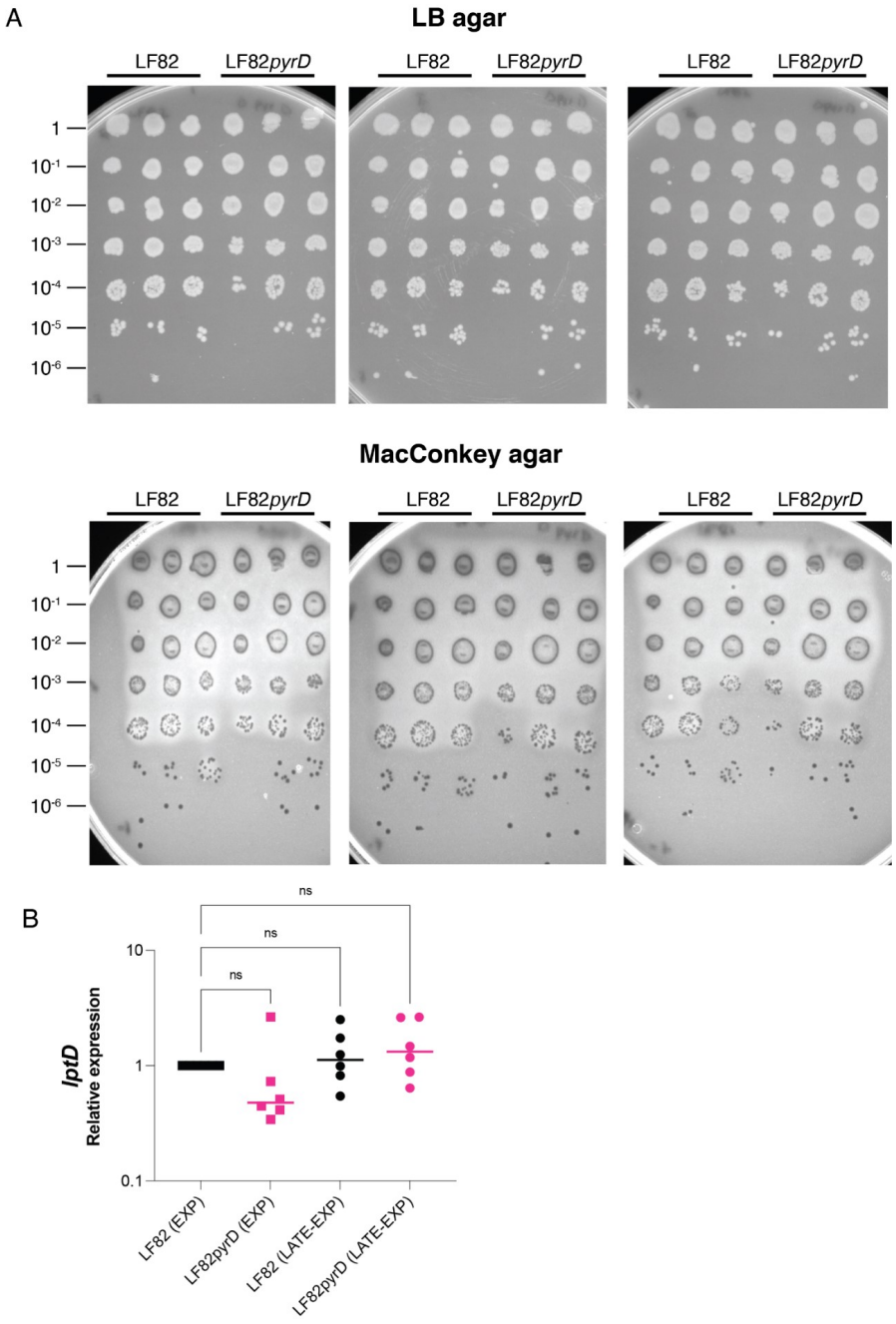


Figure S3. (**A**) Plating efficiency of LF82 and its LF82*pyrD*::Tn5 mutant derivative on Luria Agar (LA) and MacConkey media. Overnight cultures grown in LB medium were adjusted to OD600nm=1.0 and serially diluted 1:10 in microtiter plates to a final dilution of 10^{-6}. Dilutions were replicated on both LA and MacConkey. (**B**) Determination of gene expression levels in the LF82 vs the LF82*pyrD*::Tn5 (LF82*pyrD*) strain by qRT-PCR. Results of six independent experiments are shown. ns, not significant, one-way ANOVA with Tukey's test for multiple comparisons.

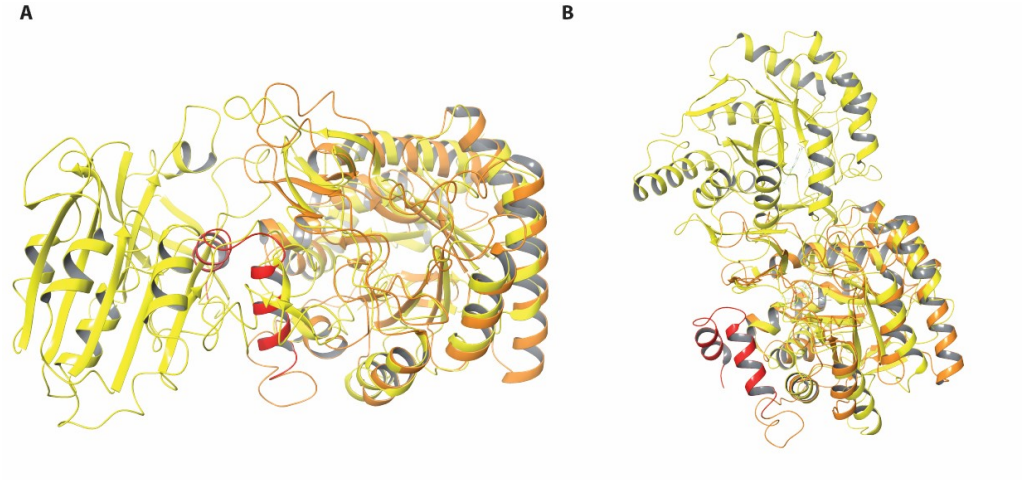


Figure S4. (A) Crystal structure of *Lactococcus lactis* DHOD type B (pdb 1EP2) where PyrD domain is superimposed to the crystal structure of EcDHOD (pdb 1F76). (**B**) Crystal structure of *Lactococcus lactis* DHOD type A (pdb 2DOR) where one of the PyrD domains is superimposed to with the crystal structure of EcDHOD (pdb 1F76).

5. References

"The library is open hunty, and in the great tradition of Paris is Burning: Reading. Is. Fundamental!"

RuPaul

1. Camilloni, C. & Sutto, L. Lymphotactin: How a protein can adopt two folds. *Journal of Chemical Physics* **131**, 54103 (2009).

2. Sutto, L., Mereu, I. & Gervasio, F. L. A Hybrid All-Atom Structure-Based Model for Protein Folding and Large Scale Conformational Transitions. *undefined* **7**, 4208–4217 (2011).

3. Sutto, L. & Camilloni, C. From A to B: A ride in the free energy surfaces of protein G domains suggests how new folds arise. *Journal of Chemical Physics* **136**, 185101 (2012).

4. Benson, M. D. *et al.* Amyloid nomenclature 2020: update and recommendations by the International Society of Amyloidosis (ISA)

nomenclature committee. *https://doi.org/10.1080/13506129.2020.1835263* **27**, 217–222 (2020).

5. Iadanza, M. G., Jackson, M. P., Hewitt, E. W., Ranson, N. A. & Radford, S. E. A new era for understanding amyloid structures and disease. *Nature Reviews Molecular Cell Biology 2018 19:12* **19**, 755–773 (2018).

6. Chiti, F. & Dobson, C. M. Protein Misfolding, Amyloid Formation, and Human Disease: A Summary of Progress Over the Last Decade. *Annu Rev Biochem* **86**, 27–68 (2017).

7. Knowles, T. P. J., Vendruscolo, M. & Dobson, C. M. The amyloid state and its association with protein misfolding diseases. *Nat Rev Mol Cell Biol* **15**, 384–396 (2014).

8. Ke, P. C. *et al.* Half a century of amyloids: past, present and future. *Chem Soc Rev* **49**, 5473–5509 (2020).

9. Nassar, R., Dignon, G. L., Razban, R. M. & Dill, K. A. The Protein Folding Problem: The Role of Theory. *J Mol Biol* **433**, 167126 (2021).

10. Tasaki, M. *et al.* Age-related amyloidosis outside the brain: A state-of-the-art review. *Ageing Res Rev* **70**, 101388 (2021).

11. Reitz, C., Brayne, C. & Mayeux, R. Epidemiology of Alzheimer disease. *Nature Reviews Neurology 2011 7:3* **7**, 137–152 (2011).

12. Bloom, G. S. Amyloid-β and tau: the trigger and bullet in Alzheimer disease pathogenesis. *JAMA Neurol* **71**, 505–508 (2014).

13. Poewe, W. *et al.* Parkinson disease. *Nature Reviews Disease Primers 2017 3:1* **3**, 1–21 (2017).

14. Lau, L. M. L. de *et al.* Epidemiology of Parkinson's disease. *Lancet Neurol* **5**, 525–535 (2006).

15. Koike, H. & Katsuno, M. Transthyretin Amyloidosis: Update on the Clinical Spectrum, Pathogenesis, and Disease-Modifying Therapies. *Neurol Ther* **9**, 317–333 (2020).

16. Ruberg, F. L., Grogan, M., Hanna, M., Kelly, J. W. & Maurer, M. S. Transthyretin Amyloid Cardiomyopathy: JACC State-of-the-Art Review. *J Am Coll Cardiol* **73**, 2872–2891 (2019).

17. Kanatsuka, A., Kou, S. & Makino, H. IAPP/amylin and β-cell failure: implication of the risk factors of type 2 diabetes. *Diabetology International 2018 9:3* **9**, 143–157 (2018).

18. Raimundo, A. F., Ferreira, S., Martins, I. C. & Menezes, R. Islet Amyloid Polypeptide: A Partner in Crime With Aβ in the Pathology of Alzheimer's Disease. *Front Mol Neurosci* **13**, 35 (2020).

19. Vergaro, G. *et al.* Atrial amyloidosis: mechanisms and clinical manifestations. *Eur J Heart Fail* (2022) doi:10.1002/EJHF.2650.

20. Zanoni, P. & von Eckardstein, A. Inborn errors of apolipoprotein A-I metabolism: implications for disease, research and development. *Curr Opin Lipidol* **31**, 62–70 (2020).

21. Tasaki, M. *et al.* A novel age-related venous amyloidosis derived from EGF-containing fibulin-like extracellular matrix protein 1. *J Pathol* **247**, 444–455 (2019).

22. World Health Organization and Alzheimer's disease International. DEMENTIA A public Helath Priority. *World Health Organization* 1–4 (2012).

23. Robbins, J. Transthyretin from discovery to now. *Clin Chem Lab Med* **40**, 1183–1190 (2002).

24. Plante-Bordeneuve, V. Transthyretin familial amyloid polyneuropathy: an update. *Journal of Neurology 2017 265:4* **265**, 976–983 (2017).

25. Said, G. & Planté-Bordeneuve, V. Familial amyloid polyneuropathy. *www.thelancet.com/neurology* **10**, (2011).

26. Muzio, L. *et al.* Retromer stabilization results in neuroprotection in a model of Amyotrophic Lateral Sclerosis. *Nat Commun* **11**, 3848 (2020).

27. Dobson, C. M. Protein misfolding, evolution and disease. *Trends Biochem Sci* **24**, 329–332 (1999).

28. Merlini, G. & Bellotti, V. Molecular Mechanisms of Amyloidosis. *New England Journal of Medicine* **349**, 583–596 (2003).

29. Sinnige, T. Molecular mechanisms of amyloid formation in living systems. *Chem Sci* **13**, 7080–7097 (2022).

30. Westermark, P., Sletten, K., Johansson, B. & Cornwell, G. G. Fibril in senile systemic amyloidosis is derived from normal transthyretin. *Proc Natl Acad Sci U S A* **87**, 2843–2845 (1990).

31. Merlini, G. & Westermark, P. The systemic amyloidoses: clearer understanding of the molecular mechanisms offers hope for more effective therapies. *J Intern Med* **255**, 159–178 (2004).

32. Portales-Castillo, I., Yee, J., Tanaka, H. & Fenves, A. Z. Beta-2 Microglobulin Amyloidosis: Past, Present, and Future. *Kidney360* **1**, 1447 (2020).

33. Merlini, G. *et al.* Systemic immunoglobulin light chain amyloidosis. *Nature Reviews Disease Primers 2018 4:1* **4**, 1–19 (2018).

34. Stevens, F. J. & Kisilevsky, R. Immunoglobulin light chains, glycosaminoglycans, and amyloid. *Cellular and Molecular Life Sciences CMLS 2000 57:3* **57**, 441–449 (2000).

35. Yan, S. du *et al.* Receptor-dependent cell stress and amyloid accumulation in systemic amyloidosis. *Nature Medicine 2000 6:6* **6**, 643–651 (2000).

36. Hartley, D. M. *et al.* Protofibrillar intermediates of amyloid beta-protein induce acute electrophysiological changes and progressive neurotoxicity in cortical neurons. *J Neurosci* **19**, 8876–8884 (1999).

37. Goodchild, S. C. *et al.* β2-microglobulin amyloid fibril-induced membrane disruption is enhanced by endosomal lipids and acidic pH. *PLoS One* **9**, (2014).

38. McLaurin, J., Chemistry, A. C.-J. of B. & 1996, undefined. Membrane disruption by Alzheimer β-amyloid peptides mediated through specific binding to either phospholipids or gangliosides: implications for. *ASBMB*.

39. Tsigelny, I. F. *et al.* Role of α-synuclein penetration into the membrane in the mechanisms of oligomer pore formation. *FEBS J* **279**, 1000–1013 (2012).

40. Jackson, M., biochemistry, E. H.-E. in & 2016, undefined. Cellular proteostasis: degradation of misfolded proteins by lysosomes. *portlandpress.com*.

41. Salminen, A., Kaarniranta, K., Kauppinen, A., … J. O.-P. in & 2013, undefined. Impaired autophagy and APP processing in Alzheimer's disease: The potential role of Beclin 1 interactome. *Elsevier*.

42. Winklhofer, K., (BBA)-Molecular, C. H.-B. et B. A. & 2010, undefined. Mitochondrial dysfunction in Parkinson's disease. *Elsevier*.

43. Uttara, B., Singh, A. v., Zamboni, P. & Mahajan, R. T. Oxidative Stress and Neurodegenerative Diseases: A Review of Upstream and Downstream Antioxidant Therapeutic Options.

44. Yankner, B. A. *et al.* Neurotoxicity of a Fragment of the Amyloid Precursor Associated with Alzheimer's Disease. *Science (1979)* **245**, 417–420 (1989).

45. Baldwin, A. J. *et al.* Metastability of native proteins and the phenomenon of amyloid formation. *J Am Chem Soc* **133**, 14160–14163 (2011).

46. Tartaglia, G. G., Pechmann, S., Dobson, C. M. & Vendruscolo, M. Life on the edge: a link between gene expression levels and aggregation rates of human proteins. *Trends Biochem Sci* **32**, 204–206 (2007).

47. Tartaglia, G. G. *et al.* Prediction of Aggregation-Prone Regions in Structured Proteins. *J Mol Biol* **380**, 425–436 (2008).

48. Neudecker, P. *et al.* Structure of an intermediate state in protein folding and aggregation. *Science (1979)* **336**, 362–366 (2012).

49. Fernandez-Escamilla, A. M., Rousseau, F., Schymkowitz, J. & Serrano, L. Prediction of sequence-dependent and mutational effects on the

aggregation of peptides and proteins. *Nature Biotechnology 2004 22:10* **22**, 1302–1306 (2004).

50. Richardson, J. S. & Richardson, D. C. Natural β-sheet proteins use negative design to avoid edge-to-edge aggregation. *Proc Natl Acad Sci U S A* **99**, 2754–2759 (2002).

51. Broome, B. M. & Hecht, M. H. Nature disfavors sequences of alternating polar and non-polar amino acids: implications for amyloidogenesis. *J Mol Biol* **296**, 961–968 (2000).

52. Pawar, A. P. *et al.* Prediction of 'aggregation-prone' and 'aggregation-susceptible' regions in proteins associated with neurodegenerative diseases. *J Mol Biol* **350**, 379–392 (2005).

53. Glickman, M. H. & Ciechanover, A. The ubiquitin-proteasome proteolytic pathway: Destruction for the sake of construction. *Physiol Rev* **82**, 373–428 (2002).

54. Ross, C. A. & Poirier, M. A. Protein aggregation and neurodegenerative disease. *Nature Medicine 2004 10:7* **10**, S10–S17 (2004).

55. Bence, N. F., Sampat, R. M. & Kopito, R. R. Impairment of the ubiquitin-proteasome system by protein aggregation. *Science* **292**, 1552–1555 (2001).

56. Dasari, A. K. R. *et al.* Transthyretin Aggregation Pathway toward the Formation of Distinct Cytotoxic Oligomers. *Scientific Reports 2019 9:1* **9**, 1–10 (2019).

57. Stroud, J. C., Liu, C., Teng, P. K. & Eisenberg, D. Toxic fibrillar oligomers of amyloid-β have cross-β structure. *Proc Natl Acad Sci U S A* (2012) doi:10.1073/pnas.1203193109.

58. Fusco, G. *et al.* Structural basis of membrane disruption and cellular toxicity by α-synuclein oligomers. *Science* **358**, 1440–1443 (2017).

59. Campioni, S. *et al.* A causative link between the structure of aberrant protein oligomers and their toxicity. *Nature Chemical Biology 2010 6:2* **6**, 140–147 (2010).

60. Tokuda, T. *et al.* Detection of elevated levels of α-synuclein oligomers in CSF from patients with Parkinson disease. *Neurology* **75**, 1766–1770 (2010).

61. Reixach, N., Deechongkit, S., Jiang, X., Kelly, J. W. & Buxbaum, J. N. Tissue damage in the amyloidoses: Transthyretin monomers and nonnative oligomers are the major cytotoxic species in tissue culture. *Proc Natl Acad Sci U S A* **101**, 2817–2822 (2004).

62. Sun, Y. *et al.* Nucleation of β-rich oligomers and β-barrels in the early aggregation of human islet amyloid polypeptide. *Biochim Biophys Acta Mol Basis Dis* (2019) doi:10.1016/j.bbadis.2018.11.021.

63. Dasari, A. K. R. *et al.* Disruption of the CD Loop by Enzymatic Cleavage Promotes the Formation of Toxic Transthyretin Oligomers through a Common Transthyretin Misfolding Pathway. *Biochemistry* **59**, 2319–2327 (2020).

64. Adams, D., Koike, H., Slama, M. & Coelho, T. Hereditary transthyretin amyloidosis: a model of medical progress for a fatal disease. *Nature Reviews Neurology 2019 15:7* **15**, 387–404 (2019).

65. Bemporad, F. & Chiti, F. Protein Misfolded Oligomers: Experimental Approaches, Mechanism of Formation, and Structure-Toxicity Relationships. *Chem Biol* **19**, 315–327 (2012).

66. Schaefer, A., Naser, D., Siebeneichler, B., Tarasca, M. v. & Meiering, E. M. Methodological advances and strategies for high resolution structure determination of cellular protein aggregates. *Journal of Biological Chemistry* **298**, 102197 (2022).

67. Chiti, F., biology, C. D.-N. chemical & 2009, undefined. Amyloid formation by globular proteins under native conditions. *nature.com* (2009) doi:10.1038/nchembio.131.

68. Chatani, E. & Yamamoto, N. Recent progress on understanding the mechanisms of amyloid nucleation. *Biophys Rev* **10**, 527–534 (2018).

69. Nelson, P., Alafuzoff, I., … E. B.-… of N. & & 2012, undefined. Correlation of Alzheimer disease neuropathologic changes with cognitive status: a review of the literature. *academic.oup.com*.

70. Otzen, D. & Riek, R. Functional amyloids. *Cold Spring Harb Perspect Biol* **11**, (2019).

71. Dannies, P. S. Concentrating hormones into secretory granules: layers of control. *Mol Cell Endocrinol* **177**, 87–93 (2001).

72. Yan, Z., Yin, M., Chen, J. & Li, X. Assembly and substrate recognition of curli biogenesis system. *Nature Communications 2020 11:1* **11**, 1–10 (2020).

73. Hervás, R. *et al.* Molecular Basis of Orb2 Amyloidogenesis and Blockade of Memory Consolidation. *PLoS Biol* **14**, e1002361 (2016).

74. Saad, S. *et al.* Reversible protein aggregation is a protective mechanism to ensure cell cycle restart after stress. *Nat Cell Biol* **19**, 1202–1213 (2017).

75. Li, J. *et al.* The RIP1/RIP3 Necrosome Forms a Functional Amyloid Signaling Complex Required for Programmed Necrosis. *Cell* **150**, 339–350 (2012).

76. Blake, C. & Serpell, L. Synchrotron X-ray studies suggest that the core of the transthyretin amyloid fibril is a continuous β-sheet helix. *Structure* **4**, 989–998 (1996).

77. Gallardo, R., Ranson, N. A. & Radford, S. E. Amyloid structures: much more than just a cross-β fold. *Curr Opin Struct Biol* **60**, 7–16 (2020).

78. Sunde, M. *et al.* Common core structure of amyloid fibrils by synchrotron X-ray diffraction. *J Mol Biol* **273**, 729–739 (1997).

79. Knowles, T. P. *et al.* Role of intermolecular forces in defining material properties of protein nanofibrils. *Science* **318**, 1900–1903 (2007).

80. Raimondi, S. *et al.* Comparative study of the stabilities of synthetic in vitro and natural ex vivo transthyretin amyloid fibrils. *J Biol Chem* **295**, 11379–11387 (2020).

81. Arosio, P., Knowles, T. P. J. & Linse, S. On the lag phase in amyloid fibril formation. *Physical Chemistry Chemical Physics* **17**, 7606–7618 (2015).

82. Sulatsky, M. I. *et al.* Effect of the fluorescent probes ThT and ANS on the mature amyloid fibrils. *https://doi.org/10.1080/19336896.2020.1720487* **14**, 67–75 (2020).

83. Reinke, A. A. & Gestwicki, J. E. Insight into Amyloid Structure Using Chemical Probes. *Chem Biol Drug Des* **77**, 399–411 (2011).

84. Kayed, R. *et al.* Fibril specific, conformation dependent antibodies recognize a generic epitope common to amyloid fibrils and fibrillar oligomers that is absent in prefibrillar oligomers. *Mol Neurodegener* **2**, 1–11 (2007).

85. Pickles, S. *et al.* Mitochondrial damage revealed by immunoselection for ALS-linked misfolded SOD1. *Hum Mol Genet* **22**, 3947–3959 (2013).

86. Singh, A., Upadhyay, V., Singh, A. & Panda, A. K. Structure-Function Relationship of Inclusion Bodies of a Multimeric Protein. *Front Microbiol* **11**, 876 (2020).

87. Masino, L. *et al.* Domain architecture of the polyglutamine protein ataxin-3: a globular domain followed by a flexible tail. *FEBS Lett* **549**, 21–25 (2003).

88. Jiménez, J. L. *et al.* The protofilament structure of insulin amyloid fibrils. *Proc Natl Acad Sci U S A* **99**, 9196–9201 (2002).

89. Fitzpatrick, A. W. P. *et al.* Atomic structure and hierarchical assembly of a cross-β amyloid fibril. *Proc Natl Acad Sci U S A* **110**, 5468–5473 (2013).

90. Goldsbury, C., Aebi, U. & Frey, P. Visualizing the growth of Alzheimer's Aβ amyloid-like fibrils. *Trends Mol Med* **7**, 582 (2001).

91. Sawaya, M. R. *et al.* Atomic structures of amyloid cross-beta spines reveal varied steric zippers. *Nature* **447**, 453–457 (2007).

92. Krone, M. G. *et al.* Role of water in mediating the assembly of Alzheimer amyloid-beta Abeta16-22 protofilaments. *J Am Chem Soc* **130**, 11066–11072 (2008).

93. Amyloid fibril structures. https://www.nature.com/collections/baccbedebj.

94. Science | AAAS. https://www.science.org/action/doSearch?AllField=amyloid+cryo-EM.

95. Cao, Q., Boyer, D. R., Sawaya, M. R., Ge, P. & Eisenberg, D. S. Cryo-EM structure and inhibitor design of human IAPP (amylin) fibrils. *Nat Struct Mol Biol* **27**, 653–659 (2020).

96. Yang, Y. *et al.* Cryo-EM structures of amyloid-b 42 filaments from human brains. *Science (1979)* **375**, 167–172 (2022).

97. Hervas, R. *et al.* Cryo-EM structure of a neuronal functional amyloid implicated in memory persistence in Drosophila. *Science (1979)* **367**, 1230–1234 (2020).

98. Swuec, P. *et al.* Cryo-EM structure of cardiac amyloid fibrils from an immunoglobulin light chain AL amyloidosis patient. **10**, 1–9 (2019).

99. Radamaker, L. *et al.* Cryo-EM reveals structural breaks in a patient-derived amyloid fibril from systemic AL amyloidosis. *Nat Commun* **12**, 1–8 (2021).

100. Schmidt, M. *et al.* Cryo-EM structure of a transthyretin-derived amyloid fibril from a patient with hereditary ATTR amyloidosis. *Nat Commun* **10**, 1–9 (2019).

101. Jaroniec, C. P. *et al.* High-resolution molecular structure of a peptide in an amyloid fibril determined by magic angle spinning NMR spectroscopy. *Proc Natl Acad Sci U S A* **101**, 711–716 (2004).

102. Iadanza, M. G. *et al.* The structure of a β2-microglobulin fibril suggests a molecular basis for its amyloid polymorphism. *Nat Commun* **9**, (2018).

103. Falcon, B. *et al.* Novel tau filament fold in chronic traumatic encephalopathy encloses hydrophobic molecules. *Nature 2019 568:7752* **568**, 420–423 (2019).

104. Falcon, B. *et al.* Structures of filaments from Pick's disease reveal a novel tau protein fold. *Nature 2018 561:7721* **561**, 137–140 (2018).

105. Fitzpatrick, A. W. P. *et al.* Cryo-EM structures of tau filaments from Alzheimer's disease. *Nature 2017 547:7662* **547**, 185–190 (2017).

106. Yamasaki, T. R. *et al.* Parkinson's disease and multiple system atrophy have distinct -synuclein seed characteristics. *Journal of Biological Chemistry* **294**, 1045–1058 (2019).

107. Gibbons, G. S., Lee, V. M. Y. & Trojanowski, J. Q. Mechanisms of Cell-to-Cell Transmission of Pathological Tau: A Review. *JAMA Neurol* **76**, 101–108 (2019).

108. Qiang, W., Yau, W. M., Lu, J. X., Collinge, J. & Tycko, R. Structural variation in amyloid-β fibrils from Alzheimer's disease clinical subtypes. *Nature 2017 541:7636* **541**, 217–221 (2017).

109. Vaquer-Alicea, J. & Diamond, M. I. Propagation of Protein Aggregation in Neurodegenerative Diseases. *https://doi.org/10.1146/annurev-biochem-061516-045049* **88**, 785–810 (2019).

110. Michaels, T. C. T. *et al.* Dynamics of oligomer populations formed during the aggregation of Alzheimer's Aβ42 peptide. *Nature Chemistry 2020 12:5* **12**, 445–451 (2020).

111. Cohen, S. I. A., Vendruscolo, M., Dobson, C. M. & Knowles, T. P. J. From macroscopic measurements to microscopic mechanisms of protein aggregation. *J Mol Biol* **421**, 160–171 (2012).

112. Jarrett, J. T., Berger, E. P. & Lansbury, P. T. Accelerated Publications The Carboxy Terminus of the ß Amyloid Protein Is Critical for the Seeding of Amyloid Formation: Implications for the Pathogenesis of Alzheimer's Disease* 1". *Biochemistry &copy* (1993).

113. Nelson, R. *et al.* Structure of the cross-β spine of amyloid-like fibrils. *Nature* **435**, 773 (2005).

114. Serio, T. R. *et al.* Nucleated conformational conversion and the replication of conformational information by a prion determinant. *Science (1979)* **289**, 1317–1321 (2000).

115. Jucker, M. & Walker, L. C. Pathogenic protein seeding in alzheimer disease and other neurodegenerative disorders. *Ann Neurol* **70**, 532–540 (2011).

116. Tanaka, M., Collins, S. R., Toyama, B. H. & Weissman, J. S. The physical basis of how prion conformations determine strain phenotypes. *Nature 2006 442:7102* **442**, 585–589 (2006).

117. Rodriguez Camargo, D. C. *et al.* Surface-Catalyzed Secondary Nucleation Dominates the Generation of Toxic IAPP Aggregates. *Front Mol Biosci* **8**, (2021).

118. Kakkar, V. *et al.* The S/T-Rich Motif in the DNAJB6 Chaperone Delays Polyglutamine Aggregation and the Onset of Disease in a Mouse Model. *Mol Cell* **62**, 272–283 (2016).

119. Kundel, F. *et al.* Measurement of Tau Filament Fragmentation Provides Insights into Prion-like Spreading. *ACS Chem Neurosci* **9**, 1276–1282 (2018).

120. Sang, J. C. *et al.* Direct Observation of Murine Prion Protein Replication in Vitro. *J Am Chem Soc* **140**, 14789–14798 (2018).

121. Kurochka, A. S., Yushchenko, D. A., Bouř, P. & Shvadchak, V. v. Influence of Lipid Membranes on α-Synuclein Aggregation. *ACS Chem Neurosci* **12**, 825–830 (2021).

122. Galvagnion, C. *et al.* Lipid vesicles trigger α-synuclein aggregation by stimulating primary nucleation. *Nat Chem Biol* **11**, 229–234 (2015).

123. Giehm, L. & Otzen, D. E. Strategies to increase the reproducibility of protein fibrillization in plate reader assays. *Anal Biochem* **400**, 270–281 (2010).

124. Buell, A. K. *et al.* Solution conditions determine the relative importance of nucleation and growth processes in α-synuclein aggregation. *Proc Natl Acad Sci U S A* **111**, 7671–7676 (2014).

125. Conway, K. A., Harper, J. D. & Lansbury, P. T. Accelerated in vitro fibril formation by a mutant alpha-synuclein linked to early-onset Parkinson disease. *Nat Med* **4**, 1318–1320 (1998).

126. Yang, J., Perrett, S., Wu, S., Uversky, N. & Ruggeri, F. S. Single Molecule Characterization of Amyloid Oligomers. *Molecules 2021, Vol. 26, Page 948* **26**, 948 (2021).

127. Tahirbegi, B. *et al.* Toward high-throughput oligomer detection and classification for early-stage aggregation of amyloidogenic protein. *Front Chem* **10**, 1019 (2022).

128. Young, L. M., Cao, P., Raleigh, D. P., Ashcroft, A. E. & Radford, S. E. Ion mobility spectrometry–mass spectrometry defines the oligomeric intermediates in amylin amyloid formation and the mode of action of inhibitors. *ACS Publications* **136**, 660–670 (2013).

129. Serio, T. R. *et al.* Structural mapping of oligomeric intermediates in an amyloid assembly pathway. doi:10.7554/eLife.46574.001.

130. Young, L., Tu, L., Raleigh, D., … A. A.-C. & 2017, undefined. Understanding co-polymerization in amyloid formation by direct observation of mixed oligomers. *pubs.rsc.org*.

131. Meng, F., Yoo, J. & Chung, H. S. Single-molecule fluorescence imaging and deep learning reveal highly heterogeneous aggregation of amyloid-β 42. *Proc Natl Acad Sci U S A* **119**, e2116736119 (2022).

132. Anfinsen, C. B. Principles that govern the folding of protein chains. *Science (1979)* **181**, 223–230 (1973).

133. Nelson, D. L. (David L., Cox, M. M. & Hoskins, A. A. Lehninger principles of biochemistry. 3.

134. Jackson, S. E. How do small single-domain proteins fold? *Fold Des* **3**, R81–R91 (1998).

135. Brockwell, D. J. & Radford, S. E. Intermediates: ubiquitous species on folding energy landscapes? *Curr Opin Struct Biol* **17**, 30–37 (2007).

136. Fersht, A. R. & Daggett, V. Protein Folding and Unfolding at Atomic Resolution. *Cell* **108**, 573–582 (2002).

137. Hartl, F. U. & Hayer-Hartl, M. Converging concepts of protein folding in vitro and in vivo R E V I E W P R OT E I N F O L D I N G. *Nat Struct Mol Biol* **16**, (2009).

138. Lee, R. van der *et al.* Classification of Intrinsically Disordered Regions and Proteins. (2014) doi:10.1021/CR400525M.

139. Levinthal, C. How to fold graciously. *Mossbauer spectroscopy in biological systems* **67**, 22–24 (1969).

140. Karplus, M. The Levinthal paradox: yesterday and today. *Fold Des* **2**, S69–S75 (1997).

141. Bryngelson, J. D. & Wolynes, P. G. Intermediates and barrier crossing in a random energy model (with applications to protein folding). *Journal of Physical Chemistry* **93**, 6902–6915 (1989).

142. Levinthal, C. Are there pathways for protein folding? *Journal de Chimie Physique* **65**, 44–45 (1968).

143. Dill, K. A. & Chan, H. S. From Levinthal to pathways to funnels. *Nature Structural Biology 1997 4:1* **4**, 10–19 (1997).

144. Wolynes, P. G., Eaton, W. A. & Fersht, A. R. Chemical physics of protein folding. *Proceedings of the National Academy of Sciences* **109**, 17770–17771 (2012).

145. Bryngelson, J. D. & Wolynes, P. G. Spin glasses and the statistical mechanics of protein folding. *Proc Natl Acad Sci U S A* **84**, 7524–7528 (1987).

146. Go, N. Theoretical Studies of Protein Folding. *https://doi.org/10.1146/annurev.bb.12.060183.001151* **12**, 183–210 (2003).

147. Taketomi, H., Ueda, Y. & Gō, N. STUDIES ON PROTEIN FOLDING, UNFOLDING AND FLUCTUATIONS BY COMPUTER SIMULATION. *Int J Pept Protein Res* **7**, 445–459 (1975).

148. PAULING, L., COREY, R. B. & BRANSON, H. R. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci U S A* **37**, 205–211 (1951).

149. PAULING, L. & COREY, R. B. Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proc Natl Acad Sci U S A* **37**, 235–240 (1951).

150. Lau, K. F. & Dill, K. A. A Lattice Statistical Mechanics Model of the Conformational and Sequence Spaces of Proteins. *Macromolecules* **22**, 3986–3997 (1989).

151. Dill, K. A. *et al.* Principles of protein folding — A perspective from simple exact models. *Protein Science* **4**, 561–602 (1995).

152. Dill, K. A. & Stigter, D. Modeling Protein Stability As Heteropolymer Collapse. *Adv Protein Chem* **46**, 59–104 (1995).

153. Stigter, D., Alonso, D. O. V. & Dill, K. A. Protein stability: electrostatics and compact denatured states. *Proceedings of the National Academy of Sciences* **88**, 4176–4180 (1991).

154. Lim, W. A. & Sauer, R. T. Alternative packing arrangements in the hydrophobic core of λrepresser. *Nature 1989 339:6219* **339**, 31–36 (1989).

155. Bowie, J. U., Reidhaar-Olson, J. F., Lim, W. A. & Sauer, R. T. Deciphering the Message in Protein Sequences: Tolerance to Amino Acid Substitutions. *Science (1979)* **247**, 1306–1310 (1990).

156. Giri, R. *et al.* Folding pathways of proteins with increasing degree of sequence identities but different structure and function. *Proc Natl Acad Sci U S A* **109**, 17772–17776 (2012).

157. Sulkowska, J. I., Noel, J. K. & Onuchic, J. N. Energy landscape of knotted protein folding. *Proc Natl Acad Sci U S A* **109**, 17783–17788 (2012).

158. Korzhnev, D. M., Religa, T. L. & Kay, L. E. Transiently populated intermediate functions as a branching point of the FF domain folding pathway. *Proc Natl Acad Sci U S A* **109**, 17777–17782 (2012).

159. Escusa-Toret, S., Vonk, W. I. M. & Frydman, J. Spatial sequestration of misfolded proteins by a dynamic chaperone pathway enhances cellular fitness during stress. *Nat Cell Biol* **15**, 1231–1243 (2013).

160. Buchner, J. Molecular chaperones and protein quality control: an introduction to the JBC Reviews thematic series. *J Biol Chem* **294**, 2074 (2019).

161. Chen, B., Retzlaff, M., Roos, T. & Frydman, J. Cellular strategies of protein quality control. *Cold Spring Harb Perspect Biol* **3**, 1–14 (2011).

162. Lu, J., Deutsch, C., Bernstein, H. & Johnson, A. Folding zones inside the ribosomal exit tunnel. *Nature Structural & Molecular Biology 2005 12:12* **12**, 1123–1129 (2005).

163. Woolhead, C. A., McCormick, P. J. & Johnson, A. E. Nascent Membrane and Secretory Proteins Differ in FRET-Detected Folding Far inside the Ribosome and in Their Exposure to Ribosomal Proteins. *Cell* **116**, 725–736 (2004).

164. Hartl, F. U. *et al.* Molecular chaperones in protein folding and proteostasis. *Nature* **475**, 1716–1720.e1 (2011).

165. Koopman, M. B., Ferrari, L. & Rüdiger, S. G. D. How do protein aggregates escape quality control in neurodegeneration? *Trends Neurosci* **45**, 257–271 (2022).

166. Chiti, F. *et al.* Studies of the aggregation of mutant proteins in vitro provide insights into the genetics of amyloid diseases. *Proceedings of the National Academy of Sciences* **99**, 16419–16426 (2002).

167. Saelices, L. *et al.* Uncovering the mechanism of aggregation of human transthyretin. *Journal of Biological Chemistry* **290**, 28932–28943 (2015).

168. Mangione, P. P. *et al.* Plasminogen activation triggers transthyretin amyloidogenesis in vitro. *Journal of Biological Chemistry* **293**, 14192–14199 (2018).

169. Mangione, P. P. *et al.* Proteolytic cleavage of Ser52Pro variant transthyretin triggers its amyloid fibrillogenesis. *Proc Natl Acad Sci U S A* **111**, 1539–1544 (2014).

170. Nag, S., Larsson, M., Robinson, R. C. & Burtnick, L. D. Gelsolin: The tail of a molecular gymnast. *Cytoskeleton* **70**, 360–384 (2013).

171. Marcoux, J. *et al.* A novel mechano-enzymatic cleavage mechanism underlies transthyretin amyloidogenesis. *EMBO Mol Med* **7**, 1337–1349 (2015).

172. Dommer, A. *et al.* #COVIDisAirborne: AI-Enabled Multiscale Computational Microscopy of Delta SARS-CoV-2 in a Respiratory Aerosol. *bioRxiv* 11 (2022) doi:10.1101/2021.11.12.468428.

173. Strodel, B. Amyloid aggregation simulations: challenges, advances and perspectives. *Curr Opin Struct Biol* **67**, 145–152 (2021).

174. Maximova, T., Moffatt, R., Ma, B., Nussinov, R. & Shehu, A. Principles and Overview of Sampling Methods for Modeling Macromolecular Structure and Dynamics. *PLoS Comput Biol* **12**, e1004619 (2016).

175. Fatafta, H., Samantray, S., Sayyed-Ahmad, A., Coskuner-Weber, O. & Strodel, B. Molecular simulations of IDPs: From ensemble generation to IDP interactions leading to disorder-to-order transitions. in *Progress in Molecular Biology and Translational Science* vol. 183 (2021).

176. Mutter, S. T., Turner, M., Deeth, R. J. & Platts, J. A. Metal Binding to Amyloid-β1-42: A Ligand Field Molecular Dynamics Study. *ACS Chem Neurosci* **9**, 2795–2806 (2018).

177. Fatafta, H., Kav, B., Bundschuh, B. F., Loschwitz, J. & Strodel, B. Disorder-to-order transition of the amyloid-β peptide upon lipid binding. *Biophys Chem* **280**, 106700 (2022).

178. Fatafta, H., Khaled, M., Owen, M. C., Sayyed-Ahmad, A. & Strodel, B. Amyloid-β peptide dimers undergo a random coil to β-sheet transition in the aqueous phase but not at the neuronal membrane. *Proc Natl Acad Sci U S A* **118**, e2106210118 (2021).

179. Illig, A. M. & Strodel, B. Performance of Markov State Models and Transition Networks on Characterizing Amyloid Aggregation Pathways from MD Data. *J Chem Theory Comput* **16**, 7825–7839 (2020).

180. Jahan, I. & Nayeem, S. M. Destabilization of Alzheimer's Aβ42protofibrils with acyclovir, carmustine, curcumin, and tetracycline: insights from molecular dynamics simulations. *New Journal of Chemistry* **45**, (2021).

181. Parvaee, E., Bozorgmehr, M. R. & Morsali, A. Role of repulsive forces on self-assembly behavior of amyloid β-peptide (1-40): Molecular dynamics simulation approach. *Physica A: Statistical Mechanics and its Applications* **513**, (2019).

182. Salimi, A., Li, H. & Lee, J. Y. Molecular insight into the early stage of amyloid-β(1-42) Homodimers aggregation influenced by histidine tautomerism. *Int J Biol Macromol* **184**, (2021).

183. Löhr, T., Kohlhoff, K., Heller, G. T., Camilloni, C. & Vendruscolo, M. A kinetic ensemble of the Alzheimer's Aβ peptide. *Nature Computational Science 2021 1:1* **1**, 71–78 (2021).

184. Löhr, T., Kohlhoff, K., Heller, G. T., Camilloni, C. & Vendruscolo, M. A small molecule stabilises the disordered native state of the Alzheimer's Aβ peptide. doi:10.1101/2021.11.10.468059.

185. Robustelli, P. *et al.* Molecular Basis of Small-Molecule Binding to α-Synuclein. *J Am Chem Soc* **144**, 2501–2510 (2022).

186. Smida, K., Albedah, M. A., Rashid, R. F. & Al-Qawasmi, A. R. Molecular dynamics method for targeting α-synuclein aggregation induced

Parkinson's disease using boron nitride nanostructures. *Eng Anal Bound Elem* **146**, 89–95 (2023).

187. Savva, L. & Platts, J. A. How Cu(II) binding affects structure and dynamics of α-synuclein revealed by molecular dynamics simulations. *J Inorg Biochem* **239**, 112068 (2023).

188. Yamauchi, M. & Okumura, H. Dimerization of α-Synuclein Fragments Studied by Isothermal-Isobaric Replica-Permutation Molecular Dynamics Simulation. *J Chem Inf Model* **61**, 1307–1321 (2021).

189. Ilie, I. M., Nayar, D., den Otter, W. K., van der Vegt, N. F. A. & Briels, W. J. Intrinsic Conformational Preferences and Interactions in α-Synuclein Fibrils: Insights from Molecular Dynamics Simulations. *J Chem Theory Comput* **14**, 3298–3310 (2018).

190. Brodie, N. I., Popov, K. I., Petrotchenko, E. v., Dokholyan, N. v. & Borchers, C. H. Conformational ensemble of native α-synuclein in solution as determined by short-distance crosslinking constraint-guided discrete molecular dynamics simulations. *PLoS Comput Biol* **15**, e1006859 (2019).

191. Tsigelny, I. F. *et al.* Dynamics of α-synuclein aggregation and inhibition of pore-like oligomer development by β-synuclein. *FEBS J* **274**, 1862–1877 (2007).

192. Dedmon, M. M., Lindorff-Larsen, K., Christodoulou, J., Vendruscolo, M. & Dobson, C. M. Mapping long-range interactions in α-synuclein using spin-label NMR and ensemble molecular dynamics simulations. *J Am Chem Soc* **127**, 476–477 (2005).

193. Ramis, R. *et al.* A Coarse-Grained Molecular Dynamics Approach to the Study of the Intrinsically Disordered Protein α-Synuclein. *J Chem Inf Model* **59**, 1458–1471 (2019).

194. Miyashita, N., Straub, J. E., Thirumalai, D. & Sugita, Y. Transmembrane structures of amyloid precursor protein dimer predicted by replica-exchange molecular dynamics simulations. *J Am Chem Soc* **131**, 3438–3439 (2009).

195. Dominguez, L., Meredith, S. C., Straub, J. E. & Thirumalai, D. Transmembrane fragment structures of amyloid precursor protein depend on membrane surface curvature. *J Am Chem Soc* **136**, 854–857 (2014).

196. Ilie, I. M. & Caflisch, A. Simulation Studies of Amyloidogenic Polypeptides and Their Aggregates. *Chem Rev* **119**, 6956–6993 (2019).

197. Robustelli, P., Piana, S. & Shaw, D. E. Developing a molecular dynamics force field for both folded and disordered protein states. *Proc Natl Acad Sci U S A* **115**, E4758–E4766 (2018).

198. Best, R. B., Zheng, W. & Mittal, J. Balanced protein-water interactions improve properties of disordered proteins and non-specific protein association. *J Chem Theory Comput* **10**, 5113–5124 (2014).

199. Piana, S., Donchev, A. G., Robustelli, P. & Shaw, D. E. Water dispersion interactions strongly influence simulated structural properties of disordered protein states. *Journal of Physical Chemistry B* **119**, 5113–5123 (2015).

200. Huang, J. *et al.* CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *nature.com*.

201. Joshi, S. Y. & Deshmukh, S. A. A review of advancements in coarse-grained molecular dynamics simulations. *https://doi.org/10.1080/08927022.2020.1828583* **47**, 786–803 (2020).

202. Huang, W. *et al.* Coarse grain models and the computer simulation of soft materials. *Journal of Physics: Condensed Matter* **16**, R481 (2004).

203. Bereau, T. & Deserno, M. Generic coarse-grained model for protein folding and aggregation. *Journal of Chemical Physics* **130**, (2009).

204. West, B., Brown, F. L. H. & Schmid, F. Membrane-Protein Interactions in a Generic Coarse-Grained Model for Lipid Bilayers. *Biophys J* **96**, 101 (2009).

205. Marrink, S. J., de Vries, A. H. & Mark, A. E. Coarse Grained Model for Semiquantitative Lipid Simulations. *Journal of Physical Chemistry B* **108**, 750–760 (2004).

206. Akkermans, R. L. C. & Briels, W. J. A structure-based coarse-grained model for polymer melts. *J Chem Phys* **114**, 1020 (2000).

207. Arash, B., Park, H. S. & Rabczuk, T. Tensile fracture behavior of short carbon nanotube reinforced polymer composites: A coarse-grained model. *Compos Struct* **134**, 981–988 (2017).

208. Souza, P. C. T. *et al.* Martini 3: a general purpose force field for coarse-grained molecular dynamics. *Nature Methods 2021 18:4* **18**, 382–388 (2021).

209. Machado, M. R. *et al.* The SIRAH 2.0 Force Field: Altius, Fortius, Citius. *J Chem Theory Comput* **15**, 2719–2733 (2019).

210. Marrink, S. J., Risselada, H. J., Yefimov, S., Tieleman, D. P. & de Vries, A. H. The MARTINI force field: Coarse grained model for biomolecular simulations. *Journal of Physical Chemistry B* **111**, 7812–7824 (2007).

211. Darré, L., MacHado, M. R., Dans, P. D., Herrera, F. E. & Pantano, S. Another coarse grain model for aqueous solvation: WAT FOUR? *J Chem Theory Comput* **6**, 3793–3807 (2010).

212. Darré, L. *et al.* SIRAH: A structurally unbiased coarse-grained force field for proteins with aqueous solvation and long-range electrostatics. *J Chem Theory Comput* **11**, 723–739 (2015).

213. Liwo, A., He, Y. & Scheraga, H. A. Coarse-grained force field: general folding theory. *Physical Chemistry Chemical Physics* **13**, 16890–16901 (2011).

214. Maupetit, J., Tuffery, P. & Derreumaux, P. A coarse-grained protein force field for folding and structure prediction. *Proteins: Structure, Function, and Bioinformatics* **69**, 394–408 (2007).

215. Kar, P., Gopal, S. M., Cheng, Y. M., Predeus, A. & Feig, M. PRIMO: A transferable coarse-grained force field for proteins. *J Chem Theory Comput* **9**, 3769–3788 (2013).

216. Lu, L., Izvekov, S., Das, A., Andersen, H. C. & Voth, G. A. Efficient, regularized, and scalable algorithms for multiscale coarse-graining. *J Chem Theory Comput* **6**, 954–965 (2010).

217. Pretti, E. & Shell, M. S. A microcanonical approach to temperature-transferable coarse-grained models using the relative entropy. *J Chem Phys* **155**, 094102 (2021).

218. Wang, J. *et al.* Machine Learning of Coarse-Grained Molecular Dynamics Force Fields. *ACS Cent Sci* **5**, 755–767 (2019).

219. Nguyen, H. D. & Hall, C. K. Molecular dynamics simulations of spontaneous fibril formation by random-coil peptides. *Proceedings of the National Academy of Sciences* **101**, 16180–16185 (2004).

220. Auer, S., Meersman, F., Dobson, C. M. & Vendruscolo, M. A Generic Mechanism of Emergence of Amyloid Protofilaments from Disordered Oligomeric Aggregates. *PLoS Comput Biol* **4**, e1000222 (2008).

221. Dima, R. I. & Thirumalai, D. Exploring protein aggregation and self-propagation using lattice models: Phase diagram and kinetics. *Protein Science* **11**, 1036–1049 (2002).

222. Abeln, S., Vendruscolo, M., Dobson, C. M. & Frenkel, D. A Simple Lattice Model That Captures Protein Folding, Aggregation and Amyloid Formation. *PLoS One* **9**, e85185 (2014).

223. Šaric, A. *et al.* Physical determinants of the self-replication of protein fibrils. *Nature Physics 2016 12:9* **12**, 874–880 (2016).

224. Cheon, M. *et al.* Structural Reorganisation and Potential Toxicity of Oligomeric Species Formed during the Assembly of Amyloid Fibrils. *PLoS Comput Biol* **3**, e173 (2007).

225. Derreumaux, P. & Mousseau, N. Coarse-grained protein molecular dynamics simulations. *J Chem Phys* **126**, (2007).

226. Ma, Y. W., Lin, T. Y. & Tsai, M. Y. Fibril Surface-Dependent Amyloid Precursors Revealed by Coarse-Grained Molecular Dynamics Simulation. *Front Mol Biosci* **8**, (2021).

227. Yu, H., Han, W., Ma, W. & Schulten, K. Transient β-hairpin formation in α-synuclein monomer revealed by coarse-grained molecular dynamics simulation. *J Chem Phys* **143**, 243142 (2015).

228. Go, N. & Taketomi, H. Respective roles of short- and long-range interactions in protein folding. *Proceedings of the National Academy of Sciences* **75**, 559–563 (1978).

229. Takada, S. Gō model revisited. *Biophys Physicobiol* **16**, 248–255 (2019).

230. Šali, A., Shakhnovich, E., biology, M. K.-J. of molecular & 1994, undefined. Kinetics of protein folding: A lattice model study of the requirements for folding to the native state. *Elsevier*.

231. Onuchic, J. N., Luthey-Schulten, Z. & Wolynes, P. G. Theory of Protein Folding: The Energy Landscape Perspective. *Annu Rev Phys Chem* **48**, 545–600.

232. Estácio, S. G., Fernandes, C. S., Krobath, H., Faísca, P. F. N. & Shakhnovich, E. I. Robustness of atomistic Gō models in predicting native-like folding intermediates. *J Chem Phys* **137**, 85102 (2012).

233. Clementi, C., Nymeyer, H. & Onuchic, J. N. Topological and energetic factors: What determines the structural details of the transition state ensemble and 'en-route' intermediates for protein folding? An investigation for small globular proteins. *J Mol Biol* **298**, 937–953 (2000).

234. Karanicolas, J. & Iii, C. L. B. The origins of asymmetry in the folding transition states of protein L and protein G. *Protein Science* **11**, 2351–2361 (2002).

235. Karanicolas, J. & Brooks, C. L. Improved Gō-like Models Demonstrate the Robustness of Protein Folding Mechanisms Towards Non-native Interactions. *J Mol Biol* **334**, 309–325 (2003).

236. Okazaki, K. I., Koga, N., Takada, S., Onuchic, J. N. & Wolynes, P. G. Multiple-basin energy landscapes for large-amplitude conformational motions of proteins: Structure-based molecular dynamics simulations. *Proc Natl Acad Sci U S A* **103**, 11844–11849 (2006).

237. Best, R. B., Chen, Y. G. & Hummer, G. Slow protein conformational dynamics from multiple experimental structures: The helix/sheet transition of Arc repressor. *Structure* **13**, 1755–1763 (2005).

238. Ganguly, D. & Chen, J. Topology-based modeling of intrinsically disordered proteins: Balancing intrinsic folding and intermolecular interactions. *Proteins: Structure, Function, and Bioinformatics* **79**, 1251–1266 (2011).

239. Knott, M. & Best, R. B. Discriminating binding mechanisms of an intrinsically disordered protein via a multi-state coarse-grained model. *J Chem Phys* **140**, 175102 (2014).

240. Wojtczak, A., Cody, V., Luft, J. R. & Pangborn, W. Structures of human transthyretin complexed with thyroxine at 2.0 Å resolution and 3′,5′-dinitro-N-acetyl-L-thyronine at 2.2 Å resolution. *Acta Crystallogr D Biol Crystallogr* **52**, 758–765 (1996).

241. Liang, Y., Ore, M. O., Morin, S. & Wilson, D. J. Specific disruption of transthyretin(105-115) fibrilization using 'stabilizing' inhibitors of transthyretin amyloidogenesis. *Biochemistry* **51**, 3523–3530 (2012).

242. Gordon, A. H., Gross, J., O'Connor, D. & Pitt-Rivers, R. Nature of the Circulating Thyroid Hormone - Plasma Protein Complex. *Nature 1952 169:4288* **169**, 19–20 (1952).

243. Raz, A. & Goodman, D. S. The Interaction of Thyroxine with Human Plasma Prealbumin and with the Prealbumin-Retinol-binding Protein Complex. *Journal of Biological Chemistry* **244**, 3230–3237 (1969).

244. Hyung, S. J., Deroo, S. & Robinson, C. v. Retinol and retinol-binding protein stabilize transthyretin via formation of retinol transport complex. *ACS Chem Biol* **5**, 1137–1146 (2010).

245. Liz, M. A. *et al.* A Narrative Review of the Role of Transthyretin in Health and Disease. *Neurol Ther* **9**, 395–402 (2020).

246. Andrade, C. A peculiar form of peripheral neuropathy; familiar atypical generalized amyloidosis with special involvement of the peripheral nerves. *Brain* **75**, 408–427 (1952).

247. Araki, S. Type I familial amyloidotic polyneuropathy (Japanese type). *Brain Dev* **6**, 128–133 (1984).

248. Sousa, A., Anderson, R., Drugge, U., Holmgren, G. & Sandgren, O. Familial Amyloidotic Polyneuropathy in Sweden: Geographical Distribution, Age of Onset, and Prevalence. *Hum Hered* **43**, 288–294 (1993).

249. Tawara, S., Nakazato, M., Kangawa, K., Matsuo, H. & Araki, S. Identification of amyloid prealbumin variant in familial amyloidotic polyneuropathy (Japanese type). *Biochem Biophys Res Commun* **116**, 880–888 (1983).

250. Yamamoto, H. & Yokochi, T. Transthyretin cardiac amyloidosis: an update on diagnosis and treatment. *ESC Heart Fail* **6**, 1128–1139 (2019).

251. Benson, M. D. & Kincaid, J. C. The molecular biology and clinical features of amyloid neuropathy. *Muscle Nerve* **36**, 411–423 (2007).

252. Kelly, J. W. *et al.* Transthyretin Quaternary and Tertiary Structural Changes Facilitate Misassembly into Amyloid. *Adv Protein Chem* **50**, 161–181 (1997).

253. Mccutchen, S. L., Lai, Z., Miroy, G. J., Kelly, J. W. & Colon, W. Comparison of Lethal and Nonlethal Transthyretin Variants and Their

Relationship to Amyloid Disease1. *Biochemistry* **34**, 13527–13536 (1995).

254. Hammarström, P., Jiang, X., Hurshman, A. R., Powers, E. T. & Kelly, J. W. Sequence-dependent denaturation energetics: A major determinant in amyloid disease diversity. *Proc Natl Acad Sci U S A* **99**, 16427–16432 (2002).

255. Ueda, M. *et al.* Clinicopathological features of senile systemic amyloidosis: an ante- and post-mortem study. *Modern Pathology 2011 24:12* **24**, 1533–1544 (2011).

256. Shiozaki, T., Sato, N., Hayashi, T., Kobayashi, K. & Asamura, H. Wild-type ATTR amyloidosis may be associated with unexpected death among the elderly. *Leg Med* **41**, 101634 (2019).

257. Westermark, P., Bergström, J., Solomon, A., Murphy, C. & Sletten, K. Transthyretin-derived senile systemic amyloidosis: clinicopathologic and structural considerations. *https://doi.org/10.1080/13506129.2003.12088568* **10**, 48–54 (2020).

258. Carvalho, A., Rocha, A. & Lobato, L. Liver transplantation in transthyretin amyloidosis: issues and challenges. *Liver Transpl* **21**, 282–292 (2015).

259. Ericzon, B. G. *et al.* Liver Transplantation for Hereditary Transthyretin Amyloidosis: After 20 Years Still the Best Therapeutic Alternative? *Transplantation* **99**, 1847–1854 (2015).

260. Benson, M. D. Liver transplantation and transthyretin amyloidosis. *Muscle Nerve* **47**, 157–162 (2013).

261. Adams, D. *et al.* Patisiran, an RNAi Therapeutic, for Hereditary Transthyretin Amyloidosis. *N Engl J Med* **379**, 11–21 (2018).

262. Benson, M. D. *et al.* Inotersen Treatment for Patients with Hereditary Transthyretin Amyloidosis. *N Engl J Med* **379**, 22–31 (2018).

263. Hammarström, P., Wiseman, R. L., Powers, E. T. & Kelly, J. W. Prevention of transthyretin arnyloid disease by changing protein misfolding energetics. *Science (1979)* **299**, 713–716 (2003).

264. Sebastião, M. P., Lamzin, V., Saraiva, M. J. & Damas, A. M. Transthyretin stability as a key factor in amyloidogenesis: X-ray analysis at atomic resolution. *J Mol Biol* **306**, 733–744 (2001).

265. Coelho, T. *et al.* Long-term effects of tafamidis for the treatment of transthyretin familial amyloid polyneuropathy. *J Neurol* **260**, 2802–2814 (2013).

266. Coelho, T. *et al.* Tafamidis for transthyretin familial amyloid polyneuropathy: a randomized, controlled trial. *Neurology* **79**, 785–792 (2012).

267. Maurer, M. S. *et al.* Tafamidis Treatment for Patients with Transthyretin Amyloid Cardiomyopathy. *New England Journal of Medicine* **379**, 1007–1016 (2018).

268. Blake, C. C. F., Geisow, M. J., Oatley, S. J., Rérat, B. & Rérat, C. Structure of prealbumin: Secondary, tertiary and quaternary interactions determined by Fourier refinement at 1.8 Å. *J Mol Biol* **121**, 339–356 (1978).

269. Naylor, H. M. & Newcomer, M. E. The structure of human retinol-binding protein (RBP) with its carrier protein transthyretin reveals an interaction with the carboxy terminus of RBP. *Biochemistry* **38**, 2647–2653 (1999).

270. Zanotti, G. *et al.* Structural and mutational analyses of protein-protein interactions between transthyretin and retinol-binding protein. *FEBS J* **275**, 5841–5854 (2008).

271. Monaco, H. L., Rizzi, M. & Coda, A. Structure of a complex of two plasma proteins: Transthyretin and retinol-binding protein. *Science (1979)* **268**, 1039–1041 (1995).

272. Zanotti, G. & Berni, R. Plasma Retinol-Binding Protein: Structure and Interactions with Retinol, Retinoids, and Transthyretin. *Vitam Horm* **69**, 271–295 (2004).

273. Johnson, S. M. *et al.* Native state kinetic stabilization as a strategy to ameliorate protein misfolding diseases: A focus on the transthyretin amyloidoses. *Acc Chem Res* **38**, 911–921 (2005).

274. Cornwell, G. G., Sletten, K., Olofsson, B. O., Johansson, B. & Westermark, P. Prealbumin: its association with amyloid. *J Clin Pathol* **40**, 226–231 (1987).

275. Bergström, J. *et al.* Amyloid deposits in transthyretin-derived amyloidosis: cleaved transthyretin is associated with distinct amyloid morphology. *J Pathol* **206**, 224–232 (2005).

276. Suhr, O. B., Lundgren, E. & Westermark, P. One mutation, two distinct disease variants: unravelling the impact of transthyretin amyloid fibril composition. *J Intern Med* **281**, 337–347 (2017).

277. Ihse, E. *et al.* Amyloid fibrils containing fragmented ATTR may be the standard fibril composition in ATTR amyloidosis. *http://dx.doi.org/10.3109/13506129.2013.797890* **20**, 142–150 (2013).

278. Gustavsson, Å., Engström, U. & Westermark, P. Normal transthyretin and synthetic transthyretin fragments from amyloid-like fibrils in vitro. *Biochem Biophys Res Commun* **175**, 1159–1164 (1991).

279. Serpell, L. C. *et al.* Examination of the structure of the transthyretin amyloid fibril by image reconstruction from electron micrographs. *J Mol Biol* **254**, 113–118 (1995).

280. Lim, K. H. *et al.* Solid-State NMR Studies Reveal Native-like β-Sheet Structures in Transthyretin Amyloid. *Biochemistry* **55**, 5272–5278 (2016).

281. Gonzalez-Duarte, A. & Ulloa-Aguirre, A. A Brief Journey through Protein Misfolding in Transthyretin Amyloidosis (ATTR Amyloidosis). *International Journal of Molecular Sciences 2021, Vol. 22, Page 13158* **22**, 13158 (2021).

282. Thylen, C. *et al.* Modifications of transthyretin in amyloid fibrils: analysis of amyloid from homozygous and heterozygous individuals with the Met30 mutation. *EMBO J* **12**, 743–748 (1993).

283. Sant'Anna, R. *et al.* Repositioning tolcapone as a potent inhibitor of transthyretin amyloidogenesis and associated cellular toxicity. *Nature Communications 2016 7:1* **7**, 1–13 (2016).

284. Quintas, A., Vaz, D. C., Cardoso, I., Saraiva, M. J. M. & Brito, R. M. M. Tetramer Dissociation and Monomer Partial Unfolding Precedes Protofibril Formation in Amyloidogenic Transthyretin Variants. *Journal of Biological Chemistry* **276**, 27207–27213 (2001).

285. Mendes Sousa, M. M., Cardoso, I., Fernandes, R., Guimarães, A. & João Saraiva, M. Deposition of Transthyretin in Early Stages of Familial Amyloidotic Polyneuropathy: Evidence for Toxicity of Nonfibrillar Aggregates. *Am J Pathol* **159**, 1993–2000 (2001).

286. Foss, T. R., Wiseman, R. L. & Kelly, J. W. The pathway by which the tetrameric protein transthyretin dissociates. *Biochemistry* **44**, 15525–15533 (2005).

287. Gasperini, R. J., Klaver, D. W., Hou, X., Aguilar, M. I. & Small, D. H. Mechanisms of transthyretin aggregation and toxicity. *Subcell Biochem* **65**, 211–224 (2012).

288. Kingsbury, J. S. *et al.* Detailed structural analysis of amyloidogenic wild-type transthyretin using a novel purification strategy and mass spectrometry. *Anal Chem* **79**, 1990–1998 (2007).

289. JARVIS, J. A., KIRKPATRICK, A. & CRAIK, D. J. 1H NMR analysis of fibril-forming peptide fragments of transthyretin. *Int J Pept Protein Res* **44**, 388–398 (1994).

290. Long, J. M. & Holtzman, D. M. Alzheimer Disease: An Update on Pathobiology and Treatment Strategies. *Cell* **179**, 312–339 (2019).

291. McKhann, G. M. *et al.* The diagnosis of dementia due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's & Dementia* **7**, 263–269 (2011).

292. Alzheimer, Association, sciencestaff & alzorg. 2019 Alzheimer's disease facts and figures. *Alzheimer's & Dementia* **15**, 321–387 (2019).

293. Glenner, G. G. & Wong, C. W. Alzheimer's disease: Initial report of the purification and characterization of a novel cerebrovascular amyloid protein. *Biochem Biophys Res Commun* **120**, 885–890 (1984).

294. Xia, W. γ-Secretase and its modulators: Twenty years and beyond. *Neurosci Lett* **701**, 162–169 (2019).

295. Voytyuk, I., de Strooper, B. & Chávez-Gutiérrez, L. Modulation of γ- and β-Secretases as Early Prevention Against Alzheimer's Disease. *Biol Psychiatry* **83**, 320–327 (2018).

296. Wei, W., Nguyen, L., Kessels, H., … H. H.-N. & 2010, undefined. Amyloid beta from axons and dendrites reduces local spine number and plasticity. *nature.com*.

297. Kamenetz, F., Tomita, T., Hsieh, H., Neuron, G. S.- & 2003, undefined. APP processing and synaptic function. *Elsevier*.

298. Suzuki, N. *et al.* An Increased Percentage of Long Amyloid β Protein Secreted by Familial Amyloid β Protein Precursor (βApp717) Mutants. *Science (1979)* **264**, 1336–1340 (1994).

299. Citron, M. *et al.* Excessive production of amyloid beta-protein by peripheral cells of symptomatic and presymptomatic patients carrying the Swedish familial Alzheimer disease mutation. *Proceedings of the National Academy of Sciences* **91**, 11993–11997 (1994).

300. de la Vega, M. P. *et al.* The Uppsala APP deletion causes early onset autosomal dominant Alzheimer's disease by altering APP processing and increasing amyloid β fibril formation. *Sci Transl Med* **13**, (2021).

301. What Happens to the Brain in Alzheimer's Disease? | National Institute on Aging. https://www.nia.nih.gov/health/what-happens-brain-alzheimers-disease.

302. Gordon, B. A. *et al.* Spatial patterns of neuroimaging biomarker change in individuals from families with autosomal dominant Alzheimer's disease: a longitudinal study. *Lancet Neurol* **17**, 241–250 (2018).

303. disease, J. H.-J. of A. & 2006, undefined. Alzheimer's disease: the amyloid cascade hypothesis: an update and reappraisal. *content.iospress.com* (2006).

304. Hardy, J. A. & Higgins, G. A. Alzheimer's disease: The amyloid cascade hypothesis. *Science (1979)* **256**, 184–185 (1992).

305. Selkoe, D. J. & Hardy, J. The amyloid hypothesis of Alzheimer's disease at 25 years. *EMBO Mol Med* **8**, 595–608 (2016).

306. Musiek, E. S. & Holtzman, D. M. Three dimensions of the amyloid hypothesis: time, space and 'wingmen'. *Nature Neuroscience 2015 18:6* **18**, 800–806 (2015).

307. Farrer, L. A. *et al.* Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease: a meta-analysis. *jamanetwork.com*.

308. Saunders, A., Strittmatter, W., Neurology, D. S.- & 1993, undefined. Association of apolipoprotein E allele ϵ4 with late-onset familial and sporadic Alzheimer's disease. *AAN Enterprises*.

309. Jack, C. R. *et al.* Introduction to the recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Elsevier* **7**, 257–262 (2011).

310. Sperling, R. A. *et al.* Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on. *Elsevier* **7**, 280–292 (2011).

311. Wozniak, M., Mee, A., A, R. I.-T. J. of P. & 2009, undefined. Herpes simplex virus type 1 DNA is located within Alzheimer's disease amyloid plaques. *Wiley Online Library* **217**, 131–138 (2009).

312. Jamieson, G. A., Maitland, N. J., Wilcock, G. K., Yates, C. M. & Itzhaki, R. F. Herpes simplex virus type 1 DNA is present in specific regions of brain from aged people with and without senile dementia of the Alzheimer type. *J Pathol* **167**, 365–368 (1992).

313. Jamieson, G. A., Maitland, N. J., Wilcock, G. K., Craske, J. & Itzhaki, R. F. Latent herpes simplex virus type 1 in normal and Alzheimer's disease brains. *J Med Virol* **33**, 224–227 (1991).

314. Carbone, I. *et al.* Herpes virus in Alzheimer's disease: relation to progression of the disease. *Elsevier*.

315. Readhead, B., Haure-Mirande, J., Neuron, C. F.- & 2018, undefined. Multiscale analysis of independent Alzheimer's cohorts finds disruption of molecular, genetic, and clinical networks by human herpesvirus. *Elsevier*.

316. Eimer, W. A. *et al.* Alzheimer's Disease-Associated β-Amyloid Is Rapidly Seeded by Herpesviridae to Protect against Brain Infection. *Neuron* **99**, 56-63.e3 (2018).

317. Spitzer, P., Condic, M., Herrmann, M., reports, T. O.-S. & 2016, undefined. Amyloidogenic amyloid-β-peptide variants induce microbial agglutination and exert antimicrobial activity. *nature.com*.

318. Kumar, D. K. V. *et al.* Amyloid-β peptide protects against microbial infection in mouse and worm models of Alzheimer's disease. *Sci Transl Med* **8**, (2016).

319. Bourgade, K. *et al.* β-Amyloid peptides display protective activity against the human Alzheimer's disease-associated herpes simplex virus-1. *Biogerontology* **16**, 85–98 (2015).

320. Soscia, S. J. *et al.* The Alzheimer's disease-associated amyloid beta-protein is an antimicrobial peptide. *PLoS One* **5**, (2010).

321. Moir, R., Lathe, R., Dementia, R. T.-A. & & 2018, undefined. The antimicrobial protection hypothesis of Alzheimer's disease. *Elsevier*.

322. Ezzat, K., Pernemalm, M., Pålsson, S., … T. R.-N. & 2019, undefined. The viral protein corona directs viral pathogenesis and amyloid aggregation. *nature.com*.

323. Saji, N. *et al.* Analysis of the relationship between the gut microbiome and dementia: a cross-sectional study conducted in Japan. *Scientific Reports 2019 9:1* **9**, 1–9 (2019).

324. Vogt, N. M. *et al.* Gut microbiome alterations in Alzheimer's disease. *Scientific Reports 2017 7:1* **7**, 1–11 (2017).

325. Cummings, J., Lee, G., Ritter, A., Sabbagh, M. & Zhong, K. Alzheimer's disease drug development pipeline: 2019. *Alzheimers Dement (N Y)* **5**, 272–293 (2019).

326. Vermunt, L. *et al.* Duration of preclinical, prodromal, and dementia stages of Alzheimer's disease in relation to age, sex, and APOE genotype. *Alzheimer's & Dementia* **15**, 888–898 (2019).

327. Gó Mez-Isla, T. *et al.* Profound loss of layer II entorhinal cortex neurons occurs in very mild Alzheimer's disease. *Soc Neuroscience* (1996).

328. Cummings, J. *et al.* Alzheimer's disease drug development pipeline: 2022. *Alzheimer's & Dementia: Translational Research & Clinical Interventions* **8**, e12295 (2022).

329. Rabinovici, G. D. Controversy and Progress in Alzheimer's Disease — FDA Approval of Aducanumab. *New England Journal of Medicine* **385**, 771–774 (2021).

330. Swanson, C. J. *et al.* A randomized, double-blind, phase 2b proof-of-concept clinical trial in early Alzheimer's disease with lecanemab, an anti-Aβ protofibril antibody. *Alzheimers Res Ther* **13**, (2021).

331. Riederer, F. Donanemab in early Alzheimer's Disease. *Journal fur Neurologie, Neurochirurgie und Psychiatrie* **22**, 142–143 (2021).

332. Masters, C. L. *et al.* Amyloid plaque core protein in Alzheimer disease and Down syndrome. *Proceedings of the National Academy of Sciences* **82**, 4245–4249 (1985).

333. Hou, L. *et al.* Solution NMR studies of the A beta(1-40) and A beta(1-42) peptides establish that the Met35 oxidation state affects the

mechanism of amyloid formation. *J Am Chem Soc* **126**, 1992–2005 (2004).

334. Masters, C., Spring, D. S.-C. & 2012, undefined. Biochemistry of amyloid β-protein and amyloid deposits in Alzheimer disease. *perspectivesinmedicine.cshlp.org* doi:10.1101/cshperspect.a006262.

335. Gremer, L. *et al.* Fibril structure of amyloid-β(1–42) by cryo–electron microscopy. *Science (1979)* **358**, 116–119 (2017).

336. Kollmer, M. *et al.* Cryo-EM structure and polymorphism of Aβ amyloid fibrils purified from Alzheimer's brain tissue. *Nature Communications 2019 10:1* **10**, 1–8 (2019).

337. Iwatsubo, T. *et al.* Visualization of Aβ42(43) and Aβ40 in senile plaques with end-specific Aβ monoclonals: Evidence that an initially deposited species is Aβ42(43). *Neuron* **13**, 45–53 (1994).

338. Lau, H. H. C., Ingelsson, M. & Watts, J. C. The existence of Aβ strains and their potential for driving phenotypic heterogeneity in Alzheimer's disease. *Acta Neuropathologica 2020 142:1* **142**, 17–39 (2020).

339. Meyer-Luehmann, M. *et al.* Exogenous induction of cerebral β-amyloidogenesis is governed bf agent and host. *Science (1979)* **313**, 1781–1784 (2006).

340. Thal, D. R., Rüb, U., Orantes, M. & Braak, H. Phases of Aβ-deposition in the human brain and its relevance for the development of AD. *Neurology* **58**, 1791–1800 (2002).

341. Nguyen, P. H. *et al.* Amyloid oligomers: A joint experimental/computational perspective on Alzheimer's disease, Parkinson's disease, type II diabetes, and amyotrophic lateral sclerosis. *Chemical Reviews* vol. 121 Preprint at https://doi.org/10.1021/acs.chemrev.0c01122 (2021).

342. Cohen, S. I. A. *et al.* Proliferation of amyloid-β42 aggregates occurs through a secondary nucleation mechanism. *Proc Natl Acad Sci U S A* **110**, 9758–9763 (2013).

343. Noel, J. K., Whitford, P. C., Sanbonmatsu, K. Y. & Onuchic, J. N. SMOG@ctbp: Simplified deployment of structure-based models in GROMACS. *Nucleic Acids Res* **38**, 657–661 (2010).

344. Noel, J. K. *et al.* SMOG 2: A Versatile Software Package for Generating Structure-Based Models. *PLoS Comput Biol* **12**, e1004794 (2016).

345. Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1–2**, 19–25 (2015).

346. Vanommeslaeghe, K. *et al.* CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields. *J Comput Chem* **31**, 671–690 (2010).

347. Piana, S., Lindorff-Larsen, K. & Shaw, D. E. How robust are protein folding simulations with respect to force field parameterization? *Biophys J* **100**, (2011).

348. Essmann, U. *et al.* A smooth particle mesh Ewald method. *J Chem Phys* **103**, 8577 (1998).

349. Bussi, G., Donadio, D. & Parrinello, M. Canonical sampling through velocity rescaling. *J Chem Phys* **126**, 14101 (2007).